UNIVERSITÀ DI PISA

Distributed Analysis and Mining of a Bank Loan Status Dataset: A Written Report

Data Understanding, Preparation & Classification using Machine Learning Methods in an Apache Spark/Hadoop Cluster Computing Framework

Ferri Lorenzo (607828)
email lorenzoferri1995@gmail.com

Ilic Ema (602796)
email ema.ilic9@gmail.com

Pappolla Roberta (534109)
email r.pappolla@studenti.unipi.it

Distributed Data Analysis and Mining (687AA), Academic Year 2020/2021

Contents

1 Context **1**
1.1 Dataset & the Structure of the Report 1

2 Data Preparation & Feature Engineering **1**

3 Data Understanding **1**
3.1 Variable Interpretations and Types 1
3.2 Statistics and Exploarative Analysis 1
3.2.1 Numerical Variables 1
3.2.2 Categorical Variables 2

4 Classification and Dimensionality Reduction **3**
4.1 Dimensionality Reduction with PCA 3
4.2 Classification 3
4.2.1 Data Preparation for Numeric Classifiers 3
4.2.2 Data Preparation for Mixed Categorical and Numeric Classifiers 3
4.2.3 Results 3

1 Context

The goal of the Distributed Data Analysis and Mining course and the project is to simulate the Big Data Analytics process. Given that the huge quantities of data generated nowadays cannot be successfully processed by a regular personal computer, the Apache Spark/Hadoop Framework is used to tackle the issue. The idea behind it is the following: dividing a large problem into smaller ones that are concurrently solved by many single processors. This is called the MapReduce paradigm.

1.1 Dataset & the Structure of the Report

The dataset was taken from the Kaggle website (aggiungere una footnote alla dataset). The goal of the Analysis is to predict whether the person in question will default on paying the debt in question or not, based on the features of this debtor/loan combination available in the dataset. The initial training and the test set were of sizes 17.9 MB and 1.69 MB, respectively, and were comprised of 18 columns and (Lorenzo?) rows. Even though the datasets seemed clean at a first glance, after the careful analysis it was clear that some serious data preparation was needed before the classification due to numerous missing values and outliers. This will be tackled in the second (Data Preparation) chapter of this report. The following chapter (Data Understanding), on the other hand, discusses the final, cleaned dataset which was used as an input for the classification methods (Chapter 4).

2 Data Preparation & Feature Engineering

3 Data Understanding

3.1 Variable Interpretations and Types

Annual_Income: Annual income of the person in question (numerical continuous).
Bankruptcies: Number of bankruptcies the relevant person has experienced so far (numerical discrete).
Credit_Score: the lower the Credit Score, the more likely the debtor is to default (numerical continuous).
Current_Credit_Balance: the amount of money that a client of a financial institution has in his or her account (numerical continuous).
Current_Loan_Amount: Loan Amount is the amount the borrower promises to repay, as set forth in the loan contract (numerical continuous).
Debt_Income_Rate: (numerical continuous)
Home_Ownership: Type of home the debtor is in posses of (categorical).
Installment_Rate: Installment Rate (numerical continuous).
Loan_Status: The label in output: Whether the Loan has been paid off or not(categorical).
Maximum_Open_Credit: An open credit is a financial arrangement between a lender and a borrower that allows the latter to access credit repeatedly up to a specific maximum limit (numerical continuous).
Monthly_Debt: Monthly debt payments are any payments you make to pay back a creditor or lender for money you borrowed (numerical continuous).
Months_since_last_delinquent: Number of months since the last law offense or violation (if the person never committed such an act, the value is -1) (numerical discrete).
Number_of_Credit_Problems: Credit problems include lack of enough credit history, denied credit application, fraud and identity theft (numerical discrete).
Number_of_Open_Accounts: The open account definition is an account which remains to be paid. Open account is also known as an account payable by the bearer (numerical discrete).
Purpose: Purpose of the debt (categorical).
Tax_Liens: A tax lien is a lien imposed by law upon a property to secure the payment of taxes. A tax lien may be imposed for delinquent taxes owed on real property or personal property, or as a result of failure to pay income taxes or other taxes (numerical discrete).
Term: Time the debtor has to pay off the debt. Categorical variable takin on only two values: Short or Long Term (categorical).
Years_in_current_job: For how many years has the person been employed at the current position (numerical discrete).
Years_of_Credit_History: (numerical discerete)
cluster_label: (numerical discrete)

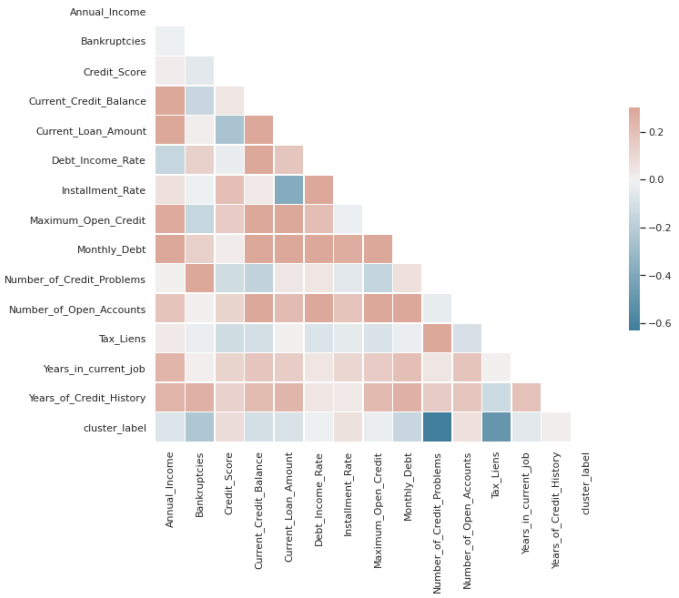
3.2 Statistics and Exploarative Analysis

Due to the total lack of Credit Score and Annual Income values (Lorenzo, erano queste le colonne?) in the test set, it was decided not use it in the analysis. The clean training set contained 72997 values, of which 50783 (70%) rows had the dependent binary variable 'Loan_Status' as 'Fully Paid' and 22214 (30%) values as 'Charged Off'. Thus, the dataset was not considered imbalanced and no further balancing methods were used.

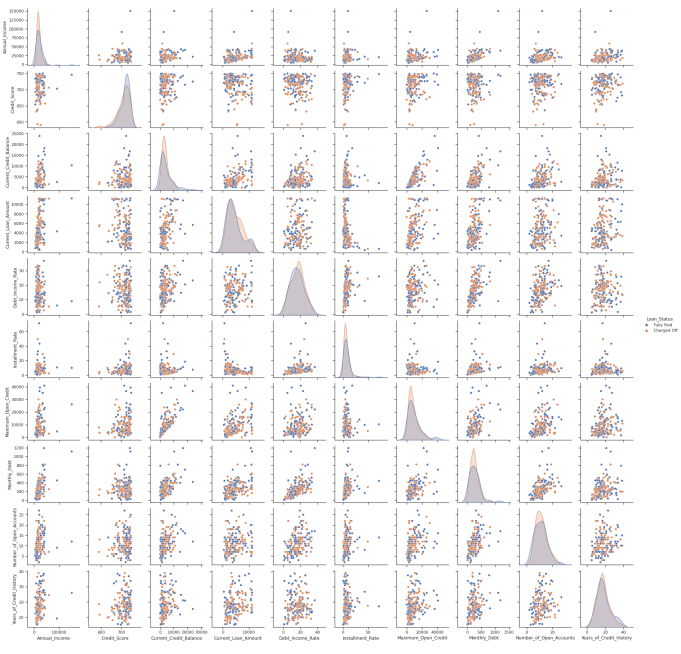
3.2.1 Numerical Variables

The statistical information relevant for each of the numerical values, such as average, median, number of distinct values, minimum and maximum value can be observed in the spark dataframe tables in notebook '2_Data_Understanding_&_Statistics'. For the sake of visualizing a Big Data scenario, Spark Dataframe was sampled to approximately 0.3% and the output classes were balanced to approximately 50:50. In fig-

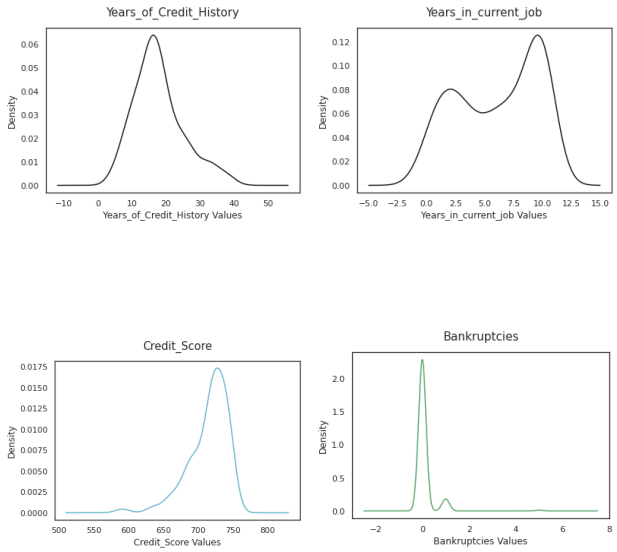
ure below a correlation heatmap matrix can be observed displaying the correlations between different attributes. As no pair of attributes had a particularly high correlation, there was no need for eliminating any attributes.



Another interesting visualization is this scattermatrix, representing the sampled plot of the most prominent numerical attributes.



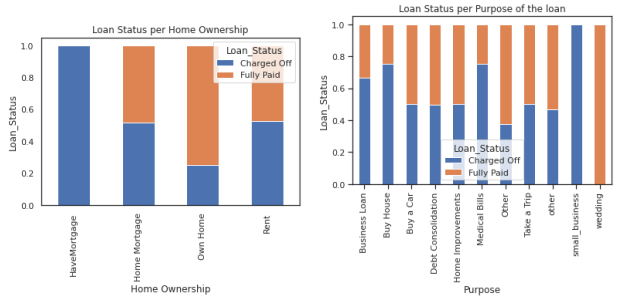
Additionally, four prominent density plots were chosen and included in the report, but for the sake of the conciseness of the report, nly two will be discussed. Namely, we can observe how both 'Years_in_Current_Job' and Bankruptcies have a bimodal distribution. From the small sample of only 178 values, most debtors had no previous bankruptcies, whereas a very small number had one bankruptcy, and an even smaller number of debtors had five bankruptcies. In this case, these are considered outliers. The peak at the Years_in_Current_Job equal to 10 is explained by the fact that in the original dataset, this attribute used to be a categorical string, and as mentioned in data preparation, all of the '10+ years' values were mapped to a number 10. Thus, one could say that the true mode is the one at approximately 2 years.



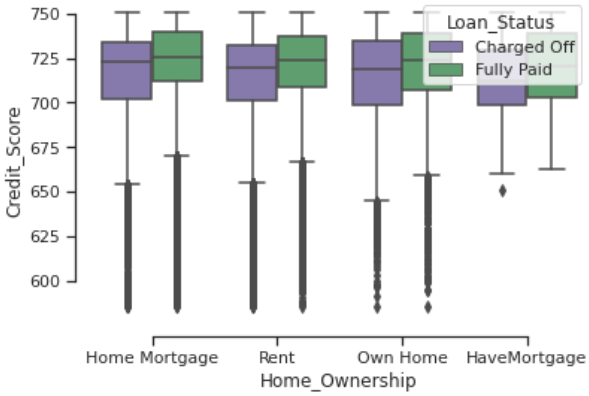
Finally, boxplots provided some meaningful insight into the outlier distribution. E.g.

3.2.2 Categorical Variables

On the figures below, the bar charts representing different types of home ownership and purposes can be observed with respect to the class variable. It is surely surprising to see that in the smaple selected, everyone who took debt for the sake of wedding paid it off, whereas essentially everyone who took out the debt for starting a business defaulted.



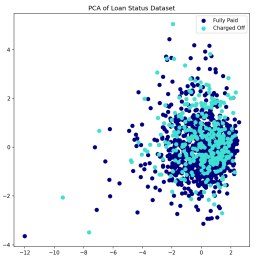
Aditionally, the boxplots below goes more in-depth on the distribution of Credit Score per different Home Ownership categories with respecct to the Loan Status variable, as well as its outliers.

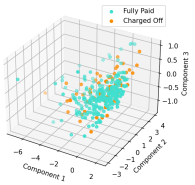


4 Classification and Dimensionality Reduction

4.1 Dimensionality Reduction with PCA

Dimensionality Reduction method called Principal Component Analysis was applied in order to see whether the Dataset can be reduced to k components preserving the highest variance without significant loss in accuracy when a classifier is applied. The results of 2-dimensional and 3-dimensional PCA analysis are displayed in figures below.





The results in terms of accuracy were impressive. Namely, it was seen that the number of dimensions can be cut down to a half (from 14 down to 7) at the cost of just one percent of accuracy! As much of a revelation that this is, it was decided that the further analysis will be continued with the non-reduced dataset.

4.2 Classification

The classification was conducted in ml.pyspark for the most part, even though also mllib was tested. Grid search 4-fold cross validation was used in order to find the right parameters for all of the classifiers. It is worthwhile noting that a great handicap of the cross-validation in pySpark was the fact that it doesn't provide the option of knowing the ideal selected hyperparameters for a machine learning model, like ScikitLearn does.

4.2.1 Data Preparation for Numeric Classifiers

Only the numeric columns were isolated, and the 'Loan_Status' column was added to this new dataset. This column was encoded using the StringIndexer, and the numeric features were vectorized using the VectorAssembler. Each model was put in its own pipeline. as the Support Vector Machines model exploits distances, standardizer was fed to its pipeline.

4.2.2 Data Preparation for Mixed Categorical and Numeric Classifiers

Using the string indexer, all of the categorical non-numeric featres were encoded in the numeric ones. Thus, the numeric features were vectorized together with the new encoded categorical features. Another addtional and final step in the pipeline was the 'label converter', which serves for decoding the lables back to the original values.

4.2.3 Results

Classifier	Accuracy
Logistic Regression (MLlib)	0.50
Logistic Regression (ml.pyspark)	0.65
Support Vector Machines	0.56
Naive Bayes Classifier	0.67
Decision Tree Classifier	0.65
Random Forest Classifier	0.67