**Milestone Two**

DAT 690

Marek Krawczyk

9/11/2022

Understanding data is the first step in making a correct analysis. Taking stock of variables, dimensions, and types of data can help us understand what kind of plan will be necessary to reach the business objective.

In our case, the process begins with collection of data. Sourcing the data from internal database, we find ourselves in possession of pure data with 32 variables and 1000 observations. It is important to note that first variable acts as an index without any significance. The last variable is our target variable, Default. Next comes understanding of the data types we wish to analyze. The data consists of Categorical, Binary and Nominal data types. We will treat each of them uniquely for better performance of the algorithm.

Once we have a good understanding of what we are working with, we can proceed to preliminary data cleaning, formatting and filling in nulls. This very data set does not have any missing values, or precarious outliers. Utilization of a parser is a fast and easy way to check for nulls in the data set. To check for outliers, we can utilize the Z-score test. In this case we can set it for 3 standard deviations which will cover 99.8% of the data.

Now that we have preliminarily cleaned, and set up our data, we can go ahead and run some preliminary regressions to find the correlative coefficient of independent variables to our target variable. From there we can set up our streamlined data, and choose how to encode it to make the forecast most accurate.

For the sake of efficiency, we chose to run a Random Forest Regression, to find the correlative significance.

Figure 1: RF results

Table

Description automatically generated

The results from our preliminary analysis, are quite interesting. With just a quick cleaning, no unique encoding, we were able to achieve an accuracy of 75.6% in predicting if the customer will default. Some variables seem to have a very weak correlation to our target variable, so we will be pruning some of these outliers.

Variables:   
MALE\_DIV,

GURANTOR,

FOREIGN,

EDUCATION,

MALE\_MAR\_or\_WID,

POP\_UNKN\_NONE,

USED\_CAR,

CO-APPLICANT,

FURNITURE,

RENT,

RADIO/TV,

Will be removed from our data set, as they just create more noise and confusion than proper data.

With a cleaner data set, we can take a look at proper preparation of our data set, mainly to create a non-biased response from our regression when it comes to all the types of variables.