Credit Card Default Prediction

Adrian Sandoval-Vargas

Computer Science, Oakland University

**Abstract**

This study report was aimed to encompass *“The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”* by I-Cheng Yeh and Che-hui Lien [1] using statistical analysis, data mining techniques, and limiting the classification algorithm to Random Forest. The data used was 30,000 credit card holders from Taiwan in a period of 6 months [2]. In this study we will investigate different which data segmentation and data preprocessing approach yields the best accuracy.

**Introduction**

This report focuses of the use of the Random Forest [3] Classifier for the primary reason that our data is structured in such a way that it fits the supervised schema for the ensembled Random Forest Classifier. The data was obtained from the University of California – Irvine in the form of an excel spread sheet. The classifier should classify if credit card holders will default or not based on this data. I will first preprocess the data and clean the data with any outliers in specific area of interest and provide some visuals on this cleaning process. I will keep the original cleaned data and create segmentation of the data into different groups that best depicts a boundary. Once data is segmented, I will be discussing the different approaches to deal with data imbalance. After dealing with the data imbalance, I will conduct a series of validations of F1, Recall, and Accuracy scores to determine which data segments performed the best. In addition, I will experiment with Normalization and Principle Component Analysis to specific segments and discuss the effects on the recall and accuracy.

The paper will be laid out in XXX sections. In section 1, I will introduce the data and discuss and visualize key features. In section 2, I will …

1 Dataset Analysis

The dataset was obtained in the form of an excel spread sheet from UCI. It was collected by a group of researchers in Taiwan over a period of 6 months from credit card holders. The dataset contains index id, 23 variables, and its target (1 = default, 0 = did not default). The 23 variables are a mixture categorical and numerical discrete values. The variables are as follows:

LIMIT\_BAL

SEX (1 = male, 2 = female)

EDUCATION (1 = grad school, 2 = university, 3 = high school, 4 = other)

MARRIAGE (1 = married, 2 = single, 3 = other)

AGE

PAY\_0, PAY\_2, …, PAY\_6

BILL\_AMT1, BILL\_AMT2, … , BILL\_AMT6

PAY\_AMT1, PAY\_AMT2, … , PAYAMT\_6

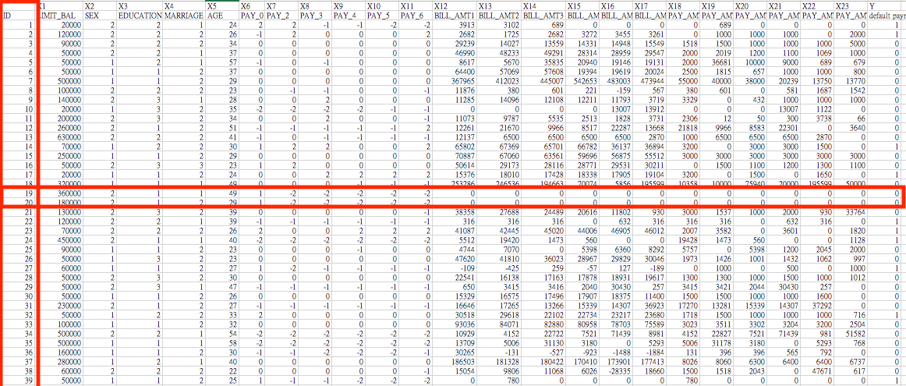
On a visual inspection we do not need the ‘ID’ column as it only marks the index of the record. We also see records in our dataset that contain -2 for the “Pay\_X” (where X means 1,…,6 ) columns that have a default value of 0 and 1 which do not make any sense to keep in our dataset.

Figure 1: the column ‘ID’ is irrelevant to the dataset. Scenario where the data contains -2’s and 0’s and has a default value of 0.

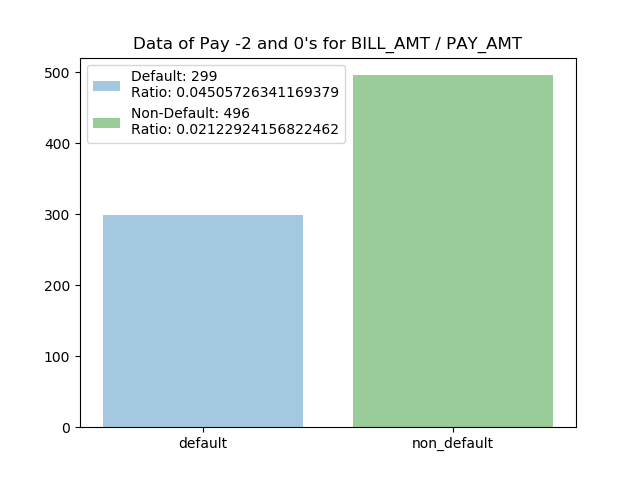
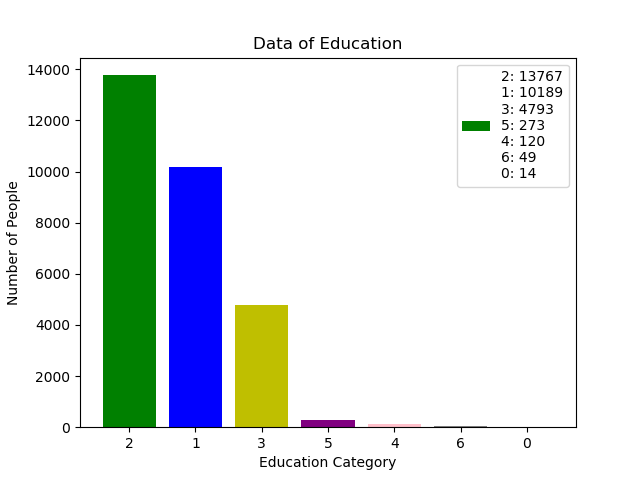
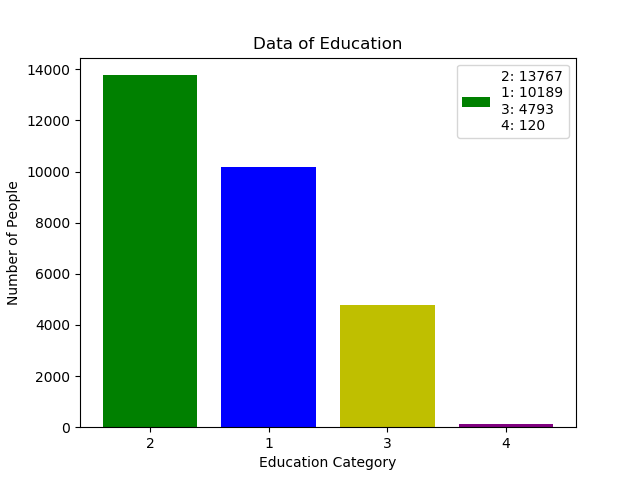
Upon writing python code to observe all incidents where all these values occur, we get

Figure 2: the amount of Default and Non-Default records what contain -2 in the PAY and 0 in

BILL\_AMT and PAY\_AMT and its ratio to it’s respective target in the overall dataset.

Since these values are not going to constitute any value to our classifier it is best to remove them. There are also values in our dataset where they were not specified by the dataset description that we must drop.

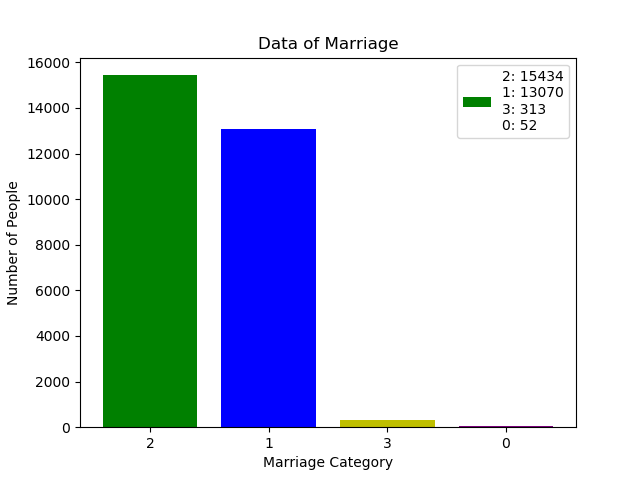
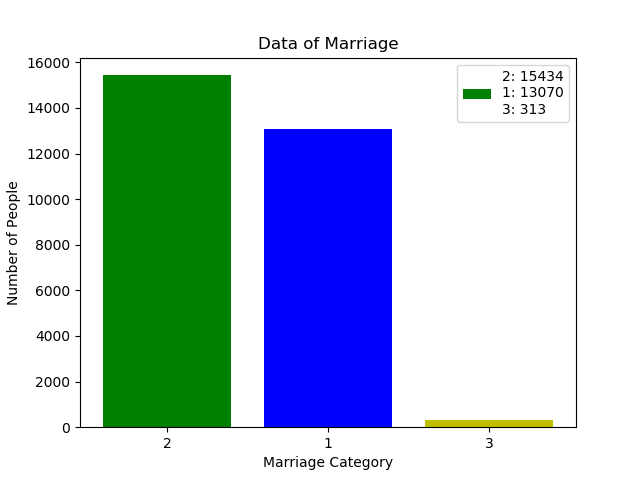


Figure 3: removal of the unspecified values for EDUCATION and MARRIAGE

In general, for any supervised ensemble classifier it is best to check if there is any imbalance in our dataset.

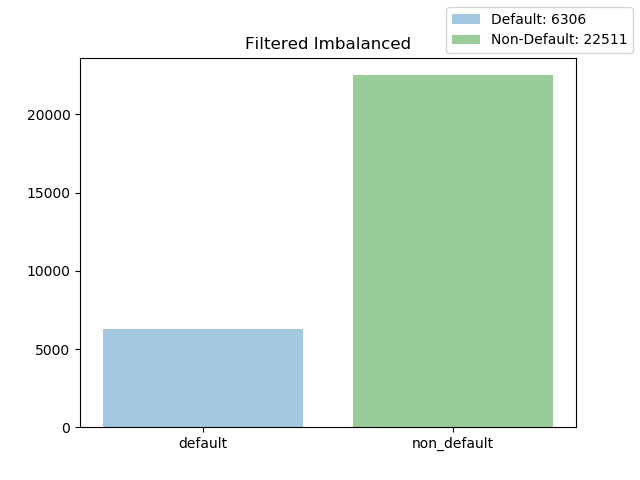


Figure 4: the amount of Default’s vs Non-Default’s in our dataset

Due to this imbalance of approximately 28% to 72% if we were to train our Random Forest Classifier it will automatically assume Non-Default due to the amount of data that is Non-Default. In order to have a good dataset to train our classifier, we must do further analysis on the dataset. Since the domain is regarding humans it is crucial that we focus statistical analysis on AGE and LMIT\_BAL due to its realistic limits. For example, while it is possible that a 78-year-old will have a credit card account, they are in the minority which could indicate a possible outlier. That is what we will analyze in AGE vs LIMIT\_BAL.

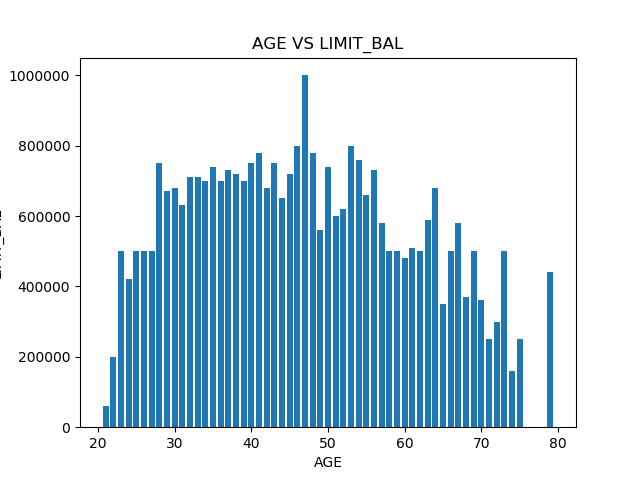


Figure 5: histogram of AGE vs LIMIT\_BAL

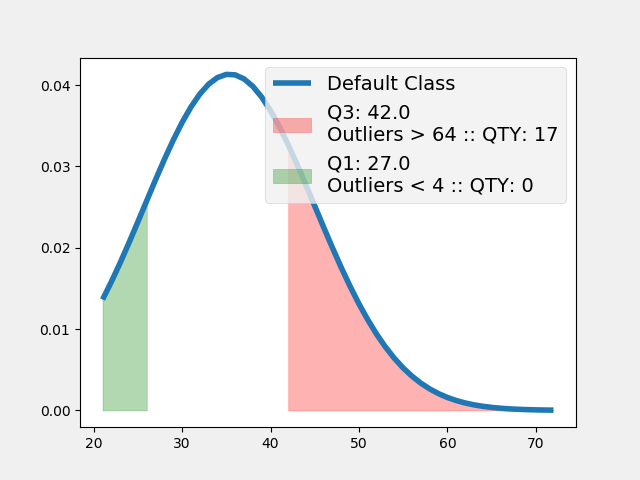
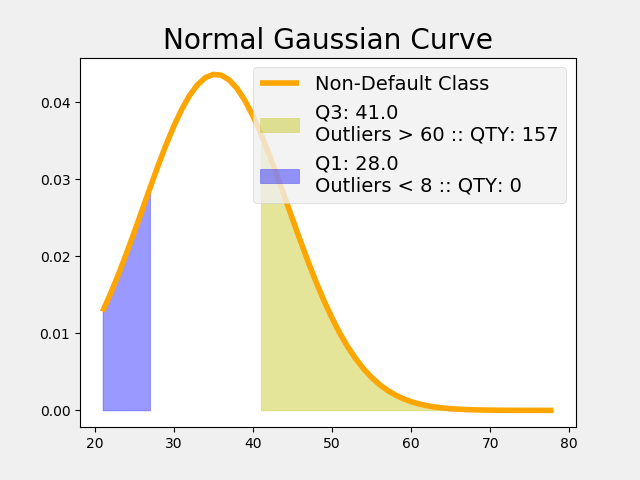
We can see that there is slight skew to the right. So, we must clean up the data by using statistical analysis. Processing the AGE vs LIMIT\_BAL into a standard normal curve we get: 

Figure 6: Normal Guassian Curve of default and non-default targets.

Given q3 and q1 we can determine the IQR and then apply the standard q1 – (1.5\*IQR) and q3 + (1.5\*IQR) to obtain the outliers (shown in figure 6). The exact process was repeated for

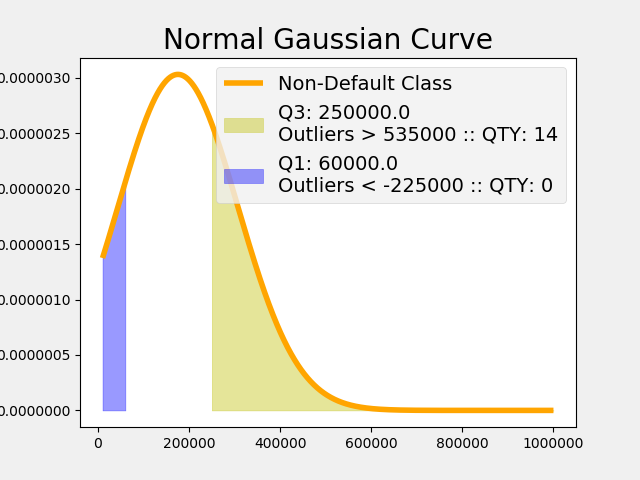
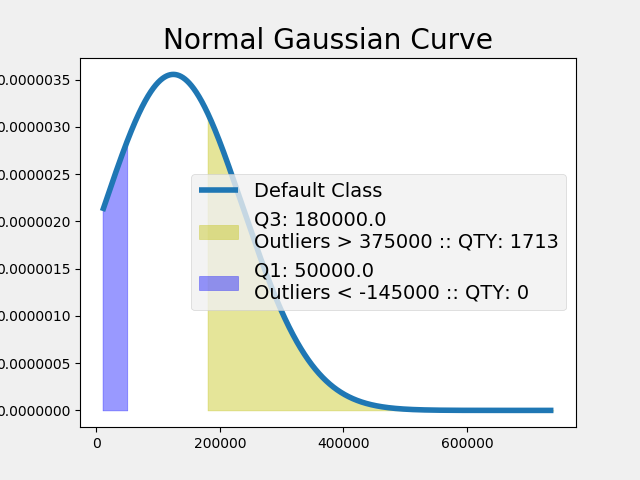
LIMIT\_BAL and the normal distribution curve is as follows:

Figure 7: LIMTI\_BAL Normal Guassian Curve of default and non-default targets.

This analysis results in the following imbalanced dataset.

