Machine Learning – Project Report

Group 18

- Beatrice Serafini ID:3150395
- Matteo Nesiti ID: 3171078
- Stiven Menekshi ID: 3193511
- Ettore Marku ID: 3189837

INTRODUCTION

The goal of this project is to be able to predict the probability of an individual experiencing financial distress in the next two years by analysing patterns within the data to understand the factors influencing the financial distress outcome.
A lot of banks use algorithms to rate loans in order to protect themselves and their customers from defaulting, especially important in our highly inflationary world. To achieve our goal, we have access to two datasets “train.csv" and “test.csv" (composed of features that describe characteristics of potential creditors).

DATA DESCRIPTION

The data used in this project comes from two main datasets: "train.csv" and "test.csv". These datasets provide valuable information for predicting probability of financial distress in the next two years. The “train.csv” is the primary dataset on which we have performed the training and the "test.csv" is the primary dataset on which we have evaluated the machine learning models.
On these datasets we can find some features, expressed as categorical, numerical and continuous values, that describe the main characteristics of potential creditors, such as:

- SeriousDlqin2yrs (**Target Variable**)
- RevolvingUtilizationOfUnsecuredLines
- Age
- NumberOfTime30- 59DaysPastDueNotWorse
- DebtRatio
- MonthlyIncome
- NumberOfOpenCreditLinesAndLoans
- NumberOfTimes90DaysLate
- NumberRealEstateLoansOrLines
- NumberOfTime60-89DaysPastDueNotWorse
- NumberOfDependents

EXPLANATORY DATA ANALYSIS

The first step, in our analysis was understanding the dataset, its features and thinking about ways to get insights into its intricacies. Some of our initial findings can be found below:

Missing Values

- Both in the “train.csv” and “test.csv” we found missing values. 20% of Individuals in our dataset have their ‘MonthlyIncome’ feature missing and 2.6% of them have missing ‘NumberOfDependents’ features.

Unbalanced Dataset

- Through our analysis we realized that out target variable ‘SeriousDlqin2yrs’ was imbalanced. We came to this conclusion as more than 93.3% of the data we got from ‘SeriousDlqin2yrs’ was 0 and 6.7% of the data was 1. If left unmodified this would lead to model bias towards the majority class.

Revolving Utilization Of Unsecured Lines

- 97.8% of values of this variable are between 0 and 1 with a well-defined right-skewed distribution, as such most people have credit limits that are higher than their total balances. Although for individuals, who pose a higher level of risk it's possible for them to exhibit balances that are higher than the credit limit as they might pose a higher risk profile. Values between 1 and 10 make up 2% of the dataset. Values beyond 10 are considered outliers and they make up less than 0.2% of our data.

Debt Ratio

- 79.4% of the values of this feature are between 0 and 2. This is also logically plausible as debt is usually lower than monthly income and in some cases might be a bit higher. The remaining 20% have a much higher values, with median equal to 1201.

Age

- This feature seems to be right skewed, with a median of 51. While, analysing the data we came across an individual whose age was equal to -1. We decided to substitute this value with the median of the feature 'Age'.

Number Of Open Credit Lines And Loans

- This variable is highly skewed to right without many outliers.

Number Real Estate Loans Or Lines

- Highly skewed to the right, as most people have between 0 and 2 real estate loans or lines, the median for this variable is 1 and we have instances of values over 20.

Number Of Dependents

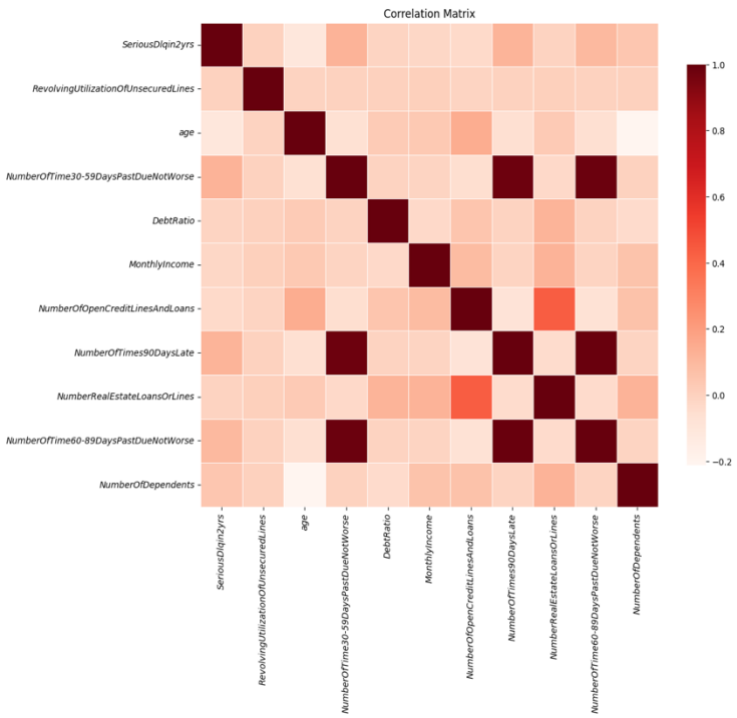
- Skewed to the right, as most individuals have between 0 to 3 dependents.

Number of N Days Past Due

- In the dataset, we found that for 202 individuals the number of times for which they were past due was 98 or 96 times. This logically does not make sense as it would mean that a person would be late on 98 or 96 loan payments at the same time. Out of all these cases, 54% of them had a serious delinquency.
- **Correlation**

Correlation

- All features that include Number of N Days Past Due have a high correlation with each other. The correlation between them is extremely close to 1 at 0.98 and 0.99.
- Number of Real Estate Loans or Lines has a relatively high correlation with Number of Open Credit Lines and Loans. The correlation is 0.43.
- Number of Dependents has a negative correlation with Age at -0.21.
- All other features do not exhibit a high correlation with each other.



METHODOLOGY

For the purpose of understanding the quality of our work we have created two different datasets. Benchmark Dataset, which is left untouched and Final Dataset, in which we will perform modifications. Data analysis helped us understand the weak points in Benchmark Dataset, as such allowing us to create our methodology thesis for improving Final Dataset. Both "train.csv" and "test.csv" exhibited many missing values and outliers, which if left untouched would hinder the performance of our models. We deemed that the best first step in our methodology was handling these discrepancies through Data Cleaning and Feature Engineering. Even though we could have implemented these methods on all the outliers and missing values, we decided to move forward with a targeted approach. Removing and imputing variables, where we thought made the most sense based on logical reasoning and the feature's importance to predicting 'SeriousDlqin2yrs'. Furthermore, we also implemented for both Datasets pre-processing in which we tried to tackle balancing, multicollinearity, overfitting and standardization problems. The final step in our methodology was model selection, where we choose the best 3 models with the highest AUC Score. As closing remarks, we will highlight why certain models work better perform than others.

DATA CLEANING

Number of N Days Past Due

- In explanatory data analysis, we highlighted the fact that some features in Number of N Days Past Due exhibited values equal to 96 or 98. This does not make logical sense as such it is likely that these values are errors in the dataset. At the beginning of our analysis, we thought that it would be beneficial to remove these values, but we realized that 54% of these cases had 'SeriousDlqin2yrs' equal to 1. This means that these data are very informative for predicting serious delinquency in the upcoming 2 years and these cases are also present in the test as such we cannot remove.

Monthly Income and Debt Ratio

- In explanatory data analysis, we highlighted the fact 20% of data coming from feature 'DebtRatio' were higher than 2. If a person has 'DebtRatio' higher than 2 it means that their debt expenses are more than 2 times higher their 'MonthlyIncome'. We also discovered that the median of individuals that have a 'DebtRatio' bigger than 2 was 1201. Due to these two factors, we deemed the cases that 'DebtRatio' was higher than 2 highly unusual. We started analysing if individuals that had these strange 'DebtRatio' data had other features, which exhibited strange behaviours. The variable, which immediately caught our attention was 'MonthlyIncome as it is the denominator of 'DebtRatio'. When we digged deeper into this variable we did in fact find that 90% of individuals that had 'DebtRatio' bigger than 2 had 'MonthlyIncome' missing. We assumed that since 'MonthlyIncome' was missing the 'DebtRatio' for these individuals was only a representation of their monthly debt expense. As such we created a model, which predicted 'MonthlyIncome' based on other features. Then, for the individuals that had missing 'MonthlyIncome' features we divided 'DebtRatio' to the newly discovered 'MonthlyIncome'. In this way, we filled the missing 'MonthlyIncome' values and corrected 'DebtRatio' that had values above 2. It is also worth noting that after the modifications on 'MonthlyIncome' and 'DebtRatio' features, we tried to maintain the same percentage of 'DebtRatio' values between 0 and 2.

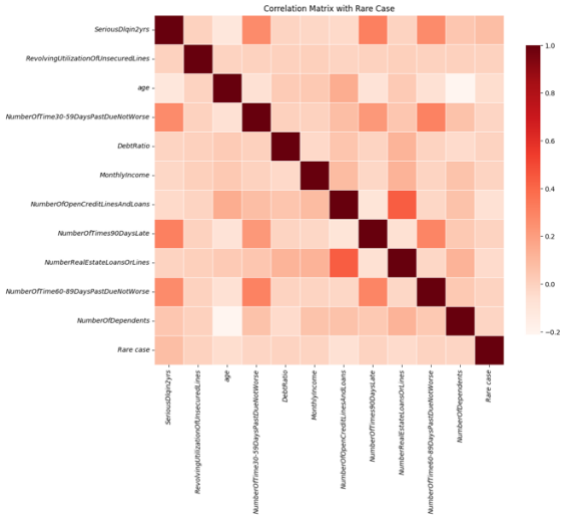
Number Of Dependents

- In explanatory data analysis, we highlighted the fact that 2.6% of 'NumberOfDependents' values were missing. We decided to fill the missing values with the mean of the feature. Due to the small proportion of the missing values in 'NumberOfDependents', we deemed it not necessary to use other more complicated methods to fill these values.

FEATURE ENGINEERING

New Feature – Rare case

- In explanatory data analysis, we highlighted the fact that for features regarding Number of N Days Past Due there were some cases equal to 98 or 96, which were highly unusual. Furthermore, we discovered that these cases explained the high correlation between the Number of N Days Past Due features. To solve the high correlation problem and at the same time retain their importance in predicting SeriousDlqin2yrs, we decided to create a new categorical feature RareCase and at the same time fill all the previous unusual values with the mean of their respective features. In the newly created categorical feature RareCase we tried to capture if the individual had previously exhibited unusual values if yes they would get a 1 if not they would get 0. After implementing this solution, we both decreased correlations between Number N Days Past Due features and increased the accuracy of our models.



PRE-PROCESSING

Balancing

- We realized at the beginning of the analysis that the data was imbalanced. Without fixing the balancing problem the model will be highly biased towards the majority class. As such we started researching ways in which we could reduce the imbalance. Two of the most relevant ways we found were Oversampling and Undersampling. The ratio between our majority and minority class is 14 to 1. The minority class is considerably smaller than the majority class as such it would be better to use Undersampling. Due to the big difference in classes we run the risk of creating high bias if we use Oversampling, as it would mean that we expand the minority class multiple times increasing the bias with each of them. While, in Undersampling we shrink the majority class to be more comparable to the minority class.

Standard Scalar and PCA

- In theory, we wanted to use in our model was PCA. We wanted to use PCA due to the fact that it directly helps in controlling overfitting and multicollinearity problems. It is especially beneficial considering that we have values in Number of Days Past Due that have a high correlation, which usually leads to multicollinearity. Implementing PCA can help in fixing the multicollinearity problem. For PCA to have the highest impact we need to incorporate Standard Scalar before running PCA, to normalize the features of the dataset. This is due to the fact that PCA is highly affected by variance in the data, as such using a method which lowers variance proves beneficial to our pre-processing. In practice, we knew before deciding to implement PCA that it wouldn't have given us useful results. This is due to the facts that our dataset is already limited in the number of features as such reducing them further would not be beneficial.

MODEL SELECTION

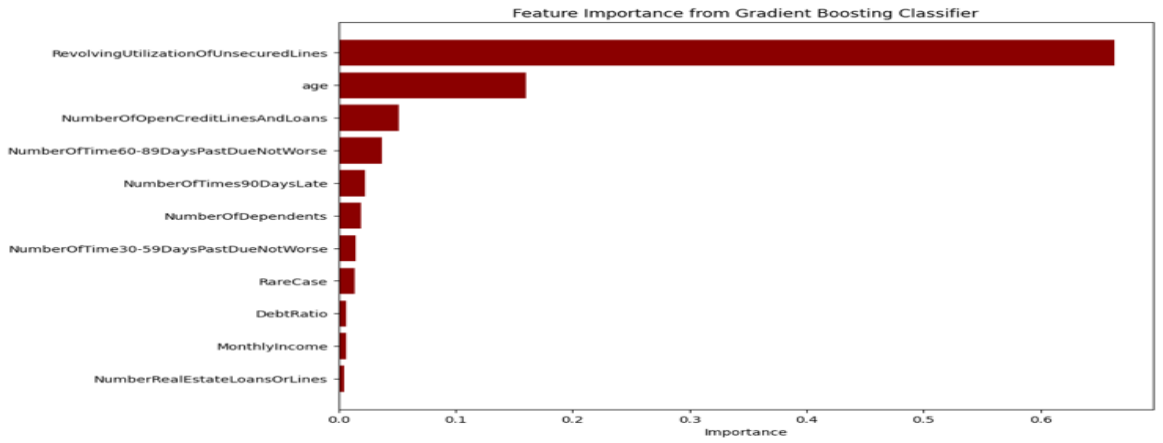
In order to understand if our assumptions were correct we used two datasets, which were then used for our model selection. Benchmark Dataset, in which we dropped all missing values and all outliers. Final Dataset, in which we have used all of the methods mentioned above including Data Cleaning, Feature Engineering and Pre – Processing. These two datasets were then used for running a variety of models with very different results (AUC scores and Accuracy). It is important to note that for all our models their performance increased from the Benchmark Dataset to the Final Dataset. In regards to the Final Dataset, we would like to highlight two models Gradient Boosting and Naive Bias, with the highest AUC score and the highest accuracy respectively. For our Gradient Boosting model we discovered that the 3 features with the highest importance were 'RevolvingUtilizationOfUnsecuredLines', 'Age' and 'NumberOfTimes60-89DaysPastDueNotWorse'. While, for Naive Bias the features with the highest importance were 'RevolvingUtilizationOfUnsecuredLines','NumberOfDependents' and 'NumberOfTimes90DaysLate'.

The new 'RareCase' feature, even though it reduced correlation between features it has not shown a high importance for neither of our chosen model, being in the bottom half of importance for both.

CONCLUSION

This analysis present some limitations such as the low feature importance and the unbalanced data. Hence the performances could be improved finding more data to balance the dataset and performing feature engineering to come up with more relevant features.

Overall this project demonstrates the potential of Machine Learning in predicting Serious Deliquency.



	Benchmark Dataset		Final Dataset	
Model Name	AUC Score	Accuracy	AUC Score	Accuracy
Logistic Regression	0.514	0.694	0.855	0.805
XGB Classifier	0.719	0.649	0.839	0.768
Random Forest	0.739	0.672	0.847	0.776
Gradient Boosting	0.746	0.699	0.857	0.792
AdaBoosting	0.737	0.701	0.843	0.784
SVM	0.757	0.720	0.848	0.793
K - Nearest Neighbors	0.709	0.654	0.804	0.768
Naive Bias	0.763	0.705	0.828	0.871
Decision Trees	0.614	0.587	0.688	0.688