Data Mining for Business (BUDT758T-0507)

Project Title: Predicting Credit card Fraud using R

Team Members: Saswati Mohanty, Vishal Vindyala, Sara Kamal, Sujeeth Ganta, Urshi Senthilnathan

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

Typed Name as Signature
Saswati Mohanty
Vishal Vindyala
Sara Kamal
Sujeeth Ganta
Urshi Senthilnathan

II. Executive Summary

Please summarize in one or two paragraphs your principal findings. You should explain – using plain, non-technical, language – what your research was about, and what the significance of your study is. After that, describe your key findings. Explain what is novel and interesting about your study, and what purpose it serves.

We found our dataset on Kaggle - records of credit transactions over two days including both regular and fraudulent transactions. Fraudulent transactions are classified as "Class = 1" while regular transactions are classified as "Class=0". The dataset contains many variables - information that was collected about each credit transaction to help determine any correlation between certain factors related to the transaction and the transaction's status as regular or fraudulent. Each variable is assigned a name and the real names of the factors measures are unavailable.

Group 19 hoped to use data mining techniques to understand if any of the unlabeled variables have a stronger correlation with fraudulent transactions. If we can help predict fraudulent transactions accurately, we can minimize the misclassification costs as well as perhaps prevent and reduce fraudulent transactions.

From the four models we used described in more detail in this report, Naive Bayes had the highest recall. This model also has the highest misclassification cost.

III. Data Description

Our data source came from the website Kaggle.com. The dataset contains transactions made by credit cards in September 2013 by European cardholders.This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The data has variables named V1-V28 that are unlabeled, these numerical input variables are the result of a PCA

transformation. Features V1, V2, … V28 are the principal components obtained with Principal Component Analysis transformation, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. All variables are numeric except for Class which assigns a discrete class to variables to identify fraudulent transactions.

The sample size is 284,807 transactions and the number of variables is 31.

The number of credit card fraud cases has increased exponentially with the advancement of many modern technologies, globalization, and the rapid expansion of economies. Credit card fraud has huge costs for users of many credit cards and the issuers of these cards. The financial companies pay billions of dollars annually in credit card fraud costs. For many companies, losses involving transaction fraud amount to more than 10% of their total expenses. The concern with these massive losses leads companies to constantly seek new solutions to prevent, detect and eliminate fraud. Therefore, we feel this is a very important area of concern and decided to apply our classroom training to try to identify fraudulent transactions from non fraudulent transactions efficiently. We attempted to run multiple models that are capable of detecting fraud in credit card operations, thus seeking to minimize the risk and loss of the business. The biggest challenge is to create a model that is very sensitive to fraud, since most transactions are legitimate, making detection difficult.

Data:
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

III. Research Questions

Describe the questions that you plan to investigate using your data. Use the terminology of the business area that is relevant (not statistical hypotheses). These should include questions that are related to classification, prediction, etc.

- Datasets with credit card transactions are very imbalanced with almost 99.5% of non-fraudulent transactions, training a model on identifying fraudulent transactions would require different sampling techniques. Identifying the key technique will be the first step.
- Furthermore, finding the best machine learning algorithm which can correctly identify the most fraudulent transactions in the dataset mitigates this prevailing menace in payment systems.
- Credit card transactions are the emerging means of e-commerce payments and financial institutions want to encourage this payment method. Frauds introduce a great challenge and untimely response to the problem can cost the financial institution their customers as well as revenue. Thus identifying the revenue losses due to these fraudulent transactions is our focus.
- All the data throughout the distribution is highly varied with some transactions more important than the other, so coming up with different cost-sensitive machine learning techniques based on calculated cost cases (as in above point) is the next step.
- Finally summing up with the best machine learning model will not only have a high fraud detection rate but also would have the most cost effective misclassification cost measure.

IV. Methodology

We used multiple techniques to see what relationships we could establish between independent variables and the dependent class variable. Listed below are the following data mining techniques with our rationale:

1. **Classification Trees:**

   Classification trees involve trying to narrow in on the most important predictors for the dependent variable. We created unpruned trees as well as prunes them at 15 and 5 to see what was the most accurate. All the trees had the same accuracy and a low error rate.

2. **K Nearest Neighbors:**

   We use the default k at k=5, This implies that we tell R to look at the 5 nearest neighbors of each transaction and classify it as fraudulent or non fraudulent. Since the majority of the transactions are non fraudulent, KNN tends to assign the majority class to all predictions.

3. **Naive Bayes:**

   Naive Bayes is a model that performs most efficiently with a large set of data, even if number of observations is less than the number of predictors. Luckily, we had quite a large number of observations albeit an imbalanced dataset.

   Naive Bayes can only be used for classification. We partitioned the dataset into training and test datasets - a 75-25 split.

   The assumption of independence is in fact "naive" since its unlikely that all 31 columns are independent to each other. Since most of our variables are unlabeled, we went ahead with this assumption.

   But due to lower threshold level there are also high number of false positives which might lead to increased misclassification costs if costs are associated to false positives

   We use a cutoff value of 0.4 to classify fraudulent activities with a train test split of 75%-25%.

   The lift chart lets us compare our model against random guessing and other models.From the lift chart the naive bayes performs way better than random guessing and other models used in this project. The model performs much better than the reference line for random guessing. The tradeoff being sensitivity and specificity is good but there is risk of over fitting.

4. **Association Rules:**

   We also looked into association rules to search for any relationship between certain variables or confirm that the rules included variables that other models had also highlighted as strong predictors.

   Unfortunately, the model produced 3304 rules with support, lift, and confidence values that were not indicative of reliable rules. Since the variables are also unlabeled, it was hard to draw any conclusions about the relationship between factors affecting transactions.

5. **Boosting**

Boosting works on any dataset to "upweight" misclassified data and hence accounts for misclassification costs. The process is repeated continuously till the model has a weighted sum of individual classifiers. Better predictors are of course assigned a higher weight. We have used 500 trees and tree depth of 4 to get the metrics here

6. **Logistic Regression**

   The Logistic regression model uses a logistic function of the dependent variable to calculate the probability of the dependent variable being =1 . The glm function in R allows us to model the logit - the function of our dependent variable by converting the probability into odds and then taking the log of odds.

   The logistic regression helped us identify the most significant variables out of all the independent variables and is able to take both numerical and categorical predictors.. Most of these variables are in common with those identified in the classification tree model. The model had very high accuracy but a very low recall. .

   Logistic regression does struggle with a large number of predictors which we can see with the low recall we calculated in this case. We would need to use regularization methods to improve performance.

Calculating Misclassification costs

- Card fraud transactions stored by financial issuers are very small compared to legitimate transactions, which results in a high imbalance credit card dataset.
- The process of determining misclassification costs for fraudulent transactions depends on the nature of the transaction as well as the financial institution.
    - Factors to consider include losses of lenders' earnings due to misclassifying a potential fraudulent transaction as a regular one. The amount of the transaction may also play a role since the loss to the financial institution is equal to the amount of money lost.
- Banks have in the past represented the "costs of misclassification by the administrative cost *Ca*, which is related to analyzing the transactions and contacting card holders. In the cases of false negative and true negative, the associated costs both equal to *Ca* because the card holder will have to be contacted. However, in the case of false positives, due to

the fact that frauds are not detected, the cost is defined as a hundred times *Ca*."

- If they misclassify a non-fraudulent transaction as fraudulent, they tend to associate minimal to no penalty with it.

Cost Matrices:

CostMatrix1:

```
> Costmatrix
          0        1
0  0.00000 88.34962
1 88.34962  0.00000
```

CostMatrix2:

```
> Costmatrix2
          0        1
0     0.00 88.34962
1 25691.16  0.00000
```

CostMatrix3:

```
> Costmatrix3
          0 1
0     0.00 0
1 25691.16 0
```

V. Results and Finding (varies considerably in length depending on study)

Name of Model	Accuracy	Recall	MC mean	MC mean & max	MC zero and mean
Classification Trees	0.9992135	0.7317	4947.57	849840.321	847808.3
Classification Trees Pruned at 5	0.9992135	0.7317	4947.578	849840.321	847808.3
Classification Trees Pruned at	0.9992135	0.7964	4938.578	593812.217	590896.7

15					
Logistic regression	0.999073	0.54471	5830.887	1439588.456	1438705
Boosting	0.9991	0.766	5124.279	721352.48	719100.23
KNN	0.9983849	0.065	10160.21	2954483	.2954483
Naïve Bayes	0.9752	0.7804	155,872.93	823662.87	693661.3

We then used under-sampling to have a 50:50 distribution of the dataset with 492 fraud and 492 non-fraudulent data. Then we ran the Naive Bayes classifier as it was the best classifier we got and did a 75:25 split in test and train to achieve the following performance measures:

```
> (Acc = (cm[1,1]+cm[2,2])/sum(cm))
[1] 0.902439
> (recall = (cm[2,2]/(cm[2,1]+cm[2,2])))
[1] 0.8292683
```

We can see a **4% increase** in the recall as compared to the Naive Bayes classifier in the Base classifying step.

VI. Conclusion

With the increasing reliance on electronic payments in recent times, business owners and e-commerce companies along with other major stakeholders such as card manufacturers, payment operators, and customers (the cardholders), credit cards face increasing security threats. That calls for a regular need to increase the efficiency and effectiveness of the previous fraud detection schemes with the aim of building reliable and secure payment systems.

Recommendations:

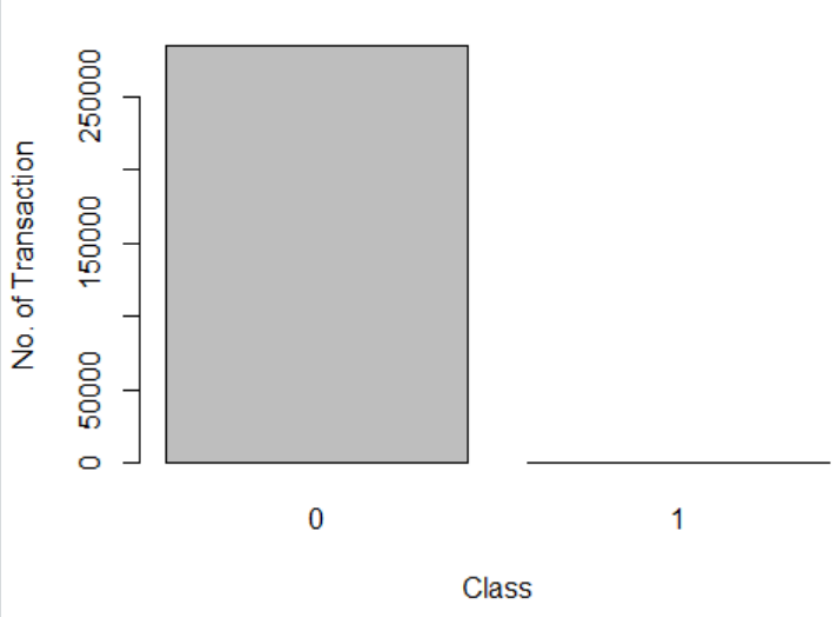
1) A better Data pre-processing tool which uses more data than our best model (984 observations data) to model the classifier as this is a very minute fraction of the data.
2) The under-sampling algorithm can be chosen more precisely as to be more logical rather than a random selection of transactions from the Non-fraudulent class.

3) Then, Account for the misclassification costs uniquely by varying the costs for each transaction as the amount for each transaction rather than a mean or max used in our use cases. This can be done by using the appropriate classifier with a varying cost matrix.
4) Finally, use ensemble methods to achieve the best predictive sensitivity ,recall and accuracy for these imbalanced datasets.

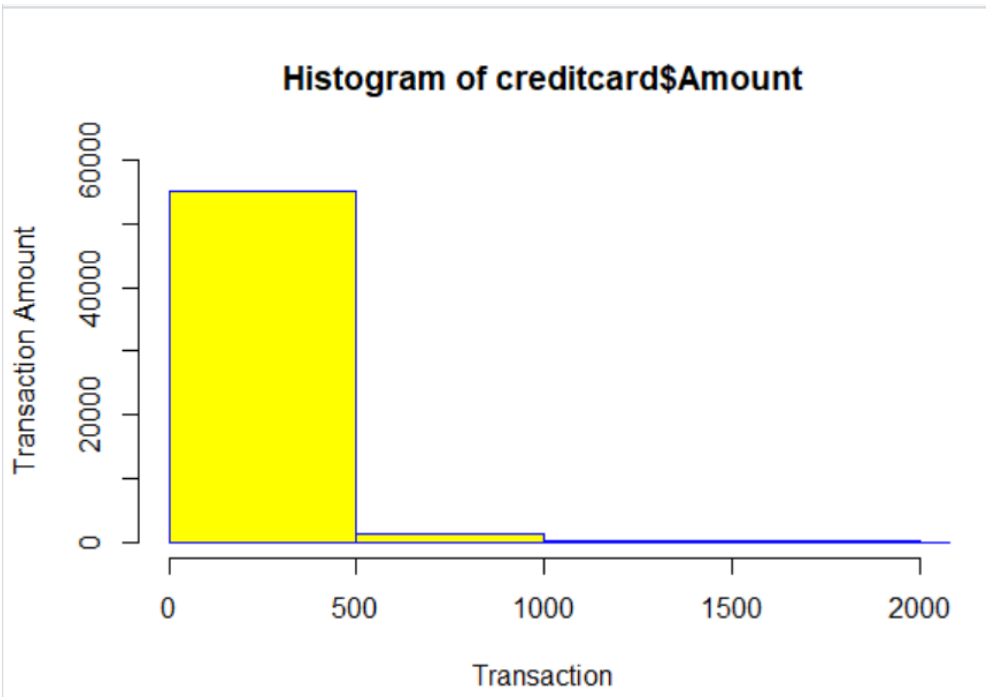
VII. Appendix (Any additional information to be submitted):

Figures:

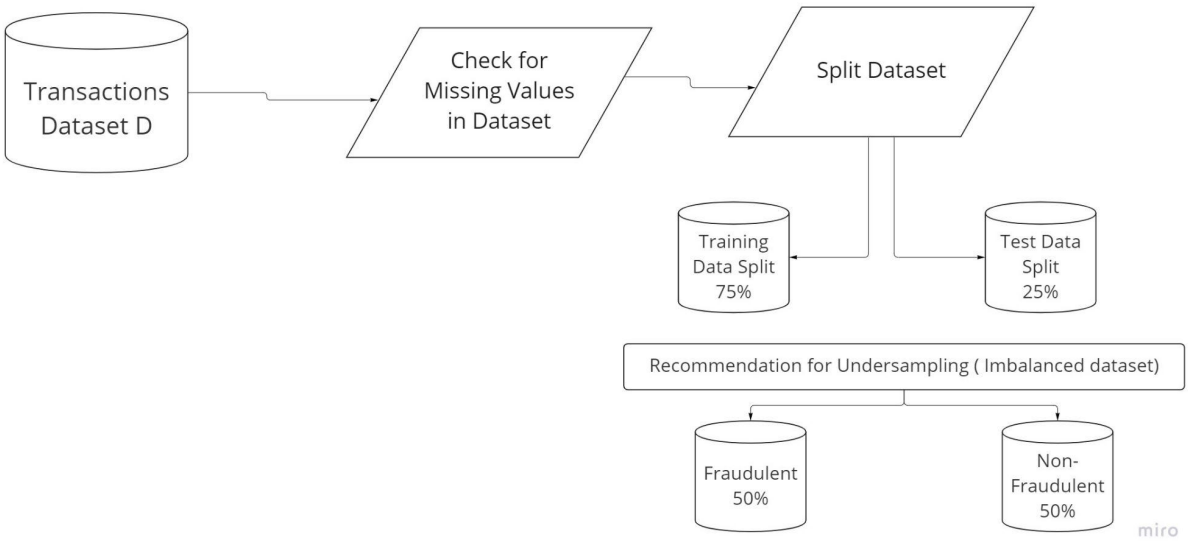
Figure(i): Class Distribution of Credit card dataset which suggests very imbalance classes



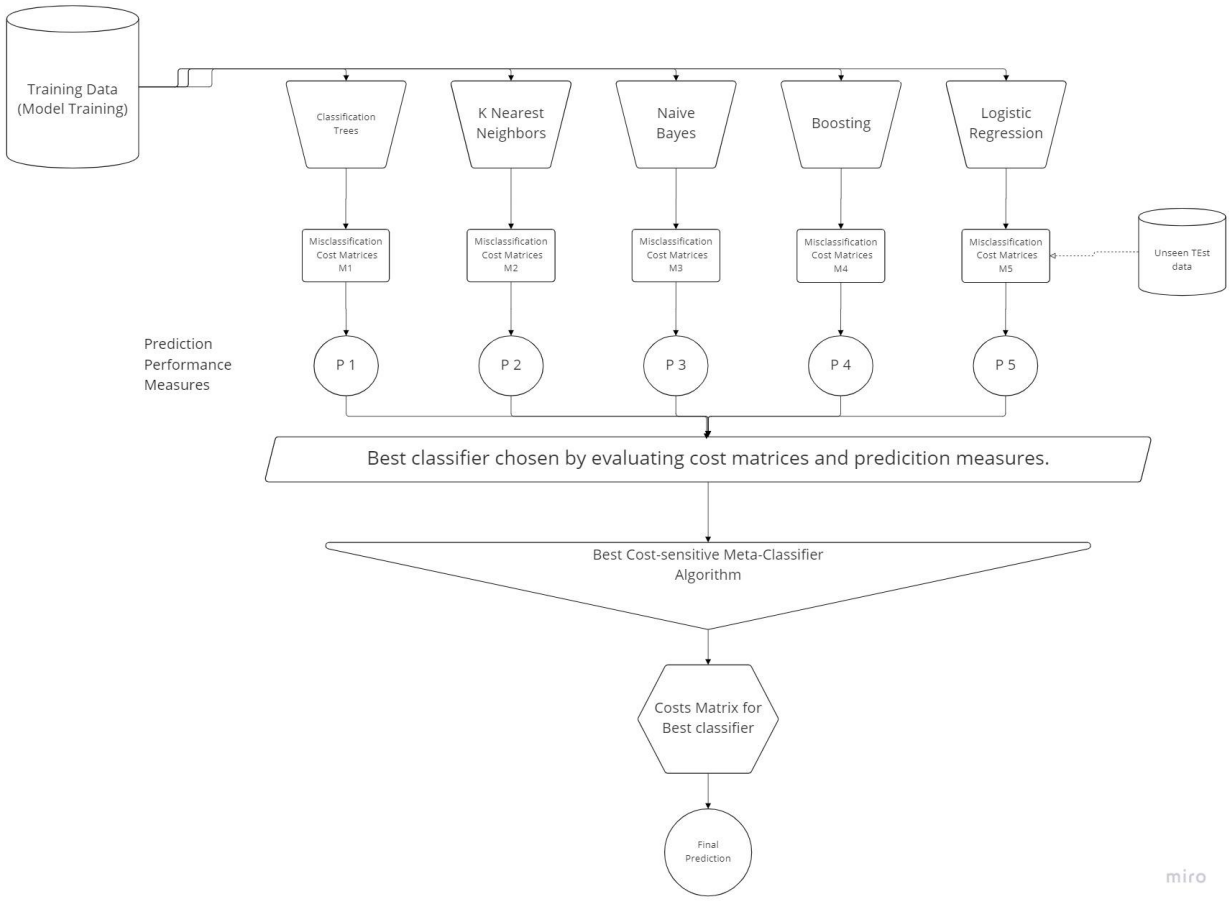
Figure(ii): Histogram of Amount for each transactions



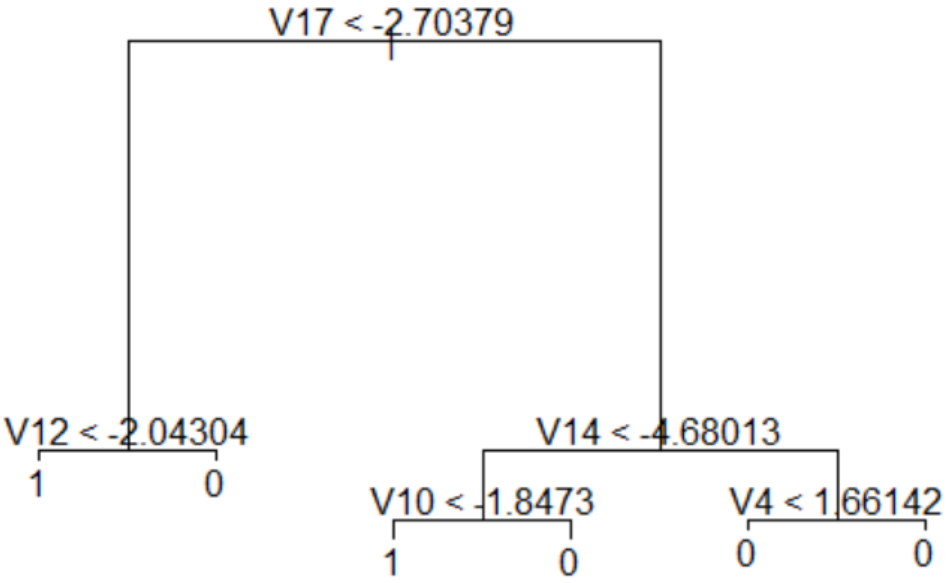
Figure(iii) Flowchart of Data Preprocessing of the Dataset



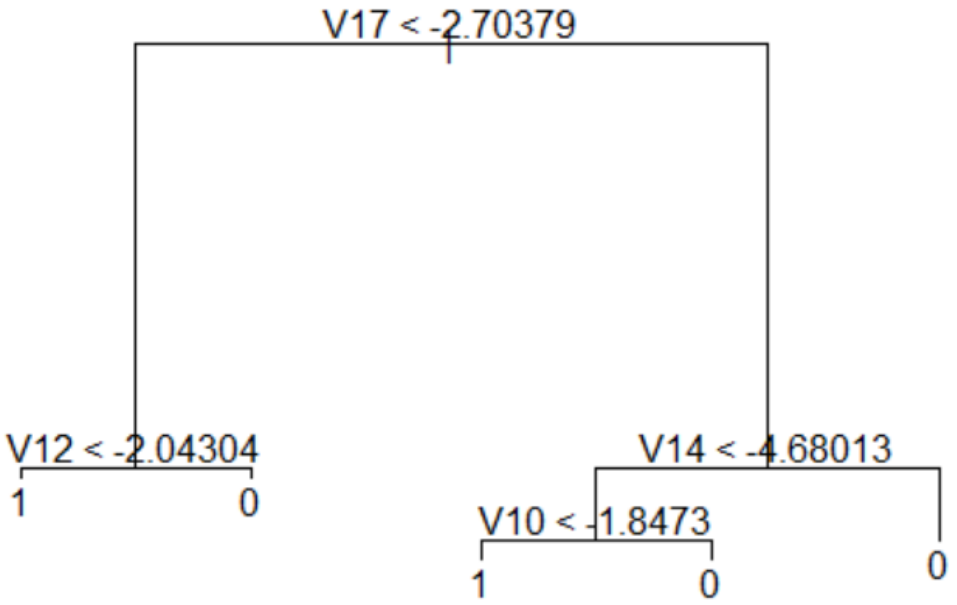
Figure(iv): Flowchart of Proposed Model Architecture



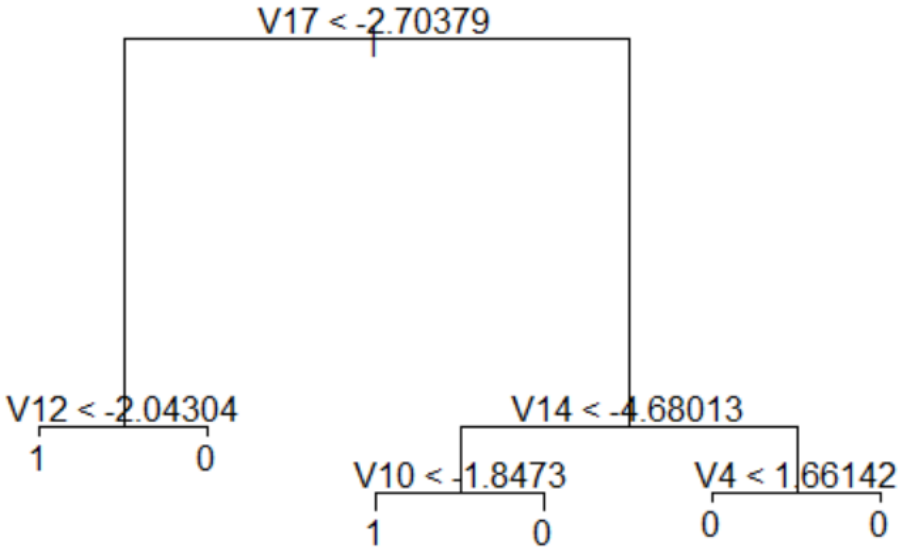
Figure(v) Classification Tree Model



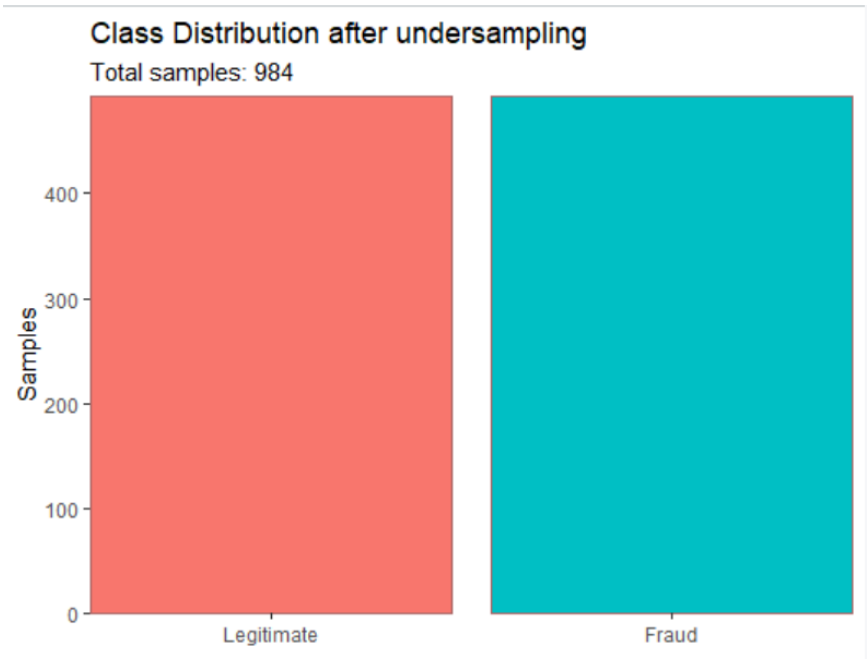
Figure(vi): Pruned Tree Model



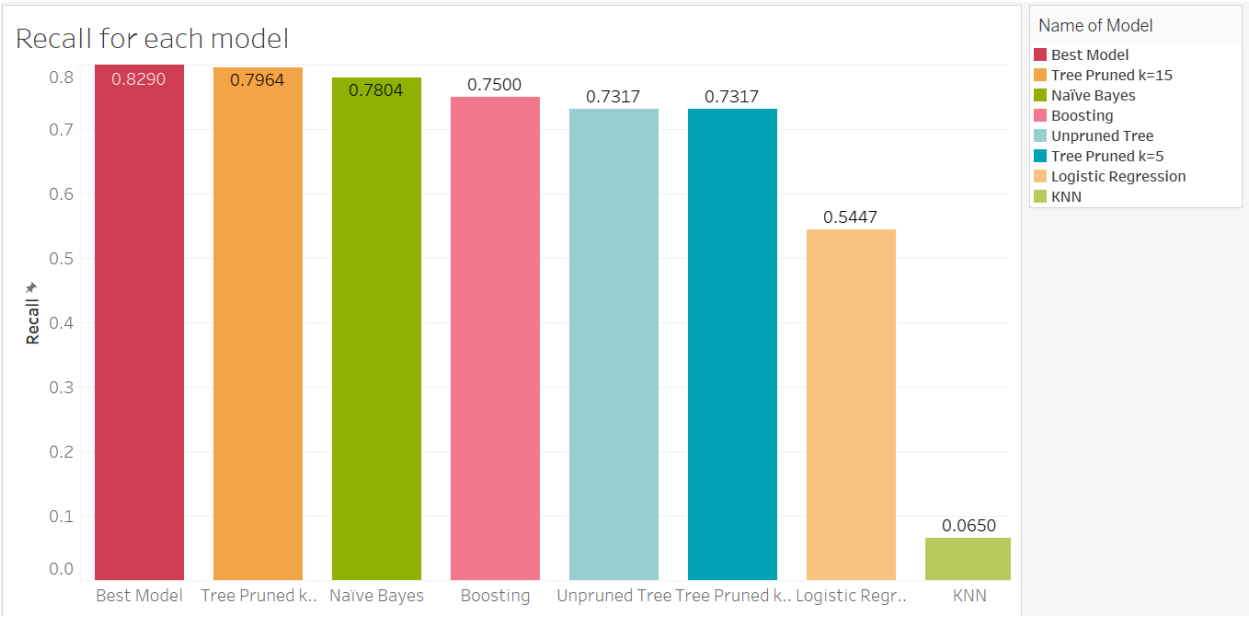
Figure(vii) : Best Pruned Tree for Classification Trees Base Classifier



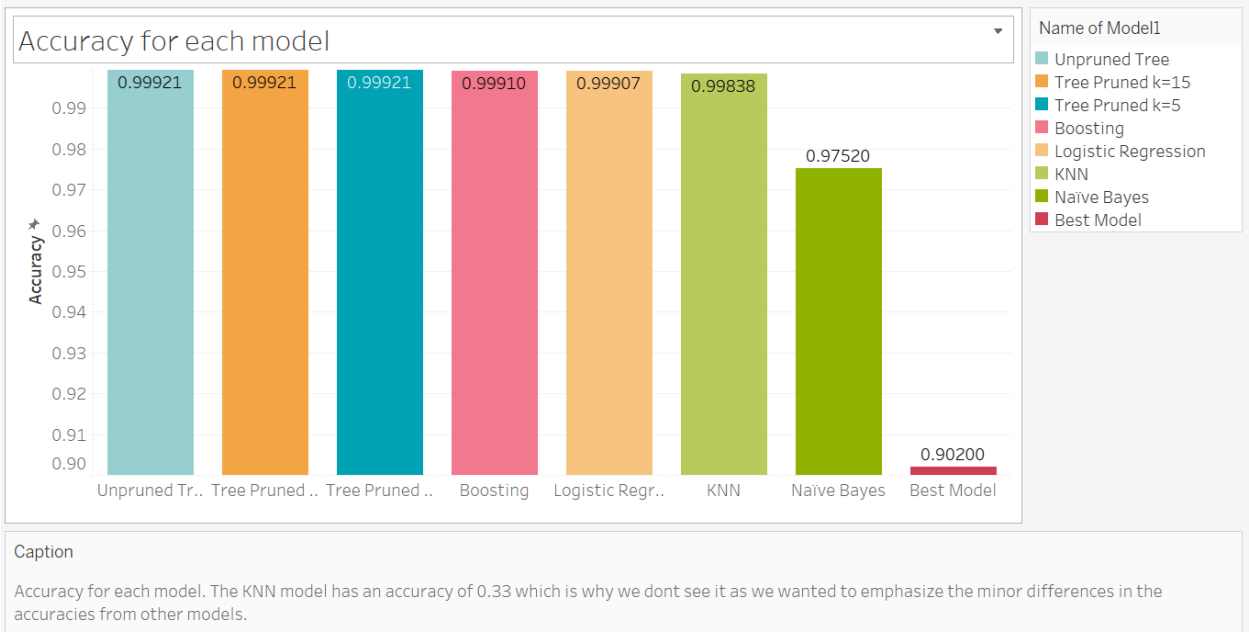
Figure(viii): Class distribution after Undersampling



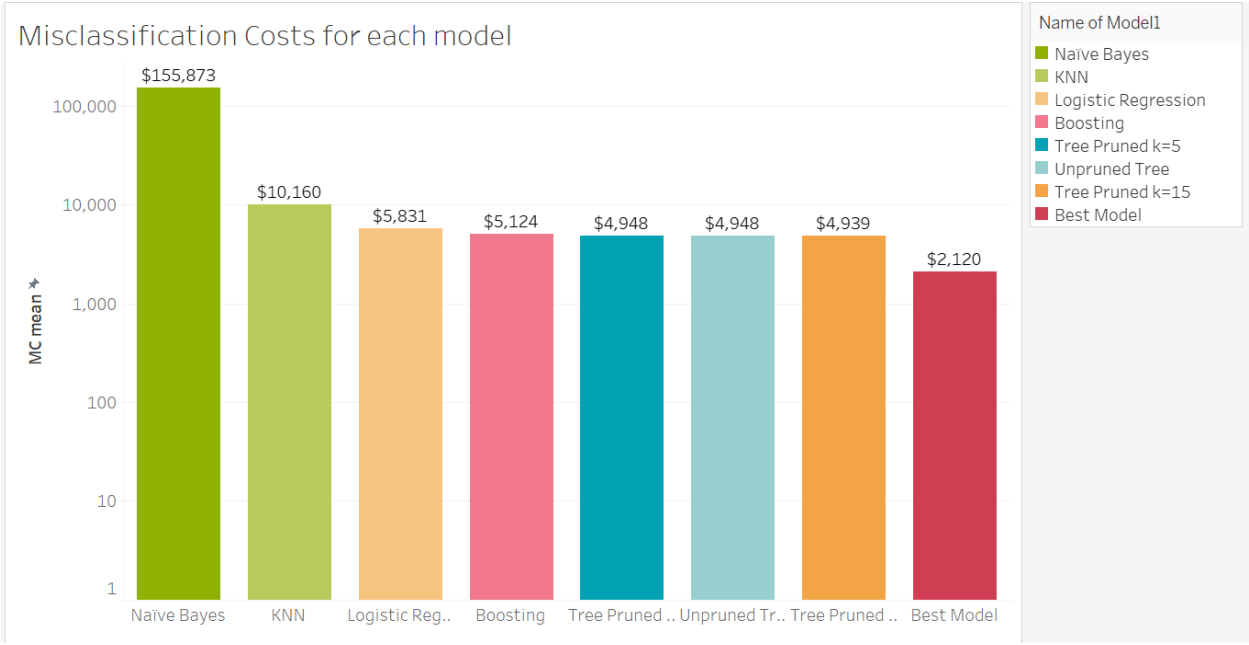
Figure(ix): Tableau Visualization of Recall



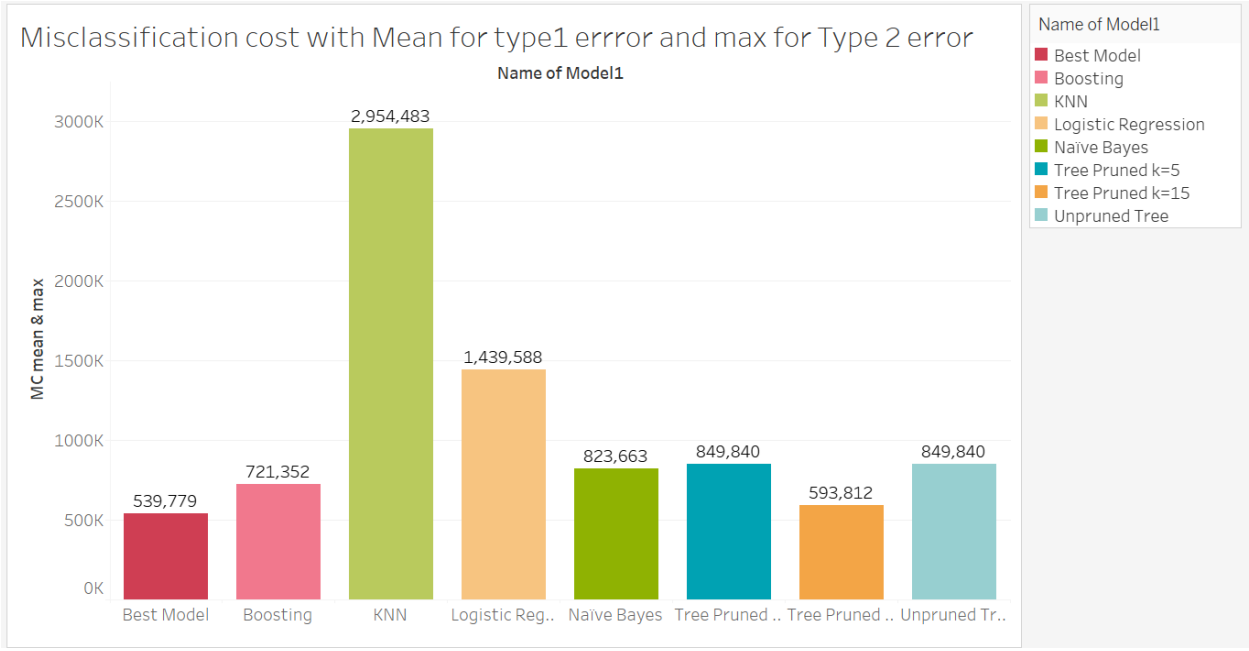
Figure(x): Tableau Visualization of Accuracy



Figure(ix): Tableau Visualization of Misclassification costs with Means for both Type 1 and Type 2 errors



Figure(x): Tableau Visualization of Misclassification costs with Means for Type 1 and Max for Type 2 errors



Figure(ix): Tableau Visualization of Misclassification costs with Zero for both Type 1 and Max for Type 2 errors(as they are the ones we want to reduce as its from the fraud activities that were not classified correctly)