ASSIGMENT 1: CLASSIFICATION

Rodrigo Vecino Haro & Asís Márquez

1. PROBLEM STATEMENT.

During this report, we will **analyze the performance of different machine learning models** explained in class for the following problem. We have a dataset, provided by machine learning´s professors, of randomized selection of mortgage-loan-level data collected from the portfolios underlying U.S. residential mortgage-backed securities (RMBS) securitization portfolios and provided by International Financial Research. Our algorithms will aim to predict if the loan has been paid by the borrower or not.

2. PREPROCESSING AND FEATURE LEARNING.

Firstly, we are going to take a first glance to our data set. It is composed of the 10 following attributes, with the data type in brackets:

1. Age (numeric): time passed from origination to observation.
2. Time to maturity (numeric): time left until end of the mortgage.
3. Balance (numeric): outstanding balance at observation time.
4. LTV (numeric): Loan-to-Value ratio at observation time.
5. Interest rate (numeric): interest rate at observation time, in %
6. Single family (numeric): Single-family home = 1, otherwise = 0.
7. Balance Orig (numeric): outstanding balance at origin time.
8. FICO Orig (numeric): FICO[1] score at origination time, in %.
9. LTV Orig (numeric): Loan-to-value ratio at origination time, in %.
10. Interest Rate Orig (numeric): Interest rate at origination time, in %.

With the expected output, our algorithm would predict:

11. Default (numeric): Class variable [0: the borrower has paid. 1: the borrower has default payment]

The dataset contains 1045 performance observations of residential U.S. mortgage borrowers. During the data preprocessing, we looked for **missing values**, and **duplicate values**, and **defined the categorical values**. We haven´t found any duplicate values, however, we have found 8 missing values, so we decided to eliminate the observations with any missing values. For the correct performance of our algorithms, we decided to define the output (**Default** column) **as a factor**, moreover, we have changed the "0" for "Y" and the "1" for "N", as it have been explained before, the 0 meant the borrower has paid and the 1 the borrower has default payment. We have done this change, to avoid problems between continuous variables and categorical values.

Outliers

[1] A FICO score is a type of credit score created by the Fair Isaac Corporation to assess credit risk and determine whether to extend credit.

Secondly, we check for outliers, which are the observations that are exceptionally far from the mainstream data. The outliers can cause our model to malfunction. To identify the outliers, we had plotted the data on boxplots:

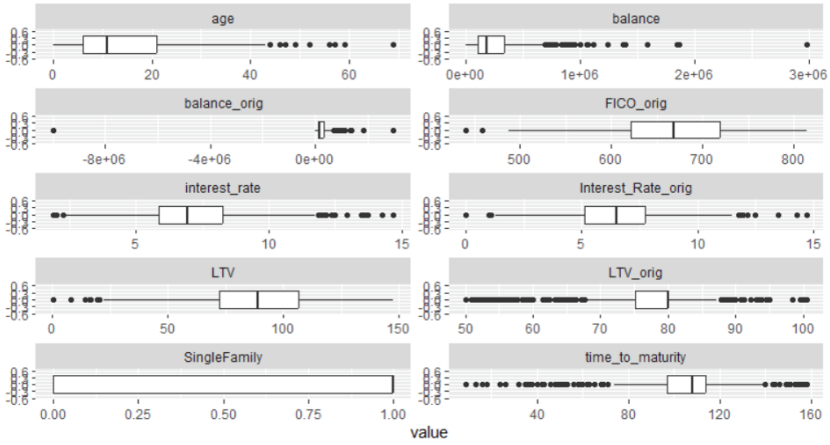


Figure 1: boxplots

As we can see in the boxplots, they are a big number of outliers in our data. So, we decided to delete in one of our trainings the outliers that differ the most from the mean, trying to affect the less as possible the number of observations, the delete outliers are the ones highlighted above:

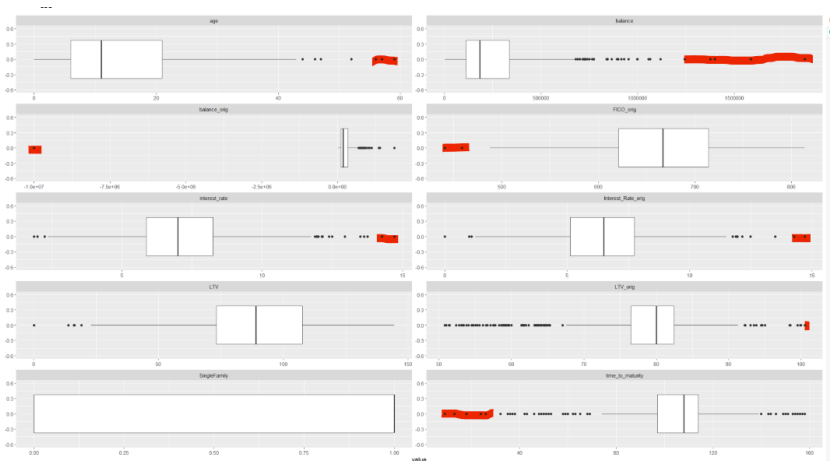


Figure 2: Delete outliers in P1 dataset.

Class imbalances

As it is sawn in the plot below our dataset have a big imbalanced, which means a distribution of the classes in the dataset is significantly skewed. This can cause a poor performance of our models in because they will tend to be biased towards the majority class. Furthermore, we will resample the dataset.

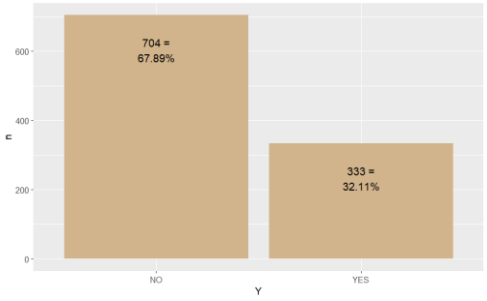


Figure 3: Class imbalances plot

We will try to use a training dataset with balances class to compare which training set works better. In the training test P2 we have the following class balances:

Figure 4: Class imbalances plot correction

Collinearity

Refers to a situation where two or more independent variables in a regression model are highly correlated with each other. Collinearity can lead to unstable and unreliable regression coefficients, making it difficult to determine the effect of each independent variable on the dependent variable.

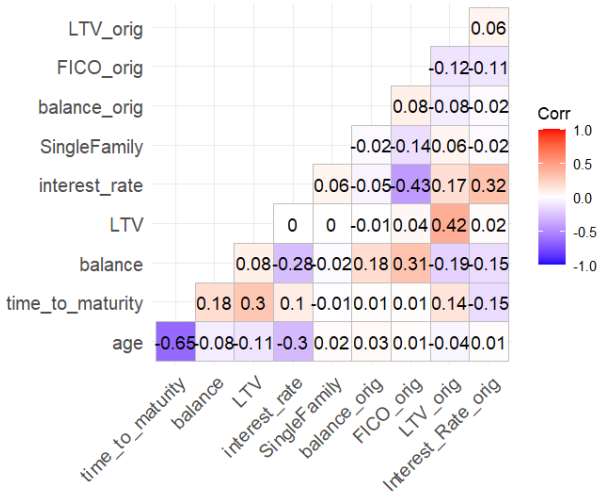


Figure 5: correlation matrix

Observing the results, the collinearity between age and time_to_maturity is in the edge, assuming 0,7 as the limit for a big collinearity. This is way, it was decided to do a different training dataset P3, to compare the results.

To summarize, 4 datasets have been selected from the preprocessing section to train the algorithms:

- P0: is the "raw" training set deleting only the observations with missing values.
- P1: is the training dataset deleting the observations with big outliers.
- P2: is the training dataset correcting the data imbalances, having the same amount of positive and negative observations.
- P3: is the training dataset deleting the time_to_maturity feature because a correlation with the feature "age".

3. MODEL SELECTION & EVALUATION

We run the following algorithms with the parameters detailed above:

1. **Logistic regression**: Is a statistical model used to analyze the relationship between a binary dependent variable and one or more independent variables. The output provides coefficients for each independent variable, which can be used to make predictions about the probability of the dependent variable being one of the two possible outcomes.
   For the logistic regression model, we indicate all other variables in the dataset should be used as predictors.
2. **K-Nearest Neighbors Algorithm (KNN):** Is a machine learning algorithm that can be used for classification or regression tasks, which classify new observations by comparing their distance to k nearby labeled observations in the training set. The optimal value of k can be determined through cross-validation.
   The turning parameter for these algorithms have been chosen by using accuracy to select the optimal model using the largest value, selecting the K value from a numeric sequence of values from 3 to 120 in increments of 4.
3. **Decision tree**: Is a machine learning algorithm that can be used for classification or regression tasks, which classify new observations by comparing their distance to k nearby labeled observations in the training set. The optimal value of k has been determined through cross-validation, with the following statements:
   a. 5 as the minimum number of observations in node to keep cutting.
   b. 5 as the minimum number of observations in a terminal node.
   c. The complexity parameter has been chosen using accuracy to select the optimal model using the largest value, iterating from a numeric sequence of values from 0 to 0.05 in increments of 0.0005.
4. **Random forest:** Is a machine learning algorithm that combines multiple decision trees to improve accuracy and reduce overfitting. The algorithm randomly selects subsets of the training data and features to create multiple decision trees and aggregates the predictions of each tree to make the final prediction. The optimal model has been determined through cross-validation, and with the following statements:
   a. 200 trees to grow.
   b. Tuning parameter, which in this case is the number of variables randomly sampled as candidates at each split. Has been chosen using accuracy to select the optimal model using the largest value, iterating from a numeric sequence of values from 1 to number of columns minus 1 in increments of 1.
5. **Support Vector Machine (SVM) – Linear**: Is a machine learning algorithm that can be used for classification or regression tasks. The algorithm finds the

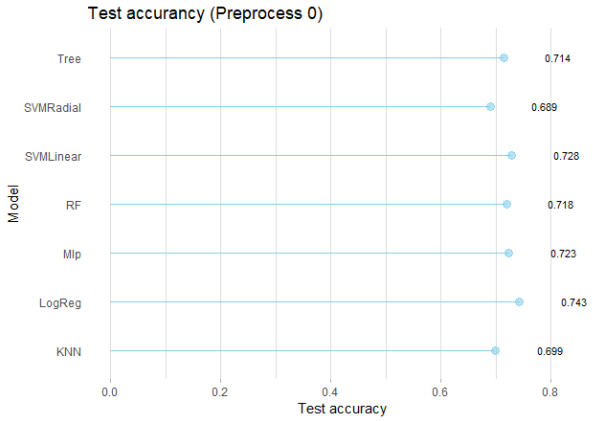
hyperplane that separates the classes with the largest margin. The algorithm can also use kernel functions to map the data to a higher-dimensional space, where it may be easier to find a separating hyperplane. The optimal model has been determined through cross-validation, and with the following statements:

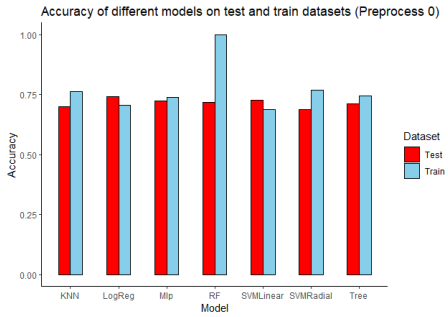
   a. The turning parameter has been chosen using accuracy to select the optimal model using the largest value, interesting from a iterating from all possible combinations of the following values 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100 & 1000.

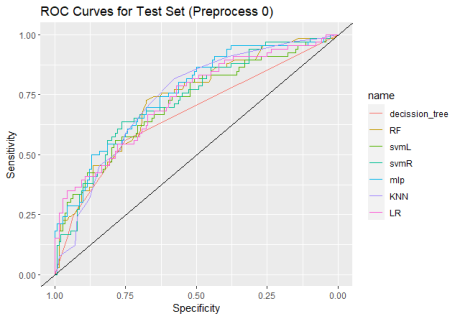
6. **Support Vector Machine (SVM) – Radial:** Is the same algorithm as the SVM – Linear, but in this case the algorithm finds the radial hyperplane that separates the classes with the largest margin. The optimal model has been determined through cross-validation, and with the following statements:
   a. The turning parameter has been chosen using accuracy to select the optimal model using the largest value, interesting from a iterating from all possible combinations of the values specified in the vectors seq(0.1,1000,length.out = 8), which are 8 numbers that range from 0.01 to 1000, and seq(0.01,50,length.out = 4), which are 4 numbers that range from 0.01 to 50.
7. **Multilayer Perceptron (MLP):** Is a type of artificial neural network that can be used for classification or regression tasks. The algorithm uses backpropagation to adjust the weights between the nodes to minimize the error between the predicted and actual values. The optimal model has been determined through cross-validation, and with the following statements:
   a. The size of the MLP has been chosen using accuracy to select the optimal model using the largest value, iterating from a numeric sequence of values from 5 to 25 in increments of 5.
   b. The decay of the parameters has been chosen using accuracy to select the optimal model using the largest value, iterating from the following values 10^(-9), 0.0001, 0.001, 0.01, 0.1 and 1.
   c. 250 is the maximum number of intersections that have been chosen.

For the evaluation we have take into consider the following metrics: accuracy of the test set, to measure precision, the difference between the accuracy of the training and test set for measuring under-over fitting, and the ROC curves to measure sensitivity and specificity of the test set, obtaining the following results for the different preprocessing datasets:

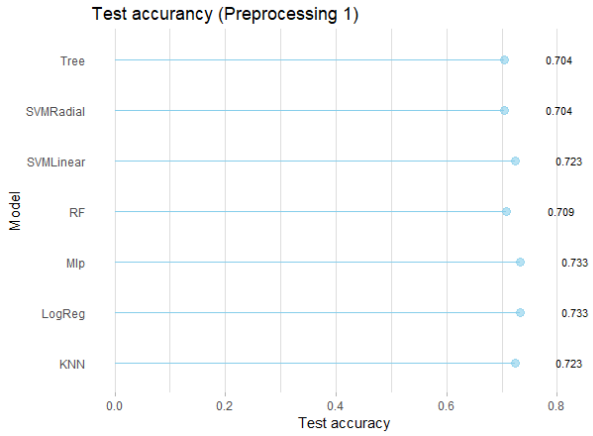
1. Preprocessing 0 dataset:

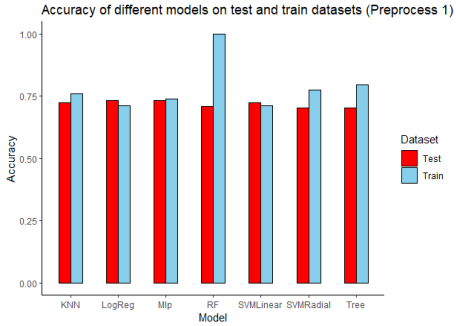


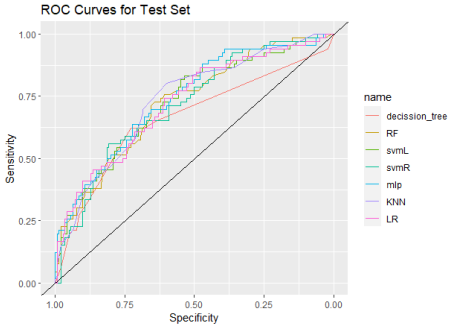




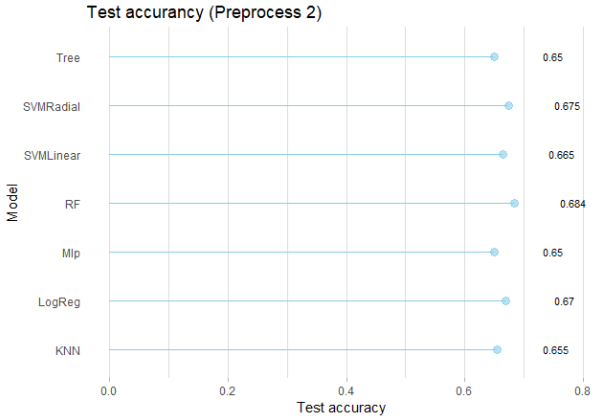
2. Preprocessing 1 dataset:

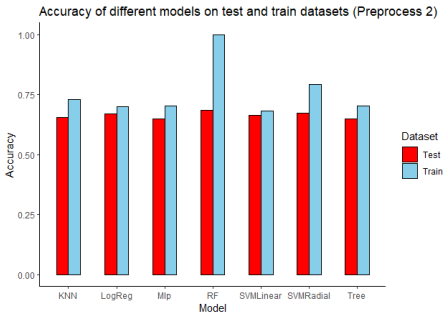


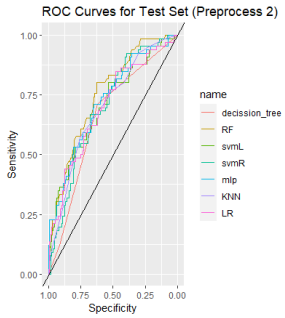




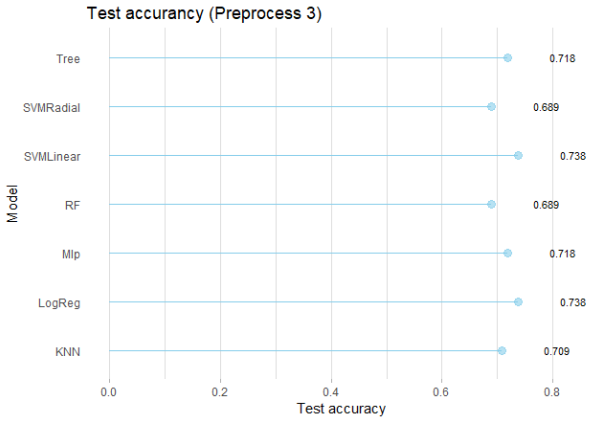
3. Preprocessing 2 dataset:

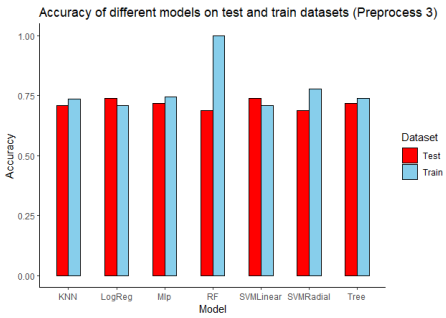


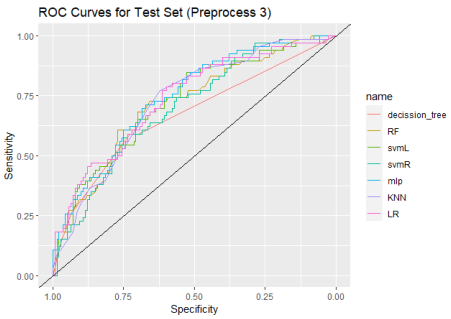




4. Preprocessing 3 dataset:







4. CONCLUSIONS

During this report, we have analyzed the performance of different machine learning models explained in class for the following problem. The dataset that we have used is a randomized selection of mortgage-loan-level data collected from the portfolios underlying U.S. residential mortgage-backed securities (RMBS) securitization portfolios and provided by International Financial Research. Our algorithms will aim to predict if the loan has been paid by the borrower or not.

The dataset has been trained using six different machine learning models, namely, Logistic Regression, KNN, SVM Linear, SVM Radial, Decision Tree, and Multilayer Perceptron. The performance of these models has been evaluated on four different training sets, P0, P1, P2, and P3.
The four different training sets used for training the models has different characteristics. P0 is the raw training set, P1 has outliers removed, P2 has corrected data imbalances, and P3 has time_to_maturity feature removed.

Firstly, we determined accuracy as the most relevant metric for model selection in our problem setting. We then compared the accuracy of various algorithms on different training datasets and found that logistic regression performed the best, with an accuracy of 0.743, with the P0 training set. SVM linear in training set P3 also had a good performance, with the same accuracy as logistic regression with the same training set.

Moreover, we observed that SVM, Logistic Regression, and MLP consistently performed well across all data
sets. In terms of training datasets, we found that the dataset without any preprocessing performed the best. The large number of observations in this dataset proved to be more effective than other preprocessing techniques such as class balance, outlier deletion, or feature selection based on correlation.

We also found that the dataset with the highest level of overfitting was the one with class balances, which makes sense because it was the one with more deleted observations. However, in our analysis, the Random Forest algorithm performed well on the training set but did not generalize well on the test set.

Finally, we used the ROC curve to evaluate the sensitivity and specificity of the models. We found that MLP, KNN, and LR had good performance, while decision tree had the worst performance.

Overall, our findings suggest that logistic regression is the best algorithm for our problem, but SVM and MLP can also be considered as good alternatives. Moreover, a dataset with a large number of observations without preprocessing proved to be the most effective.