Ca' Foscari
University
of Venice

Master's Degree Programme in
Data Analytics for Business and Society

Road-Based Tourism Mobility in Italy: Predictive Modelling and Behavioural Profiling through Explainable AI

Master's Thesis

Supervisor: Dr. Lorenzo Schiavon
Graduand: Daniele Barbato
Student ID: 883661
Academic Year: 2025/2026

Contents

Introduction

Tourism represents one of the most important economic and cultural activities in Italy, playing a central role in shaping mobility patterns, local economies, and territorial attractiveness. The country has historically been characterised by a strong tourism vocation, supported by its cultural heritage, natural landscapes, and diversified tourism offer. At the same time, tourism also constitutes a relevant component of the Italian economy, generating employment opportunities and contributing to the development of both urban and rural areas.

In this context, domestic tourism and road-based mobility have traditionally occupied a particularly important position. The prevalent use of private cars has made road travel one of the most flexible and accessible forms of tourism in Italy. This form of mobility allows travellers to organise their journeys more autonomously, going in less accessible destinations, and adapt travel itineraries based on personal preferences and constraints.

The relevance of road-based tourism became even more marked during the Covid-19 pandemic, which profoundly altered global tourism dynamics and travel behaviour. Restrictions on international mobility, health concerns, and the growing preference for more independent and proximity-based forms of travel encouraged the adoption of flexible mobility solutions. In this scenario, trips with private road-based vehicles emerged as one of the most resilient tourism patterns, particularly in countries such as Italy, where domestic tourism and the use of car already played a major role before the pandemic.

Despite the central role of tourism in Italy, the literature has only partially explored how Italian tourists organise and experience their travel mobility patterns. Existing studies often focus on tourism demand, destination choice, or transport systems separately, while fewer contributions analyse how socio-demographic characteristics and trip features influence the way individuals move during their holidays.

In particular, relatively limited attention has been devoted to understanding the factors associated with different forms of tourist mobility and the behavioural characteristics linked to more flexible and autonomous travel patterns. This aspect has become increasingly relevant in recent years, especially after the transformations in tourism behaviour observed during and after the Covid-19 pandemic.

For this reason, the present study aims to analyse tourist mobility patterns in Italy using microdata from the Istituto Nazionale di Statistica (ISTAT, 2023) survey *Viaggi e Vacanze*. The analysis focuses on the characteristics associated with different ways of organising and experiencing travel, with specific attention devoted to road-based tourism as one of the most widespread forms of mobility in the Italian context. Based on this data, we investigate mobility choices by first classifying trips into private road-based trips and other type of trips categories, a step that enables a focused analysis of the factors shaping travel behaviour.

The contribution is both substantive and methodological. Substantively, the thesis aims to provide a population-level characterization of road-based trip tourism in Italy over nearly a decade that includes a major exogenous shock. By anchoring the construct to the main mode field of the survey and documenting edge cases and robustness, it replaces ad hoc labels with a reproducible definition. To the best of our knowledge, a data-driven approach based on survey microdata is still relatively uncommon in studies that focus on road-based travel within national tourism contexts.

The findings speak to several audiences. For destination managers and DMOs (Destination Management Organizations), knowing the sociodemographic profile and spatial footprint of road-based trips helps target information, parking policies, and visitor-flow management, especially in peak seasons and sensitive areas. For transport planners, comparative evidence on the prevalence of car-based trips across regions and over time can inform infrastructure maintenance, intermodal integration, and roadside services. For hospitality and attractions, differences in duration and activity mixes across modes can guide pricing, packaging, and scheduling. Finally, the predictive models provide a forward-looking lens: by quantifying which traits and contexts most increase the probability of a road-based trips, they suggest where marginal changes (e.g., in access or information) might have the greatest impact.

In short, by combining official microdata with advanced machine learning methods and contemporary interpretability techniques, this thesis provides a structured, data-driven portrait of road-based trip mobility in Italy and clarifies how it differs from other forms of travel. The purpose is not only to describe, but to articulate insights that are actionable for stakeholders seeking to better understand and manage tourism mobility patterns in Italy.

The aim of this thesis is to better understand how Italians organise and experience their holidays, with particular attention to tourism mobility patterns and the role of road-based travel within the Italian tourism context. More specifically, the analysis investigates the socio-demographic and trip-related characteristics associated with different forms of tourist mobility, in order to identify the main factors shaping how individuals move during their vacations. Particular attention is also devoted to model interpretability and to the changes in tourism mobility behaviours observed after the Covid-19 pandemic.

The thesis is structured as follows. Chapter 1 introduces the research context, discussing tourism mobility patterns in Italy, the role of road-based travel, and the transformations in travel behaviour observed during and after the Covid-19 pandemic. Chapter 2 presents the ISTAT *Viaggi e Vacanze* dataset, describes the construction of the analytical dataset and the road-based indicator, and provides a descriptive overview of the main socio-demographic and trip-related characteristics of travellers. Chapter 3 displays the empirical strategy and the methodological framework used in the analysis, including both traditional statistical models and machine learning techniques. Chapter 4 presents the empirical results, focusing on predictive performance, behavioural interpretation through explainable artificial intelligence tools, and the main characteristics associated with road-based tourism in Italy. Finally, Chapter 5 concludes the thesis by summarising the main findings, discussing the limitations of the study, and suggesting potential avenues for future research.

Chapter 1

Background

1.1 Tourism and mobility in Italy

Tourism has long represented a cornerstone of the Italian economy, culture, and identity.

According to the United Nations World Tourism Organization (UNWTO, 2021) estimates, Italy has consistently ranked among the most visited countries worldwide in recent years, attracting tens of millions of international tourists every year. The country thus maintains a central position in global tourism flows, attracting travellers with its unique combination of historical cities, art, architecture, gastronomy, religious sites, and natural landscapes. From the Alps in the north to the beaches of Sicily in the south, Italy offers a diverse array of experiences that appeal to a wide spectrum of international and domestic visitors

This attractiveness is also reflected in the travel behaviour of Italian residents. According to data from the Italian National Institute of Statistics (ISTAT, 2023) "Trips and Holidays" survey, which collects detailed information on trips taken by residents of Italy, a substantial share of recorded trips between 2015 and 2023 were directed toward domestic destinations, with a smaller but notable portion involving international travel. Figure 1.1 illustrate these trends showing the distribution of trips across broad destination areas (Europe, Africa, Asia, and the Americas) and across Italian regions.

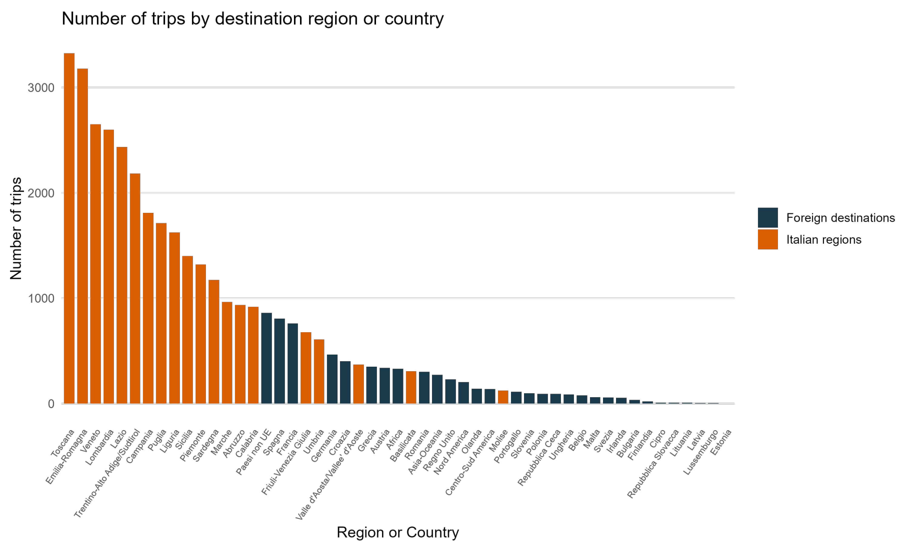


Figure 1.1: Number of trips by destination region or country.

Depending on the source and on whether indirect and induced effects are included, its contribution before the Covid-19 pandemic was generally estimated to range between 5% and 13% of national GDP, and between 6% and 15% of total employment(ISTAT, 2023; WTTC, 2024). Beyond its aggregate contribution, tourism sustains many local economies, particularly in areas that depend heavily on seasonal flows. Regions such as Veneto, Tuscany, Emilia-Romagna, Lazio and Lombardy consistently attract the largest volumes of visitors, while rural villages, mountain areas and inland towns have also experienced growth thanks to the rising interest in alternative and niche forms of travel.

However, tourism in Italy is far from a uniform phenomenon. It embraces a wide variety of motivations and practices: leisure holidays on the coast or in mountain resorts; cultural trips to cities of art and UNESCO heritage sites (as of 2025, Italy holds the record for the most UNESCO, 2025, World Heritage Sites in the world, with 61 sites in total: 55 cultural and 6 natural properties); religious pilgrimages to sanctuaries and sacred destinations; eno-gastronomic itineraries dedicated to regional wines and traditional cuisine; wellness and spa retreats; as well as sports-related and adventure tourism in alpine and rural areas. Alongside these, luxury travel and organised package tours coexist with low-cost and independent forms of mobility, reflecting the diversity of demand across socio-economic groups and generations. In recent years, new modalities such as experiential tourism, sustainable or "slow" tourism, and proximity-based travel have also gained ground, encouraged both by shifting values and by external shocks such as the pandemic (Dimanche and Andrades, 2024).

Within this broader tourism landscape, road-based mobility plays a particularly important role in Italy. The widespread use of private vehicles allows travellers to organise trips with greater flexibility, autonomy, and accessibility compared to more infrastructure-dependent forms of transport. In particular, road-based travel is often associated with domestic and short-to-medium distance tourism, where travellers can more easily adapt itineraries, destinations, and travel timing according to personal preferences and household needs.

The relevance of road-based tourism has also increased alongside broader changes in travel behaviour, including the growing demand for flexible, proximity-based, and more independent forms of tourism. In this sense, road-based mobility represents an important component of contemporary Italian tourism patterns and a key dimension for understanding how Italian residents organise and experience their holidays.

This pattern is also confirmed by evidence from the ISTAT survey. As shown in Figure 1.2, the share of holidays undertaken through road-based mobility remained relatively stable around 60–65% during the pre-pandemic years, increased sharply to over 76% in 2020, and then gradually declined while remaining above pre-pandemic (2019) levels.

These dynamics suggest that private road-based mobility became particularly important during the Covid-19 shock, when restrictions on international travel and concerns related to shared transport encouraged more flexible and autonomous forms of tourism mobility. At the same time, the persistence of relatively high levels even after the pandemic confirms the structural relevance of road-based travel within Italian tourism patterns.

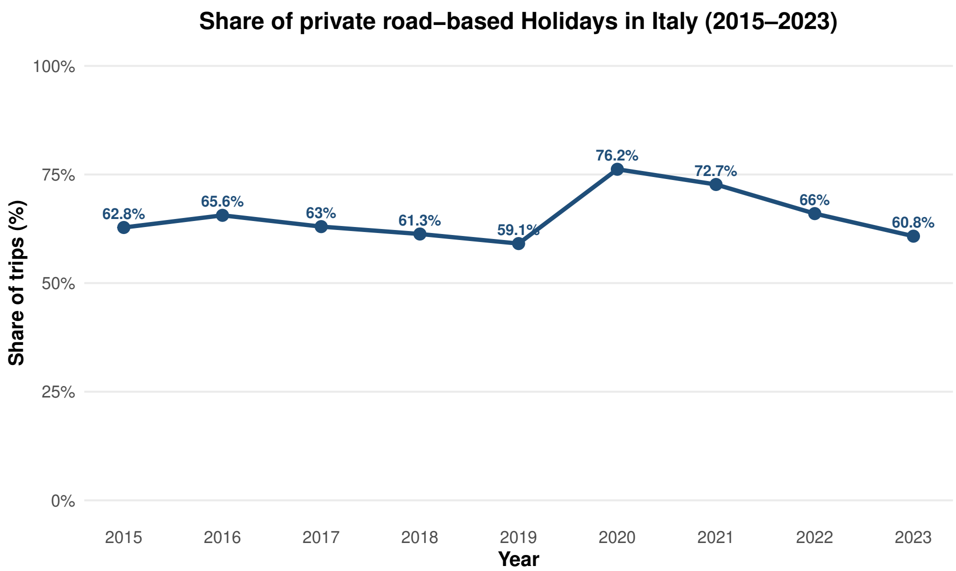


Figure 1.2: Share of road-based holidays in Italy, 2015–2023.

Road-based trip also embodies a cultural dimension. We can trace its historical roots back to the early 20th century, particularly in the United States, where the proliferation of the car and road infrastructure gave birth to the first large-scale travel-by-car movements.

The growing relevance of road-based tourism is closely connected to the historical diffusion of private mobility and mass motorisation during the twentieth century. The expansion of road infrastructure and the increasing accessibility of private vehicles progressively transformed tourism mobility patterns, making car-based travel one of the most widespread forms of domestic tourism.

Early forms of car-based tourism emerged in the United States during the first decades of the twentieth century, following the rapid expansion of automobile ownership and highway infrastructure. As noted by (Belasco, 1979), travel by car progressively evolved beyond a simple transport solution and became associated with broader transformations in leisure practices, autonomy, and spatial mobility.

These transformations progressively extended to Europe during the post-war economic expansion, when increasing motorisation and infrastructure development facilitated the diffusion of domestic tourism and private mobility (Salomon, 1985).

In Italy, the rapid development of the motorway system and the increasing accessibility of private car ownership played a central role in integrating road-based mobility into tourism behaviour. The expansion of road infrastructure facilitated travel towards coastal, rural, and mountain destinations, progressively making private mobility a structural component of domestic tourism patterns.

From a broader sociological perspective, these changes also reflected wider transformations in leisure behaviour and tourism practices. As observed by (Urry, 1990), modern tourism is deeply connected to the organisation of leisure time and to the separation between work and recreational activities. Within this framework, private mobility emerged as a flexible and accessible form of tourism, allowing individuals and households to organise travel more autonomously and independently from traditional tourism structures.

Overall, the historical diffusion of private mobility contributed to making road-based tourism a structural component of contemporary tourism systems. Understanding these transformations is therefore important for interpreting tourism mobility patterns and the growing importance of flexible and short-distance forms of travel.

1.2 The impact of Covid-19 on tourism mobility

The Covid-19 pandemic significantly alters tourism and mobility patterns around the world, generating important changes in travel behaviour, destination choice, and transport preferences. International travel restrictions, lockdowns, and concerns about health risks significantly reduced traditional tourism flows and altered the way people organised their vacations.

In particular, several studies documented a sharp reduction in long-distance travel, a decline in the use of public transport, and an increasing preference for more flexible and autonomous forms of mobility, including private vehicles, walking, and cycling (Lee and Eom, 2024; Barbieri et al., 2021). The spread of remote working further contributed to changing everyday mobility patterns, reducing systematic commuting and increasing local and proximity-based travel.

These transformations were also evident in the Italian tourism context. During the pandemic, domestic and proximity tourism became increasingly important, while alternative accommodation types such as agritourisms, campsites, and small-scale hospitality structures often showed greater resilience compared to traditional urban tourism. More generally, the pandemic accelerated several pre-existing trends towards more decentralised, flexible, and independent forms of tourism mobility. As a result, road-based tourism represented one of the most resilient forms of tourism mobility during the Covid-19 shock.

Several studies also suggest that at least part of these behavioural changes may persist beyond the pandemic period, reflecting broader transformations in mobility preferences and travel organisation (OECD, 2022; Urry, 1990). Understanding whether these changes represent temporary adaptations or longer-term structural shifts is therefore particularly relevant for interpreting contemporary tourism mobility patterns.

For this reason, the present thesis analyses tourism mobility in Italy across both the pre- and post-pandemic periods, with particular attention devoted to the role of road-based travel and to the behavioural characteristics associated with different forms of tourism mobility.

1.3 Research objectives

Understanding tourism mobility behaviour represents a central issue in contemporary tourism research, particularly in periods characterised by strong economic and social disruptions such as the Covid-19 pandemic. Changes in individual preferences, mobility constraints, transport accessibility, and risk perceptions have significantly transformed the way people organise and experience travel, generating new tourism dynamics and mobility patterns.

Within this context, the present thesis investigates tourism mobility patterns in Italy, with particular attention devoted to the role of road-based travel within domestic and international tourism behaviour. The objective of the study is to better understand how Italian residents organise their holidays, how different socio-demographic and trip-related characteristics influence mobility choices, and how these behavioural patterns evolved before and after the Covid-19 pandemic.

More specifically, the analysis focuses on the characteristics associated with road-based tourism, interpreted as a broader form of tourism mobility linked to the use of private road transport. The study combines descriptive analysis, predictive modelling, and explainable artificial intelligence techniques in order to identify the main determinants associated with different forms of tourism mobility and to explore the behavioural structure underlying road-based travel.

As discussed by Guidotti et al. (2018) and Barredo Arrieta et al. (2020a), explainable artificial intelligence (XAI) refers to a set of methodologies designed to improve the transparency and interpretability of machine learning models. These approaches allow researchers not only to evaluate predictive performance, but also to understand how specific variables contribute to model predictions. In the context of tourism research, XAI techniques provide valuable tools for interpreting complex behavioural patterns while maintaining a high degree of predictive flexibility.

The empirical analysis relies on microdata from ISTAT's national survey *Trips and Holidays* (*Viaggi e Vacanze*), which collects cross-sectional information on the travel behaviour of Italian residents over the period 2015–2023. The dataset includes a wide range of socio-demographic characteristics, such as age, gender, education, employment condition, and region of residence, together with detailed trip-level information including travel duration, accommodation type, seasonality, destination, and transport mode.

The thesis addresses three main research questions. First, it investigates which socio-demographic and trip-related characteristics are more strongly associated with road-based tourism mobility through the use of machine learning and explainable artificial intelligence techniques. Second, it examines how tourism mobility patterns changed before and after the Covid-19 pandemic.

To address these objectives, the study combines descriptive analysis with both traditional statistical models and more flexible machine learning approaches. These methods are used not only to classify and predict road-based tourism behaviour, but also to identify and interpret the factors that most strongly influence tourism mobility choices among Italian travellers.

Chapter 2

Trips and holidays Italian data

The analysis in this thesis is based on microdata from the *Trips and Holidays* (Viaggi e Vacanze) survey conducted annually by the Italian National Institute of Statistics (ISTAT, 2023). The survey provides a nationally representative picture of tourism participation among Italian residents and constitutes a key statistical source for analysing tourism demand in Italy.

A distinctive feature of the survey is its ability to jointly capture socio-demographic and economic characteristics of individuals alongside detailed information on the trips they undertake. This dual structure allows individual profiles to be linked to observed travel outcomes, making the data particularly suitable for analyses that focus on heterogeneity in tourism behaviour across population groups. In particular, it enables the identification of which characteristics of people are more likely to engage in private road-based tourism, rather than providing an examination of travel patterns.

The availability of data across multiple years makes it possible to construct a consistent analytical dataset covering the period 2015-2023. This time span provides the empirical context for examining differences in travel behaviour across individuals, including those observed around the COVID-19 pandemic, without implying a fully dynamic or longitudinal analysis.

2.1 The ISTAT "Trips and Holidays" survey

ISTAT *'Trips and Holidays'* survey is one of the few data collections capable of systematically capturing not only the volume of tourism flows, but also their underlying socio-demographic and behavioural determinants. Its strength lies in its ability to provide a comprehensive and continuous picture of how Italians travel, both within the country and abroad, over an extended period of time.

The comparability and reliability of the ISTAT survey are further strengthened by its alignment with the European statistical framework on tourism. In particular, Regulation (EU) No. 692/2011 established a harmonised system for the collection of tourism statistics across European Union Member States by defining common concepts, classifications, and variables. The regulation requires national statistical institutes to collect detailed information on both tourism demand and tourism supply, including trips, overnight stays, tourism expenditure, and accommodation capacity.

More recently, Regulation (EU) 2019/1681 updated and refined these provisions in order to better capture emerging tourism trends and new mobility behaviours. As a result, the ISTAT *Trips and Holidays* survey provides data that are not only nationally representative, but also fully comparable within the broader European statistical system (European Commission, 2019).

The survey aims to quantify and characterise tourism flows of Italian residents, both domestically and abroad, and to describe travel modes and socio-demographic profiles of both tourists and non-tourists.

The observation period covers all months of the calendar year. For each reference period and for each household member included in the sample, the survey collects information on trips completed in the reference month. The breadth of collected variables allows researchers to distinguish holiday travel from business travel and to classify holidays into short breaks (1–3 nights) and long holidays (4+ nights). In addition, the survey documents several trip attributes, including destination, purpose, organisation of the trip, accommodation type, main mode of transport, duration, and seasonality.

Since 2014, the travel module has been incorporated into ISTAT's broader Household Budget Survey (*Indagine sulle spese delle famiglie*), and it has been collected on a monthly basis primarily through Computer-Assisted Personal Interviewing (CAPI), a survey technique based on face-to-face interviews supported by digital questionnaires. The Household Budget Survey combines an initial and a final interview with a 14-day expenditure diary, within which the travel module is administered. This integration ensures that travel behaviour can be analysed jointly with socio-demographic characteristics and household conditions within a coherent and nationally representative sampling framework.

The household sample is selected using population registers. In recent years the sampling frame has progressively shifted from local municipal registers (LAC) to the National Register of the Resident Population (ANPR, 2025), which is currently the primary source for sample selection. The survey involves around 33,000 households distributed across approximately 540 municipalities, ensuring national representativeness while enabling disaggregation by region and key socio-demographic dimensions.

The Covid-19 pandemic also affected fieldwork procedures in 2020. During part of March and throughout the second quarter, face-to-face interviewing was temporarily replaced by Computer-Assisted Telephone Interviewing (CATI), a data collection technique based on telephone-administered questionnaires. In addition, data collection was suspended in April due to lockdown restrictions. Given the exceptionally low incidence of observed travel during that period, the near-zero level of recorded trips during the lockdown month can reasonably be considered a realistic representation of the underlying phenomenon.

Finally, ISTAT (2023) provides survey weights to support inference to the resident population. These weights account for the sampling design, non-response adjustments, and calibration to population totals. In this thesis, weights are employed to obtain a nationally representative analysis.

2.1.1 Weights

Survey weights are defined at the individual level and reflect the number of resident individuals represented by each sampled respondent. ISTAT provides final survey weights designed to allow inference from the sample to the resident population in order to be nationally representative. These weights incorporate the complex sampling design of the survey, adjustments for total non-response, and a calibration procedure to known population totals derived from external sources.

In particular, the calibration aligns the sample with the population distribution by region, sex, age, household structure, typology of municipality, labour-market condition, and month of observation. As a result, weighted estimates reproduce the demographic and territorial structure of the Italian resident population. In this thesis, survey weights are employed in descriptive analyses to produce population-representative estimates of the prevalence and distribution of road-based trip participation across socio-demographic groups.

To clarify the role of survey weights, consider a simple illustrative example. Let w_i denote the final survey weight associated with individual i, reflecting the number of individuals in the population represented by that observation. Let y_i be a binary indicator equal to 1 if individual i engaged in at least one road-based trip during the reference year, and 0 otherwise. The weighted estimate of the population prevalence of road-based trip participation is given by:

$$\hat{P}_{\text{roadbased}} = \frac{\sum_{i=1}^{N} w_i\, y_i}{\sum_{i=1}^{N} w_i}. \tag{2.1}$$

In this expression, each sampled individual contributes to the estimate proportionally to the size of the population group they represent. When survey weights are not applied, all observations contribute equally, implicitly assuming a simple random sample. This example illustrates how survey weights adjust descriptive estimates to reflect the underlying population structure, without affecting the definition of the individual-level outcome variable.

2.2 Data structure

The ISTAT Trips and Holidays microdata are organised in two distinct but complementary datasets, reflecting the dual nature of the information collected by the survey.

The first dataset called *Individui* is structured at the individual level and contains socio-demographic and household characteristics of all sampled respondents. Each observation corresponds to an individual interviewed in a given reference year and includes variables such as age, gender, education, employment status, household composition, and region of residence. This dataset represents the natural starting point for analyses aimed at profiling travellers and identifying population groups with different tourism behaviours.

The second dataset called *Viaggi* is structured at the trip level. Each row corresponds to a single journey undertaken by a respondent within the reference period and records detailed information on travel characteristics, including destination, duration, purpose of travel, accommodation type, and main mode of transport. As a result, individuals who undertook multiple trips within the same year appear multiple times in this dataset.

While the *Viaggi* dataset provides detailed information on tourism mobility and travel behaviour, the analysis developed in this thesis specifically focuses on the relationship between travel modes and the socio-demographic and trip-related characteristics associated with road-based tourism. The dataset allows the integration of mobility-related information with a broad set of individual and travel characteristics, making it possible to analyse how different profiles and trip attributes are associated with the choice of road-based travel.

2.2.1 Construction of the road-based mobility indicator

Road-based tourism is operationalised at the trip level, which constitutes the statistical unit of the empirical analysis. Each observation corresponds to a single journey undertaken by an individual within a given reference year. Defining the outcome at the trip level allows the analysis to focus directly on the transport mode adopted during the journey and on the socio-demographic and trip-related characteristics associated with road-based mobility choices.

The classification of road-based tourism relies on the main mode of transport adopted during the journey. More specifically, the analysis uses the variable `MEZZO` contained in the *Viaggi* dataset, which records the primary transport mode associated with each observed trip.

The use of the main transport mode as a classification measure is consistent with both the structure of the ISTAT survey and the objective of the analysis. The variable `MEZZO` identifies the primary transport mode associated with each journey, thus capturing the main mobility choice underlying the trip. Since the study focuses on the mobility of road-based tourism, this variable provides a clear and consistent way to distinguish trips based on privately used road transport from those relying on alternative transport modes. In addition, the use of a standardized survey variable ensures comparability and reproducibility across observations and survey years.

Moreover, this operationalisation aligns with the behavioural focus of the analysis, as the main transport mode reflects the key constraint and enabling factor shaping travel flexibility, route choice, and destination accessibility. In this sense, the proxy captures the core dimension that distinguishes road-based travel from other forms of tourism.

Some limitations of this operational definition should nevertheless be acknowledged. First, the main mode of transport represents only the predominant segment of the journey and may not fully capture multimodal trips involving different transport modes at different stages of travel. Second, the classification does not account for the intensity or duration of road use within the trip, potentially leading to minor ambiguity in borderline cases. Third, some tourism mobility characteristics are not observed in the dataset, including for example itinerary complexity and the number of intermediate stops.

The classification is based exclusively on the primary transport mode recorded in the survey. As a result, trips undertaken mainly through private road transport are classified as road-based even when the stay itself is relatively stationary, such as holidays in seaside or mountain destinations. On the other hand, journeys involving substantial road mobility after the main transport segment may not be fully captured. As an example, trips that include air transport followed by the use of local car are classified according to the primary transport mode reported in the dataset.

A trip is defined as road-based if its primary transport mode is a private car, rental car, camper, or motorcycle. All other transport modes – such as rail, air, coach/bus, ship/ferry, or other – are classified as non road-based trips. Formally, the trip-level indicator is defined as:

$$\text{roadbased}_{it} = \begin{cases} 1 & \text{if the main mode of transport of trip } t \text{ undertaken by individual } i \\ & \quad \in \{\text{Private car, Motorcycle, Camper, Rental car}\} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

The indicator constitutes the dependent variable of the empirical analysis and provides a direct representation of transport mode choice at the trip level. Modelling the outcome at this level of observation allows the analysis to capture variation in mobility behaviour across trips, while accounting for both socio-demographic characteristics and trip-specific attributes.

This specification is particularly suitable for the objectives of the study, as it focuses on the determinants associated with road-based tourism mobility rather than on aggregate travel behaviour. As a result, the modelling framework can exploit both socio-demographic heterogeneity across individuals and variation in trip characteristics within individuals.

2.3 Descriptive analysis

The final analytical dataset was constructed by integrating socio-demographic information from the *Individui* dataset with the trip-level records contained in the *Viaggi* dataset, providing a coherent framework for analysing the determinants of private road-based participation in Italy. The merge was performed using the unique respondent identifier (`progind`) and the reference year (`annrif`), ensuring that each journey is associated with the characteristics of the individual who undertook it.

The resulting dataset is structured at the trip level and counts 36903 observations. Each row corresponds to a single journey completed by an Italian resident and includes both travel attributes (such as destination, duration, purpose, accommodation type, and main mode of transport) and individual characteristics (including age, gender, education, labour-market status, and region of residence).

This structure implies that individuals who undertook multiple journeys within the same year appear multiple times in the dataset. This repetition arises from the nature of the analysis, where each trip is considered as a separate observation rather than focusing on individuals.

For the purposes of the empirical analysis, the dataset is restricted to trips involving at least one overnight stay, in line with the survey definition of tourism. Furthermore, variables were harmonised across years by recoding and aggregating categories when necessary, ensuring that socio-demographic characteristics and trip-related variables were measured consistently across all survey editions from 2015 to 2023.

The next section provides a descriptive overview of the dataset, examining the distribution of road-based trip participation across socio-demographic and then across trip-level characteristics.

Unless otherwise specified, all descriptive statistics and graphical representations are weighted using the final survey weights provided by ISTAT.

2.3.1 Distribution of Trips Across Years

Figure 2.1 illustrates the distribution of observed trips across survey years in the analytical dataset.

A visible decline in the number of recorded trips can be observed during the Covid-19 period, reflecting the substantial disruption that the pandemic caused to travel behaviour and tourism mobility. In the subsequent years, the number of trips gradually increases again, suggesting a partial recovery of tourism activity.

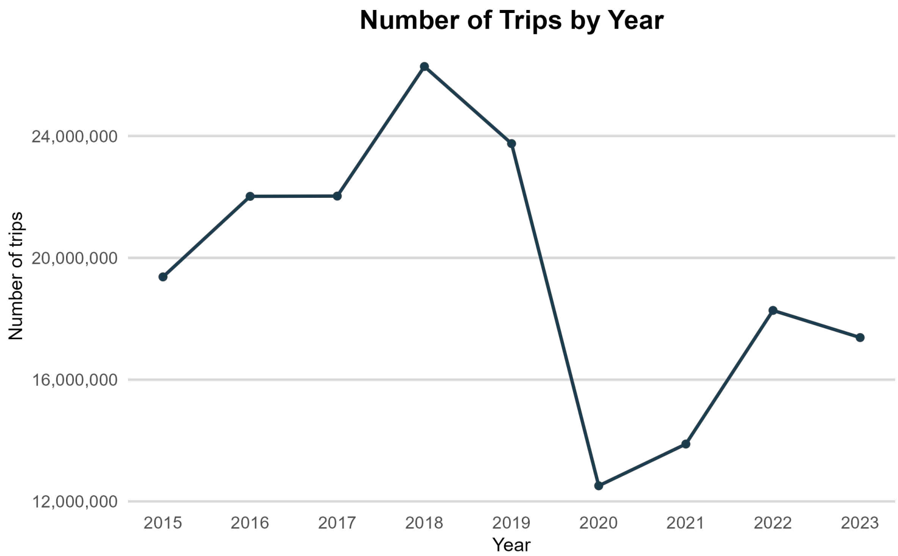


Figure 2.1: Weighted number of trips by survey year (2015–2023). The horizontal axis reports the survey year, while the vertical axis shows the estimated total number of trips.

Overall, the temporal distribution of observations confirms that the dataset provides adequate coverage of both pre-pandemic and post-pandemic travel behaviour, allowing for the exploration of potential changes in tourism mobility patterns over time. This pattern reflects the strong sensitivity of tourism activity to external shocks and confirms the relevance of distinguishing between pre- and post-pandemic periods in the empirical analysis.

2.3.2 Transport mode distribution

Figure 2.2 illustrates the distribution of the main transport modes used in the observed trips. Private car travel clearly represents the dominant transport mode, accounting for more than sixty percent of the journeys included in the sample. Air travel and train transport constitute the second and third most frequent modes, although their shares remain substantially lower compared to private road transport.

All remaining transport modes account for relatively limited proportions of total trips. Tourist coaches, maritime transport, rental cars, public buses, campers, and motorcycles appear only marginally within the overall distribution.

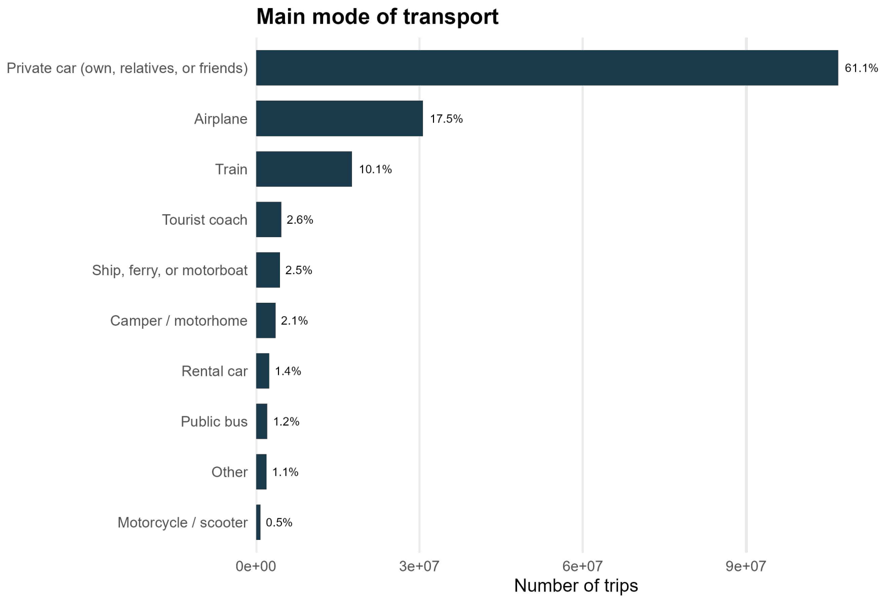


Figure 2.2: Distribution of transport modes used for trips in the sample. The horizontal axis shows the estimated number of trips, while the vertical axis reports the main mode of transport.

The predominance of private road transport reflects some structural characteristics of tourism mobility in Italy. The geographical fragmentation of destinations, the relevance of domestic tourism, and the widespread presence of small coastal, rural, and mountain locations often make private mobility a more flexible and accessible option compared to rail or air transport. In many cases, private vehicles allow travellers to reach destinations that are less directly connected to major transport infrastructures while also providing greater autonomy in terms of schedules, route planning, and travel organisation.

From an empirical perspective, this distribution is particularly relevant for the objectives of the study, as it highlights the central role of road-based mobility within Italian tourism behaviour. At the same time, the observed imbalance across transport modes implies that the dependent variable is not uniformly distributed across classes. This aspect should be taken into account when interpreting model predictions and classification performance, especially in relation to the ability of the models to correctly identify less frequent transport categories.

Figure 2.3 illustrates the distribution of trips according to whether they are classified as road-based trips or other types of journeys according to definition in Equation 2.2. Road-based trips represent the majority of the observed trips in the dataset, accounting for a substantially larger share compared to other types of travel.

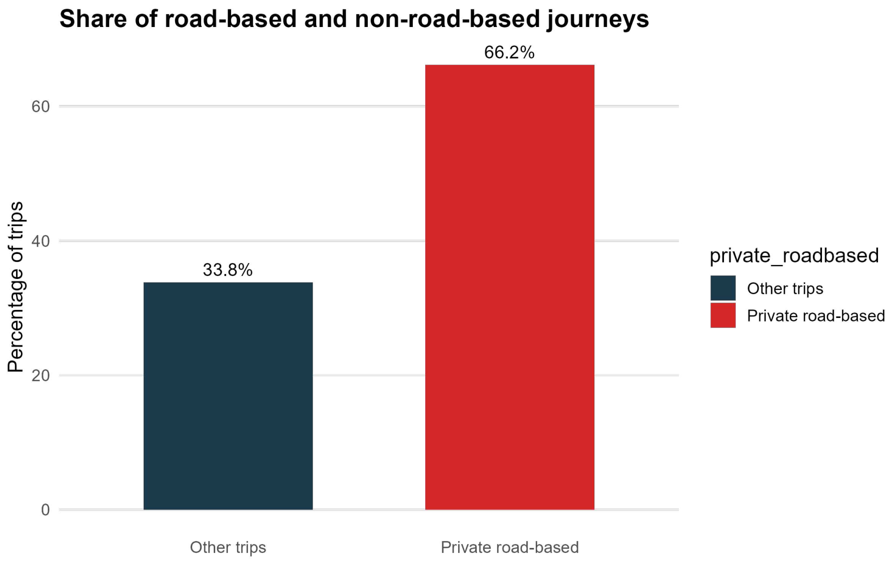


Figure 2.3: Distribution of road-based and other trips in the analytical dataset.

This result reflects the strong predominance of road-based mobility in the Italian tourism context, where private vehicles are frequently used for domestic travel. The high share of road-based trips in the dataset further justifies the focus of this study on the characteristics and determinants of this specific form of tourism mobility.

2.3.3 Age distribution by trip type

Understanding the socio-demographic characteristics of travellers represents an important step in the descriptive exploration of the dataset. In particular, age may influence several aspects of travel behaviour, including destination choice, travel duration, and preferred modes of transport.

Moreover, comparing the age profile of individuals undertaking road-based trips with that of travellers involved in other types of journeys may provide useful insights into potential differences in tourism mobility behaviour across population groups. Although this section focuses only on descriptive evidence, these patterns may later inform the empirical analysis presented in the following chapters.

Figure 2.4 compares the age distribution of travellers undertaking road-based trips with that of individuals using other transport modes. The comparison provides a first descriptive overview of potential demographic differences associated with road-based tourism mobility.

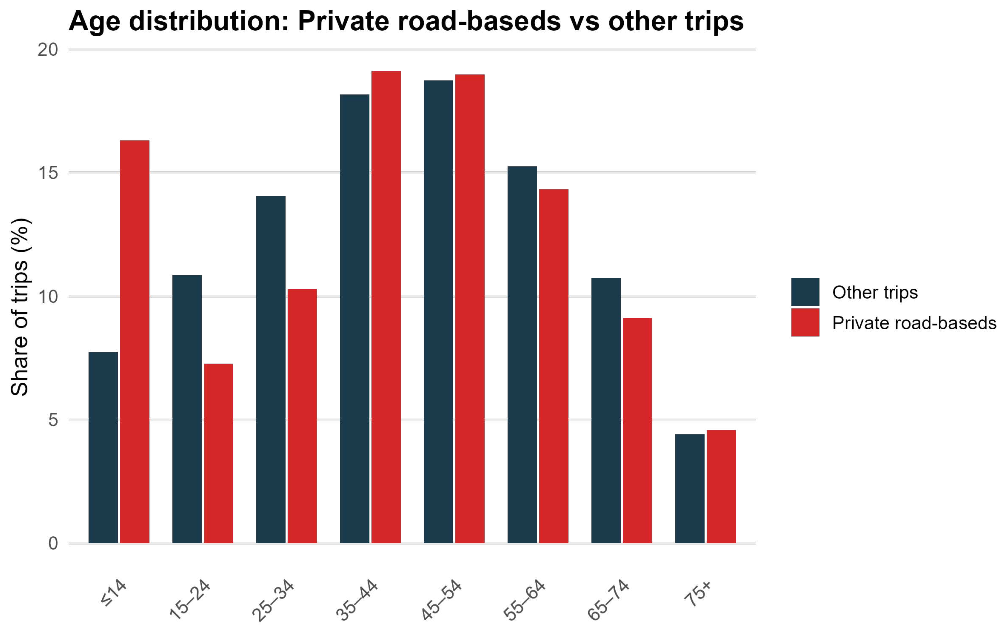


Figure 2.4: Age distribution of travellers: comparison between road-based and non-road-based trips.

The figure shows that travel activity is largely concentrated among adult and middle-aged individuals in both categories. In particular, travellers between 35 and 64 years old account for the largest shares of observed trips, highlighting the central role of these age groups within overall tourism mobility patterns.

At the same time, some differences emerge between road-based and non-road-based travel. Road-based trips seems to be relatively more common among middle-aged individuals, in particular between 35 and 54 years old, whereas younger travellers – especially those aged between 15 and 24 years – represent a comparatively smaller share of road-based mobility. This pattern may partially reflect differences in vehicle availability, economic resources, or travel autonomy across age groups.

The youngest age category is substantially more represented among private road-based trips. This pattern likely reflects the greater prevalence of family travel within this segment, as children typically travel together with their parents. Consistent with this interpretation, the age groups most commonly associated with parenthood (35–54 years old) also display a relatively high incidence of road-based trips. Taken together, these patterns provide preliminary evidence of a close association between family travel and the use of private road-based transport.

However, these descriptive differences should be interpreted with caution, as they may also reflect the influence of other socio-demographic and trip-related factors. The role of these characteristics will therefore be investigated more formally through the predictive models presented in the following chapters.

For older age groups, the differences between the two categories appear more limited, indicating a relatively similar age composition among senior travellers.

Although purely descriptive, these patterns provide an initial indication that age may be associated with differences in tourism mobility behaviour. This aspect will be explored more formally in the predictive models presented in the following chapters.

2.3.4 Gender distribution by transport mode

Gender is one of the key socio-demographic characteristics commonly examined in tourism research, as travel behaviour may differ between men and women in terms of mobility patterns, travel preferences, and mode of transport. Analysing the gender composition of travellers therefore helps identify whether road-based participation is associated with specific demographic profiles or whether it is broadly distributed across the population.

In the context of this study, comparing the gender distribution between road-based and other trips provides an initial indication of whether the choice of travelling by private vehicle is influenced by gender-related factors. Figure 2.5 illustrates the share of trips undertaken by men and women across the two categories of journeys.

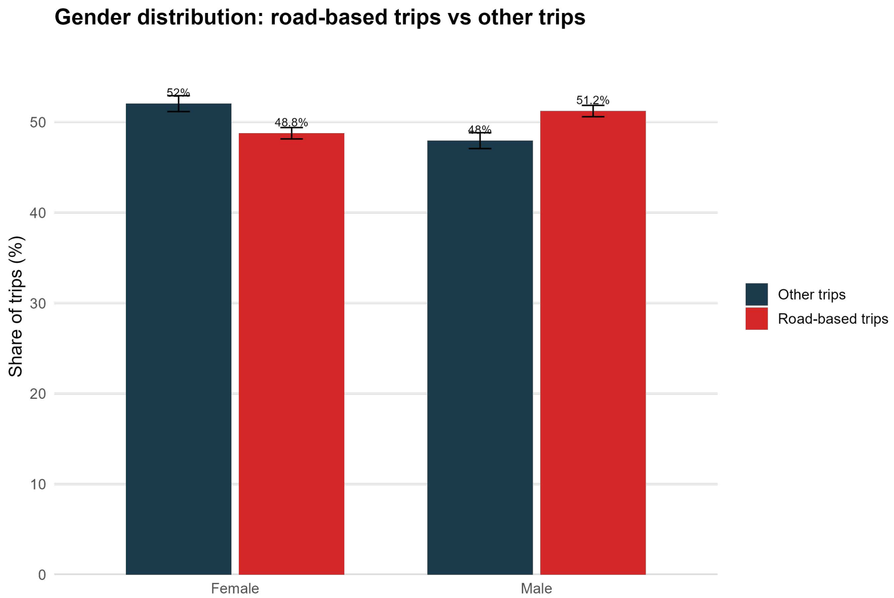


Figure 2.5: Gender distribution of travellers: comparison between road-based trips and other trips.

The figure shows that the gender composition of travellers is relatively balanced across the two categories of trips. Men account for a slightly larger share of road-based trips, whereas women appear marginally more represented among non-road-based journeys.

Although, however, the observed differences remain limited, a Pearson's Chi-squared test indicates the presence of a statistically significant association between gender and trip type ($\chi^2 = 35.19$, $p < 0.001$). Although the differences in proportions are relatively small, the large sample size produces highly precise estimates and statistically robust results.

From a substantive perspective, the evidence suggests that gender alone is unlikely to represent a major determinant of tourism mobility behaviour. Other socio-demographic and trip-related characteristics are therefore likely to play a more important role in explaining the adoption of road-based transport.

2.3.5 Education level of travellers

Education level represents an important socio-demographic characteristic in tourism studies, as it is often associated with differences in travel behaviour, mobility patterns, and destination choices. Analysing the educational composition of travellers therefore helps provide a clearer overview of the socio-demographic profiles represented in the dataset.

Comparing the educational distribution of travellers undertaking private road-based trips with that of individuals involved in other types of journeys may also provide useful insights into potential differences in tourism mobility behaviour across population groups. Figure 2.6 illustrates the distribution of trips by education level across the two categories.

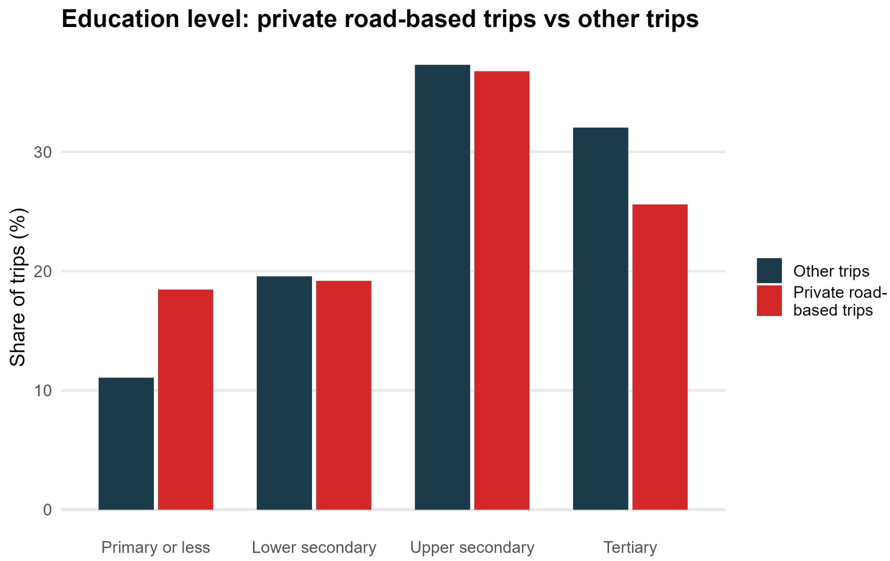


Figure 2.6: Education level of travellers: comparison between private road-based trips and other trips.

The figure shows that travellers with upper secondary and tertiary education account for the largest share of trips in both categories, reflecting the overall educational composition of the travelling population. Nevertheless, some differences emerge between private road-based and non-road-based trips.

In particular, individuals with tertiary education appear relatively more represented among non-road-based trips, whereas private road-based mobility shows a comparatively larger share of travellers with lower educational attainment. This pattern may reflect differences in travel organisation, destination accessibility, and transport preferences across socio-demographic groups.

At the same time, it is important to interpret with caution the relatively high share of travellers with primary education or less among private road-based trips. As highlighted in the previous age analysis, children represent a large proportion of road-based trips, and their inclusion within the lowest education category likely contributes to this result.

Overall, the evidence suggests that education level may be associated with differences in tourism mobility behaviour, although part of the observed relationship is likely influenced by the demographic (in particular by the age) composition of travelling households.

2.3.6 Regional distribution of road-based trip participation

Geographical context could represent an important factor influencing travel behaviour. Differences in territorial accessibility, infrastructure availability, proximity to tourist destinations, and local mobility habits may contribute to shaping how individuals organise their trips and which transport modes they choose. For this reason, examining the spatial distribution of road-based trips across regions provides a useful descriptive perspective on the dataset.

Figure 2.7 illustrates the share of road-based trips by region of residence of travellers. For each region, the percentage is calculated as the weighted proportion of trips classified as road-based trips over the total number of trips undertaken by residents of that region. It is important to note that the map does not represent the absolute number of trips originating from each region. Instead, it reports the share of road-based trips among all trips undertaken by residents of each region. This distinction is important because regions with a larger population or higher travel frequency don't necessarily display higher values in the map; instead, the figure reflects the relative propensity to undertake road-based trips.

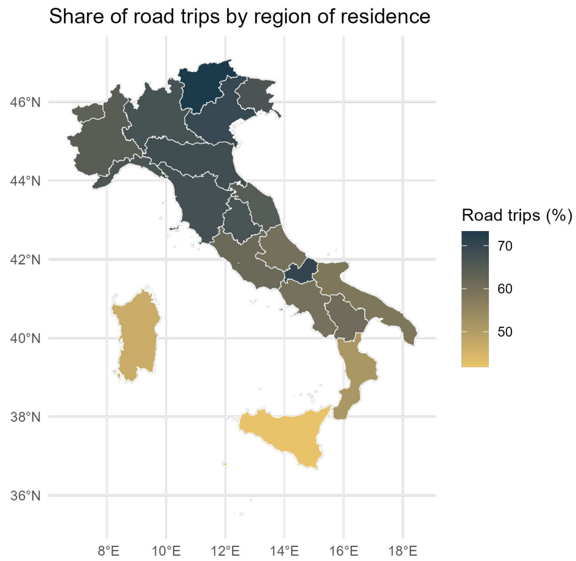


Figure 2.7: Share of road-based trips by region of residence in Italy. Regions are shaded according to the percentage of trips classified as road-based trips, with darker colours indicating higher shares.

The resulting map provides a first geographical overview of the prevalence of road-based trips across the Italian territory. Although the representation is purely descriptive, it highlights the existence of some regional variation in tourism mobility behaviour.

In most Italian regions, road-based trips account for the majority of observed journeys, often exceeding sixty percent of total trips. Particularly high shares emerge in several northern and central regions, including Veneto, Trentino-Alto Adige, Emilia-Romagna, and Toscana. By contrast, lower proportions are observed in the island regions, especially Sicilia and Sardegna, where alternative transport modes appear relatively more prevalent.

These geographical differences are probably influenced by a combination of infrastructural, territorial, and accessibility-related factors. Regions woth good road infrastructure, easy access to domestic tourist destinations and extensive local mobility networks are more likely to rely on private road-based transport. Island regions instead may rely more heavily on air and maritime transport due to the structural constraints associated with long-distance mobility and insular accessibility.

In general, the descriptive evidence presented in this section provides a broader overview of the socio-demographic and geographical composition of tourism mobility patterns in Italy. The analysis now moves from traveller characteristics to the main features of the trips themselves, focusing on variables such as trip duration, destination, and other travel-related attributes.

2.3.7 Trip duration: road-based trip vs other-trip

Trip duration represents an important dimension of travel behaviour, as different transport modes are often associated with different lengths of stay. To explore this aspect, trips were grouped into four duration categories: 1–3 days, 4–7 days, 8–14 days, and 15 or more days.

The Figure 2.8 is presented as a 100% stacked bar chart in order to emphasise the composition of trip duration within each group. This makes the comparison between road-based trips and other trips more immediate, as the reader can directly observe the relative importance of short, medium, and long stays in the two categories.

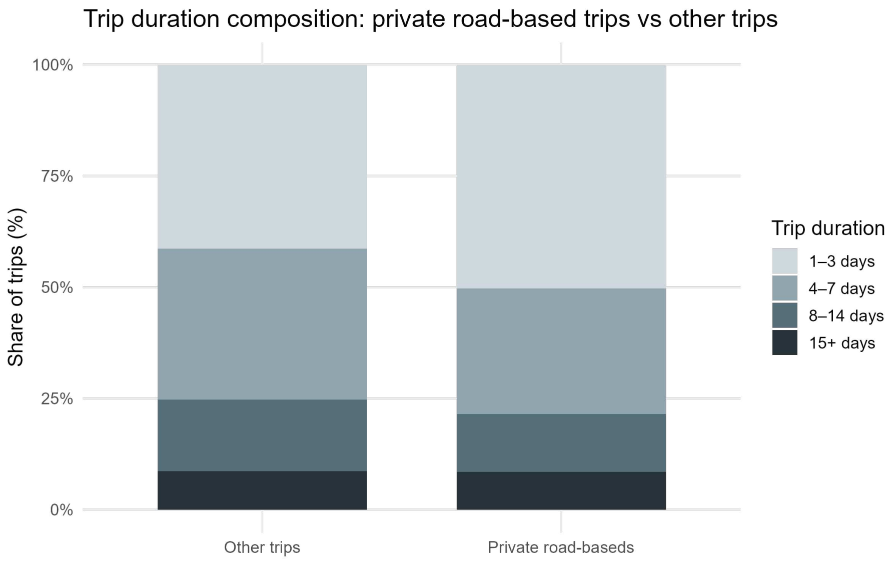


Figure 2.8: Trip duration composition: Private road-based trips vs other trips.

This pattern suggests that road-based trips are more frequently associated with shorter and more flexible travel arrangements, which are easier to organise using private vehicles. Longer trips, by contrast, may involve more structured planning and therefore a greater reliance on alternative transport modes such as air travel.

2.3.8 Accomodation

Accommodation choice reflects both the type of trip and the level of planning involved. Different forms of accommodation may also be associated with specific transport modes, particularly in the case of road-based trips.

To improve readability, detailed accommodation categories were aggregated into broader groups. In particular, similar types of accommodation were combined into categories such as hotel and similar, second home, friends or relatives, rental accommodation, and camping. Less frequent and heterogeneous categories were grouped into a residual category labelled *"Other"*.

Figure 2.9 shows the distribution of accommodation types for road-based trips and other trips. The comparison shows clear differences in accommodation preferences between the two groups, providing insight into the nature of road-based travel.

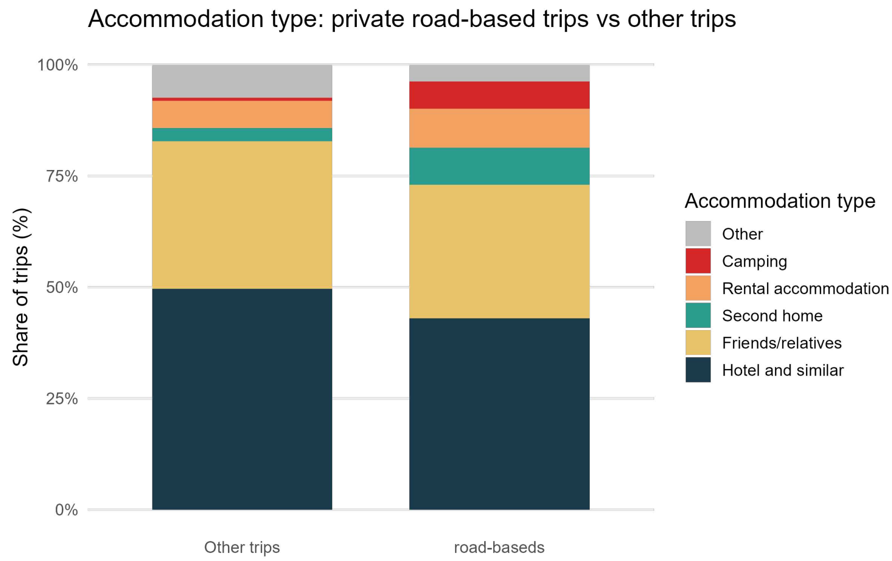


Figure 2.9: Distribution of accommodation types by trip type. The bars report the share of trips by accommodation category for road-based trips and other trips, expressed as percentages.

Overall, road-based trips seem to be more frequently associated with flexible and informal accommodation types, for example second homes, camping, or staying with friends and relatives, whereas other trips may show a higher reliance on traditional accommodation structures such as hotels. A particularly remarkable pattern concerns camping, which appears to be almost negligible among other trips, while being more present within road-based trips. This suggests that camping is strongly associated with road-based travel, likely due to the greater flexibility offered by private transport.

2.3.9 Seasonality

Seasonality represents a key dimension of tourism behaviour, as travel activity is often concentrated in specific periods of the year. To explore this aspect, trips are grouped into four quarters (Q1–Q4) based on their starting period.

Figure 2.10 shows the distribution of trips across quarters, distinguishing between private road-based trips and other types of journeys. An evident seasonal pattern emerges, with the highest concentration of trips emerging during the summer period (Q3).

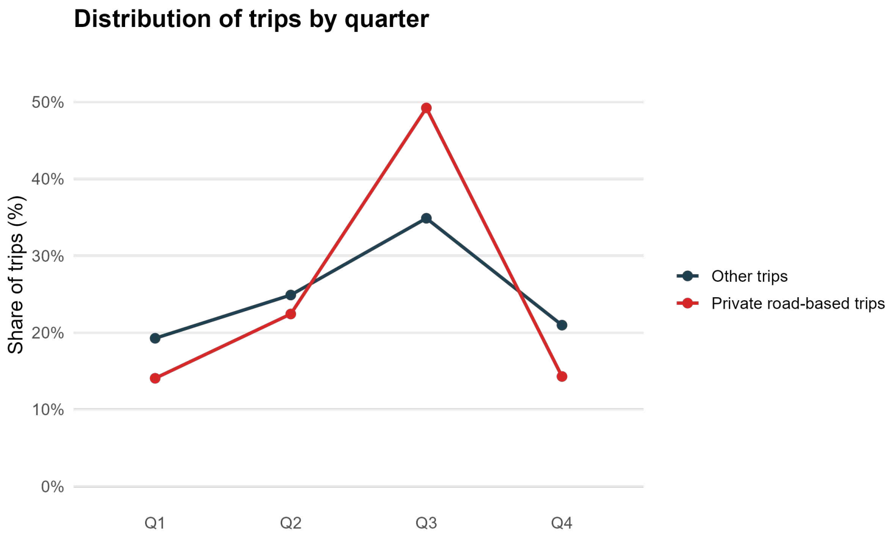


Figure 2.10: Seasonal distribution of trips: comparison between private road-based and non-road-based trips.

Private road-based trips are particularly concentrated during Q3, representing a substantially larger share of journeys compared to the remaining quarters. By contrast, non-road-based trips appear more evenly distributed throughout the year.

Several factors may contribute to this seasonal concentration. First, alternative transport modes – particularly air travel – often become significantly more expensive during peak summer months, potentially increasing the relative attractiveness of private road transport. In addition, the summer period coincides with school holidays, when family travel becomes more frequent. In this context, private vehicles may represent a more flexible and economically convenient solution for households travelling with children, as somewhat already emerged in the age analysis

More generally, the evidence suggests that seasonality plays an important role in shaping tourism mobility choices, particularly in relation to the adoption of private road-based transport during peak travel periods.

2.3.10 Destination

The geographical distribution of destinations provides an additional perspective on travel behaviour and allows for a comparison between domestic and international trips. To ensure a clear and interpretable representation, Italian regions are analysed individually, while foreign destinations are grouped into two broader categories: Europe and Other. The category Europe includes all European countries, whereas Other comprises destinations located outside both Italy and Europe.

Figure 2.11 presents the weighted distribution of trips by destination, distinguishing between road-based trips and other trips. The figure highlights substantial differences between domestic and international travel patterns.

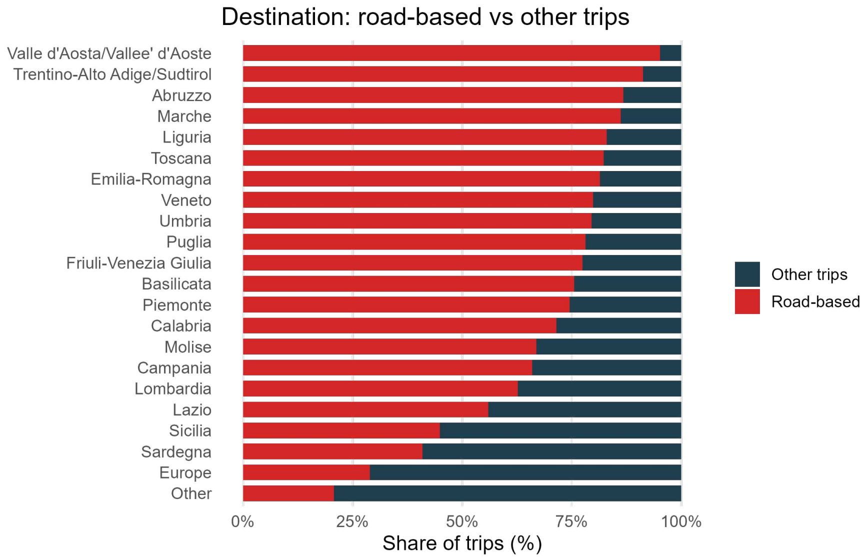


Figure 2.11: Distribution of destinations by trip type. The figure reports the share of private road-based and non-road-based trips across destination regions

Trips directed to foreign destinations appear to be predominantly classified as other trips, reflecting a lower reliance on private road transport. This pattern is consistent with the greater use of alternative modes of transportation, such as air travel, for international journeys. By contrast, trips within Italian regions show a higher share of road-based trips, suggesting that private transport plays a more central role in domestic tourism.

2.4 Limitations of the data

Despite the richness and representativeness of the ISTAT "Trips and Holidays" survey, some limitations should be acknowledged when interpreting the results of this study.

First, the data are based on self-reported information collected through household surveys. As with most survey-based datasets, responses may be affected by recall bias or misreporting, particularly for variables related to trip characteristics such as duration, expenditure, or purpose.

Second, the survey has a cross-sectional structure, meaning that individuals are observed at specific points in time rather than being followed longitudinally. As a result, the analysis does not allow for tracking behavioural changes at the individual level over time, but only for comparing aggregate patterns across years.

Another limitation regards the availability of information for specific periods. In particular, data collection was disrupted during April 2020 due to the COVID-19 lockdown, and it wasn't possible to impute missing observations with sufficient reliability. Although the scarcely presence of travel during that period is consistent with real-world conditions, this gap should be considered when interpreting temporal trends.

Finally, estimates referring to small population subgroups or relatively rare travel behaviours should be interpreted with caution due to their higher variability.

Even with these limitations, the dataset offers a comprehensive and reliable basis for analysing tourism behaviour in Italy.

2.5 Summary of descriptive evidence

The descriptive analysis shows important patterns in tourism mobility behaviour and provides useful insights for the empirical analysis developed in the following chapters.

1. First, private road-based mobility are the dominant form of tourism transport within the dataset, accounting for the majority of observed trips. This pattern is consistent with the structure of Italian tourism demand, which is largely characterised by domestic and short-to-medium distance travel. From a modelling perspective, the predominance of private road-based trips also implies a moderate imbalance in the dependent variable, an aspect that should be considered when evaluating classification performance and prediction errors across models.

2. Second, the descriptive evidence suggests the existence of heterogeneous mobility behaviours across socio-demographic groups and travel characteristics. Differences emerge across age groups, educational levels, seasonality patterns, and geographical areas, indicating that transport mode choice is associated with multiple dimensions of tourism behaviour. At the same time, these descriptive patterns should be interpreted cautiously, as they may reflect the combined influence of several socio-demographic and trip-related factors that require a multivariate analytical framework.

Overall, the descriptive evidence provides preliminary indications of systematic differences between private road-based and non-road-based mobility patterns, thereby motivating the use of predictive models to formally analyse the determinants associated with transport mode choice.

The next chapter introduces the methodological framework adopted in the study. In particular, it presents the empirical strategy, the statistical models, and the machine learning techniques employed to analyse the factors associated with private road-based tourism mobility.

Chapter 3

Empirical strategy

This chapter presents the empirical strategy used to analyse the determinants of private road-based tourism in Italy. Building on the descriptive overview of the dataset provided in Chapter 2, the focus now shifts from the general characterization of travel behaviour to a systematic investigation of the factors associated with road-based trip participation.

3.1 Problem definition

The empirical analysis aims to identify the socio-demographic and trip-related characteristics that influence the probability that a journey is undertaken using a private or rental vehicle. In this context, road-based trips are defined as trips carried out using private (or rental) road transport modes, while all remaining journeys are classified as non-road-based.

The modelling framework compares interpretable statistical approaches with more flexible machine-learning methods in order to capture both linear relationships and more complex patterns in the data. This combined approach makes it possible not only to evaluate predictive performance, but also to explore the behavioural mechanisms underlying travel choices and to identify the variables that contribute most strongly to the prediction of road-based tourism.

3.1.1 Target Variable

The objective of the empirical analysis is to identify the factors associated with private road-based participation. In order to address this question, the modelling framework is formulated as a binary classification problem in which the dependent variable indicates whether a given journey is undertaken using private or rental road-based transport.

The target variable used in the analysis is therefore a binary indicator identifying private road-based journeys. As described in Chapter 2, the classification is based on the main transport mode reported for each trip in the ISTAT *Viaggi e Vacanze* microdata.

More specifically, road-based trips include journeys undertaken using private cars, rental cars, campers or motorhomes, and motorcycles or scooters. These transport modes share a common feature: they rely on privately controlled road vehicles that allow travellers a high degree of autonomy in route choice, travel timing, and stop flexibility. In contrast, trips undertaken using alternative transport modes – such as trains, airplanes, ships, or organised coach services – are classified as non-road-based trips.

Based on this classification, the value of the target variable is:

$$Y_i = \begin{cases} 1 & \text{if trip } i \text{ is classified as a road-based} \\ 0 & \text{if trip } i \text{ is classified as a not road-based} \end{cases}$$

With this formulation, the applications of classification algorithms is particularly appropriated. The models developed in the following sections therefore aim not only to predict the probability of road-based participation, but also to identify and interpret the socio-demographic and trip-related key factors that drive this choice, using a range of explanatory variables derived from the ISTAT dataset.

Traveller Characteristics

The first group of explanatory variables captures the socio-demographic profile of the individual undertaking the trip. These characteristics are expected to play a key role in shaping travel behaviour and, in particular, in influencing the propensity to undertake a road-based trip.

The analysis includes the following variables:

- **Gender**: binary categorical variable distinguishing between male and female travellers. It captures differences in travel behaviour related to mobility patterns, travel purposes, and household responsibilities.

- **Age group**: ordinal categorical variable grouping individuals into ordered age classes representing different stages of the life cycle. It may influence both travel preferences and transport choices.

- **Marital status**: provides information on the family structure of the individual, which may affect travel organisation, group composition, and preferences for flexibility.
- **Education level**: is often associated with socio-economic status, travel experience, and preferences for specific types of tourism.
- **Employment status**: provide information on an individual's employment situation, which may influence time availability, income, and travel constraints.
- **Professional position**: includes more detailed occupatioinal categories, which may reflect differences in socio-economic conditions and lifestyles.
- **Region of residence**: provides detailed geographical information, allowing the model to capture local heterogeneity in transport availability, car dependency, and proximity to tourist destinations.

Overall, these variables allow for a detailed profiling of individuals and enable the identification of systematic differences in road-based trip behaviour across population groups.

Trip Characteristics

The second group of variables captures the characteristics of the trip itself. These attributes are directly related to how the journey is organised and are expected to play a central role in shaping transport mode choice, particularly the decision to undertake a road-based trip.

In this analysis the main trip-related variables included are:

- **Trip duration**: measured in days and grouped into categories, this variable indicates the leght of the trip. Longer trips may require different planning and may be more compatible with flexible transport modes such as private vehicles.
- **Seasonality (trimester)**: indicates the period of the year in which the trip takes place, capturing seasonal patterns related to weather conditions, holiday periods, and tourism demand.
- **Year**: accounts for time dynamics and structural changes in travel behaviour, including potential shifts related to post-Covid tourism patterns.

- **Type of accommodation**: includes different forms of lodging (e.g., hotels, rented apartments, campsites, or staying with relatives/friends). This variable captures the degree of flexibility and independence of the trip, which may be closely related to road-based trip behaviour.
- **Destination**: an additional variable identifying the geographical destination of the trip was initially considered in the analysis, distinguishing between domestic and international destinations as well as broader destination areas. However, this variable was ultimately excluded from the final models because it introduced an excessively deterministic component into the classification task. In particular, some destination categories were almost mechanically associated with the absence of road-based trip behaviour (e.g., long-haul intercontinental destinations), artificially inflating predictive performance and reducing the behavioural interpretability of the results. Excluding this variable allows the analysis to focus more clearly on the socio-demographic and travel-related characteristics associated with road-based tourism.

These variables capture the structural and contextual dimensions of the trip, complementing the socio-demographic characteristics of travellers. Their inclusion allows the model to identify how different types of journeys are associated with a higher or lower propensity to be undertaken as road-based trips.

Together with traveller characteristics, these variables constitute the set of explanatory features used in the modelling framework. A detailed summary of them, including their statistical classification and category labels, is reported in Appendix A.

Covariate preprocessing

The explanatory variables included in the analysis comprised a mixture of nominal categorical variables, ordinal categorical variables, and grouped quantitative variables. Missing values for travel purpose and professional position were retained in the analysis by introducing a dedicated *Missing* category, thereby avoiding the exclusion of observations from the sample.

Particular attention was devoted to the treatment of variables originally measured on a continuous scale but available in grouped form within the survey. Age and trip duration were therefore classified as grouped quantitative variables. Although these variables were represented through intervals rather than exact numerical values, their categories preserve a meaningful quantitative ordering and reflect increasing levels of the underlying characteristic. Consequently, they were modelled as ordered factors rather than purely nominal categories.

Education level and seasonality were treated as ordinal categorical variables. Both variables possess a natural ranking, but the distances between adjacent categories cannot be assumed to be equal. Representing these variables through numerical scores would therefore impose arbitrary spacing and potentially restrictive functional assumptions. Treating them as ordered factors preserves their ranking while allowing the models to capture non-linear relationships with the probability of undertaking a private road-based trip.

The treatment of covariates depended on the modelling approach adopted. Factor variables were automatically transformed into indicator variables through R's internal coding mechanism for the logistic regresson model. Classification Trees and Random Forests instead handled factor variables directly without needing additional transformations. For the XGBoost model, all factor variables were converted into a sparse matrix of binary indicators through one-hot encoding using the `sparse.model.matrix()` procedure, allowing the algorithm to process categorical information efficiently.

No standardisation was applied to the covariates. Standardisation is generally required when continuous predictors measured on different scales are included in the analysis, particularly for algorithms that are sensitive to the magnitude of the input variables. In the present study, the explanatory variables consisted exclusively of factors representing nominal, ordinal, or grouped quantitative categories. Furthermore, the tree-based algorithms employed in the analysis are inherently insensitive to monotonic transformations of the input variables, making variable standardisation unnecessary within the adopted modelling framework.

3.1.2 Modelling approach

The analysis is formulated as a supervised binary classification problem, where the outcome variable indicates whether a trip is classified as a road-based trip.

Let Y_i denote the transport mode choice for trip i, and let X_i represent a vector of explanatory variables capturing both traveller characteristics and trip attributes. The modelling objective is to estimate the conditional probability:

$$P(Y_i = 1 \mid X_i) = F(X_i\beta),$$

that is, the probability that a trip is undertaken as a road-based trip given the observed features.

From a statistical learning perspective, the problem consists in approximating an unknown function $f(X_i)$ that maps the input features into a probability space.

Different modelling approaches provide alternative approximations of this function, varying in terms of flexibility, interpretability, and predictive performance.

A key aspect of the modelling strategy concerns the bias–variance trade-off. Simpler models, such as logistic regression, impose stronger functional assumptions and are less prone to overfitting, but may fail to capture complex relationships in the data. More flexible machine-learning models, such as tree-based methods, are able to model non-linearities and interactions but may exhibit higher variance, making predictions more sensitive to fluctuations in the training data and increasing the risk of overfitting if the models are not properly regularised.

To explore this trade-off, the analysis adopts a comparative approach, implementing multiple models with increasing levels of complexity. Model performance is evaluated on a validation set using standard classification metrics, like accuracy and ROC-AUC, with the scope of identifying the best balance between predictive accuracy and interpretability. From an optimisation perspective, each model is trained by minimising a loss function computed on the training data.

This approach allows not only to improve predictive performance, but also to provide a more comprehensive understanding of the factors associated with road-based trip behaviour.

3.2 Classification methods

Within this framework, several alternative specifications can be considered, differing in terms of their underlying assumptions, flexibility, and predictive capabilities.

The discussion begins with binary regression models, which provide a statistically grounded and highly interpretable benchmark. It then considers more flexible approaches, including tree-based methods and ensemble techniques such as random forests and gradient boosting algorithms, which can capture complex non-linear relationships and interactions among variables.

This progression allows for a systematic comparison between models with different levels of complexity, highlighting the trade-off between interpretability and predictive performance.

3.2.1 Binary Linear Regression Models

Binary linear regression models provide a natural framework for analysing outcomes that take only two possible values. The aim of these models is therefore to estimate the conditional probability that a given journey is undertaken as a road-based trip, given linear combination of observable characteristics.

Let $Y_i \in \{0, 1\}$ denote the binary outcome associated with trip i, where $Y_i = 1$ if the trip is classified as a road-based trip and $Y_i = 0$ otherwise. Let $X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})'$ be the vector of explanatory variables. In general terms, binary linear regression models specify the relationship between the conditional expectation of Y_i and the covariates as:

$$P(Y_i = 1 \mid X_i) = F(X_i'\beta),$$

where $F(\cdot)$ is a cumulative distribution function mapping real values into the unit interval, and β is the vector of parameters to be estimated.

This formulation ensures that predicted probabilities remain between 0 and 1, overcoming the limitations of linear probability models. In this context binary regression models are notably useful because they allow the analysis to focus directly on the probability of observing a road-based trip, and in the mean time providing a formal framework to assess how individual and trip-level characteristics are associated with transport mode choice.

In this study, binary linear regression is used as the first family of models within the broader classification framework. These models are especially relevant because they provide a statistically grounded and relatively interpretable benchmark. Compared with more flexible machine learning methods, binary regression models impose a more structured functional form, which may reduce flexibility but aids interpretation and inference.

Logistic Regression

The most widespread binary regression model is logistic regression (see Hosmer et al., 2013 for a comprehensive discussion of binary response models). In this framework, the conditional probability of observing a road-based trip is modelled through the logistic cumulative distribution function:

$$P(Y_i = 1 \mid X_i) = \Lambda(X_i'\beta) = \frac{1}{1 + e^{-X_i'\beta}}.$$

Equivalently, the model can be expressed in terms of log-odds:

$$\log\left(\frac{P(Y_i = 1 \mid X_i)}{1 - P(Y_i = 1 \mid X_i)}\right) = X_i'\beta. \tag{3.1}$$

The logistic function maps any real-valued input into the interval $(0, 1)$, producing an S-shaped curve that allows unit-interval outputs between the explanatory variables and the predicted probability of the outcome.

Figure 3.1 illustrates the shape of the logistic function and its role in mapping binary outcomes into predicted probabilities.

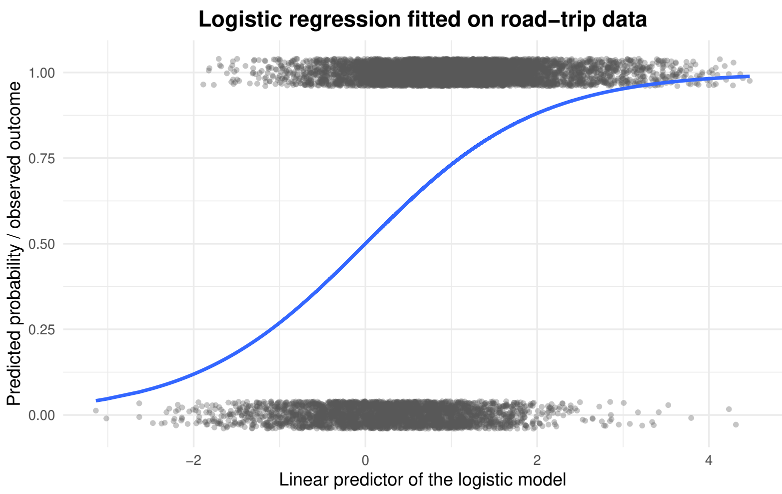


Figure 3.1: Illustrative example of a logistic regression model. Observations are binary (0 or 1), while the fitted curve represents the estimated probability of the outcome as a function of the explanatory variable.

The S-shaped form of the curve implies that the effect of the explanatory variables on the predicted probability is inherently non-linear. In particular, changes in the covariates have the largest impact around intermediate probability values, while their effect becomes progressively smaller as the probability approaches 0 or 1.

A key feature of logistic regression is that, despite this non-linearity in the probability space, the model remains linear in the log-odds as shown in Equation (3.1).

Each coefficient β_j therefore measures the change in the log-odds of undertaking a road-based trip associated with a one-unit change in the corresponding explanatory variable, holding all other variables constant. For ease of interpretation, these coefficients are often exponentiated to obtain odds ratios, which provide a more intuitive measure of the effect of each variable.

From an estimation perspective, logistic regression is typically estimated by maximum likelihood. Assuming that the binary outcome follows a Bernoulli distribution, the likelihood function is constructed as the joint probability of observing the sample data. Then the parameter vector β is obtained by maximising this likelihood, that is, by identifying the set of coefficients that best matches the predicted probabilities with the observed outcomes.

Logistic regression provides a transparent framework for analysing how traveller characteristics and trip attributes are associated with the probability of undertaking a road-based trip. For this reason, it plays a central role as baseline model in this study. Given the large sample size and the predominantly categorical nature of the explanatory variables, logistic regression offers an interpretable benchmark against which more flexible machine-learning models can be compared. In particular, logistic regression allows direct inference on model parameters and predicted probabilities, facilitating the interpretation of how explanatory variables are associated with the likelihood of undertaking a road-based trip. However, its parametric structure imposes a linear relationship in the log-odds space, which may limit its ability to capture complex non-linear patterns and interactions among variables.

Logistic regression was selected among alternative binary regression model for several reasons.

First, logistic regression represents the most commonly used and established specification for binary outcomes, giving a direct and intuitive interpretation of results in terms of log-odds and odds ratios. This makes it particularly appropriate for analysing how explanatory variables are associated with the probability of undertaking a road-based trip.

Second, while alternative specifications such as the probit model (see Greene, 2018 for a presentation of probit regression in econometric applications) rely on different distributional assumptions, empirical results are typically very similar in practice. The main differences concern the scaling of coefficients rather than substantive interpretation. For this reason, logistic regression is often preferred due to its simpler interpretability and more straightforward communication of results.

Third, compared to more complex extensions of binary regression, such as models including extensive interaction terms or variable selection procedures, a standard logistic specification provides a clear and parsimonious framework and avoids unnecessary complications in interpretation. Stepwise procedures and regularisation methods extend this framework by addressing, respectively, the need for parsimony and the risk of overfitting.

Moreover, logistic regression provides a useful benchmark against which the performance of more advanced machine-learning models can be evaluated. Given the large number of observations available in the dataset and the predominance of categorical explanatory variables, the model is able to estimate stable relationships between travellers' characteristics and the probability of undertaking a road-based trip. Nevertheless, its relatively restrictive functional form may not fully account for the heterogeneity of travel behaviour, particularly when the effects of individual characteristics vary across different trip contexts.

3.2.2 Classification Trees

Classification trees (Breiman et al., 1984) represent a flexible, non-parametric alternative to binary regression models. While these latter approaches assume a specific functional form linking the covariates to the probability of the outcome (e.g., through a logistic function), classification trees adopt a fundamentally different strategy based on recursive partitioning of the feature space.

From a statistical perspective, classification trees can be interpreted as estimators of the conditional probability function $P(Y = 1 \mid X)$ based on local averages within each partition of the feature space, rather than through a global parametric specification.

In logistic regression, the conditional probability of the outcome implies a smooth and continuous relationship between the predictors and the probability of the event. In contrast, classification trees approximate this relationship through a sequence of discrete decision rules, partitioning the feature space into disjoint regions and producing a piecewise-constant representation of the conditional probability function:

$$P(Y_i = 1 \mid X_i) = \sum_{m=1}^{M} c_m \, \mathbb{1}(X_i \in R_m)$$

where R_m denotes the regions generated by the tree partitions and c_m represents the predicted probability associated with each terminal node.

The core idea of a classification tree is to divide the feature space into a set of rectangular regions by recursively splitting the data. Starting from the root node (which contains the entire dataset), the algorithm choose a variable x_j and a threshold c to perform a split of the form: $x_j \leq c$.

This rule partitions the observations into two subsets, which are thensplit again in the same manner. The process continues until a stopping criterion is reached, such as a minimum node size or a maximum tree depth.

Figure 3.2 presents a simplified example of a classification tree and illustrates the partitioning process underlying tree-based methods. Starting from the root node, the data are sequentially divided according to a series of decision rules based on explanatory variables. Each internal node represents a splitting criterion, while the terminal nodes (leaves) correspond to the final predicted class.

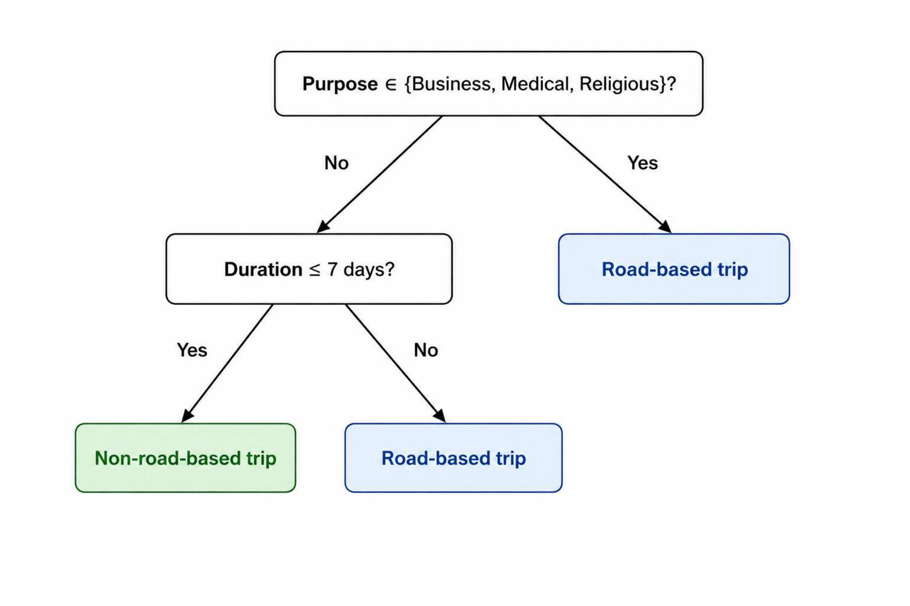


Figure 3.2: Classification tree for road-based trip prediction

At each step, the combinaton of feature and the optimal split are chosen by maximising the reduction in node impurity. For classification problems, one of the most commonly used impurity measures is the Gini index:

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2,$$

where p_k is the proportion of observations belonging to class k in node t. The algorithm selects the split that leads to the greatest decrease in impurity, thus creating increasingly homogeneous subgroups.The final nodes of the tree, known as terminal or leaf nodes, define regions of the feature space in which the predicted class is constant.

Figure 3.3 provides a further graphical illustration of this mechanism. The feature space, defined by trip duration and year, is partitioned into rectangular regions through a series of binary splits. Each region corresponds to a leaf node and is associated with a predicted class. This highlights a key characteristic of classification trees: the ability to model complex, non-linear decision boundaries through simple, interpretable rules.

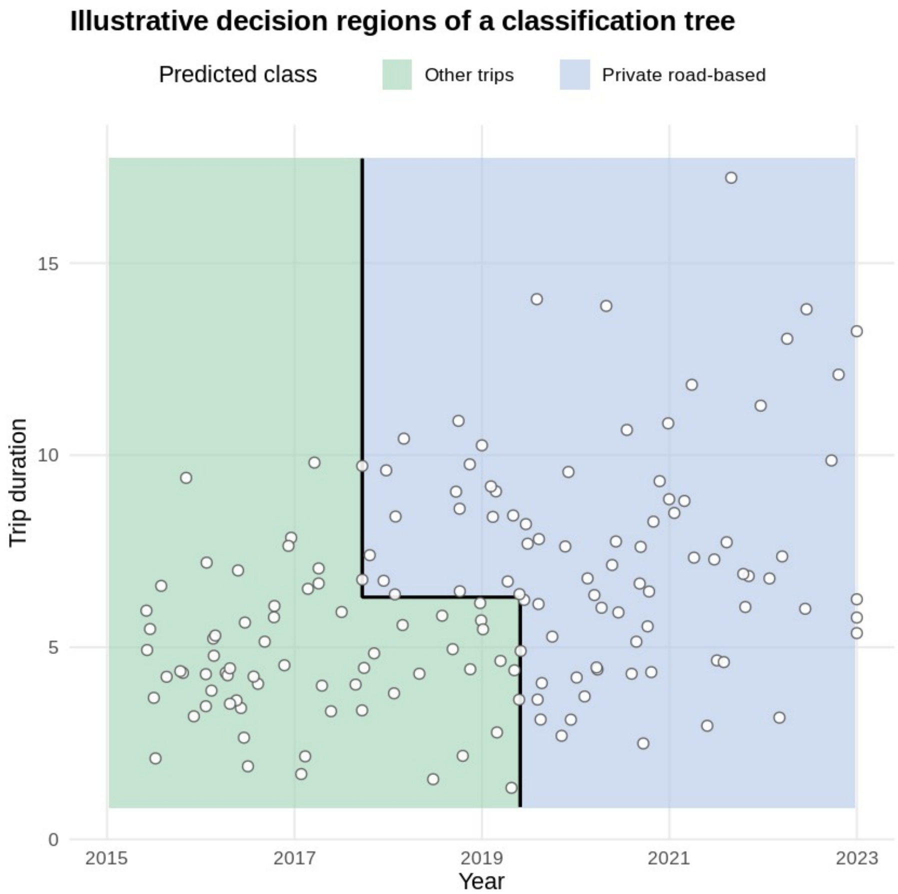


Figure 3.3: Illustrative decision regions of a classification tree. The feature space is recursively partitioned into two rectangular regions, both associated with a predicted class.

Compared to logistic regression, classification trees offer several advantages. They do not require any assumption about the functional form of the relationship between predictors and the outcome allowing naturally capture interactions and non-linearities without the need for explicit specification. This makes them particularly useful in complex settings where the true relationship is unknown. For instance,we aspect marital status influence how trip duration affects the probability of choosing a road-based mode of transport. The relationship between traveller characteristics and transport choices is therefore unlikely to be purely additive. Classification trees can capture these interaction effects automatically by recursively partitioning the feature space.

However, this flexibility comes at a cost. Classification trees are characterised by high variance, meaning that small changes in the training data can generate substantially different tree structures. As a consequence, fully grown trees often overfit the training sample, capturing noise rather than underlying behavioural patterns and resulting in poor out-of-sample performance.

To mitigate these issues, also in the case of this thesis, pruning techniques are employed. Instead of using the fully grown tree, a smaller subtree is selected by introducing a complexity parameter that penalises the number of terminal nodes. This results in a trade-off between model fit and simplicity, improving the model's out-of-sample performance.

Classification trees are particularly useful because they can identify distinct traveller profiles associated with road-based tourism through a sequence of intuitive decision rules. For example, combinations of trip duration, travel purpose, age, and destination characteristics may define groups of travellers with markedly different probabilities of undertaking a road-based trip. This hierarchical structure facilitates the exploration of behavioural segmentation patterns that may be less evident in traditional regression models.

Nevertheless, several characteristics of the dataset suggest that the predictive performance of a single classification tree may be limited. First, road-based tourism decisions are likely influenced by complex interactions among numerous socio-demographic and trip-related variables. Different combinations of traveller characteristics may lead to similar transport choices, while similar travellers may exhibit different mobility patterns depending on trip-specific attributes. As a consequence, a single tree structure may struggle to capture the full complexity of the underlying decision process, forcing heterogeneous behavioural patterns into a limited number of hierarchical rules and potentially overlooking weak but jointly relevant effects.

Second, the dataset contains several categorical variables with multiple levels, increasing the number of possible splits and potentially generating highly fragmented terminal nodes. This fragmentation may reduce the stability of the estimated tree and increase sensitivity to random fluctuations in the training sample.

Third, classification trees may struggle to represent smooth and gradually changing relationships between predictors and the outcome. Since predictions are generated through a sequence of discrete splits, complex non-linear patter ns often require a large number of partitions to be approximated accurately. Consequently, modelling smooth behavioural dynamics may lead to increasingly complex tree structures, raising the risk of overfitting and reducing generalisability.

In the context of this thesis, classification trees represent an intermediate step between parametric models and more advanced ensemble methods. They provide greater flexibility than binary regression models while maintaining a relatively high degree of interpretability. At the same time, their limitations motivate the use of ensemble techniques, such as random forests and boosting, which aim to improve predictive accuracy by aggregating multiple trees whil controlling complexity

For these reasons, while classification trees remain valuable for exploratory analysis and behavioural profiling, more advanced ensemble methods are expected to provide superior predictive performance by combining information from multiple trees and reducing variance.

3.2.3 Ensemble Methods

Ensemble methods represent a major development in predictive modelling, based on the idea that combining multiple weak or unstable learners can produce a model with superior predictive accuracy and greater robustness than any individual model taken in isolation (Breiman, 2001; Zhou, 2012).

In the context of classification problems, these approaches are particularly valuable when the relationship between the explanatory variables and the outcome is complex, non-linear, and characterised by interactions that may be difficult to capture through traditional parametric models.

By aggregating their predictions, ensemble methods can reduce the risk of overfitting, improve generalisation to unseen data, and enhance predictive stability. In particular, they are often introduced as an extension of decision trees (Dietterich, 2000).

Among the most widely used ensemble learning methods are random forests and boosting-based approaches. Random forest relies on the aggregation of many decorrelated trees built on bootstrap samples, mainly aiming to reduce variance. Boosting methods, by contrast, construct models sequentially, with each new learner focusing on the errors made by the previous ones, thereby improving predictive performance through an iterative refinement process. Gradient boosting and one of its more advanced implementation, XGBoost, belong to this latter family and are especially appreciated for their ability to achieve high predictive accuracy in complex classification tasks.

In this study, ensemble methods are included in order to complement both binary linear regression models and single-tree classifiers, allowing for a broader comparison between more interpretable approaches and more flexible machine learning algorithms.

Random Forest

Random forest is an ensemble learning method introduced by Breiman (2001), consisting of a collection of tree-structured classifiers $h_1(x), h_2(x), \ldots, h_B(x)$. Each tree is trained on a randomized version of the dataset through bootstrap sampling and random feature selection. The final prediction is obtained by aggregating the outputs of all trees.

In a classification setting, the prediction of the random forest model is obtained through majority voting:

$$\hat{f}_{RF}(x) = \text{mode}\{h_1(x), h_2(x), \ldots, h_B(x)\},$$

where B denotes the number of trees in the ensemble.

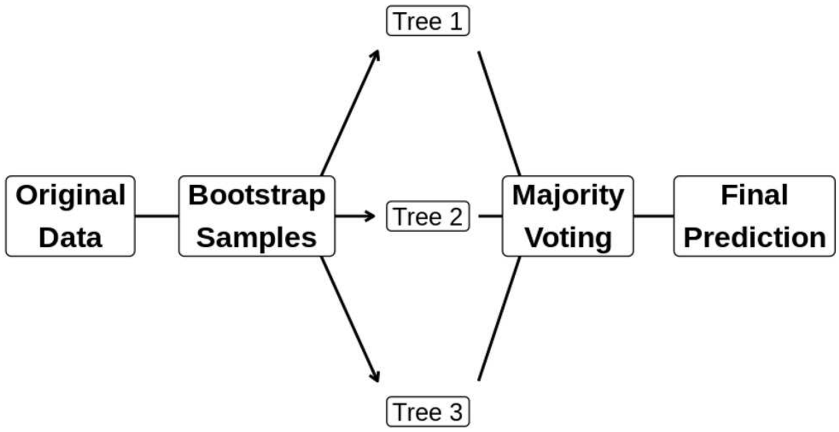


Figure 3.4: Schematic representation of the classification random forest algorithm. Each classification tree is built on a bootstrap sample of the data and uses a random subset of predictors at each split. The final prediction is obtained by aggregating the outputs of all trees.

The key idea underlying random forests is that combining multiple decision trees can substantially improve predictive performance compared to a single tree. While individual trees are flexible and capable of capturing non-linear relationships, they tend to exhibit high variance. By averaging across many trees, random forests reduce this variance and produce more stable predictions.

A fundamental theoretical result established by Breiman (2001) is that the generalization error converges as the number of trees increases. Formally, if $h(x, \theta)$ denotes a generic tree in the forest, the prediction converges to its expected value:

$$\lim_{B \to \infty} \hat{f}_{RF}(x) = \mathbb{E}_\theta[h(x, \theta)],$$

The above result implies that the variance of the ensemble decreases as the number of trees increases, leading to more stable predictions. Unlike a single classification tree, whose predictive performance may deteriorate as model complexity increases, Random Forests benefit from the addition of further trees. As the ensemble grows, prediction variance tends to decrease and performance typically stabilises, while the risk of overfitting remains limited.

Breiman showed that the predictive performance of a Random Forest depends on a balance between two key properties: the predictive strength of individual trees and the correlation among their predictions.

- *Strength*: the ability of individual trees to correctly classify observations;
- *Correlation*: the degree of similarity between the predictions produced by different trees.

In general, stronger trees improve predictive accuracy, whereas lower correlation increases the benefits of aggregation. Random Forests achieve this balance through two sources of randomness:

- **Bootstrap sampling (*bagging*)**: each tree is trained on a different sample drawn with replacement from the original dataset;
- **Random feature selection**: at each split, only a random subset of predictors is considered. The number of predictors included in this subset is controlled by the parameter *mtry*. Larger values of *mtry* generally increase the strength of individual trees by allowing more candidate predictors to be evaluated at each split. However, trees also become more similar, increasing correlation among their predictions. Conversely, lower values of *mtry* increase differences across trees, although excessively small values may weaken the predictive ability of individual trees.

These mechanisms make trees less similar to each other, reducing the correlation of their prediction errors while maintaining an acceptable level of predictive strength. The resulting reduction in variance typically outweighs the modest loss in individual tree accuracy, leading to improved predictive performance.

The effectiveness of Random Forests therefore arises from combining a large number of reasonably accurate but diverse trees, allowing the ensemble to achieve lower variance and more stable predictions than a single decision tree.

An additional advantage of Random Forests is the availability of internal evaluation tools. In particular, the out-of-bag (OOB) error provides an approximately unbiased estimate of the generalization error. Out-of-bag observations are those not included in the bootstrap sample used to train a given tree and can therefore be used as a validation set for that tree. This characteristic minimises the necessity of an independent validation dataset.

Lastly, Random Forests provide measures of variable importance helping to quantify the contribution of each predictor to model performance. A common approach is permutation importance, where the values of a predictor are randomly shuffled and the resulting decrease in predictive accuracy is measured. While these measures identify the most influential predictors, they don't show the direction or the specific nature of their effects on the outcome.

In the context of this thesis, Random Forest is particularly well suited to modelling road-based tourism behaviour. The choice of transport mode is likely influenced by a complex combination of socio-demographic characteristics and trip-related attributes, including age, travel purpose, education, professional status, trip duration, and destination characteristics. in addition, the effect of these variables is unlikely to be homogeneous among all travellers. For example, the influence of age may differ according to travel purpose, while the propensity to undertake a road-based trip may depend on specific combinations of trip duration, destination type, and family-related characteristics. Such interactions and non-linear relationships are difficult to capture through a simple parametric specification.

Random Forests are able to accommodate this complexity by automatically capturing interactions and non-linear relationships without requiring them to be specified a priori. Furthermore, the method performs particularly well in settings characterised by numerous categorical predictors and heterogeneous behavioural patterns, both of which are prominent features of the present dataset. Given the large number of observations available and the potentially complex decision process underlying transport mode choice, Random Forests are expected to provide a substantial improvement in predictive accuracy relative to simpler models.

At the same time, some limitations should be taken into account. Although Random Forests generally achieve higher predictive accuracy than individual trees and traditional regression models, they sacrifice a substantial degree of interpretability. Variable importance measures help identify the most important predictors, but they don't directly reveal the direction, magnitude, or specific nature of their effects. For this reason, additional interpretability tools are required to better understand the behavioural mechanisms underlying road-based tourism choices.

Boosting and Gradient Boosting

Boosting is an ensemble learning technique that constructs a strong predictive model by combining a sequence of weak learners, typically shallow decision trees or decision stumps (Friedman, 2001). Unlike random forests, where trees are built independently, boosting methods generate models sequentially, with each new learner focusing on the errors made by the previous ones. This sequential dependency allows boosting to incrementally enhance predictive performance by reducing bias.

Formally, let $f_m(x)$ denote the model at iteration m. Boosting methods build the final model as an additive expansion:

$$f_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x),$$

where $h_m(x)$ represents the weak learner at iteration m, and γ_m is a weight parameter that determines its contribution to the final model.

A key improvement in this context is represented by gradient boosting, which formulates the boosting procedure as a numerical optimisation problem. In particular, the method aims to minimise a differentiable loss function $L(y, f(x))$ by iteratively adding new learners that approximate the negative gradient of the loss function.

At each iteration, the pseudo-residuals are computed as:

$$r_{im} = - \left. \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right|_{f=f_{m-1}},$$

which represent the direction of steepest descent in the function space. A new base learner $h_m(x)$ is then fitted to these residuals, and the model is updated as:

$$f_m(x) = f_{m-1}(x) + \nu \gamma_m h_m(x),$$

where $\nu \in (0, 1]$ is the learning rate, controlling the contribution of each new learner and helping to prevent overfitting.

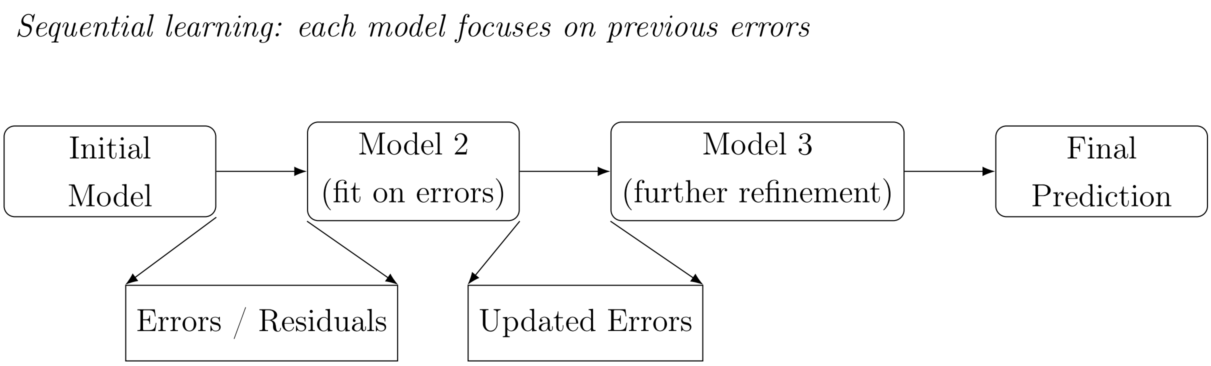


Figure 3.5: Conceptual illustration of the boosting algorithm. Models are built sequentially, with each new learner focusing on correcting the errors of the previous ones. The final prediction is obtained by combining all models in an additive manner.

The sequential nature of boosting leads to a reduction in bias, as the model progressively adapts to difficult observations. However, this also increases the risk of overfitting, particularly if the number of iterations is too large or if the learning rate isn't adequately controlled. For this reason, regularisation techniques like shrinkage (learning rate), early stopping, and tree constraints are typically used.

In many application, boosting methods achieve higher predictive accuracy compared to random forest, but at the cost of reduced interpretability and increased sensitivity to hyperparameter tuning. Because the final prediction results from the sequential aggregation of a potentially large number of trees, the contribution of individual predictors cannot be directly inferred from the model structure. Unlike logistic regression, boosting methods do not provide readily interpretable coefficients, and unlike a single decision tree, they do not yield a transparent set of decision rules. As a consequence, understanding how specific covariates influence the predicted probability of road-based tourism generally requires dedicated post-hoc interpretation techniques.

The combination of sequential learning and flexible tree structures makes boosting methods particularly effective in complex classification problems, where predictive performance is the primary objective.

3.2.4 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost, Chen and Guestrin, 2016) is an advanced implementation of gradient boosting that combines high predictive performance with explicit regularization. The model builds an additive ensemble of decision trees in a sequential manner, where each new tree is trained to improve the current model by minimizing a regularized objective function.

The prediction for observation i is given by:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i), \qquad f_m \in \mathcal{F},$$

where each f_m is a regression tree and $\mathcal{F}$ denotes the space of all possible trees.

Unlike bagging-based methods such as random forests, trees in XGBoost are built sequentially, with each new tree focusing on correcting the errors made by the current ensemble.

The learning process is driven by the minimization of a regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m),$$

where $l(\cdot)$ is a differentiable loss function and $\Omega(f_m)$ is a penalty term controlling model complexity.

At iteration t, the model is updated as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i),$$

where η is the learning rate.

Instead of directly fitting residuals, XGBoost minimizes the objective function using a second-order Taylor expansion. Let

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \qquad h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$$

denote the first- and second-order derivatives of the loss function with respect to the current prediction. The gradients, denoted by g_i, indicate how the loss function changes with respect to the current prediction, whereas the Hessians, denoted by h_i, measure the local curvature of the loss function around that prediction.

Using these quantities, the objective at iteration t can be approximated as

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

This quadratic approximation provides a simplified optimisation problem on which the next tree is fitted. Each tree is constructed by selecting splits that maximise the reduction in the approximated objective function, known as the *gain*. The gain is computed using the aggregated gradients and Hessians associated with the observations contained in the candidate child nodes.

A split is accepted only when the improvement exceeds a threshold controlled by the regularisation parameter γ. This mechanism prevents unnecessary tree growth, discourages overly complex tree structures, and contributes to reducing overfitting.

From an optimisation perspective, the learning process can be viewed as a sequence of updates. At each boosting iteration, the original loss function is locally approximated by a quadratic function around the current prediction, and a new tree is fitted to identify the update that most effectively reduces the objective function. By repeatedly fitting trees to these local approximations, the ensemble progressively moves towards regions associated with lower values of the loss function, as illustrated in Figure 3.6. This regularised optimisation procedure is one of the key factors underlying the strong predictive performance and robustness of XGBoost.

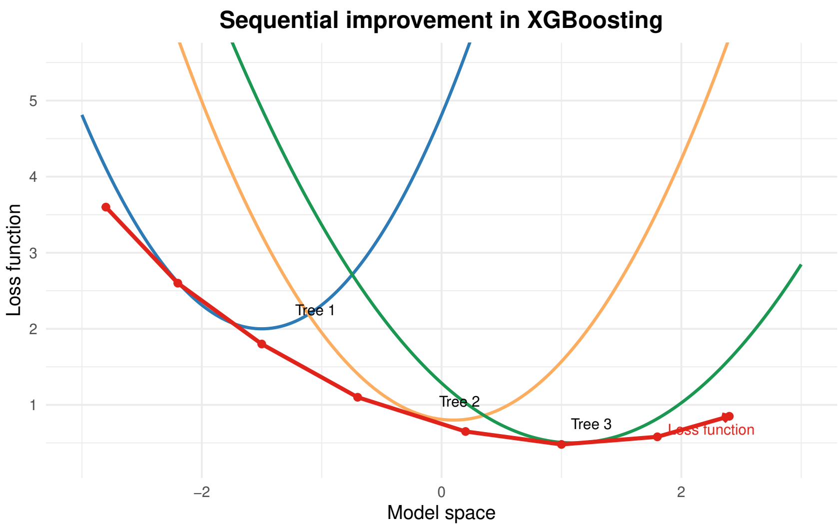


Figure 3.6: Conceptual illustration of the XGBoost optimisation process. At each iteration, the loss function is locally approximated by a quadratic function based on first- and second-order derivatives. A new tree is then fitted to this approximation, progressively reducing the value of the objective function.

Regularization and overfitting control

Due to its sequential boosting structure and high flexibility, XGBoost may be prone to overfitting if model complexity is not properly controlled. For this reason, the algorithm incorporates several regularization mechanisms aimed at improving generalization performance and reducing excessive adaptation to the training data.

The main strategies include:

- **Explicit regularization (γ and λ):** XGBoost incorporates a regularization term that directly penalises model complexity. A common specification is

  $$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.2}$$

  where T denotes the number of terminal leaves, w_j represents the prediction score associated with leaf j, and γ and λ are regularization parameters controlling model complexity.

  The parameter γ imposes a cost for each additional leaf, favouring simpler tree structures and limiting unnecessary splits. As a result, candidate splits are accepted only when they provide a sufficient improvement in model fit to justify the increase in complexity. The parameter λ penalises large leaf weights, reducing the influence of individual terminal nodes and preventing overly aggressive predictions. Together, these penalties improve generalisation performance and reduce the risk of overfitting.

- **Shrinkage (learning rate, η)**: each newly added tree contributes only partially to the final prediction, slowing the learning process and improving model stability;

- **Tree depth limitation**: restricting the maximum depth of trees prevents the model from learning overly specific patterns and complex interactions that may not generalize well out of sample;

- **Subsampling of observations**: at each iteration, only a random subset of observations is used to train the new tree, introducing additional randomness into the learning process and helping to reduce variance and improve generalisation performance;

- **Subsampling of features**: only a subset of predictors is considered when constructing each tree, increasing model diversity and reducing the risk that predictions become excessively dependent on a small number of dominant variables.

These mechanisms allow XGBoost to achieve a balance between predictive accuracy and model complexity, making it particularly effective in high-dimensional and non-linear classification problems.

Strengths and limitations

XGBoost is widely considered as one of the most effective machine learning algorithms for structured and tabular data. Its strong predictive performance stems from the combination of sequential learning, regularisation mechanisms, and efficient optimisation procedures. By iteratively focusing on the errors made by the current ensemble, the algorithm is able to progressively reduce bias while maintaining good generalisation performance through complexity control.

Another important advantage is its ability to automatically capture complex interactions and non-linear relationships among predictors without requiring their explicit specification by the researcher. This feature is particularly valuable in settings where the underlying decision process is influenced by multiple interconnected factors and where simple linear assumptions may be overly restrictive.

In the context of this thesis, these characteristics make XGBoost particularly attractive for modelling road-based tourism behaviour. The choice of transport mode is likely determined by a combination of socio-demographic characteristics and trip-related attributes whose effects may vary across traveller profiles. For example, the influence of age may depend on travel purpose, while the effect of trip duration may differ according to destination characteristics and family-related travel behaviour. Such heterogeneous and potentially non-linear patterns are difficult to represent through traditional parametric models but can be learned automatically by XGBoost through its ensemble of sequential decision trees.

Furthermore, the large number of observations and the presence of numerous categorical predictors allow the algorithm to exploit its flexibility while limiting the instability that may arise in smaller samples. The boosting procedure can therefore identify both strong behavioural regularities and more subtle patterns associated with specific traveller segments, potentially leading to superior predictive performance.

Despite these advantages, XGBoost presents some important limitations. First, the model is considerably less interpretable than traditional statistical approaches such as logistic regression and even less transparent than a single classification tree. Since predictions are generated through the aggregation of a large number of sequential decision trees, the relationship between explanatory variables and the outcome is not directly observable. Consequently, the direction, magnitude, and behavioural interpretation of individual covariate effects cannot be inferred directly from the model structure and require dedicated explainability techniques.

Second, the algorithm relies on a relatively large number of hyperparameters, including the learning rate, tree depth, regularization parameters, and subsampling rates. Obtaining stable and reliable results therefore requires careful tuning and validation procedures.

Among the models considered in this thesis, XGBoost is the most complex, but also the one able to capture the widest range of patterns in the data. Its ability to model non-linearities, interaction effects, and heterogeneous behavioural patterns makes it particularly suitable for distinguishing road-based from non-road-based trips. This flexibility, however, is accompanied by lower interpretability and increased model complexity

3.3 Models evaluation and comparison

The evaluation and comparison of predictive models is a crucial step in empirical analysis, particularly when multiple modelling approaches are considered. In this study, different methods – including parametric models, tree-based algorithms, and ensemble techniques – are employed to analyse the determinants of road-based trip behaviour. Given their structural differences, a coherent and rigorous evaluation framework is required to ensure comparability and interpretability of results.

From a methodological perspective, model assessment involves several key dimensions. First, it is necessary to evaluate the predictive performance of each model on unseen data, in order to assess its generalisation ability and avoid overfitting. Second, it is important to consider different evaluation metrics, as no single measure is sufficient to capture all aspects of classification performance. Finally, beyond predictive accuracy, understanding the internal functioning of the models is essential, especially when dealing with complex machine learning algorithms.

In this context, the evaluation framework adopted in this study is based on three main components: (i) a structured data splitting strategy, (ii) the use of multiple performance metrics, and (iii) the application of explainable artificial intelligence techniques to interpret model behaviour.

3.3.1 Dataset splitting

To ensure a reliable evaluation of model performance, the dataset is divided into two main subsets: a training set and a test set. The training set is used to estimate the models, while the test set is reserved for out-of-sample evaluation. This approach allows for an unbiased assessment of predictive accuracy, as the test data are not used during model estimation.

The dataset split is performed at the individual level rather than at the trip level. Specifically, 75% of the individuals are randomly assigned to the training set, while the remaining 25% are assigned to the test set. Consequently, all trips associated with the same individual are included entirely in the training set or in the test set.

This procedure is adopted to avoid potential information leakage between the two subsets. Since the survey structure allows the same individual to appear multiple times across different trips or years, splitting observations independently could lead to the same individual being present in both the training and test sets. Such overlap may artificially inflate predictive performance by introducing dependence between the two subsets.

In addition, for models requiring tuning or structural optimisation – such as classification trees and ensemble methods – an internal validation procedure is implemented within the training set. Specifically, the training data are further partitioned into two subsets: one used for model estimation and the other for validation purposes. This strategy is particularly important for controlling model complexity, for instance in the pruning of classification trees or in the selection of hyperparameters in ensemble models.

Overall, this multi-level splitting strategy reduces the risk of overfitting and provides a more robust and realistic evaluation of out-of-sample predictive performance.

3.3.2 Evaluation metrics

The performance of classification models is evaluated using a set of complementary metrics, each capturing a different aspect of predictive accuracy. Relying on multiple measures is particularly important in classification problems, where the distribution of the target variable may be imbalanced and different types of errors may have different implications.

The overall classification error is first considered, defined as the proportion of incorrectly classified observations. While this metric provides a general indication of model performance, it does not distinguish between different types of errors.

For this reason, the analysis also focuses on false positive (FP) and false negative (FN) rates.

The false positive rate is particularly relevant in this context, as it captures cases in which trips that are not actual road-based trips are nevertheless classified as road-based trips by the model. These situations are especially informative from a behavioural perspective, as they highlight travel profiles that share characteristics commonly associated with road-based tourism. Consequently, false positives provide insights into the overlap between road-based trips and other forms of travel behaviour.

Precision and recall are also considered. Precision measures the proportion of trips classified as road-based that are actually road-based:

$$\text{Precision} = \frac{TP}{TP + FP},$$

whereas recall (or sensitivity) measures the proportion of actual road-based trips that are correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

In the context of this thesis, recall evaluates the ability of the model to correctly detect road-based trips, while precision assesses the reliability of positive classifications.

To account for the trade-off between precision and recall, the F1-score is also computed. This measure combines both aspects into a single indicator, providing a balanced evaluation of classification performance.

Finally, the performance of the model is evaluated using the Receiver Operating Characteristic (ROC) curve. The ROC curve represents the relationship between the true positive rate (sensitivity) and the false positive rate (1 - specificity) across different classification thresholds, providing a graphical assessment of the model's discriminative ability. To summarize this information, the Area Under the ROC Curve (AUC) is also computed. The AUC provides a threshold-independent measure of the model's ability to discriminate between classes and is particularly useful for comparing models with different structures.

3.3.3 Model interpretation

While predictive performance is a key criterion for model comparison, understanding how models generate their predictions is equally important, especially in the context of behavioural analysis. Traditional parametric models offer a high degree of interpretability, but more flexible machine learning methods – such as Random Forests and XGBoost – are often regarded as black-box models. Although these algorithms frequently achieve better predictive performance, the mechanisms underlying their predictions are not directly observable.

To address this problem, explainable artificial intelligence (XAI) techniques have been developed to improve the transparency and interpretability of complex machine-learning models. As discussed by Barredo Arrieta et al. (2020b), XAI seeks to make the decision-making process of predictive algorithms more understandable and trustworthy. Among the available methods, SHapley Additive exPlanations (SHAP), proposed by Lundberg and Lee (2017), provide a unified framework for interpreting model predictions based on concepts from cooperative game theory.

Formally, let $x = (x_1, \ldots, x_M)$ denote the observation being explained and let M be the set of all explanatory variables. For a given observation x, the SHAP value associated with predictor j, denoted by $\phi_j(x)$, is defined as the average marginal contribution of that predictor across all possible subsets of explanatory variables. Let $S \subseteq M \setminus \{j\}$ denote a subset of explanatory variables that excludes predictor j. The SHAP value is computed as:

$$\phi_j(x) = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} \Big[E[f(x) \mid x_{S \cup \{j\}}] - E[f(x) \mid x_S]\Big].$$

where $E[f(x) \mid x_S]$ denotes the expected prediction conditional on the variables contained in subset S, while $E[f(x) \mid x_{S \cup \{j\}}]$ denotes the expected prediction after including predictor j. The difference between the two quantities therefore measures the marginal contribution of predictor j to the prediction for observation x.

This formulation ensures that the contribution of each predictor is evaluated fairly by averaging its effect across all possible combinations of explanatory variables. As a result, each prediction can be decomposed into additive contributions attributable to the individual covariates, allowing complex machine learning models to be interpreted in a transparent and consistent manner.

The SHAP framework satisfies several desirable theoretical properties, including local accuracy, consistency, and missingness. These properties make sure that the decomposition of predictions into individual feature contributions remains coherent and theoretically justified across different machine learning models.

One of the key advantages of SHAP is that it allows for both global and local interpretation. The global interpretation is obtained by aggregating SHAP values between observations. In particular, the average absolute SHAP value provides a measure of variable importance, allowing the identification of the predictors that contribute most strongly to model predictions. This makes it possible to identify the main drivers of road-based tourism behaviour.

Figure 3.7 illustrates a typical SHAP summary plot. Each point represents an observation, while the horizontal position corresponds to the contribution of a predictor to the model prediction. Variables are ranked according to their average absolute SHAP value, allowing the identification of the most influential predictors. The colour gradient indicates the magnitude of the feature value, facilitating the interpretation of both variable importance and effect direction.

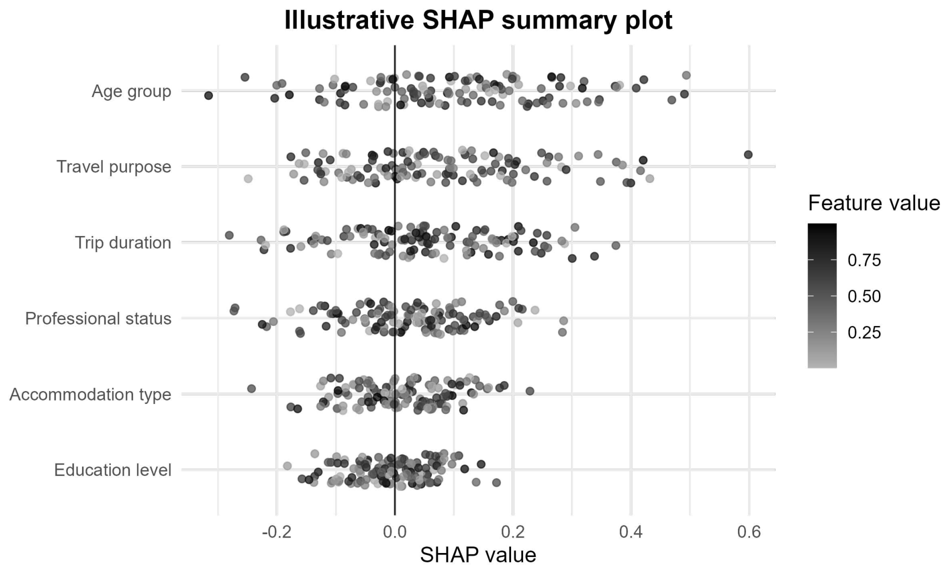


Figure 3.7: Illustrative SHAP summary plot. Variables are ranked according to their average contribution to model predictions, while individual points represent observation-specific SHAP values.

Local interpretation focuses instead on individual observations and explains why a specific prediction differs from the average prediction in the dataset. For each trip, SHAP values quantify how each explanatory variable contributes to increasing or decreasing the predicted probability of undertaking a road-based trip. This allows individual predictions to be decomposed into meaningful behavioural components.

In addition, SHAP dependence plots can be used to investigate how the effect of a given predictor varies across its range of values. These plots are particularly useful for detecting non-linear relationships and interaction effects, which are often difficult to identify using traditional parametric models.

An example of a SHAP dependence plot is reported in Figure 3.8. Such plots provide a graphical representation of how the contribution of a predictor varies across its range of values, making it possible to identify non-linear effects, threshold behaviours, and potential interactions with other explanatory variables.

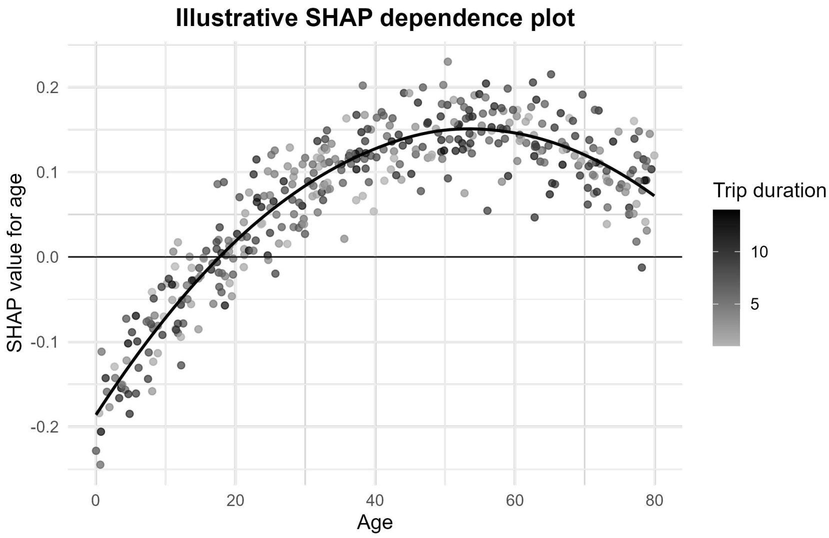


Figure 3.8: Illustrative SHAP dependence plot. The vertical axis reports the SHAP value associated with a predictor, while the horizontal axis reports the observed value of the feature.

The use of SHAP values is particularly important in the present study because the final predictive models are based on ensemble methods capable of capturing complex non-linear relationships and interactions among socio-demographic and trip-related characteristics. While these algorithms often achieve superior predictive performance, they do not provide a direct interpretation of the direction or magnitude of individual covariate effects. SHAP values address this limitation by quantifying how characteristics such as age, travel purpose, professional status, trip duration, destination characteristics, and other traveller attributes contribute to increasing or decreasing the probability of road-based tourism.

Consequently, the analysis can move beyond pure prediction and provide insights into the behavioural mechanisms underlying transport mode choices. This represents a crucial objective of the thesis, as the aim is not only to accurately classify road-based trips, but also to understand the factors that characterise road-based travellers and distinguish them from other types of tourists.

Chapter 4

Analysis and results

This chapter presents the empirical results obtained from the analysis of road-based tourism in Italy. Building on the modelling framework outlined in the previous chapter, the objective is to evaluate the performance of the proposed models and to identify the key determinants associated with the likelihood of undertaking a road-based trip. The findings provide evidence on the characteristics associated with this form of travel and offer insights into the behavioural patterns captured by the different modelling approaches.

4.1 Analytical Dataset

The empirical analysis is based on the merged dataset described in Chapter 2. Following the data cleaning, harmonisation, and preprocessing procedures, the final modelling dataset consists of 36903 observations and 14 explanatory variables. Each observation corresponds to a single trip undertaken by an individual during the period 2015–2023.

The explanatory variables include 8 nominal categorical variables, 3 ordinal categorical variables, and 2 grouped quantitative variables (age class and trip duration). In addition, the dataset contains the binary target variable indicating whether the trip was undertaken using a private road-based transport mode (1) or another transport mode (0).

Depending on the estimation method, factor variables were either handled directly or internally transformed into dummy variables, yielding a design matrix whose columns represent the explanatory features used for prediction.

No standardisation was applied to the explanatory variables. Since the predictors are represented as categorical factors or grouped quantitative variables rather than continuous numerical measurements, differences in measurement scales do not arise.

Moreover, the tree-based algorithms used in this study are inherently insensitive to monotonic transformations of the predictors, making standardisation unnecessary.

The dataset was randomly partitioned into a training set (25832 observations, 70%) and a validation set (11071 observations, 30%). The training sample was used for model estimation and hyperparameter tuning, while the validation sample was reserved exclusively for out-of-sample performance assessment.

4.2 Predictive performance of the analysed models

This section evaluates the predictive performance of the analysed classification models in identifying road-based trips from other types of travel. The objective is not only to determine which model achieves the highest predictive accuracy, but also to assess how different modelling approaches behave in terms of classification errors and discriminative ability.

The comparison includes both traditional statistical models and machine learning approaches. All models are evaluated on the same test set in order to ensure comparability of the results.

The first aspect to analyse concerns the overall predictive performance of the models. Before focusing on the interpretation of the results and the behavioural implications of the analysis, it is necessary to evaluate how accurately the different approaches are able to distinguish road-based trips from other types of travel.

Different evaluation metrics are considered in order to capture complementary dimensions of model performance. Accuracy provide an overall measure of predictive capability, summarising the proportion of correctly and incorrectly classified observations.

However, relying exclusively on overall accuracy may be insufficient in behavioural classification problems, since different types of errors may have different interpretative implications. For this reason, the analysis also examines false negative and false positive rates separately. False negatives correspond to actual road-based trips incorrectly classified as other forms of travel, whereas false positives represent non-road-based trips incorrectly identified as road-based. Analysing these errors provides additional insights into the behavioural overlap between travel categories.

The F1-score is included as a synthetic measure balancing precision and recall, allowing for a more robust evaluation of classification quality. Finally, the ROC-AUC metric evaluates the discriminative ability of the models across all possible classification thresholds and is particularly useful for comparing predictive performance independently from a specific threshold choice.

Table 4.1 reports the predictive performance of the analysed models on the test set.

Table 4.1: Predictive performance of the analysed models.

Model	**Accuracy**	**FN rate**	**FP rate**	**F1-score**	**ROC-AUC**
Logistic regression	0.715	0.628	0.107	0.805	0.735
Classification tree	0.700	0.742	**0.072**	0.803	0.706
Random Forest	0.711	**0.542**	0.158	0.794	0.747
XGBoost	**0.726**	0.616	0.097	**0.813**	**0.755**

Note: FN = False Negative; FP = False Positive; ROC-AUC = Area Under the Receiver Operating Characteristic Curve. Bold values indicate the best result for each evaluation metric.

The results show relatively similar levels of predictive performance across the analysed models, although some important differences emerge in terms of discriminative ability and error structure.

Among the considered approaches, XGBoost achieves the best overall performance. In particular, it records the highest accuracy (0.726), the highest F1-score (0.813), and the highest ROC-AUC value (0.755). These results suggest that the boosting approach is particularly effective in capturing complex and potentially non-linear relationships between socio-demographic characteristics, trip attributes, and the likelihood of undertaking a road-based trip.

Logistic regression remains relatively competitive despite its more restrictive linear structure. Its predictive performance is only moderately lower than that of the ensemble methods, achieving an accuracy of 0.715 and a ROC-AUC value of 0.735. This finding indicates that a substantial part of the relationship between traveller characteristics and transport mode choice can still be captured through relatively stable and approximately logg-odds linear effects.

The classification tree represents the weakest performing approach among the analysed models. Although its false positive rate is the lowest (0.072), indicating a relatively conservative classification behaviour, it exhibits the highest false negative rate. This suggests that the model tends to classify many actual road-based trips as other forms of travel, limiting its overall predictive capacity.

Random Forest also performs well, obtaining a ROC-AUC value of 0.747, which is close to that of XGBoost. Compared to the single classification tree, the ensemble structure significantly improves the ability to correctly identify actual road-based trips, as reflected by the lower false negative rate (0.542 versus 0.742). This confirms the advantages of aggregation methods in reducing variance and improving model robustness.

An interesting aspect concerns the trade-off between false negatives and false positives across models. Random Forest substantially reduces the false negative rate, meaning that it is more effective in identifying actual road-based trips, but this improvement comes at the cost of a higher false positive rate. By contrast, the classification tree behaves more conservatively, producing fewer false positives but failing to identify a larger share of true road-based trips. XGBoost achieves the best balance between these two types of errors, which further contributes to its superior overall performance.

More generally, the relatively limited differences in predictive accuracy across models suggest that the underlying behavioural patterns associated with road-based trips are reasonably stable and identifiable across different modelling approaches. At the same time, the superior performance of ensemble methods indicates the presence of interaction effects and non-linear relationships that are not fully captured by simpler parametric models.

In addition to the summary metrics reported in Table 4.1, it is useful to analyse the ROC curves of the different models. ROC curves provide a graphical representation of the trade-off between the true positive rate and the false positive rate across all possible classification thresholds, offering a more comprehensive evaluation of discriminative performance.

Figure 4.1 compares the ROC curves obtained for the analysed models.

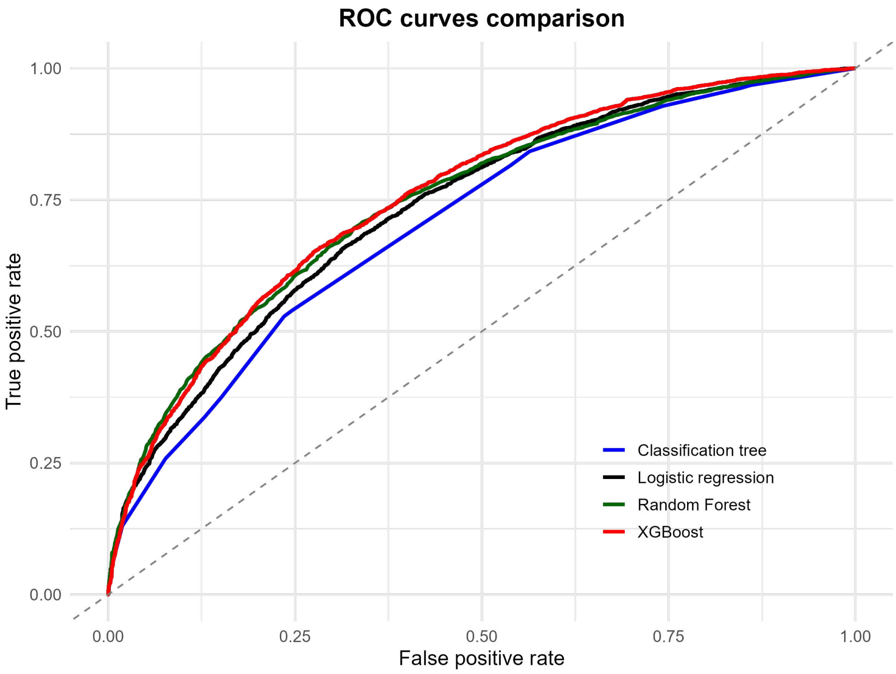


Figure 4.1: ROC curves comparison across the analysed models

The ROC curves confirm the results previously observed in Table 4.1. In particular, XGBoost and Random Forest dominate most of the ROC space, indicating a superior ability to distinguish road-based trips from other types of travel across different classification thresholds.

Among the analysed models, XGBoost achieves the best overall discriminative performance. Its curve remains consistently above the others for most threshold values, reflecting a better balance between sensitivity and specificity. Random Forest exhibits a very similar behaviour, although with slightly lower overall performance.

By contrast, the single classification tree shows the weakest predictive capacity, as its ROC curve lies below the other models in most of the domain. This result is consistent with the lower ROC-AUC value previously reported and suggests that a single tree is less capable of capturing the complexity of the underlying relationships.

Interestingly, logistic regression remains relatively competitive despite its more restrictive linear structure. Its ROC curve is close to those of the ensemble methods in several regions of the graph.

To summarise, the ROC analysis confirms that ensemble methods are better suited to capture the complex behavioural structure underlying road-based trip tourism.

4.2.1 Classification errors and model behaviour

While overall predictive performance provides useful information on the accuracy of the analysed models, it is equally important to examine the structure of the classification errors. In particular, analysing false positives and false negatives can provide additional insights into which types of trip are more difficult to classify and whether some behavioural profiles overlap across travel categories.

Particular attention is dedicated to false positives, namely trips that are classified by the model as road-based trips even though they belong to other travel categories. These observations are especially informative because they represent travel profiles that share several characteristics with road-based trips, suggesting the presence of overlapping behavioural patterns.

To further investigate these misclassifications, SHAP values are analysed for the false positive observations identified by the XGBoost model. Figure 4.2 reports the most influential features associated with these classification errors.

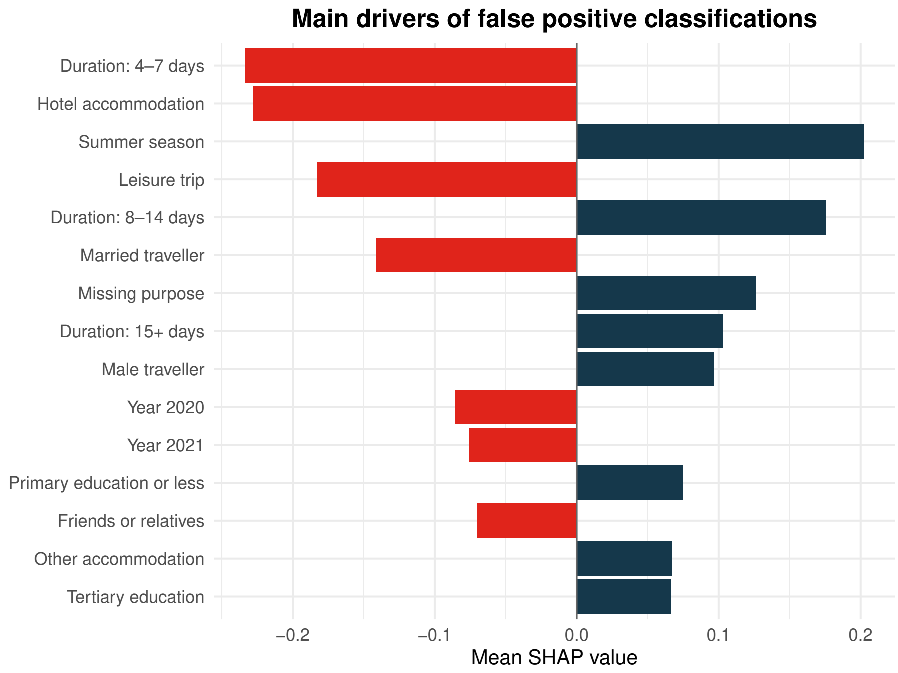


Figure 4.2: Main features contributing to false positive classifications in the XGBoost model. The figure reports the average SHAP values associated with non-road-based trip observations incorrectly classified as road-based. Higher absolute values indicate a stronger contribution of the feature to the misclassification process.

The false positive analysis indicates that the model tends to misclassify as road-based those non-road-based trips that exhibit characteristics typically associated with road-based tourism. In particular medium trip duration, hotel accommodation, summer travel periods, and leisure purpose emerge as the most influential factors driving these incorrect predictions.

These findings suggest that some travel-related characteristics are very strongly associated with road-based tourism. Consequently, when a non-road-based trip displays a combination of features that is commonly observed among road-based travellers, the model may incorrectly classify it as road-based. In this respect, the misclassifications are mainly driven by trip-specific attributes rather than by socio-demographic characteristics.

Socio-demographic variables such as marital status, gender, and education also contribute to some false positive classifications, although their influence is considerably smaller. Overall, the results indicate that the model is more likely to make classification errors when non-road-based trips resemble road-based tourism in terms of travel organisation, duration, accommodation choice, and seasonal patterns.

To provide a more detailed understanding of the mechanisms underlying false positive classifications, a representative misclassified observation is examined. Table 4.2 reports both the characteristics of the selected observation and the corresponding prediction generated by the XGBoost model. Although the trip belongs to the non-road-based category, the model assigned a probability of 0.659 to the road-based class. Since this value exceeds the classification threshold of 0.5, the observation was classified as road-based, resulting in a false positive prediction.

Table 4.2: Characteristics and prediction outcome of the selected false positive observation

Variable	**Value**
Observed class	Non-road-based
Predicted probability	0.659
Predicted class	Road-based
Year	2019
Gender	Female
Age class	45–54 years
Marital status	Married
Region of residence	Liguria
Education	Tertiary education
Professional position	Manager / Executive / Employee
Season	April–June
Trip duration	4–7 days
Accommodation type	Camping
Trip purpose	Leisure

To understand the factors that led to this misclassification, Figure 4.3 presents the local SHAP explanation for the selected observation. The decomposition illustrates how the individual contributions of the explanatory variables combine to produce the final prediction.

The explanation reveals that leisure purpose, marital status and accommodation-related characteristics provide the strongest positive contributions towards the road-based prediction. These factors are commonly associated with traveller profiles that are more likely to undertake road-based trips and therefore increase the predicted probability of belonging to that category.

At the same time, the 4–7 day duration category and the travel season exert substantial negative contributions, pushing the prediction towards the non-road-based class. Several other variables also contribute negatively, although their individual effects are relatively modest. It is important to note that SHAP values are computed for all encoded features; therefore, the observed contributions reflect not only the characteristics present in the selected trip but also the absence of alternative

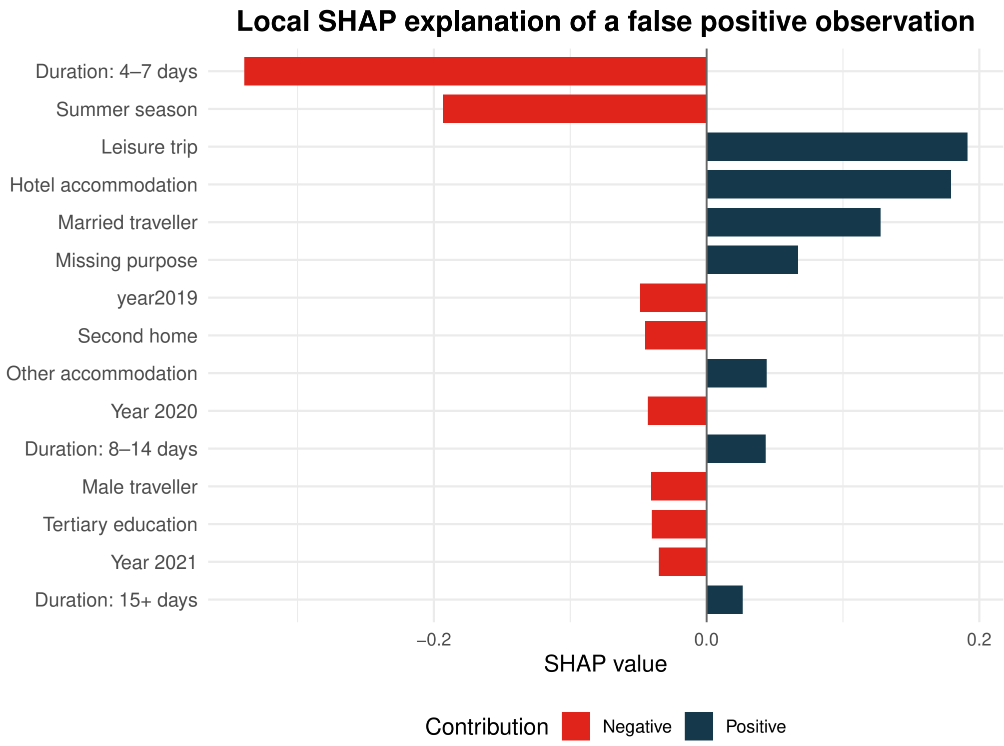


Figure 4.3: Local SHAP explanation for a representative false positive observation. Positive SHAP values increase the probability of being classified as road-based, while negative values decrease it.

categories.

Overall, the positive effects associated with leisure travel, marital status and accommodation characteristics outweigh the negative contributions. As a result, the model assigns a higher probability to the road-based class and incorrectly classifies the observation as road-based despite its true non-road-based status.

4.2.2 Models' interpretability

The comparison between traditional statistical models and machine learning approaches highlights the existence of a trade-off between predictive performance and interpretability.

Interestingly, the relatively small differences in predictive performance between logistic regression and the ensemble methods suggest that a substantial share of road-based behaviour can still be explained through stable and approximately linear relationships. At the same time, the superior performance of XGBoost indicates the presence of additional interaction effects and non-linear behavioural structures that are not fully captured by standard regression models.

For this reason, the analysis combines predictive modelling with explainable artificial intelligence techniques. In particular, the following section investigates the internal structure of the XGBoost model through SHAP values in order to identify the most relevant features and explore the behavioural patterns associated with road-based tourism.

4.3 Interpretation of model results

Predictive performance is only one aspect of the analysis. Understanding the factors associated with road-based tourism is essential to obtain meaningful behavioural insights. For this reason, the interpretation focuses on the XGBoost model, which achieved the highest predictive accuracy among all the approaches.

SHAP values are used to examine the relative importance of the explanatory variables and to investigate how different traveller and trip characteristics influence the probability of choosing a private road-based mode of transport. The analysis is conducted at both the global and local levels, allowing the identification of the main determinants of road-based tourism as well as the behavioural patterns captured by the model.

Given the good predictive performance achieved by the logistic regression model, its estimated log-odds are also examined to provide additional insights into the determinants of road-based tourism and to complement the interpretation obtained from SHAP values.

4.3.1 Global feature importance

The SHAP summary plot provides a global overview of the variables that most strongly influence the prediction of road-based tourism within the XGBoost model.

Figure 4.4 reports the SHAP summary plot obtained from the XGBoost model. Because categorical variables were encoded through dummy variables, each category enters the XGBoost model separately. The SHAP summary plot reports only the categories with the highest average absolute SHAP values, namely those contributing the most to the overall predictive performance of the model. For this reason, only some categories of the original variables appear in the figure. The variables are ranked according to their overall importance, measured through the average absolute SHAP value. Each point represents an individual observation, while the horizontal position indicates the direction and magnitude of the contribution of the variable to the prediction.

Positive SHAP values increase the predicted probability of a trip being classified as a road-based, whereas negative values reduce it. In addition, the colour represents the value assumed by the variable for each observation, allowing the relationship between feature values and model predictions to be analysed simultaneously.

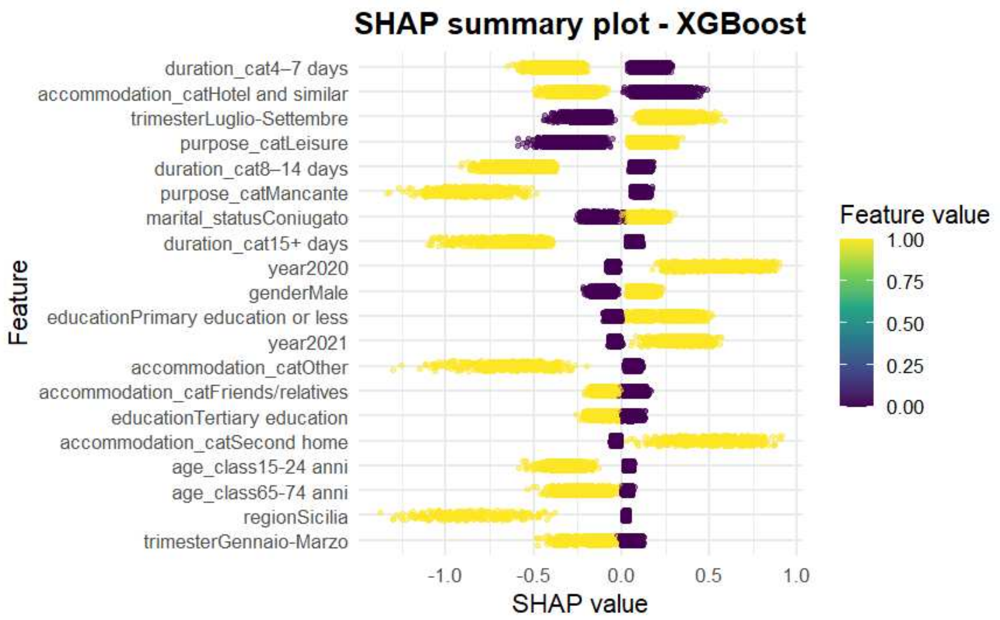


Figure 4.4: SHAP summary plot for the XGBoost model. Variables are ranked according to their average contribution to the prediction of road-based tourism.

One of the most evident results concerns the role of time-related variables. The years 2020 and 2021 show strong positive SHAP contributions, suggesting that road-based tourism became considerably more relevant during the pandemic and in its immediate aftermath. This finding is coherent with the restrictions affecting international mobility and with the increased preference for more flexible and domestic forms of travel observed during that period.

A similar pattern emerges with respect to seasonality. Trips occurring between July and September are strongly associated with higher predicted probabilities of road-based tourism, whereas trips taking place between January and March tend to contribute negatively to the prediction. The model therefore suggests that road-based tourism is predominantly linked to summer leisure mobility.

Trip duration also emerges as a central determinant. Medium-length trips, especially those lasting between 4 and 7 days, are negatively associated with road-based behaviour; also trips lasting more than 8 days tend to reduce the predicted probability of being classified as road-based trips. This suggests that road-based trips are more commonly associated with relatively short and flexible travel experiences rather than with prolonged stays.

Accommodation patterns provide additional behavioural insights. Second homes show positive contributions to the prediction, indicating that road-based tourism is often connected to relatively organised and destination-oriented forms of travel. By contrast, the category "Other" accommodation types is strongly associated with negative SHAP values. This residual category includes a heterogeneous set of accommodation arrangements, such as hostels, agritourism facilities, mountain huts, and other less common lodging solutions. The negative contribution suggests that these more diverse and less standardised accommodation choices are generally less associated with road-based tourism behaviour.

The model also captures some relevant socio-demographic and territorial differences. Travellers with primary education display positive SHAP contributions, while young individuals aged 15–24 and older travellers aged 65–74 are associated with lower probabilities of undertaking road-based trips.

A possible explanation for the positive contribution of primary education is that this category includes a substantial share of children travelling with their families. Family holidays are frequently organised using private vehicles because they offer greater flexibility, lower transportation costs per traveller, and easier management of luggage and travel schedules. Consequently, the observed effect may reflect household travel behaviour rather than the individual educational attainment of the traveller.

The negative contribution associated with individuals aged 15–24 may instead be linked to a greater reliance on alternative transport modes, such as rail or air travel, particularly for study-related, social, or international trips. Similarly, travellers aged 65–74 may face mobility constraints that reduce the attractiveness of long-distance driving, increasing the use of organised transport services or other travel arrangements. From an economic perspective, these results suggest that private road-based tourism is particularly associated with family-oriented travel patterns, where the cost advantages and flexibility of private vehicles play a central role in transport choice.

Similarly, observations associated with Sicily tend to contribute negatively to the prediction, suggesting geographical differences in mobility patterns and transport behaviour. This effect may reflect the greater relevance of air and maritime transport for an island region. From an economic perspective, the result suggests that transport choices are influenced not only by traveller characteristics but also by geographical constraints that affect the relative cost and convenience of alternative modes of transport.

Taken together, the SHAP results indicate that road-based tourism is primarily associated with short-duration summer travel and with accommodation arrangements that facilitate autonomous mobility, such as second homes. The analysis also highlights the existence of socio-demographic and territorial differences in travel behaviour. Finally, the strong positive contribution associated with the years 2020 and 2021 provides further evidence of the exceptional role played by the Covid-19 pandemic in increasing the prevalence of road-based tourism during a period characterised by mobility restrictions and heightened uncertainty.

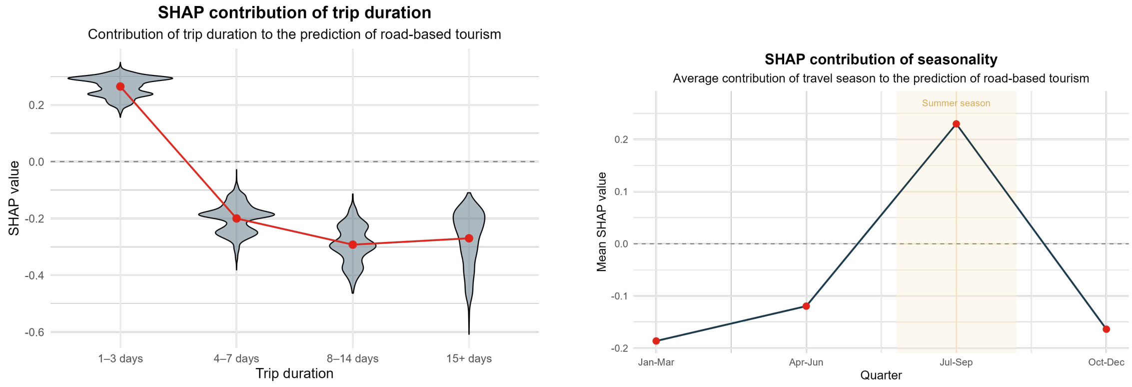
4.3.2 Behavioural patterns associated with road-based tourism

While global SHAP importance provides information on the variables that contribute most to model predictions, a deeper understanding of road-based tourism behaviour can be obtained by examining how the contribution of individual predictors varies across their categories. To this end, SHAP values are analysed for a set of key variables that emerged as particularly relevant in the previous section. Positive SHAP values indicate categories that increase the predicted likelihood of road-based tourism, whereas negative values indicate categories associated with a lower propensity for road-based travel.

The analysis focuses on four dimensions that characterise the behavioural profile of road-based tourists: trip duration, seasonality, accommodation choice, and age. Together, these variables provide insights into when road-based trips are undertaken, how they are organised, and which traveller profiles are most strongly associated with this form of tourism.

An important behavioural pattern concerns the temporal organisation of travel. In particular, both trip duration and seasonality play a central role in shaping the likelihood of road-based tourism.

Figures 4.5a and 4.5b present the average SHAP contributions associated with trip duration and seasonality.



(a) Trip duration (b) Seasonality

Figure 4.5: SHAP contributions associated with temporal travel characteristics.

The results reveal two distinct temporal patterns. First, trips lasting between one and three days exhibit a positive SHAP contribution, indicating that short-duration travel increases the predicted likelihood of road-based tourism. In contrast, all longer trip durations display negative contributions. This suggests that road-based tourism is particularly associated with short-term leisure travel and domestic mobility patterns.

A second pattern emerges with respect to seasonality. The summer quarter (July–September) is the only period associated with a clearly positive SHAP contribution, whereas all other quarters display negative values. This finding indicates that the propensity for road-based tourism is strongly concentrated during the peak summer season. The result is consistent with the central role of family travel in road-based tourism, as summer coincides with school holidays and therefore offers greater opportunities for households with children to undertake leisure trips. At the same time, the summer period is characterised by increased demand for long-distance travel, often accompanied by higher prices for alternative transport modes such as air and rail services. Under these conditions, as travelling by car allows transportation costs to be shared among multiple travellers (often making it less expensive than purchasing separate train or airline tickets for each family member), further reinforcing the association between summer travel and road-based tourism.

These findings suggest that road-based tourism is characterised by a combination of short-duration and strongly seasonal travel patterns. The concentration of positive contributions among short trips and summer travel supports the interpretation of road-based tourism as a flexible and leisure-oriented form of mobility.

Since trip duration and seasonality exhibit relatively regular and monotonic patterns and given that the logistic regression model achieved predictive performance comparable to that of the more complex machine-learning approaches, the corresponding odds ratios provide an intuitive and informative quantitative measure of the magnitude of these effects. Consistent with the SHAP analysis, longer trips are substantially less likely to be road-based than short trips. In particular, journeys lasting 15 days or more exhibit odds 65% lower than trips lasting 1–3 days. Conversely, trips undertaken during the summer quarter (July–September) display odds 62% higher than those observed during April–June, confirming the strong seasonal component of road-based tourism. Overall, the logistic regression results confirm both the direction and the magnitude of the duration and seasonal effects highlighted by the SHAP analysis. Detailed odds-ratio estimates are reported in Appendix B.

The organisational dimension of travel can be further explored by examining the contribution of different accommodation types to the prediction of road-based tourism. Figure 4.6 reports the average SHAP contribution associated with each accommodation category. Positive SHAP values indicate accommodation types that increase the predicted likelihood of road-based tourism, whereas negative values indicate categories associated with a lower propensity for road-based travel.

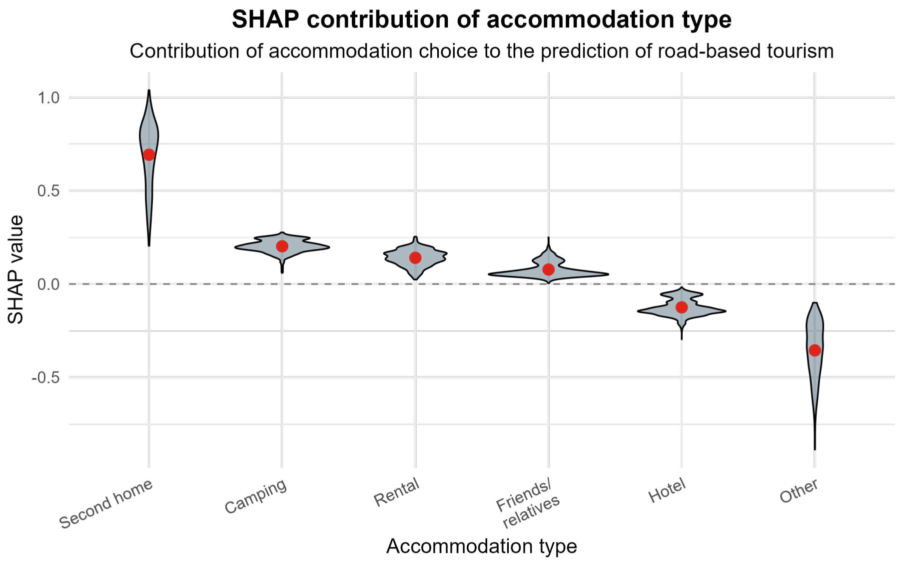


Figure 4.6: SHAP contribution of accommodation type to the prediction of road-based tourism.

A behavioural pattern related to the organisation of travel emerges from the reuslts. Second home accommodation shows by far the strongest positive SHAP contribution, indicating that travellers staying in second homes are more likely to engage in road-based tourism. Positive contributions are also observed for camping and rental accommodation, although their effect is considerably smaller.

In contrast, hotel accommodation contributes slightly negatively, while the category “Other” exhibits the strongest negative SHAP values. Accommodation with friends or relatives remains close to neutrality, suggesting a limited influence on the likelihood of undertaking a road-based trip.

These results suggest that accommodation choices play an important role in shaping road-based tourism behaviour. The strong positive contribution of second homes, together with the positive effects observed for camping and rental accommodation, indicates that road-based tourism is more frequently associated with travel arrangements that facilitate independent mobility and destination accessibility.

From an economic perspective, these types of accomodation are often linked to forms of tourism where travellers seek cost containment. In particular, second homes and camping facilities are frequently located outside major urban centres and may require private transport for convenient access. Moreover, travelling by car allows transportation costs to be shared among several travellers and facilitates the transport of luggage, equipment, or supplies, reducing the overall cost of the holiday. The combination of lower accommodation expenses and the cost advantages associated with private road transport may therefore reinforce the observed relationship between these accommodation choices and road-based tourism.

In addition to travel organisation and temporal dynamics, socio-demographic characteristics also contribute to shaping the behavioural profile associated with road-based tourism. Among these factors, age emerges as one of the most relevant predictors, highlighting substantial differences in the contribution of different demographic groups to the probability of undertaking a road-based trip.

Figure 4.7 reports the average SHAP contribution associated with each age category. Positive SHAP values indicate age groups that increase the predicted likelihood of road-based tourism, whereas negative values indicate groups associated with a lower propensity for road-based travel.

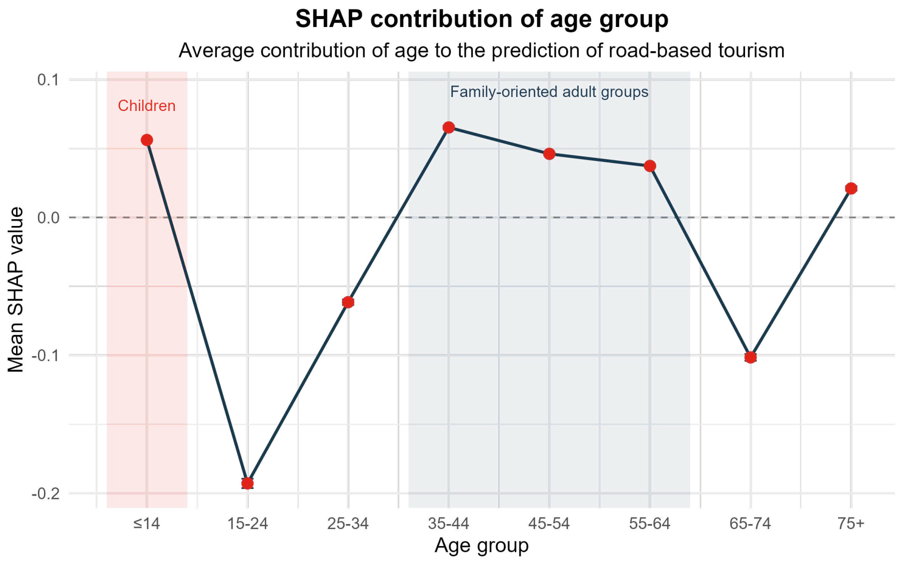


Figure 4.7: SHAP contribution of age groups to the prediction of road-based tourism. The highlighted areas identify children and family-oriented adult age groups.

The results reveal a non-linear relationship between age and road-based tourism. The strongest positive contribution is observed for children aged 14 years or younger. This finding, obviously, should not be interpreted as an individual preference for road-based travel, since children are rarely responsible for transport decisions. Rather, the result likely reflects the presence of children within travelling households and therefore captures a broader family-oriented travel pattern.

In this context, the positive SHAP contribution associated with children can be interpreted as evidence that road-based tourism is particularly common among family groups, where the flexibility, convenience, and luggage capacity offered by private vehicles make road transport especially attractive. From an economic perspective, travelling by car may also reduce transportation costs. This cost advantage may be even greater if several families travel together, allowing fuel and toll expenses to be distributed across all the travellers. Often, this results in lower costs per person than alternative transport modes such as rail or air travel.

A second cluster of positive contributions emerges among middle-aged adults, particularly between 35 and 64 years of age. These groups are the demographic categories most commonly associated with family travel organisation, greater economic stability, and higher levels of private mobility. The positive SHAP values observed for these age groups indicate that they contribute to increasing the predicted likelihood of road-based tourism.

By contrast, younger adults aged between 15 and 34 years exhibit negative SHAP contributions. This finding may reflect a greater reliance on alternative transport modes, particularly rail and air transport, which are often preferred for study-related, social, and international travel. In addition, younger individuals may have lower vehicle availability and may be more likely to travel independently rather than as part of a family group.

Also among individuals aged between 65 and 74 years, negative contributions are observed. Although many individuals in this age group are retired and generally have more flexible schedules, they prefer comfort and convenience when travelling. As a result, they may be more inclined to choose transport and tourism options that reduce the load of long-distance driving, even when these alternatives involve higher costs. For example, organised tours, cruises and package holidays often require less planning and may be perceived as more comfortable, making private road transport a less attractive option for this age group

Taken together, these patterns suggest that road-based tourism is strongly associated with family-oriented travel behaviour. The simultaneous presence of positive contributions among children and middle-aged adults points towards a household-based travel structure, where private vehicles facilitate the mobility needs of families and support flexible travel arrangements.

The logistic regression results provide further quantitative support for the age profile identified by the XGBoost model. Relative to children aged 14 years or younger, all other age groups exhibit substantially lower odds of undertaking a road-based trip. In particular, individuals aged 15–24 years and 65–74 years display odds that are 62% and 63% lower, respectively, reinforcing the interpretation that road-based tourism is strongly associated with family-oriented travel behaviour. More specifically, the odds ratio of 0.382 estimated for the 15–24 age group indicates that the odds of undertaking a road-based trip are only 38.2% of those observed for the reference category. Detailed odds-ratio estimates are reported in Appendix B.

Finally, the SHAP analysis reveals the presence of substantial territorial heterogeneity in the factors associated with road-based tourism. While the descriptive analysis presented in Chapter 2 already highlighted geographical differences in the distribution of road-based trips, the SHAP framework provides additional insights by identifying how each region contributes to the prediction of road-based tourism.

Figure 4.8 presents a geographical representation of the average regional SHAP contributions in Italy. Positive values indicate regions associated with a higher predicted likelihood of undertaking a road-based trip, whereas negative values indicate regions associated with a lower predicted likelihood.

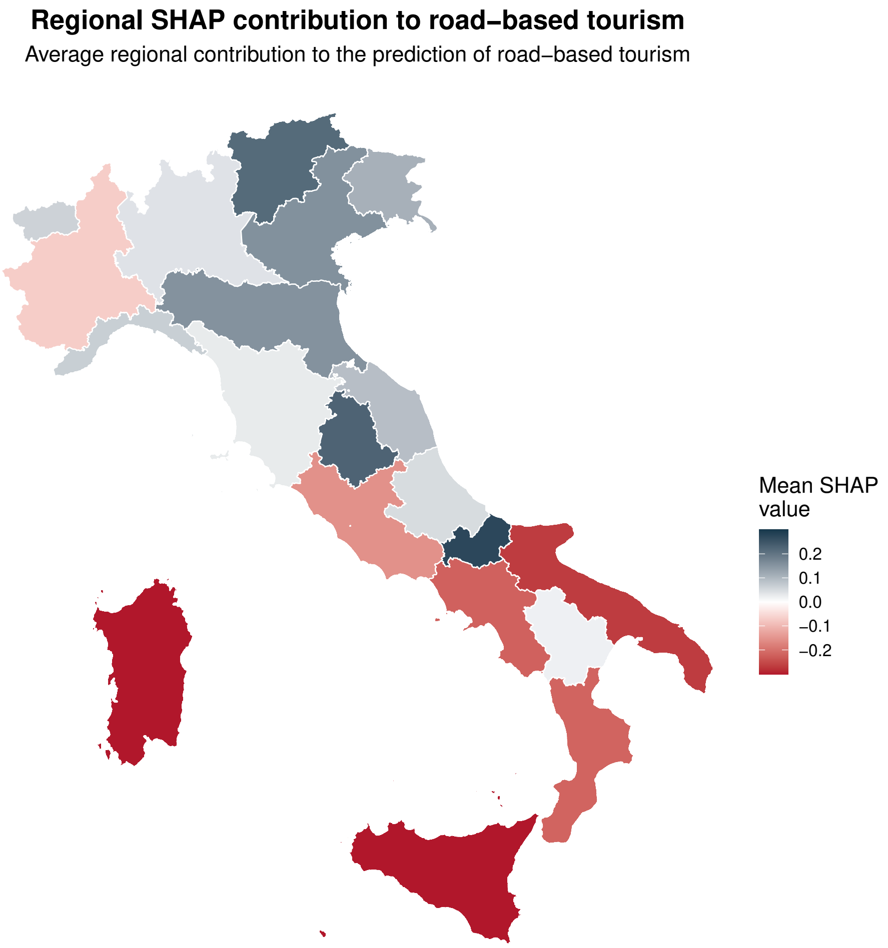


Figure 4.8: Average regional SHAP contributions to the prediction of road-based tourism.

The map reveals a clear spatial pattern. Several northern and central regions display positive SHAP contributions, indicating a higher predicted likelihood of undertaking a road-based trip. In particular, Molise, Umbria, Trentino-Alto Adige, Veneto, and Emilia-Romagna emerge as the regions with the strongest positive contributions. By contrast, the most negative contributions are concentrated in southern and insular regions, especially Sardinia and Sicily, followed by Apulia, Campania, and Calabria.

These results suggest that the propensity to undertake road-based trips is not uniformly distributed across the country. Regions connected to the mainland road network seem to be more strongly associated with road-based tourism, while insular territories exhibit substantially lower contributions. This pattern reflects broader differences in geographical accessibility, transport infrastructure, and tourism structures than a direct causal effect of regional residence itself.

The logistic regression estimates broadly confirm the geographical patterns identified by the XGBoost model. Residents of Molise, Umbria, and Trentino-Alto Adige exhibit odds of undertaking a road-based trip 39%, 38%, and 31% higher, respectively, than those of residents in Abruzzo. By contrast, the strongest negative effects are observed for Sardinia and Sicily, where the odds are 57% and 56% lower. These results provide additional quantitative evidence supporting the regional patterns highlighted by the SHAP analysis. Detailed odds-ratio estimates are reported in appendix B.

4.3.3 Non-linear effects and interaction patterns

While the previous analysis focused on the independent contribution of the main explanatory variables, road-based trip behaviour may also emerge through combinations of factors interacting in a non-linear way. In this respect, the XGBoost model allows more complex behavioural interaction structures to be explored.

The interaction heatmap provides additional evidence that road-based trip behaviour is shaped not only by isolated variables, but also by combinations of behavioural and socio-demographic characteristics interacting simultaneously. Although the absolute magnitude of the interaction effects remains moderate, several meaningful patterns emerge from the analysis.

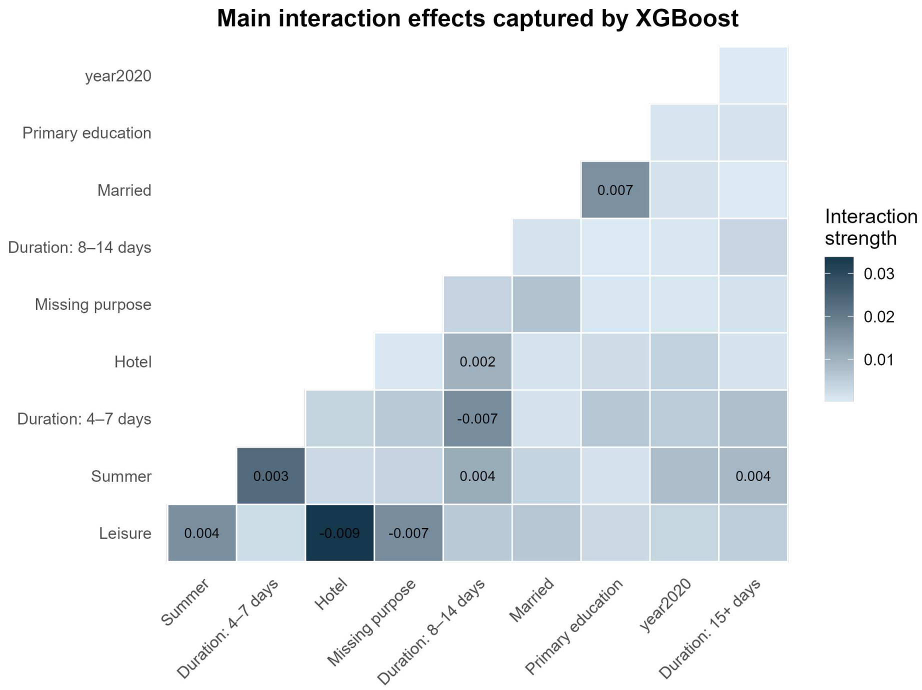


Figure 4.9: Main behavioural interaction effects captured by the XGBoost model. The diagonal is excluded in order to focus only on interactions between different explanatory variables.

The interaction heatmap highlights a limited number of meaningful interaction effects, suggesting that most of the predictive structure captured by the XGBoost model is explained by the main effects discussed previously.

The strongest interaction involves leisure travel and hotel accommodation. The negative interaction value indicates that the positive contribution associated with leisure travel is attenuated when trips involve hotel stays. This result is supported by the accommodation analysis presented earlier, which showed that hotel accommodation is less strongly associated with road-based tourism than alternative accommodation types such as second homes or camping facilities.

Additional positive interactions emerge between summer travel, leisure travel, and longer trip durations. In particular, the positive interactions between summer seasonality and longer trip durations indicate that these characteristics reinforce each other. Although longer trips are generally associated with lower probabilities of road-based tourism, their negative effect is weaker when they take place during the summer period. In the same way, the positive interaction between leisure travel and summer seasonality indicates that the effect of recreational travel becomes stronger during the peak holiday season.

The heatmap also reveals notable socio-demographic interactions. In particular, the combinations of married status with primary education and male gender shows relatively strong interaction effects. These patterns may partly reflect broader household and generational structures. Married individuals with lower educational attainment are more likely to belong to older cohorts, where private vehicle use and family-oriented travel habits tend to be more common. Similarly, the interaction between marital status and gender may capture aspects of household travel organisation that emerge only when these characteristics are considered jointly.

The interaction analysis complements the main SHAP results by showing that road-based tourism is shaped not only by the independent contribution of individual variables, but also by the way travel characteristics and socio-demographic factors combine within specific behavioural settings.

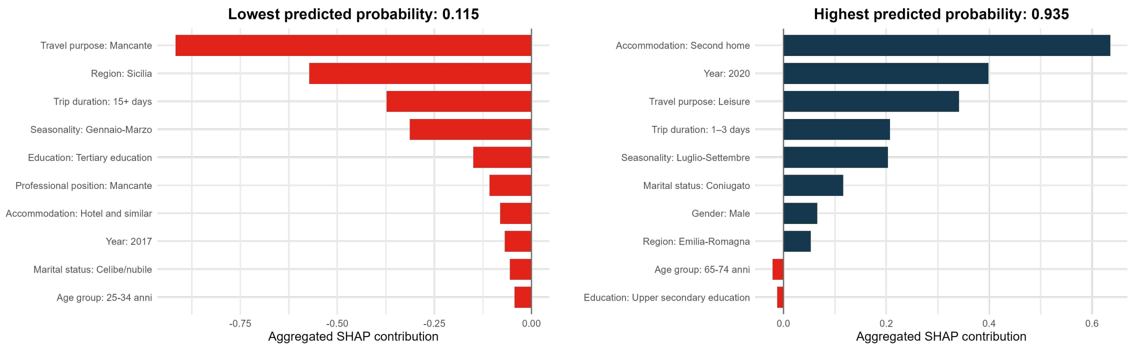
4.3.4 Behavioural profiles associated with road-based tourism

The analyses conducted in the previous sections highlighted several behavioural, socio-demographic, and territorial characteristics associated with road-based tourism. By combining the evidence emerging from global SHAP importance, feature-level SHAP contributions, and interaction effects, it is possible to outline a broader behavioural profile associated with road-based travel in Italy.

Table 4.3: Main characteristics associated with road-based tourism

Dimension	Main evidence	Behavioural interpretation
Trip duration	1–3 day trips show a positive SHAP contribution	Road-based tourism is mainly associated with short and flexible mobility patterns
Seasonality	July–September shows the strongest positive SHAP contribution	Road-based travel is closely connected to peak summer leisure mobility
Accommodation	Second homes, camping, and rental accommodation show positive SHAP contributions	These accommodation types are compatible with autonomous and self-organised travel
Age profile	Children and middle-aged adults show positive SHAP contributions	The results suggest a family-oriented travel structure based on household mobility
Territorial dimension	Mainland regions tend to show more positive SHAP contributions, while island regions show negative contributions	Geographical accessibility and integration with the road network influence road-based mobility
Interaction patterns	Leisure travel interacts with seasonality, duration, accommodation, marital status, gender, and education	Road-based tourism emerges from combinations of trip-related and socio-demographic characteristics

While the previous analyses focused on average behavioural patterns, SHAP values can also be used to explain individual predictions. Figure 4.10 presents two extreme cases: the observation with the highest predicted probability of being classified as road-based and the observation with the lowest predicted probability.



(a) Lowest predicted probability

(b) Highest predicted probability

Figure 4.10: Local SHAP explanations for the observations with the lowest and highest predicted probabilities of road-based tourism.

The highest-probability observation exhibits a predicted probability of 0.935 and is characterised by several features previously identified as favourable to road-based tourism. In particular, second-home accommodation represents the largest positive contribution, followed by leisure travel purpose, short trip duration, summer seasonality, and residence in Emilia-Romagna. The year 2020 also contributes positively, suggesting that this individual prediction reflects the post-pandemic increase in the relevance of private and flexible mobility patterns discussed earlier. These characteristics collectively reinforce the prediction and generate a highly road-oriented travel profile.

By contrast, the observation with the lowest predicted probability (0.115) is associated with a combination of characteristics that were found to be negatively related to road-based tourism. The strongest negative contributions originate from missing travel purpose, residence in Sicily, long trip duration, winter seasonality, hotel accommodation, and tertiary education. The negative contribution associated with missing travel purpose should be interpreted with caution, as it reflects unavailable information rather than a meaningful travel behaviour. Taken together, these features produce a profile that is substantially less compatible with road-based mobility.

Overall, the two local explanations provide a concrete illustration of the global SHAP findings. The high-probability case combines almost all the characteristics previously associated with road-based tourism, whereas the low-probability case displays the opposite profile. This confirms that the model does not rely on a single determinant, but combines accommodation, purpose, duration, seasonality, territorial location, and year-specific effects to generate individual predictions.

4.3.5 Interpreting complexity: machine learning versus traditional models

An important consideration emerging from the SHAP analysis concerns the extent to which the behavioural patterns identified by XGBoost differ from those that could be obtained using more traditional statistical approaches. Although the machine-learning model achieved the highest predictive performance among all models considered, many of the behavioural relationships highlighted by the SHAP analysis appear intuitive and consistent with both the descriptive evidence and the results that would likely emerge from conventional statistical modelling.

The effects identified by the SHAP analysis exhibit relatively regular and interpretable patterns. Road-based tourism is primarily associated with short-duration trips, summer travel, family-oriented age groups, and flexible accommodation arrangements. These findings suggest that a substantial part of the behavioural structure underlying road-based tourism could also be identified through more traditional approaches such as logistic regression.

At the same time, the interaction analysis reveals that travel behaviour cannot be entirely reduced to a set of independent effects. Several meaningful interactions emerge between travel purpose, seasonality, accommodation choice, trip duration, and socio-demographic characteristics. While the magnitude of these interaction effects remains smaller than that of the main SHAP contributions, they nevertheless highlight the existence of heterogeneous behavioural contexts that would be more difficult to identify and specify explicitly within a traditional modelling framework.

This finding is particularly relevant because it suggests that the superior predictive performance of machine-learning methods should not necessarily be interpreted as evidence of fundamentally different behavioural structures. Rather, the additional predictive power appears to derive from a more flexible representation of relationships that are already present in the data.

Therefore, the added value of XGBoost does not primarily lie in identifying completely different determinants of road-based tourism, but rather in its ability to capture non-linear effects and interaction patterns that cannot be easily represented within a standard logistic regression framework. Thanks to the use of SHAP values, these additional layers of complexity remain interpretable, allowing the underlying behavioural mechanisms to be explored rather than treated as a black box. Although the machine-learning model achieved the highest predictive performance, the main determinants of road-based tourism remain broadly consistent with those identified through more conventional statistical approaches. In this sense, XGBoost provides a more flexible representation of travel behaviour while preserving interpretability through SHAP-based explanations.

4.4 Road-based tourism as a resilient mobility pattern during Covid-19

The Covid-19 pandemic represented an exceptional shock for tourism mobility, affecting travel intensity, destination choices, and transport preferences. Restrictions on international travel, health concerns and uncertainty surrounding long-distance trips encouraged the adoption of more flexible and autonomous forms of mobility. In this context, road-based tourism became particularly relevant because private road transport offered greater control over travel conditions, greater compatibility with domestic and short-distance travel, and limited the contact with crowded transport environments.

To investigate the potential impact of Covid-19 on road-based tourism behaviour, the sample is partitioned into two distinct periods. The years 2015–2019 are considered the *pre-Covid* period, representing tourism behaviour before the pandemic shock, whereas the years 2020–2023 are grouped into the *post-Covid* period. The subsequent analysis compares the behavioural structure associated with road-based tourism across these two periods.

Figure 4.11 shows the evolution of road-based tourism between 2015 and 2023. The share of road-based trips was relatively stable before the pandemic, fluctuating around 61–65%. A sharp increase is observed in 2020, when the share reached almost 80%, followed by a gradual decline in 2021–2023. Although the share decreased after the pandemic peak, it remained slightly above the 2019 level. This pattern suggests that road-based tourism represented one of the most resilient forms of tourism mobility during the Covid-19 shock.

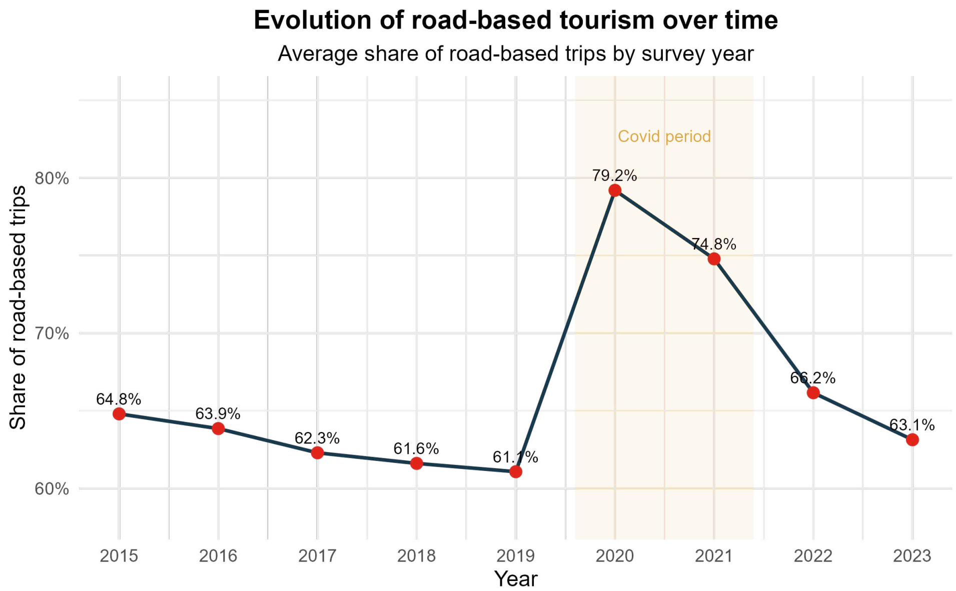


Figure 4.11: Evolution of road-based tourism over time. The figure reports the average share of road-based trips by survey year, highlighting the Covid-19 period.

Figure 4.12 compares, before and after Covid-19, the aggregated SHAP importance of the main explanatory dimensions. The figure shows that the overall ranking of predictors remained largely stable in the two periods. Trip duration and travel purpose remained the most important dimensions, followed by accommodation type and seasonality. The most visible change concerns seasonality, whose importance increased during Covid-19, suggesting a stronger concentration of road-based tourism during specific periods of the year.

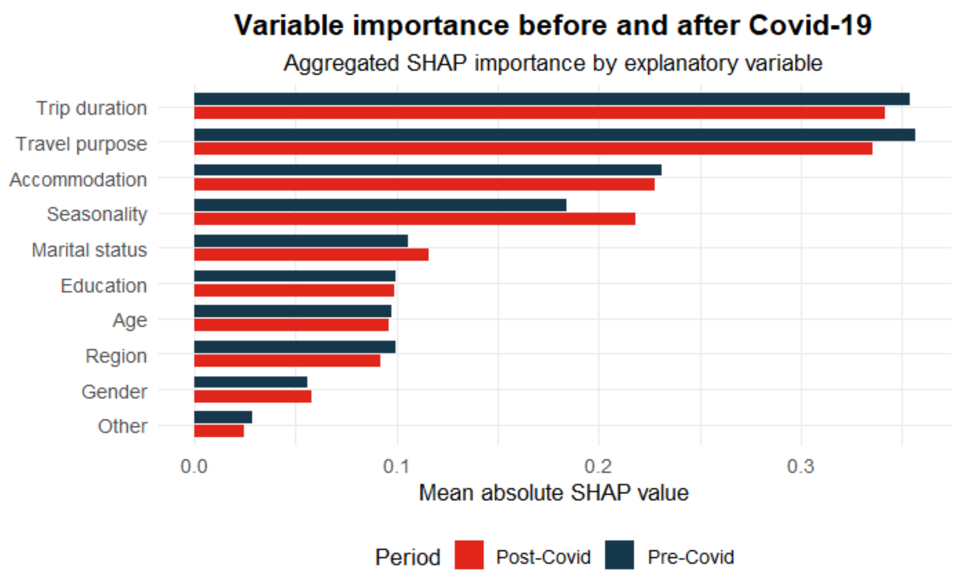
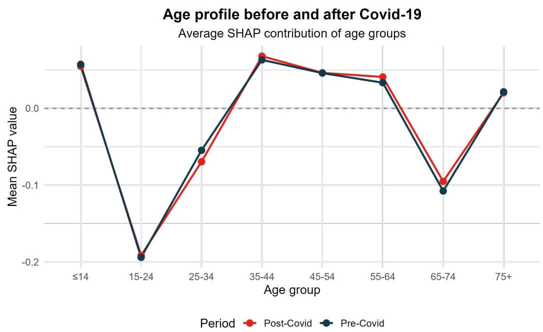
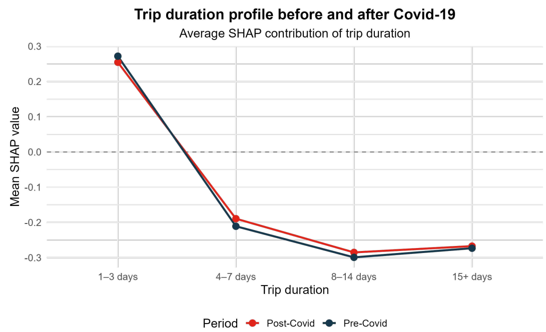


Figure 4.12: Aggregated SHAP importance before and after Covid-19. Values represent the mean absolute SHAP contribution of each explanatory dimension.

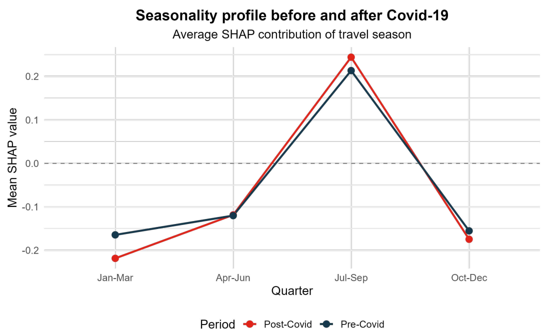
A more detailed comparison of behavioural profiles confirms this interpretation. Figure 4.13 compares the SHAP contributions associated with age, trip duration, seasonality, and accommodation type before and after Covid-19. The profiles show a high degree of overlap across the two periods, indicating that the pandemic did not substantially alter the main behavioural structure of road-based tourism.



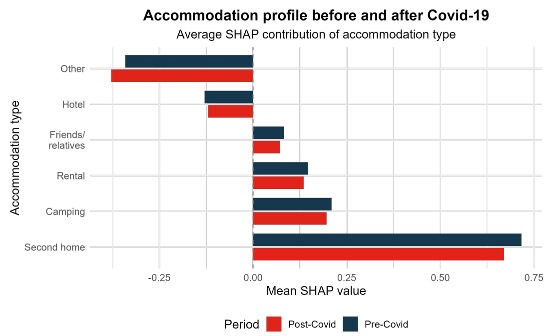
(a) Age profile



(b) Trip duration



(c) Seasonality



(d) Accommodation type

Figure 4.13: Behavioural profiles before and after Covid-19. The figures compare average SHAP contributions across the main behavioural dimensions associated with road-based tourism.

The behavioural profiles display a high degree of stability across periods. Children and middle-aged adults continue to show positive SHAP contributions, short trips remain the category most strongly associated with road-based tourism, and second homes continue to exhibit the strongest accommodation effect. These patterns suggest that the core behavioural structure of road-based tourism remained largely unchanged after Covid-19.

The main difference concerns seasonality. The positive contribution associated with the July–September period becomes stronger after Covid-19, while winter and autumn quarters remain negatively associated with road-based tourism. This suggests that the pandemic reinforced the seasonal concentration of road-based travel, making summer mobility even more central within the road-based tourism segment.

Finally, Figure 4.14 examines whether the territorial structure of road-based tourism changed after Covid-19. The dumbbell plot shows that most regional SHAP contributions remained highly stable across periods, with pre- and post-Covid values almost overlapping for most regions. The main visible differences concern Sicily and Sardinia, where the negative contribution becomes less pronounced after Covid-19. This suggests that the gap between insular and mainland regions may have reduced during the pandemic period, possibly deu to the reorientation of tourism towards domestic and proximity-based mobility.

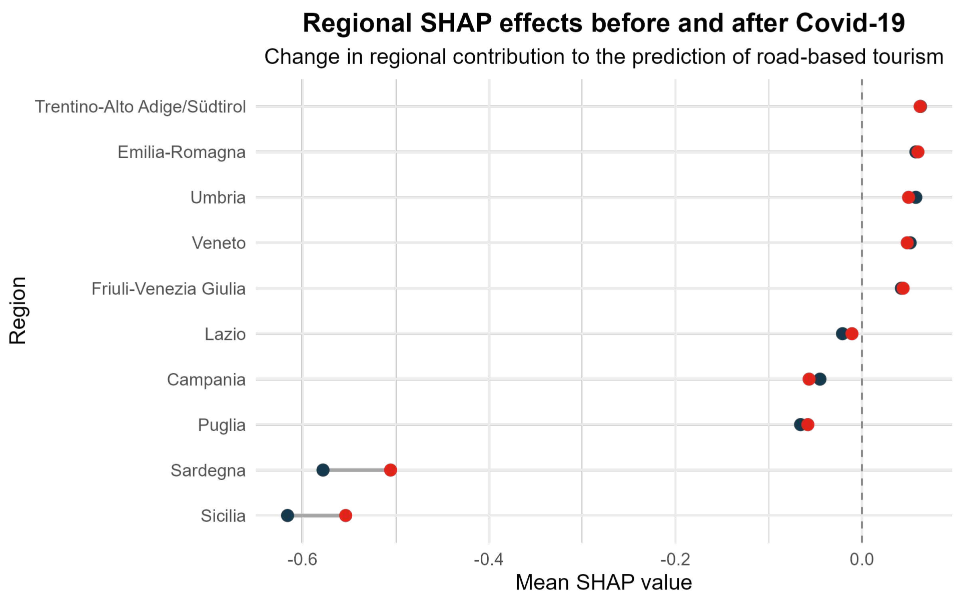


Figure 4.14: Regional SHAP contributions before and after Covid-19. The overlap between pre- and post-Covid values indicates substantial stability in the territorial structure of road-based tourism, with the main visible changes concerning Sicily and Sardinia.

Taken together, these findings suggest that Covid-19 acted primarily as an accelerator rather than as a structural transformer of road-based tourism. The pandemic sharply increased the prevalence of road-based travel, but it did not fundamentally change the characteristics of the travellers and trips most strongly associated with this form of tourism. Age structure, trip duration, accommodation preferences, and territorial effects remained remarkably stable. The main transformation concerns the stronger concentration of road-based tourism during the summer period. In this sense, Covid-19 reinforced pre-existing behavioural patterns rather than generating an entirely new mobility profile.

The results presented in this chapter provide consistent evidence that road-based tourism constitutes a distinct and recognisable behavioural pattern within the Italian tourism system. Its determinants appear remarkably stable across modelling approaches and remain largely unchanged even after the disruption generated by the Covid-19 pandemic.

Chapter 5

Discussion and conclusions

This thesis investigated the determinants of road-based tourism in Italy using microdata from the ISTAT *Trips and Holidays* survey covering the period 2015–2023. By combining traditional statistical approaches, machine-learning techniques, and explainable artificial intelligence tools, the analysis sought to identify the characteristics associated with the propensity to undertake a road-based trip and to provide a behavioural profile of this tourism segment.

Unlike studies that focus primarily on transport mode choice, the present research approached road-based tourism as a broader behavioural phenomenon. This perspective made it possible to examine not only the predictive factors associated with road-based travel, but also the underlying travel patterns, socio-demographic characteristics, and organisational features that distinguish this form of tourism from other travel behaviours.

5.1 Main findings and practical implications

The results indicate that road-based tourism represents a distinct and recognisable tourism segment within the Italian tourism system. Rather than being defined solely by the use of private road transport, it reflects a broader behavioural configuration characterised by specific travel preferences, organisational choices, and mobility patterns.

The analysis shows that travel organisation, seasonality, household-related mobility, and geographical context all contribute to shaping the propensity to undertake road-based trips. Although these dimensions influence travel behaviour in different ways, they collectively point towards a coherent form of tourism that extends beyond a simple transport mode choice.

These findings also have several practical implications for tourism stakeholders, destination managers, and policymakers interested in understanding and supporting road-based tourism in Italy.

First, the strong association between road-based tourism and family-oriented travel suggests that destinations interested in attracting this segment should pay particular attention to the needs of travellers with children. Family-friendly services, accessible accommodation, flexible travel options, and activities suitable for different age groups may be especially relevant for this category.

Second, the importance of second homes, camping, and rental accommodation highlights the relationship between road-based tourism and relatively autonomous forms of travel organisation. This suggests that tourism strategies targeting road-based travellers should not focus exclusively on traditional hotel-based tourism, but should also consider different accommodation types, such as those that are compatible with greater mobility flexibility and independent travel planning.

The results also underline the importance of seasonality. The strong concentration of road-based tourism during the summer months indicates that destinations can face significant seasonal fluctuations in demand. A better understanding of these seasonal patterns may help destinations plan tourism services, infrastructure and mobility systems more effectively during periods of high demand.

Finally, the existence of important territorial differences suggests that tourism policies should also account for regional heterogeneity. The factors associated with road-based tourism may vary according to geographical accessibility, transport infrastructure, and local tourism structures. Consequently, strategies aimed at promoting road-based tourism may require different approaches across Italian regions rather than relying on a uniform national framework.

In particular, regions characterised by lower participation in road-based tourism, such as island destinations, could benefit from strengthening those tourism services and accommodation arrangements that appear more strongly associated with road-based travel behaviour, including second homes, camping facilities, and rental accommodation. While the present analysis does not allow causal relationships to be established, enhancing the attractiveness and accessibility of these tourism products may contribute to making such destinations more compatible with the preferences of road-based travellers.

5.2 Methodological implications

An additional contribution of this study concerns the comparison between traditional statistical models and more advanced machine-learning approaches. Although XG-

Boost achieved the best predictive performance among the models considered, the improvement over logistic regression was relatively limited. This finding suggests that a substantial share of the behavioural structure underlying road-based tourism can be captured through relatively stable and interpretable relationships.

At the same time, the machine-learning approach proved valuable in identifying non-linear effects and interaction patterns that would be more difficult to specify within a traditional modelling framework. In particular, the SHAP analysis provided a clear interpretation of the model predictions, making it possible to combine predictive accuracy with behavioural interpretability.

These results highlight the potential of explainable artificial intelligence techniques for tourism research. Rather than replacing traditional statistical approaches, machine-learning methods can be considered complementary tools, allowing researchers to explore complex behavioural relationships while maintaining a meaningful level of interpretability.

5.3 Limitations and future research

Despite the contributions of the present study, several limitations should be acknowledged when interpreting the results.

A first limitation concerns the nature of the available data. The analysis relies on the ISTAT *Trips and Holidays* survey, which is based on self-reported information collected through interviews. Although this survey represents one of the most comprehensive sources of tourism data available in Italy, it does not directly observe actual mobility behaviour. As a result, some aspects of travel decision-making may be affected by reporting errors, recall bias, or incomplete information provided by respondents.

A second limitation relates to the set of variables available in the dataset. While the survey contains a rich set of socio-demographic and trip-related characteristics, several potentially relevant determinants of transport mode choice are not observed. In particular, information on fuel prices, toll costs, vehicle ownership, access to alternative transport modes, environmental attitudes, and travellers' individual preferences is not available. These factors may influence the decision to do a road-based trip and could contribute to explaining part of the residual variability not captured by the models.

A further limitation regards the definition of road-based tourism adopted in this study. Due to data availability, road-based trips are identified based on the main transport mode used during the journey. However, tourism mobility is often multimodal, and the survey does not fully capture the complexity of transport choices taking place before, during, and after the trip. Consequently, some travel behaviours may be only partially represented by the binary classification adopted in the analysis.

From a methodological perspective, machine-learning models identify patterns and associations within the data but do not allow causal relationships to be established. Although SHAP values provide valuable insights into the factors associated with road-based tourism, the results should be interpreted as predictive and descriptive rather than causal. The positive or negative contribution of a variable indicates its association with the outcome, but it does not imply that modifying that characteristic would necessarily lead to a change in travel behaviour.

Finally, the study focuses exclusively on the Italian context. Italy represents a particularly interesting case due to its strong domestic tourism market, regional heterogeneity, and widespread reliance on private mobility. Nevertheless, the behavioural patterns identified in this research may not be directly generalisable to countries characterised by different geographical conditions, transport infrastructures, and tourism systems.

Despite these limitations, the study provides a comprehensive overview of the behavioural factors associated with road-based tourism and demonstrates the potential of combining machine-learning techniques with explainable artificial intelligence tools to investigate tourism mobility patterns.

Future research could extend the present analysis in several directions. First, the availability of more detailed mobility data would allow the role of transport-related factors, such as vehicle ownership, travel costs, and accessibility to alternative modes of transport, to be investigated more thoroughly. Second, future studies could exploit longitudinal or panel data to examine how individual travel behaviour evolves over time and in response to external shocks. Third, the integration of additional data sources, such as geospatial information, mobility tracking data, or tourism expenditure records, could provide a more comprehensive representation of travel behaviour.

Another promising direction would consist in identifying distinct behavioural segments through clustering techniques. Rather than focusing exclusively on the probability of undertaking a road-based trip, future research could explore whether different traveller profiles emerge according to combinations of socio-demographic characteristics, accommodation choices, travel purposes, and mobility patterns.

Furthermore, future studies could complement the predictive approach adopted in this work with more rigorous causal frameworks. While machine-learning models are particularly effective at identifying associations and behavioural patterns, causal inference methods could help assess whether specific factors, such as accommodation type, seasonality, or transport accessibility, have a direct effect on the propensity to engage in road-based tourism.

Finally, comparative analyses across different countries or tourism systems would help assess the extent to which the behavioural patterns identified in Italy are generalisable to other contexts and mobility environments.

5.4 Conclusions

The findings of this study suggest that road-based tourism should not be viewed simply as a transport choice, but rather as a broader form of tourism behaviour. The results suggest that private road transport constitute a broader pattern of travel behaviour shaped by individual preferences, practical constraints and available opportunities.

The persistence of the behavioural patterns identified during the analysis suggests that road-based tourism is characterised by a relatively stable structure. Even the disruption generated by the Covid-19 pandemic appears to have reinforced existing tendencies rather than fundamentally altering the characteristics associated with this form of mobility.

More generally, the study highlights how transport choices and tourism behaviour are deeply interconnected. Understanding why travellers choose road-based mobility therefore requires moving beyond purely transport-related explanations and considering the wider behavioural and organisational dimensions of tourism activity. In this respect, the integration of machine-learning techniques and explainable artificial intelligence offers a useful framework for exploring complex tourism behaviours while maintaining a meaningful level of interpretability.

The evidence presented in this thesis contributes to a better understanding of contemporary tourism mobility and provides a basis for future research on the evolving relationship between travel behaviour, mobility choices, and tourism experiences.

Appendix A

Explanatory Variables

Table A.1: Explanatory variables included in the analysis

Variable	**Type**	**Categories / Labels**	**N**
Gender	Nominal categorical	Male, Female	2
Age group	Grouped quantitative	<14, 15–24, 25–34, 35–44, 45–54, 55–64, 65–74, $\geq$75	8
Marital status	Nominal categorical	Single, Married, Divorced, Widowed	4
Education level	Ordinal categorical	Primary or less, Lower secondary, Upper secondary, Tertiary	4
Employment status	Nominal categorical	Employed, Non-employed	2
Professional position	Nominal categorical	Employee, Self-employed, Student, Retired, Other	5
Region of residence	Nominal categorical	Italian regions	20
Trip duration	Grouped quantitative	1–3 days, 4–7 days, 8–14 days, 15+ days	4
Accommodation type	Nominal categorical	Hotel, Camping, Rental accommodation, Relatives/Friends, Second home, Other	6
Seasonality (trimester)	Ordinal categorical	Q1, Q2, Q3, Q4	4
Year	Temporal categorical	2015–2023	9

Appendix B

Additional Logistic Regression Results

Table B.1: Logistic regression coefficients associated with trip duration and seasonality

Trip duration			Seasonality		
Category	β	**% Change**	**Category**	β	**% Change**
4–7 days	-0.782	-54.3	January–March	-0.203	-18.4
8–14 days	-1.030	-64.3	July–September	0.483	+62.1
15+ days	-1.060	-65.4	October–December	-0.176	-16.1

Note: Logistic regression coefficients are reported relative to the reference categories (1–3 days and April–June). The percentage change in odds is computed as $(e^{\beta} - 1) \times 100$.

Table B.2: Logistic regression coefficients associated with age groups

Age group	β	Change in odds (%)
15–24 years	-0.963	-61.8
25–34 years	-0.801	-55.1
35–44 years	-0.548	-42.2
45–54 years	-0.639	-47.2
55–64 years	-0.745	-52.5
65–74 years	-0.997	-63.1
75+ years	-0.845	-57.0

Note: Coefficients are estimated from the logistic regression model. The reference category is children aged 14 years or younger. The percentage change in odds is computed as $(e^{\beta} - 1) \times 100$.

Table B.3: Logistic regression coefficients for the regions with the strongest SHAP contributions

Region	β	**Change in odds (%)**
	Positive SHAP contribution	
Trentino-Alto Adige/South Tyrol	0.273	+31.4
Emilia-Romagna	0.236	+26.7
Umbria	0.325	+38.4
Veneto	0.216	+24.1
Friuli Venezia Giulia	0.186	+20.5
Molise	0.325	+38.5
	Negative SHAP contribution	
Calabria	-0.247	-21.8
Lazio	-0.107	-10.1
Campania	-0.168	-15.5
Apulia	-0.263	-23.1
Sicily	-0.827	-56.3
Sardinia	-0.853	-57.4

Note: Logistic regression coefficients are reported relative to the reference category (Abruzzo). The percentage change in odds is computed as $(e^{\beta} - 1) \times 100$.

Bibliography

ANPR (2025). Anagrafe nazionale della popolazione residente (anpr). Ministero dell'Interno, Roma. Accessed October 1, 2025.

Barbieri, D., Lou, B., Passavanti, M., Hui, C., Hoff, I., Lessa, D., Sikka, G., Chang, K., Gupta, A., Fang, K., Banerjee, A., Maharaj, B., Lam, L., Ghasemi, N., Naik, B., Wang, F., Mirhosseini, A., Naseri, S., Liu, Z., ..., and Rashidi, T. (2021). Impact of COVID-19 pandemic on mobility in ten countries and associated perceived risk for all transport modes. *PLOS ONE*, 16(2):e0245886.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020a). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Barredo Arrieta, A. et al. (2020b). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Belasco, W. J. (1979). *Americans on the Road: From Autocamp to Motel, 1910–1945*. MIT Press, Cambridge, MA.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer.

Dimanche, F. and Andrades, L. (2024). Tourism trends: Current challenges for tourism destinations management. In *Tourism Destination Management: Theory and Practice*. Springer.

European Commission (2019). Commission delegated regulation (eu) 2019/1681 of 9 august 2019 amending regulation (eu) no 692/2011 of the european parliament and of the council concerning european statistics on tourism, as regards the transmission deadlines and the adjustment of annexes i and ii. Official Journal of the European Union, L259, 1–4. Accessed September 30, 2025.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Greene, W. H. (2018). *Econometric Analysis*. Pearson, 8 edition.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley, 3 edition.

ISTAT (2023). Viaggi e vacanze in italia e all'estero.

Lee, K.-S. and Eom, J. K. (2024). Systematic literature review on impacts of covid-19 pandemic and corresponding measures on mobility. *Transportation*, 51(5):1907–1961.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

OECD (2022). Tourism trends and policies 2022: Italy. Available at: https://doi.org/10.1787/tour-2022-en.

Salomon, I. (1985). *A Billion Trips a Day: Tradition and Transition in Urban Travel*. MIT Press, Cambridge, MA.

UNESCO (2025). World heritage list by country: Italy. `https://whc.unesco.org/en/statesparties/it`. Italy counts 61 World Heritage Sites: 55 cultural and 6 natural.

UNWTO (2021). International tourism and covid-19. Available at: https://www.unwto.org/international-tourism-and-covid-19 (accessed March 2026).

Urry, J. (1990). *The Tourist Gaze: Leisure and Travel in Contemporary Societies*. Sage, London.

WTTC (2024). Economic impact report italy 2024. Available at: https://wttc.org/research/economic-impact.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, Boca Raton, FL.