Università
Ca'Foscari
Venezia

Master's Degree

in Economics and Finance

Final Thesis

Machine Learning Applied to Credit Rating for Italian Private Companies

Supervisor
Ch.ma Prof.ssa Monica Billio

Co-Supervisor
Ch.mo Prof. Ionut Florescu (Stevens Institute of Technology)

Graduand
Luca Melzi
852176

Academic Year
2019 / 2020

Abstract

Credit Ratings have always been fundamental for investors to evaluate a company and its debt. Moreover, a company with a good credit score can issue debt at lower costs helping its growth. The Big Three Rating Agencies (S&P, Moody's, and Fitch) have focussed their efforts on building models to predict credit ratings for listed companies leaving out smaller ones not publicly traded – or private. The need for credit ratings for such companies is high as banks or financial institutions engaging in M&A but also private investor interested in investing in a private company need to know the health of said firm. In fact it might be that a company is underrated and might represent a good investment or on the other hand, it might be overrated, and the investor might spend more than the actual value. This is the case in Italy where most of the registered companies are private thus have no or very little access to attainable credit rating. The scope of this report then is to come up with a model that allows to predict accurately a credit rating for private companies in Italy. To do so two different machine learning algorithms will be used: K-Means Clustering and Convolutional Neural Network. The data is downloaded from WRDS Compustat and Bureau Van Dijk AIDA.

Summary

1. Introduction........ 1

2. Data........ 3

2.1 Data Collecting........ 3

2.2 Data Processing........ 3

3. Methodology........ 6

3.1 K-Means Clustering........ 6

3.2 Convolutional Neural Network........ 7

4. Implementation of the Models and Results........ 10

4.1 K-Means Clustering........ 10

4.2 Convolutional Neural Network........ 14

Conclusions and Future Work........ 17

Glossary........ 19

References........ 20

1. Introduction

A rating is a judgement that is made by an external and independent agency (i.e. the rating agency) based on an entity's ability to repay its debts. It is therefore an assessment of the entity's creditworthiness and summarises the quantitative and qualitative information available on the entity.

The ratings of an agency are used to calculate the cost of debt for the issuers of bonds. In other words, the rating is one of the factors that determines the cost of a loan and the resulting repayment conditions. Moreover, ratings can help investors monitor the value of their investment over time: a possible downgrade of an issuer by a rating agency represents a clear signal of alert to markets and individual investors, who can take precautions to safeguard their portfolio.

The "Big Three" (Moody's, Standard & Poor's and Fitch Group) rating agencies focus their analyses – due to the high interest the market and its players have – on the debt issued by listed companies and Governments' only, leaving out unlisted companies. However, private companies too are interested in knowing their creditworthiness when it comes to borrowing money or issuing their own debt. Banks offer such service, but different banks might have different evaluation methods thus yielding different results. In addition, a bank might tilt in its favour the rating to increase the cost of debt and earn more at the expense of the company. So, the goal of this project is to build a model that by combining financial knowledge and machine learning algorithms can efficiently assign a fair rating to a private company, specifically, Italian unlisted companies.

As mentioned above, the model we are interested in building will be a hybrid type: the inputs will be balance sheet entries that were deemed most suitable to predict a company's stability and creditworthiness, the computation of the model, instead, will be based on machine learning algorithms both unsupervised (KMeans clustering) and supervised (Convolutional Neural Network, CNN).

The data I used was downloaded from different sources. For the training of the model, the data was taken from the WRDS website, specifically, from S&P's Compustat database. Here, ratios and balance sheet entries for American listed companies were taken, as well as their daily rating. The data for Italian companies (i.e. the testing dataset) was gathered from Bureau Van Dijk's AIDA (Analisi Informatizzata

Delle Aziende) database. Given the private nature of the companies considered, no rating nor probability of default was available in the abovementioned dataset. For this reason, the ratings for Italian companies were taken from Cerved website which is the leading rating agency in Italy. Although it might have been better due to similar characteristics, industrials sectors have been disregarded and all companies were taken as part of one market. This was done due to a lack of usable data and to preserve what was available avoiding loss of too many observations.

2. Data

2.1 Data Collecting

In order to carry out the research, data had to be gathered from various sources. An initial analysis of the most suitable entries was carried out and all the voices from the balance sheet, the income statement and financial ratios were considered. In the end, a list of the following voices[1] was drawn up:

- Long-Term Debt,
- Current Ratio,
- Debt/EBITDA,
- Current Assets,
- Total Assets,
- Cash and Cash Equivalents,
- EBITDA,
- Total Inventories,
- Total Receivables,
- Total Stockholders Equity,
- Interest Expense,
- Accounts Payable

The above features were selected using the guidelines provided by the presentation "Il Rating bancario e la valutazione del Credito" from Banca Intesa Sanpaolo and Dott. Alain Devalle's presentation for the "Associazione dei Dottori Commercialisti e degli Esperti Contabili delle Tre Venezie". In these two presentations, the Bank and Devalle describe their idea of how a company should be analysed and which factors should be looked at to understand the creditworthiness of the subject. Given my relatively basic knowledge of accounting I decided to rely on the ideas contained in these presentations for this analysis. If this had not been a "hybrid" project, more features – if not all available – would have been selected as machine learning is known for its ability of working particularly well with large amounts of data.

The number of features chosen was based on a financial analysis and was limited to a small number as KMeans, which is the first algorithm tested, performs better with a small number of parameters. Of course, to the features mentioned above Credit Ratings must be added.

2.2 Data Processing

Due to the fact that data was collected from various sources, it had to be cleaned and uniformed in style.

[1] In the glossary at the end of the report there is a brief description of each voice.

The first issue with the data was the difference in frequency: Compustat was provided in monthly observations for balance sheet, income statement, and ratings. AIDA and Cerved instead, provided only yearly observations. This meant that the former datasets (Compustat's) had to be converted to a yearly frequency. This was done primarily on Python by grouping the data based on the date and taking the mean or the last observation, depending on the nature of the feature. As a second step, since the WRDS data came in two different datasets, these had to be merged based on common features such as the date of the observation and the Company ID.

In addition, the way the AIDA dataset was structured was not compatible with the layout of Compustat's data. The first issue was that the observations for a single feature for different years were inserted in different columns. This was solved by importing the data in R, transposing the various columns, and merging all the columns representing one feature into one single column ordered by Company ID and date of observation. The second step to the data management of AIDA was to change the names and the re-index the columns such that the layout was comparable to Compustat's data.

The only transformation that had to be done on the ratings downloaded from the Cerved website was to map them such that they were equivalent to S&P's ratings that were downloaded from WRDS. The mapping was done by following the table below which is available on the Cerved website:

CERVED RATING AGENCY	S&P'S
A1.1	AAA
A1.2	AA+ / AA
A1.3	AA-
A2.1	A+
A2.2	A
A3.1	A-
B1.1	BBB+ / BBB
B1.2	BBB-
B2.1	BB+ / BB
B2.2	BB-
C1.1	B+ / B
C1.2	B-
C2.1	CCC / C

Table 1: Equivalency table between Cerved and S&P Ratings

As a final step in the data management process of this project, once the ratings for Italian and American companies were on the same scale, the ratings had to be transformed from categorical values to numerical. This step was needed as the machine learning algorithms could only work with numerical outputs. In order to ease the prediction process, the S&P rating scale was reduced from 22 classes to only 6 using the following scheme:

S&P Rating	Numerical Value
AAA, AA+, AA, AA-, A+, A, A-	0
BBB+, BBB	1
BBB-, BB+	2
BB, BB-	3
B+, B, B-	4
CCC, CCC, CCC-, CC, C, D	5

Table 2: mapping ratings from S&P to numerical values.

3. Methodology

In this section the different models used will be described. First of all, it is worthwhile to distinguish between unsupervised and supervised machine learning model. Unsupervised learning algorithms can automatically discover interesting and useful patterns or clusters in unlabelled data and in recent years have gained popularity among researchers and practitioners (M. Emre Celebi, Kemal Aydin). Models using unsupervised learning must learn relationships between elements in a dataset without labelling the data. Some common implementations of unsupervised models are clustering (data observations are placed in the clusters based on their distance from the centre of the cluster to the data point), and anomaly detection (useful to find outliers).

On the other hand, supervised learning refers to algorithms that determine a predictive model using labelled data which is made up of input data as well as its output. The model improves by training through an appropriate learning algorithm (neural networks in this case) and in the meanwhile it uses optimization methods to minimize losses. Supervised learning models have two major applications in machine learning: regression (outputs that are real variables) and classification (outputs are classes in which the inputs are sorted).

3.1 K-Means Clustering

The k-means algorithm divides a set of N samples X into K disjoint clusters C, each described by the mean μ_j of the samples in the cluster. This algorithm requires that the number of clusters must be specified by the user. The means are commonly called the *cluster centroids* which are chosen such that the inertia (or *within cluster sum-of-squares*) criterion, which measures of how internally coherent clusters are, is minimised:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(\| x_i - \mu_j \|^2)$$

The algorithm is made up of three steps: the first chooses some k initial centroids from the dataset X. The second step assigns each sample to its nearest centroid. Finally, the third step computes new centroids by taking the mean value of all the samples assigned to the previous centroids. In this process, the difference between old and new centroids are computed. The algorithm repeats these last two (II and III) steps in loop after initialization until the distance value is less than a user-determined threshold or until the centroids do not move significantly.

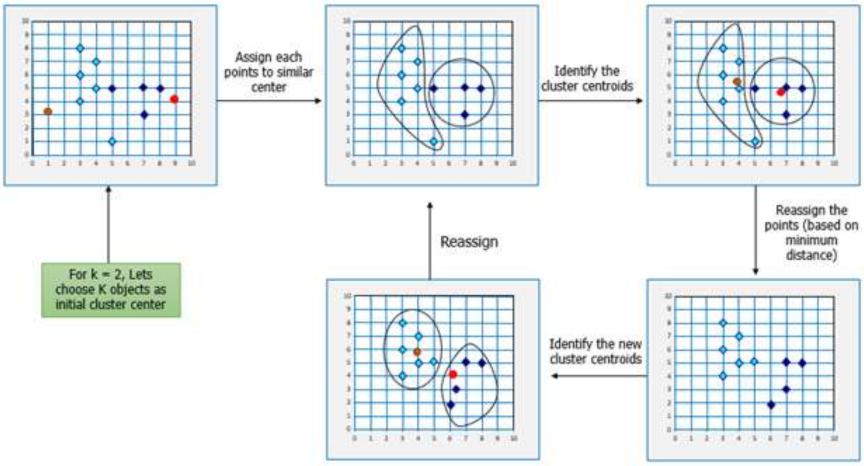

Figure 1: Simplified model for clustering algorithm

In this project, one of the challenges was to identify the optimal number of clusters. Given the large quantity of observation it was impossible to do so without any computation. To complete the task, the elbow method was used. This process fits the model with a range of k values for K and calculates the values of distortion (average of the squared distances from the cluster centres of the respective clusters) and inertia for each value of k in the given range. The elbow point is the point after which distortion or inertia start decreasing in a linear fashion. The optimal number of clusters is the value of k where the elbow is located.

3.2 Convolutional Neural Network

It is worthwhile providing a brief general description of what a neural network is and how it works before describing CNNs. A neural network algorithm is a learning system that by using a network of functions tries to transform a data input into a desired output. The idea of neural network was inspired by the way neurons in the human brain work together to understand inputs from external stimuli. The learning process starts by analysing many labelled inputs that are fed to the algorithm and using the relative outputs to learn what characteristics of the input are needed to construct the correct result. Once a sufficient number of samples have been processed, the neural network can process new inputs and return results that increase in accuracy as more samples are analysed.

A convolutional neural network (CNN) is a subset of neural networks where at least one of the layers is of convolutional type and it is mainly used for image processing. The word convolutional in mathematics stands for a mixture of two overlapping functions; in machine learning the word generally refers to a convolutional filter or layer. The filter is a matrix having the same rank as the input matrix, but a smaller shape and is one of the two actors (the second being a slice of the original matrix) taking part in a convolutional operation. The second instead is a layer of a deep neural network where a convolutional filter passes along the input matrix performing a series of convolutional operations on different slices of said matrix. A typical CNN is made up of a combination of the following layers:

- convolutional layers
- pooling layers: a layer that reduces a matrix created by a previous convolutional layer to a smaller matrix.
- dense layers: also known as fully connected layer. It is a hidden layer in which each node is connected to every node in the following hidden layer.

In order to work efficiently, the CNN, and in general a neural network, needs an activation function. An activation function takes in the weighted sum of the inputs from the previous layer and generates an output value passing it on to the next layer along with a certain bias. This step is fundamental to introduce non-linearity to the neural network that else would be comparable to a linear regression. Examples of such function are ReLu (Rectified linear unit) and Softmax[2].

[2] A description of this type of functions is provided in the glossary at the end of the report.

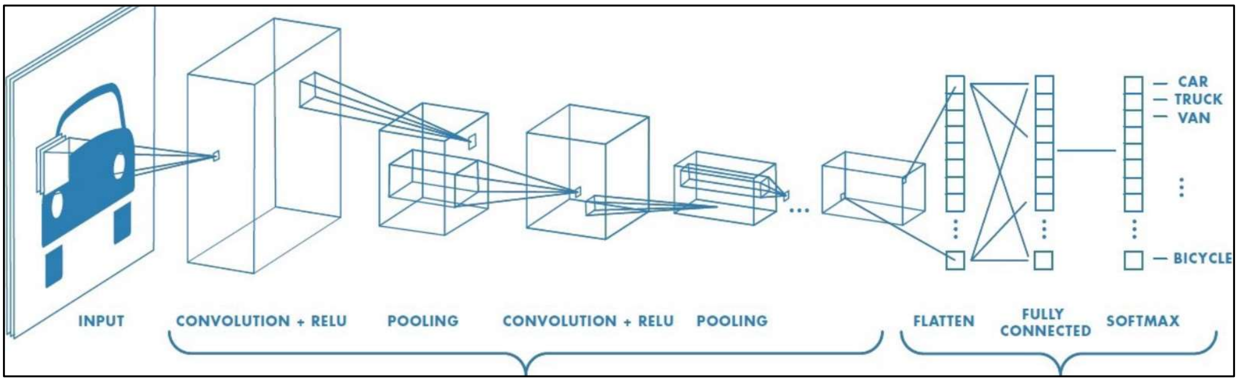

Figure 2: Simplified model for Neural Network algorithm

4. Implementation of the Models and Results

As a first step to both clustering and CNN, the Compustat data was split into testing and training in order to have an initial idea of the performance of the two algorithms. A random division was avoided as it might have caused some companies to be left out of one of the two subsets thus it was split according to the date of the observation: the subset from 2009 to 2015 is the training dataset whereas from 2016 to 2018 the testing one. The final testing will be the AIDA dataset and the whole Compustat data will serve as training.

4.1 K-Means Clustering

To proceed working with the Clustering models some important steps to optimise the data needed to be done. First, columns with a high percentage of NAs had to be dropped as clustering does not perform well with them: Interest Expense and Accounts Payable columns were dropped. Once these were dropped, the remaining rows containing NAs had to be dropped too. This caused the amount of available observations to decrease drastically. A new dataset with new data was not available so the machine had to be trained on what was left. The second step was to scale the data in the interval (0,1) using the MinMaxScaler function:

$$X_{std} = \frac{X_i - X_{\min}}{X_{max} - X_{min}}$$

Where X_{min} and X_{max} are the lowest and highest value of the given dataset. In this way, the difference in magnitude among the dataset was reduced and the data is now more manageable.

Ideally, once the model is fitted, the algorithm should sort similar observations together in the same cluster. Then each cluster would be assigned a rating based on the most widely represented rating within the cluster (i.e. voting majority): this is to say, for example, if 70% of the observations inside cluster 1 were rated 3, then that cluster would represent rating class 3.

As mentioned in the previous section, the optimal number of clusters had to be identified through the elbow point:

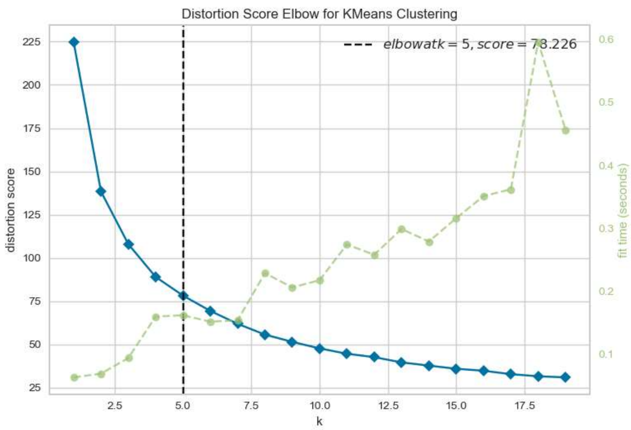

Figure 3: Elbow Point identification

Five was identified as the optimal number of clusters out of 20 different possibilities. The K-Means model was then fit using this value for the clusters. The result of the fitting is represented in the table below:

Cluster ID	Observations
0	723
1	3473
2	55
3	1116
4	146

Table 3: Results for clustering model fit

It is clear from the table that the clusters have all different sizes. To investigate even further, the distribution of the cluster is graphed:

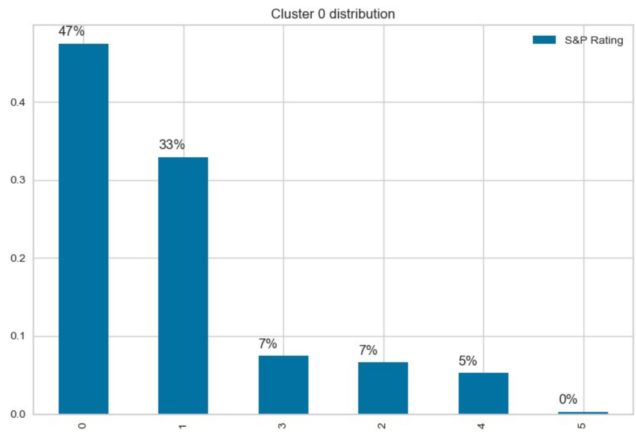

Figure 4

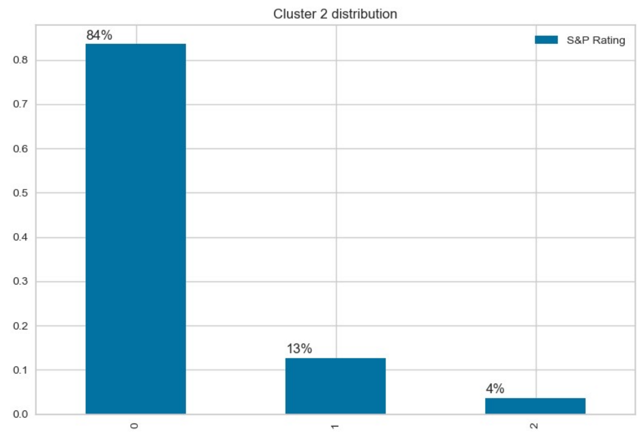

Figure 6

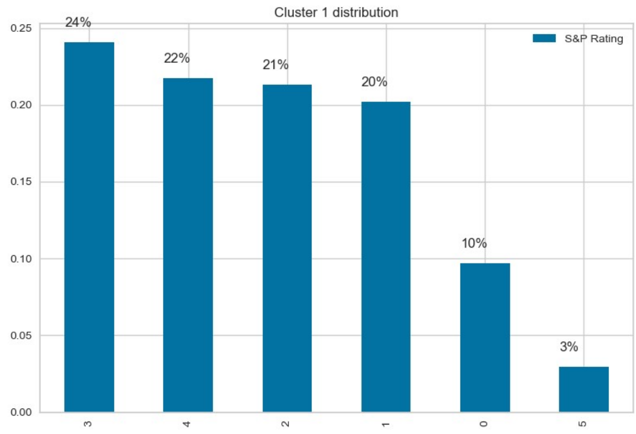

Figure 5

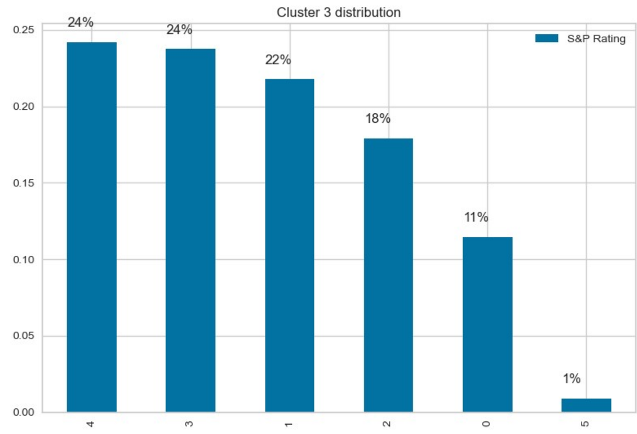

Figure 7

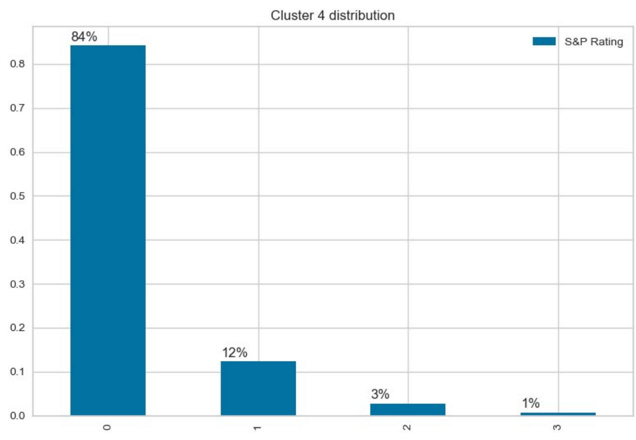

Figure 8: Distributions of the clusters created after model fitting.

From the graphs above, two different distribution schemes can be identified, one in which most of the ratings are equally distributed (clusters 1 and 3) and another in which one rating is way more represented than the others (clusters 2 and 4). Cluster 0 instead, is unique and is the only one having two ratings among the most represented.

To solve the issue with equally distributed clusters, sub-clusters were introduced. These are secondary clusters within the main cluster. The goal of these secondary clusters was to further distribute the various ratings in different clusters in the hope that a clearer majority resulted from this division. The procedure to create these sub-groups is the same as for the main ones: identify the elbow point, fit the model, and check the distribution of the secondary clusters. The elbow point is again 5, so k = 5 in our final model:

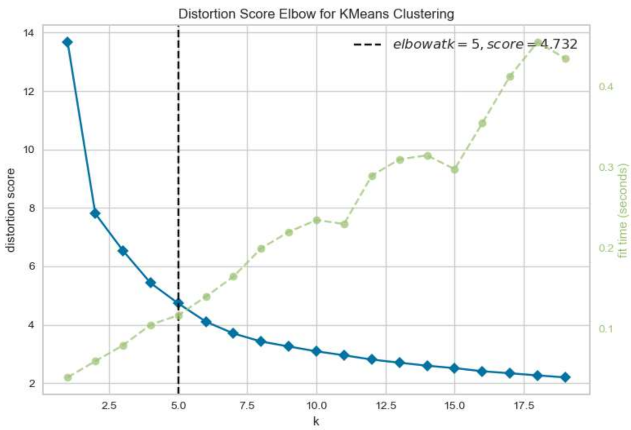

Figure 9: Elbow Point identification - sub-clustering

Surprisingly, the distribution of the ratings inside the sub-clusters did not change significantly and it was not possible to identify a clear majority within each group. This is clearer by looking at the graphs below:

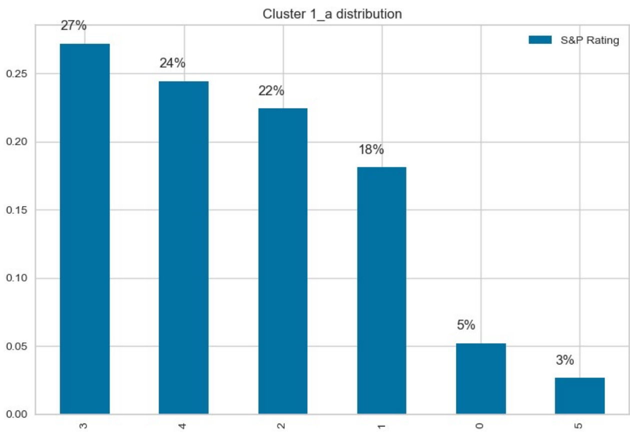

Figure 10

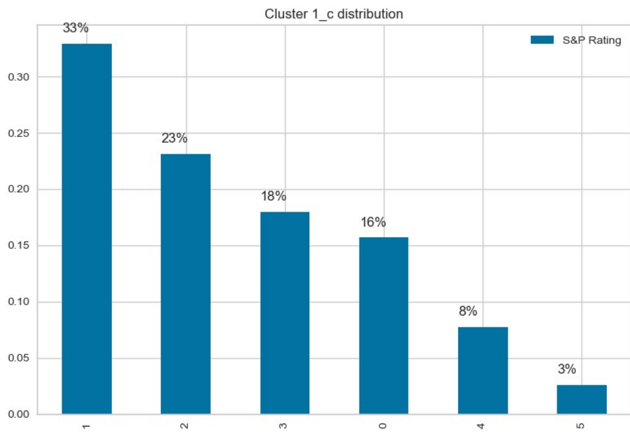

Figure 12

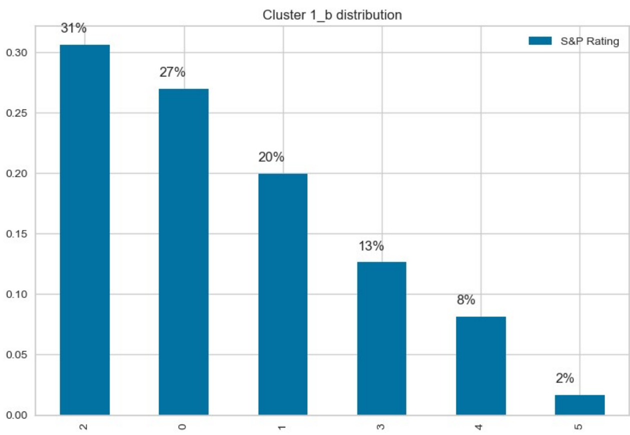

Figure 11

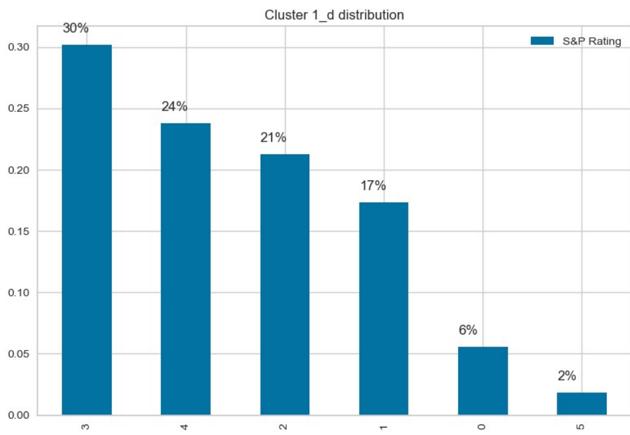

Figure 13

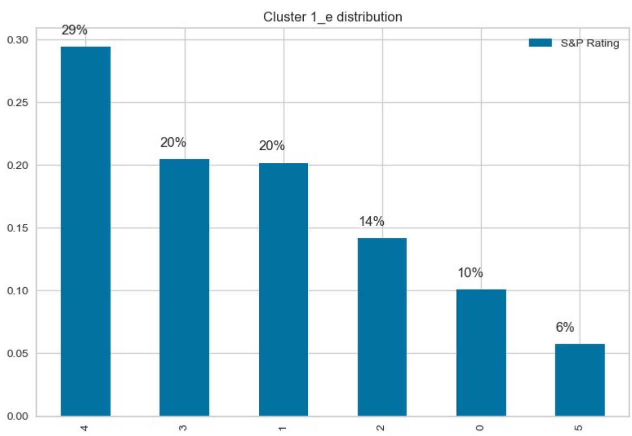

Figure 14: Distributions of the clusters created after model fitting - sub-clustering

As it is clear from the graphs above, the difference among the most represented rating and the second one in each sub-cluster is not significant to allow the assignment of the rating to that particular group.

This is a first sign that clustering is not a suitable method for the type of prediction this project was pointing to.

Moreover, it is can be seen that there is an issue also with clusters 2 and 4: both have 0 as most represented rating causing a conflict in the rating assignment. This cannot be solved easily as sub-clustering is not a viable solution. In fact, both clusters have a large difference between the first and most represented cluster and the second (~70 percentage points difference). If we were to accept this division, we would have two clusters predicting 0 as rating which is counterproductive for our purpose.

Considering all the reasons explained above, the clustering method is deemed inadequate. Due to the poor results, this method was not applied to the AIDA dataset and it was decided to follow another approach.

4.2 Convolutional Neural Network

Given the different nature of the CNN, not many preliminary actions on the data had to be carried out. The two columns that were dropped for the K-Means were dropped again due to their high percentage of NAs and the rest of the missing values were filled with an arbitrary value (-100) and then both datasets were standardised between 0 and 1. The scaling was done keeping as a benchmark the maximum and minimum values of the Compustat (training) dataset. The interval for the preliminary testing and training of the model was kept the same as in the previous section: Compustat data from 2009 to 2015 for training purposes and the remaining for testing. The final and official testing will still be carried out on the AIDA dataset.

The first step was to build the neural network layout. The first three layers are of convolutional type, the first one having 64 filters and the others with 32 and all three activated by ReLu functions. These are then followed by a flattening layer and three Dense layers, two activated by ReLu functions and the final one by Softmax. It is important having Softmax as final activation function given it returns a probability distribution of the label candidates for a particular input.

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_6 (Conv1D)            (None, 8, 64)             256
_________________________________________________________________
conv1d_7 (Conv1D)            (None, 6, 32)             6176
_________________________________________________________________
conv1d_8 (Conv1D)            (None, 4, 32)             3104
_________________________________________________________________
flatten_1 (Flatten)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               16512
_________________________________________________________________
dropout_2 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 128)               16512
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_4 (Dense)              (None, 6)                 774
=================================================================
Total params: 43,334
Trainable params: 43,334
```

Figure 15: Convolutional Neural Network summary

The trial of the model on Compustat data gave the following percentages for accuracy:

- Training dataset: 77.34%
- Testing dataset: 68.35%

These results are not particularly high but are better than the performance of the K-Means. It is worth also remembering the fact that this is a hybrid project where finance and machine learning are mixed. To increase the performance of the model more inputs should have been provided. Adding a feature selection step could also help.

The second step to this model was to run the algorithm on both Compustat and AIDA datasets, the first one as training and the latter as testing. Before doing so, since the two datasets will be fed simultaneously to the Neural Network, the features had to be named in the same way to simplify the machine's task. The columns names from Compustat were picked as example and the ones in AIDA changed. The layout of the CNN was kept the same and fitted to the new training and testing datasets. The new results are the following:

- Training score: 84.98%

- Testing score: 11.86%

The result on the testing dataset surprisingly increased to a reassuring ~85%. On the other hand, though, the testing accuracy is not as high as hoped. This can be the consequence of the fact that only a small number of companies in the AIDA dataset were provided with a rating, so the machine did not have enough observations to improve the accuracy. Another reason for this might be the large difference in magnitude that occurs between the Italian private companies and the listed American ones despite the scaling. In fact, as partial proof to this, if the datasets were scaled independently (i.e. AIDA considering its own min and max, same thing for Compustat) the results would be the following:

- Training score: 82.88%
- Testing score: 33.34%

Despite the small decrease in the accuracy of the training, the testing accuracy increased threefold. This serves as proof that the idea of transferring learning from a model trained on big, multinational, listed American companies to smaller private Italian ones might not be appropriate.

Conclusions and Future Work

Overall, both models did not perform as expected. The most disappointing model was the clustering. Despite its unsupervised nature, it did not manage to identify feasible clusters within the data. This might be caused by the fact that the rating of a company is not derived only by the few features selected and more are needed. In fact, when a rating agency is asked to release a rating for an entity, it looks at various aspects of the entity itself and carries out a broad and thorough analysis of these aspects. It might also be possible to enhance the performance of the clustering algorithm by introducing a feature selection function. Clustering in fact works best with not too many features: instead of feeding a large number of features directly to the K-Means model, one could first skim them by introducing a feature selection model such as a Random Tree or Backward/Forward Feature Selection and then take the most important features and pass them through the clustering model.

As for the Convolutional Neural Network, despite its acceptable performance, some improvements might be made. First of all, instead of limiting the number of features, it might be more productive to pass through the network more features. Like for image processing, the more "detailed" the inputs are and the higher the quality, the better the CNN performs.

In general, given the magnitude difference discussed earlier, it might also be more suitable and wiser to train the algorithm on companies that are closer in size and numbers to Italian private ones. A possible solution might be training it on Italian listed companies which despite being larger financial-wise, are not as different as American ones. However, the optimal solution, would be to test it and train it on themselves. A dataset containing many ratings for private companies is needed which could not be found with the resources available.

It is worth continuing to work on such topic as it might be of great interest not only for the private companies themselves that want an estimate of their creditworthiness but also for larger listed companies. These firms in fact, could be interested in investing in underrated private companies to develop them and re-sell them to earn a higher margin or simply acquire them for a lower cost and add them to their group. In addition, the market segment in Italy for ratings for private companies is not as developed thus there might be possibilities for profits by selling a well-behaving algorithm to banks and consulting companies.

Glossary[3]

Long-Term Debt: The item represents debt obligations due more than one year from the company's balance sheet date.

Current Ratio: Indicates a company's ability to meet short-term debt obligations. Measures whether or not a firm has enough resources to pay its debts over the next 12 months.

Debt/EBITDA

Current Assets: Cash and other assets that are expected to be realized in cash or used in the production of revenue within the next 12 months.

Total Assets: Total assets of a company at a point in time.

Cash and Cash Equivalents: Any immediately negotiable medium of exchange and funds convertible into cash within a short period of time.

EBITDA: Sum of Sales - Net (SALE) minus Cost of Goods Sold (COGS) minus Selling, General & Administrative Expense (XSGA).

Total Inventories: Merchandise bought for resale and materials and supplies purchased for use in production of revenue.

Total Receivables: Debts owed to a company by its customers for goods or services that have been delivered or used but not yet paid for.

Total Stockholders Equity: Common/ordinary and preferred/preference shareholders' interest in the company and any reserves reported in the Stockholders' Equity section.

Interest Expense: The price of obtaining loans and borrowing.

Accounts Payable: Amount owed to a CREDITOR for delivered goods or completed services.

ReLu: ReLu is a non-linear activation function that is used in multi-layer neural networks or deep neural networks. This function can be represented as:

$$\boldsymbol{f(x) = \max(0, x)} \quad (1)$$

According to equation 1, the output of ReLu is the maximum value between zero and the input value. An output is equal to zero when the input value is negative and the input value when the input is positive. Thus, we can rewrite equation 1 as follows:

$$\boldsymbol{f(x) = \begin{cases} 0, if\ x < 0 \\ x, if\ x \geq 0 \end{cases}} \quad (2)$$

Softmax: A Softmax function is a type of squashing function. Squashing functions limit the output of the function into the range 0 to 1. This allows the output to be interpreted directly as a probability. Similarly, softmax functions are multi-class sigmoids, meaning they are used in determining probability of multiple classes at once.

[3] All definitions are taken from the following the internet. Some are from the WRDS database dictionary others from websites providing IFRS definitions and deepAI.org.

References

Tsai, Chih-Fong, and Ming-Lun Chen. "Credit rating by hybrid machine learning techniques." Applied soft computing 10.2 (2010): 374-380.

Huang, Zan, et al. "Credit rating analysis with support vector machines and neural networks: a market comparative study." Decision support systems 37.4 (2004): 543-558.

Ye, Yun, Shufen Liu, and Jinyu Li. "A multiclass machine learning approach to credit rating prediction." 2008 International Symposiums on Information Processing. IEEE, 2008.

Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36.2 (2003): 451-461.

Bradley, Paul S., and Usama M. Fayyad. "Refining initial points for k-means clustering." ICML. Vol. 98. 1998.

Alsabti, Khaled, Sanjay Ranka, and Vineet Singh. "An efficient k-means clustering algorithm." (1997).

Cieslak, David A., and Nitesh V. Chawla. "Learning decision trees for unbalanced data." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2008.

Formenti, Matteo. "A method for pricing the credit valuation adjustment of unlisted companies." Journal of Risk Management in Financial Institutions 12.2 (2019): 184-194.

Banca Intesa Sanpaolo, "Il Rating bancario e la valutazione del Credito".

Alain Devalle, March 9th, 2018. "Novità OIC: profili civilistici, riflessi fiscali e sul rating bancario". Presented at the "Associazione dei Dottori Commercialisti e degli Esperti Contabili delle Tre Venezie".

Compustat Daily Updates - Fundamentals Annual. (last accessed: 04/30/2020) https://wrds-web.wharton.upenn.edu/wrds/

Analisi Informatizzata dei bilanci (AIDA). Dati Aziende non quotate. (last accessed 04/30/2020) https://aida.bvdinfo.com/