Give Me Some Credit

Kevin Loun
Math 154 Computational Statistics

July 3, 2024

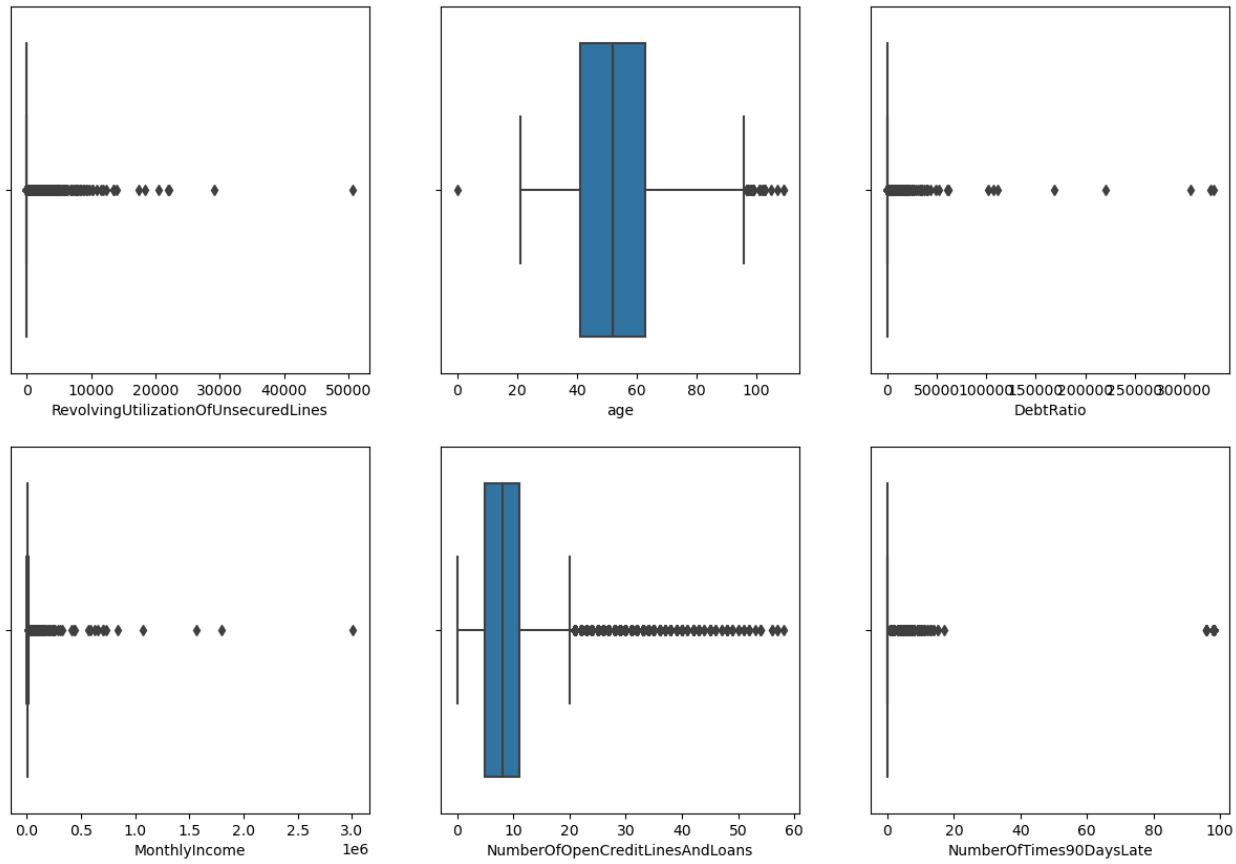
1 Introduction

Banks are crucial in market economies, controlling credit distribution and influencing investments. The availability of credit is vital for growth and stability. Banks use credit scoring algorithms to assess the likelihood of loan defaults, balancing risk management and fair credit access. The "Give Me Some Credit" Kaggle competition challenged us to refine credit scoring by predicting financial distress risk over two years. This challenge seeks to develop models that aid banks in minimizing defaults and help borrowers make informed decisions. Enhanced credit scoring models promise a fairer, more efficient financial system, benefiting both lenders and borrowers. We predicted financial delinquency by exploratory data analysis, feature engineering, model development, and evaluation, addressing the challenges of missing data, class imbalance, and outlier detection.

2 Exploratory Data Analysis

The primary dataset used in this project is sourced from the Kaggle competition "Give Me Some Credit". This dataset is rich in financial attributes, providing a comprehensive view of individual credit histories. It includes variables like age, number of dependents, income, debt ratio, and past delinquency records. EDA was conducted to gain an initial understanding of the dataset, focusing on the distribution and relationships of various variables.

We started by creating boxplot visualizations to provide a snapshot of the key financial metrics in the dataset. The presence of outliers in several variables required us to carefully handle variables to ensure they do not adversely affect the predictive models. These insights were vital for outlier treatment, transformation of skewed data, and creating derived variables for modeling.



2.1 Age

In the analysis of the 'age' variable within our dataset, certain anomalies were identified. Primarily, an observation recorded an age of zero, which is logically implausible in the context of our study, as it precludes the possibility of having a monthly income or an established loan history. Additionally, the dataset included age values exceeding 100 years, prompting further investigation.

To assess the validity of these observations, a statistical approach was adopted. We formulated a hypothesis test using the T-test method, with the following hypotheses:

$$H_0 : \text{Age observations ¿100 are not outliers}$$
$$H_a : \text{Age observations ¿100 are outliers}$$

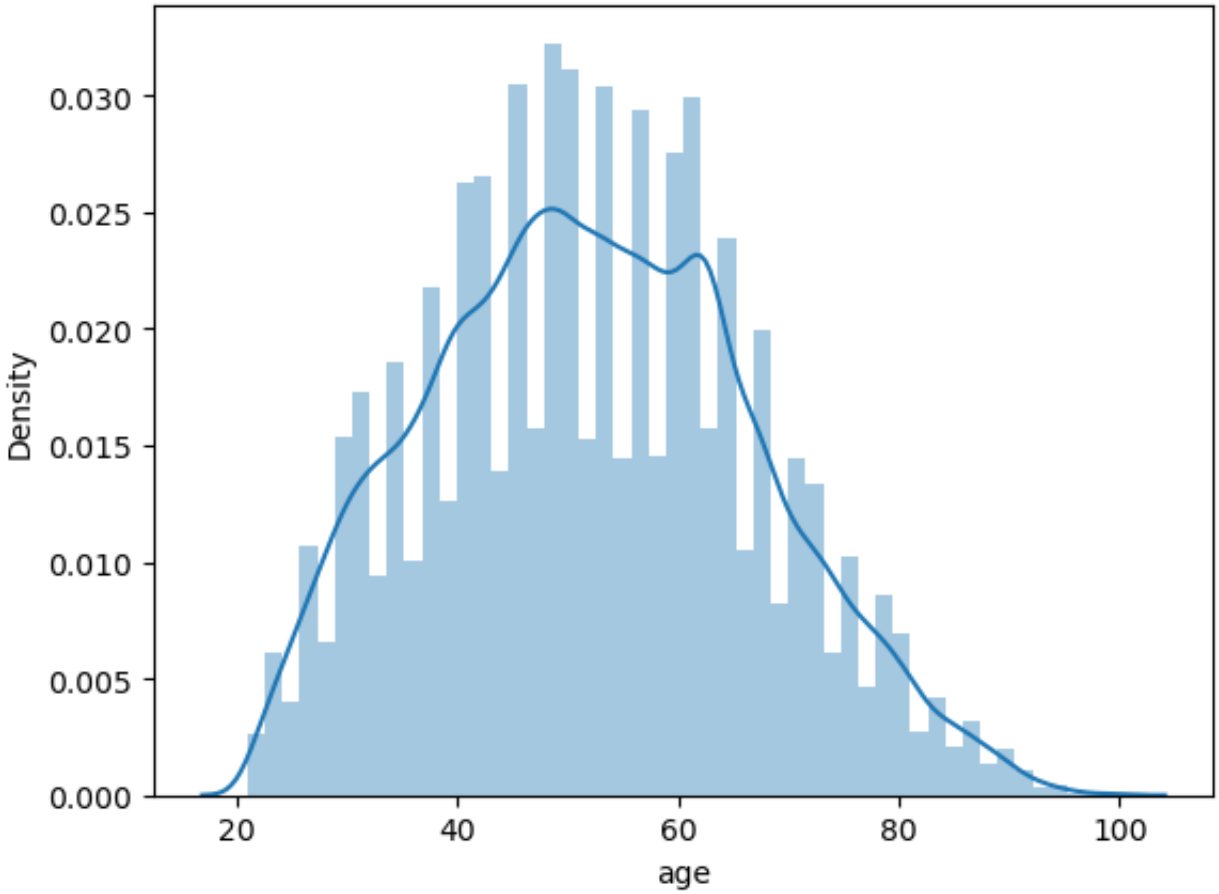
Upon conducting the T-test, we obtained a p-value of 0. This result led us to reject the null hypothesis (H_0), indicating that the age observations over 100 should be considered outliers.

The subsequent step involved deciding on an appropriate method for addressing these outliers. Initially, the removal of these observations was considered. However, after careful deliberation, this approach was not pursued. It was acknowledged that individuals over 100 years old could still provide valuable insights into the dataset. Furthermore, eliminating a significant number of observations could potentially be more detrimental to the efficacy of our model's learning process than beneficial. Therefore, an alternative strategy for managing these outliers was sought, prioritizing both the integrity of the dataset and the robustness of the model.

In addressing the identified outliers within the 'age' variable, we opted for a Winsorization technique. This method involved substituting the outlier 'age' values with the 95th percentile value of the dataset. This decision was informed by our objective to maintain the integrity of the dataset while mitigating the impact

of extreme age values.

Subsequent to this data cleansing process, the distribution of the 'age' variable exhibited a more normal appearance. This adjustment enhanced the overall quality and reliability of the dataset, ensuring a more accurate representation of the age variable for further analysis and modeling.



2.2 Monthly Income

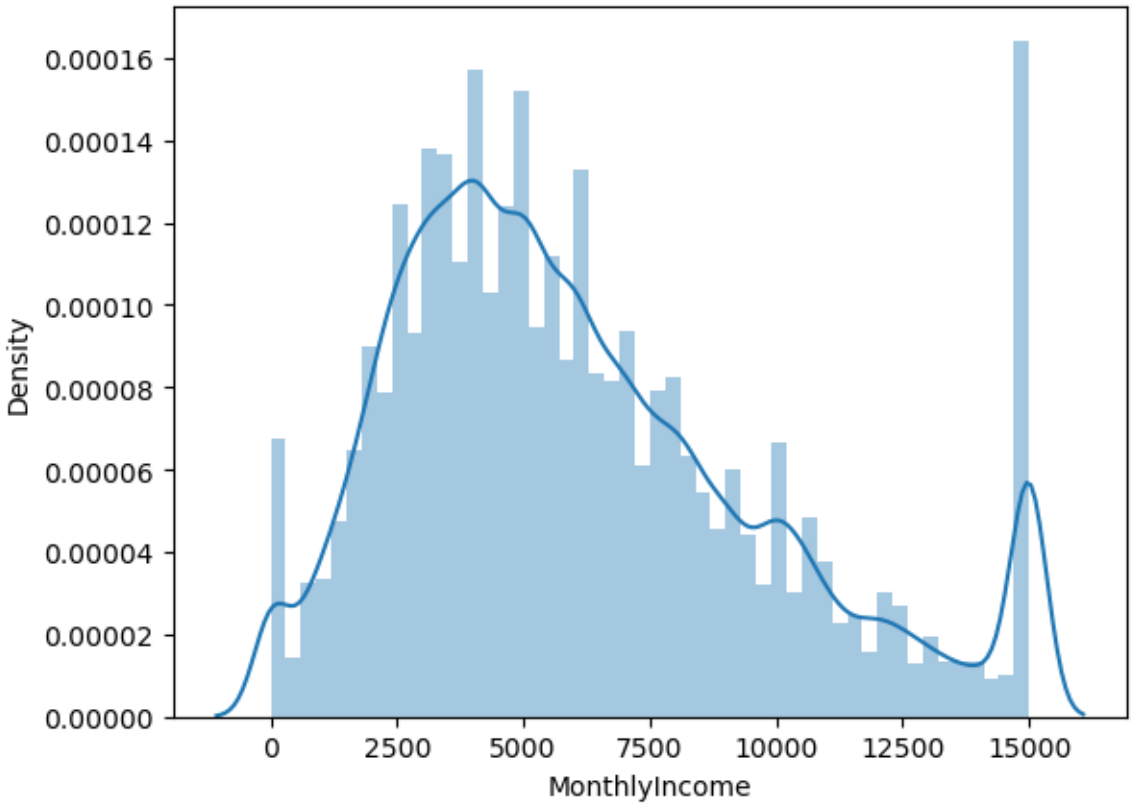
In our dataset, the 'monthly income' variable exhibited characteristics similar to those found in the 'age' variable, notably the presence of extreme outliers and a substantial number of missing values.

To address the issue of outliers, we employed the Winsorization technique. This approach was particularly suitable as high-income outliers have the potential to disproportionately influence the mean and other key statistical measures. By applying Winsorization, we effectively limited the impact of these extreme values. This adjustment resulted in a more balanced average income, significantly reducing skewness caused by unusually high values. Furthermore, it is important to consider that extremely high incomes might not accurately reflect the financial status of the majority of credit applicants. Thus, Winsorizing the data helped in aligning it more closely with the financial realities of the broader population.

For handling missing values, a segmented imputation method was chosen. This decision was based on the contextual nuances of the dataset. Initially, the plan was to impute missing values using the median monthly income. However, given the skewed nature of the data and the variability of income across different age groups, a more nuanced approach was deemed necessary. Therefore, we opted to impute missing values based on the median monthly income corresponding to the age of each observation. For instance, if an observation was missing a monthly income value and the age was 35, we imputed the median income for all

entries with age 35.

While the distribution for the ‘monthly income‘ variable did not end up being completely normal, we were still able to reduce some of the severe right skew in the data.



2.3 Dependents

For the 'dependents' variable, we encountered several atypical and missing values. To address the atypical values, we implemented a Winsorization process, capping them at the 95th percentile. This decision was based on the goal of minimizing the influence of extreme data points while preserving the dataset's overall integrity. Regarding the missing values, we elected to apply a straightforward imputation technique, substituting them with the median value. This method was deemed appropriate as the distribution of the 'dependents' variable approximated normality following the Winsorization process.

2.4 Number of Times 90 Days Late

For the examined variable, although there were no missing values, a considerable number of substantial outliers were observed, with some records showing up to 100 days late. While such extreme values are plausible within the given context, their potential to adversely affect the model's performance was a concern, particularly given their volume. Therefore, we opted to Winsorize the variable, setting a maximum limit of 3 days late, reflecting the predominant trend of observations being 0 days late. Despite this adjustment, numerous outliers remained. Nevertheless, it is important to recognize that these outliers may actually enhance the model's predictive capacity by accurately representing real-life variations in payment timeliness. The presence of outliers underscores the diversity in payment behavior, ranging from those who are consistently punctual to those who frequently incur delays.

2.5 Number Of Open Lines and Loans

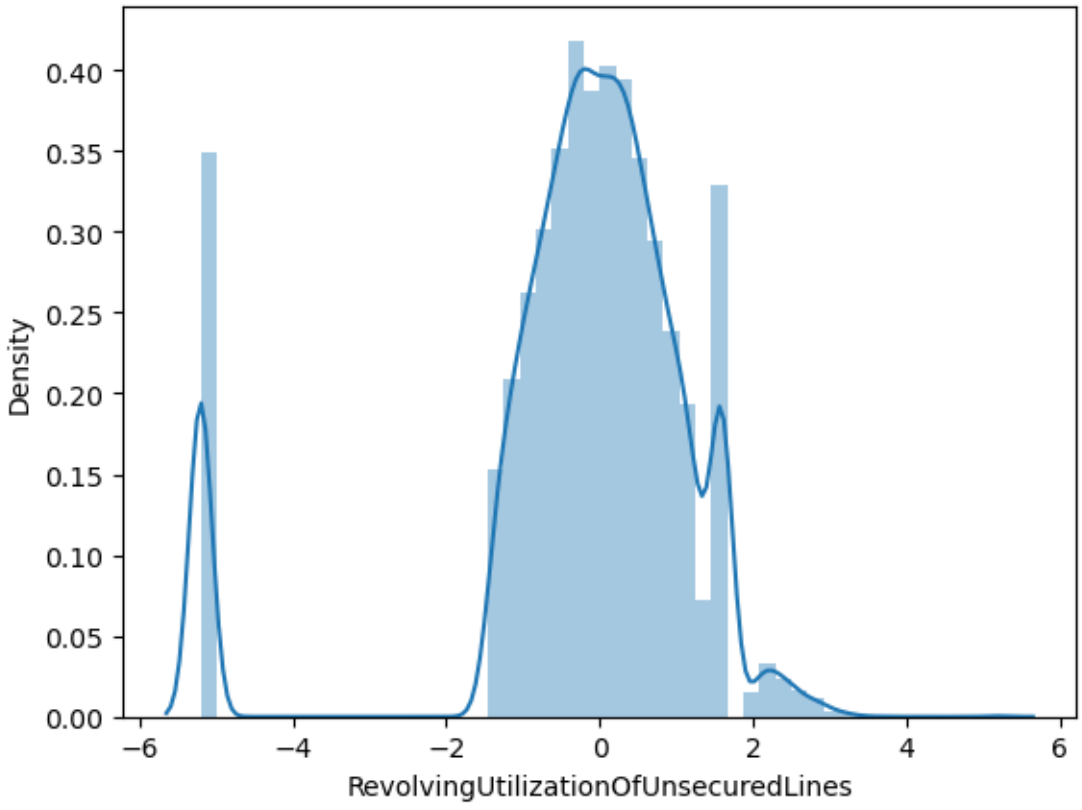
Similar to other variables with outliers we chose to winsorize this variable to the 99th percentile to reduce extreme values and also because in context those with over 24 open lines or accounts will most likely have the same result when they are applying for a credit loan. This allows the data to be approximately normal.

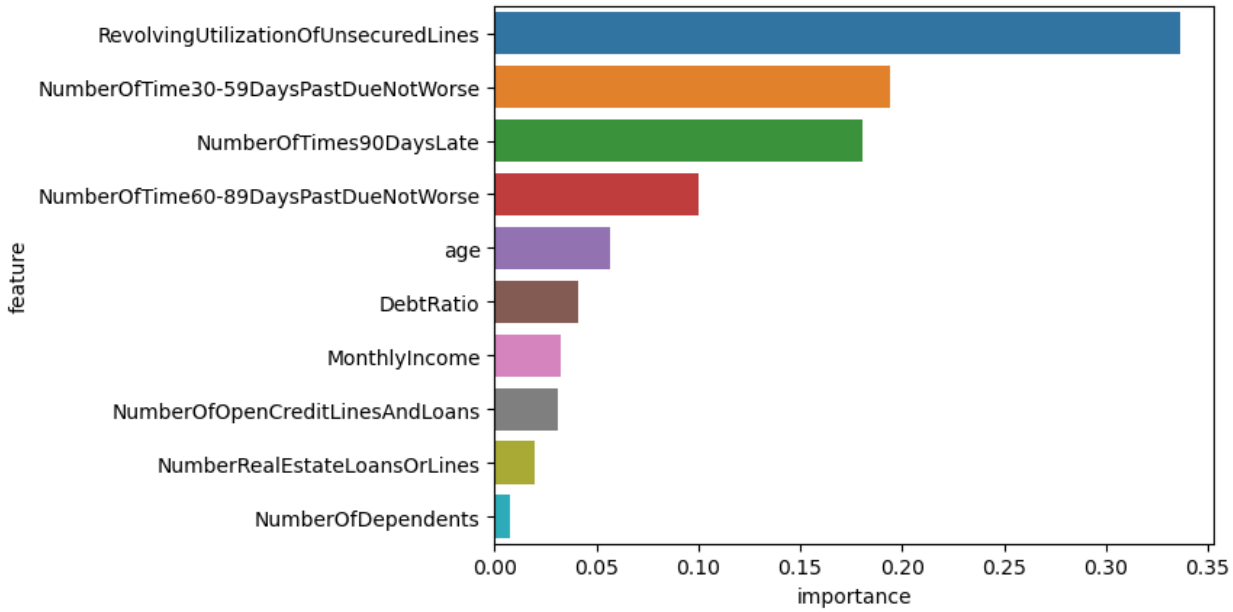
2.6 Debt Ratio

For debt ratio we noticed both outstanding outliers and that the range of the data was extremely large. As a result we chose to Winsorize to the 99th percentile and also scale the data with a robust scaler due to the presence of outliers in the data. Robust scaler both scales down the data and is more proactive against the presence of outlier points in the data. Similar to other variables, the distribution was not completely normal but in the context of the problem there are people who will have outstanding debt ratios and so these outliers may help the model.

2.7 Revolving Utilization of Unsecured Lines

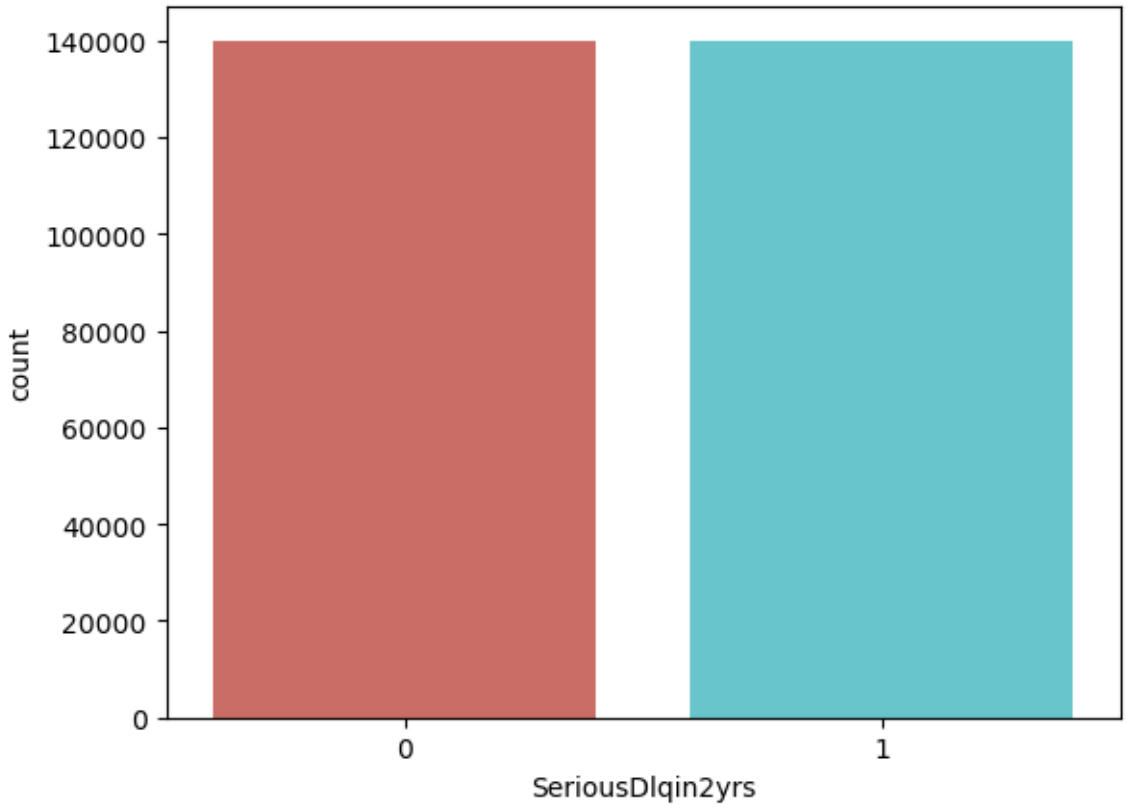
Upon examination of this particular variable, a substantial number of outliers were identified. Our initial remedial measure involved employing a Robust Scaler, which typically mitigates the impact of outliers by using more robust metrics. Despite this, the resultant distribution demonstrated minimal change, with the data range and the presence of outliers remaining largely unaltered. Consequently, we proceeded to implement a Quantile Transformer. This method is adept at managing datasets with numerous outliers, as it recalibrates the feature distribution to follow either a normal or uniform distribution, determined by quantiles. This process effectively diminishes the influence of outliers by evening out the most common values. This method was chosen because when feature importance was plotted, this variable was found to have the most impact on the model and as a result any loss of data would impact the model.





2.8 Data Imbalance

An examination of the class frequencies in the training data revealed a pronounced imbalance: rejections significantly outnumber approvals. Such a disparity, known as Data Imbalance, can skew the model's predictive performance, potentially leading to a bias towards more rejections over approvals. To counteract this imbalance, we adopted an up-sampling strategy. This technique involves random sampling with replacement from the minority class to augment the data, thereby striving for a more balanced dataset.



3 Model Building

Two types of models were chosen to attempt to classify observations in the dataset: Random Forests, and a simple Neural Network model.

3.1 Random Forests

The random forest model was chosen to attempt to classify the data into its correct categories. In order to chose the optimal parameters for the model we utilized Gaussian Process Optimization with Cross Validation to chose parameters while attempting to avoid over-fitting the data. Ultimately the best parameters found were: N-Estimators: 193, Maxdepth: 11, and Min-Leaf-Samples: 1. The AUC score achieved using this model in the Kaggle Competition was: 0.86385

3.2 Neural Network

Out of curiosity we trained a Neural Network with the following architecture:

1. Dense (256 - Relu)
2. Dense (128- Relu)
3. Batch Normalize
4. Dropout (0.5)
5. Dense (64- Relu)
6. Batch Normalize
7. Dense (32 - Relu)
8. Dropout (0.3)
9. Dense (1 - Sigmoid)

The Architecture was chosen in order to allow the network to have enough complexity to learn patterns in the data as well as to regularize itself against overfitting (Batch and Dropout Layers). The final layer contained a sigmoid function since the desired output for the competition was in probability and because this was a binary classification class rather than a multiclass. The final AUC score achieved by this model was 0.85325