**DATA 622 Homework 3**

**Loan Approval Status**

**Group 4: Abdellah Ait Elmouden | Gabriel Abreu | Jered Ataky**

**Problem Statement**

*For this assignment, we will be working with the attached dataset on loan approval status. The loan approval status is the target variable here. For all the models here, don’t forget to also provide performance statistics.*

*1. As you begin working with the dataset, at the beginning, please conduct a thorough exploratory data analysis. This step is necessary as you figure out which variables should be included in models.*

*2. Use the LDA algorithm to predict the loan approval status. Please be sure to walk through the steps you took, this includes how you decided on the key variables.*

*3. Use K-nearest neighbor (KNN) algorithm to predict the loan approval status variable. Please be sure to walk through the steps you took. This includes talking about what value for ‘k’ you settled on and why.*

*4. Use Decision Trees to predict on loan approval status.*

*5. Use Random Forests to predict on loan approval status.*

*6. Model performance: please compare the models you settled on in problem # 2- 5. Comment on their relative performance. Which one would you prefer the most? Why?*

**Background**

The two most pressing issues in the banking sector are: 1) How risky is the borrower? 2) Should we lend to the borrower given the risk? Banking processes use manual procedures to determine whether a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were many loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can be used to assess a customer's loan status and build strategies. In this project we want to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. Four machine learning models have been used for the prediction of loan approvals and to automate this process: LDA algorithm, KNN algorithm, Decision Trees and Random Forests.

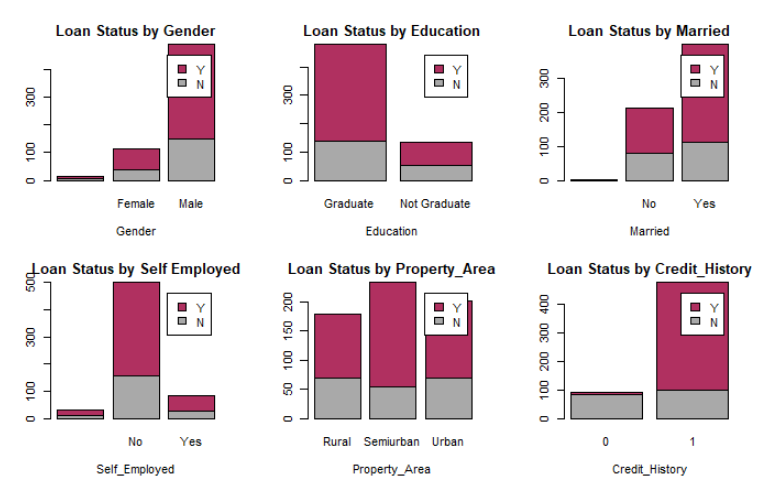
**Preliminary Analysis & Data Preparation**

The dataset contains 12 predictor variables and one response variable (Loan\_Status) over 614 separate observations. Before training any models, the data was prepared, and some initial analysis was performed. Proper data preparation is important to optimizing the performance of machine learning models, and analysis can offer us insight into how the data needs to be prepared, and perhaps what to expect out of the models.

We only needed to do some basic data preparation. The data was originally contained in a .csv file, and was imported into R. There were missing values throughout the dataset, however, less than 3% of the total dataset was missing, and no specific variable had more than 10% of its data missing. Thus, we were able to ‘impute’ (guess the missing values) using the CART method via the ‘mice’ package in R. This package helps in imputing missing values with plausible data values. These values are inferred from a distribution that is designed for each missing data point. Note that the variables Loan\_Amount\_Term and LoanAmount have a relatively small number of missing variables which can be imputed. There is an acceptable number of missing rows for the variable Credit\_History. Loan\_ID is a prime candidate to eliminate from the data, since there are 614 unique values that are not going to provide additional value to the models we will build. In addition, all data were centered and scaled prior to being fed into the machine learning algorithms; this aids in training the models.

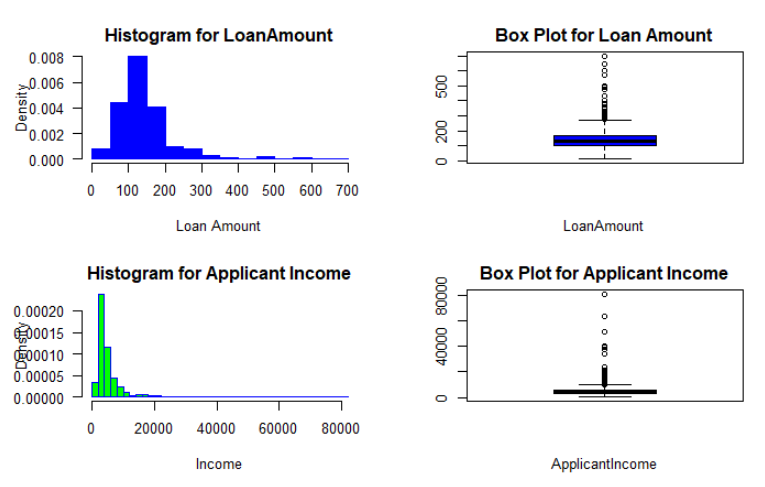
As for preliminary data analysis, we first examined the distribution of the loan status. The distribution of loan status values in the dataset (Figure 1) shows that males have more records, and more than half of the applicants' applications have been approved. There are fewer female applicants, but still more than half of their applications have been approved

**Fig. 1 Distribution of the categorical data against loan status**

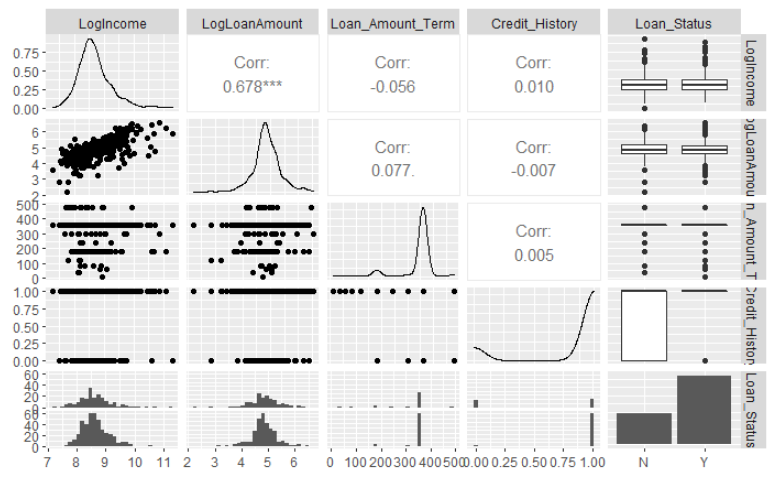


We also examined the distributions of each numerical predictor variable (Figure 2). There were several different types of distributions, including many multimodal distributions. We ultimately decided not to do any transformations on these variables aside from centering and scaling to keep our process simpler, and to make it easier to directly interpret models where possible.

**Fig. 2 Distributions of numerical predictor variables**



**Fig. 3 Correlation Matrix numerical variables including Loan\_Status**



We also briefly examine correlations between variables. The correlation matrix above (Figure 3) shows correlations between variables. The closer the absolute value of the correlation is to 1, the more correlated (either positively or negatively) two variables are. We want the predictor variables to be correlated to the response variable, but we do not want predictor variables to be correlated to each other. Generally, high correlation between different predictor variables can lead to problems for some machine learning models.

**Model Consideration and Analysis**

Multiple different models were created to predict the loan status given the 12 different predictor variables, and potentially to understand the relationships between the predictors and the loan status. Ultimately, each trained model can be used to predict the loan status of a solution where the predictor variables are known but loan status is not known. We considered the 4 different types of supervised learning models mentioned earlier. Our goal in trying different types of models was to find one that worked the best, and use that one to make predictions on future data and/or deployed the model if needed.

The dataset was split into training (80%) and testing (20%) sets; the training set is used to train the models, while the testing set is used to evaluate the models. It’s important to use data that a model was not trained on when evaluating a model, so you have an idea of what it will do with data it has not seen before. Each of the models were tuned individually with different combinations of “parameters”. Model parameters are basically the configurable, structural parts of a model, and models using different combinations of them need to be trained and tested separately to find the most appropriate ones.

The results of each tuned model (models using the best parameters) are measured by the accuracy, precision, recall and F1 score. Good models have high accuracy but need also to have great values on the other measures.

**Model Results**

**Table 1. Model accuracy metrics.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model type** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| **LDA** | **0.8033** | **0.9762** | **0.7885** | **0.8723** |
| **KNN** |  |  |  |  |
| **Decision Trees** | **0.8191** | **0.8015** | **0.9793** | **0.8815** |
| **Random Forests** | **0.8033** | **0.80** | **0.9524** | **0.8696** |

*Add about criteria of decision and why we chose such model*

**Best Model Interpretation (Add Model of choice)**

Add about the model chosen

Maybe include ROC if necessary or any graphs

**Conclusions**

**Appendix**

Please see the attached Rmd for details in code.