Credit Card and Mobile Fraud Detection using Supervised Learning Algorithms

Deepika Jindal, Yash Sharma, Himanshu Premchandani, Nipun Diwan, Matthew A. Lanham
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
djindal@purdue.edu; sharm364@purdue.edu; himanshu@purdue.edu; ndiwan@purdue.edu; lanhamm@purdue.edu

Abstract

In this study we analyze the pattern of fraudulent transactions by making use of different predictive modeling techniques. We use the results obtained from our study to predict and prevent similar fraud cases in future. We performed feature engineering to develop new variables that helped improve our prediction. We investigated three different models: logistic regression, random forest and XG boost for training our dataset and measured the performance of each model in terms of balanced accuracy, sensitivity and specificity.

Introduction

According to latest report from Javelin Strategy & Research, “*Around 15.4 million consumers were victims of identity theft or fraud last year and the total fraud costed a whooping $16 billion.*” It is therefore paramount to enhance the fraud detection mechanisms. Our objective is to predict future fraudulent activities with maximum accuracy using the available data.

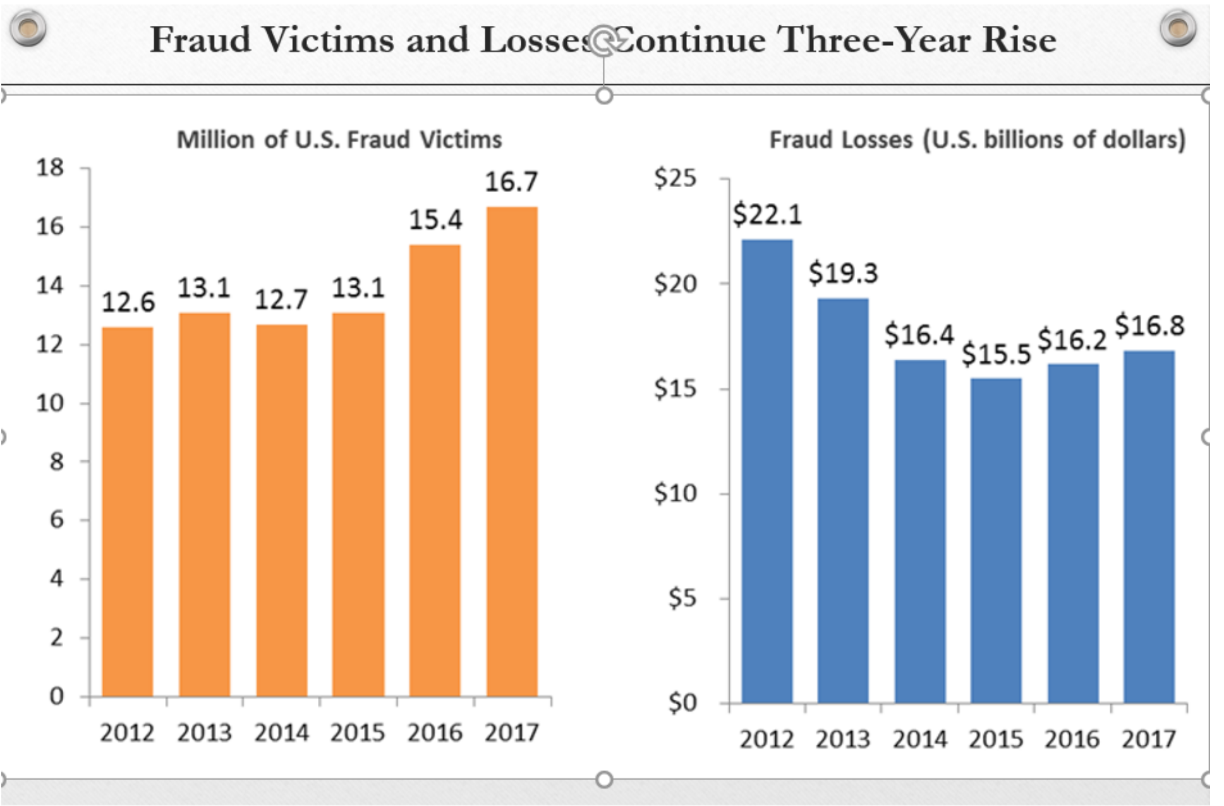
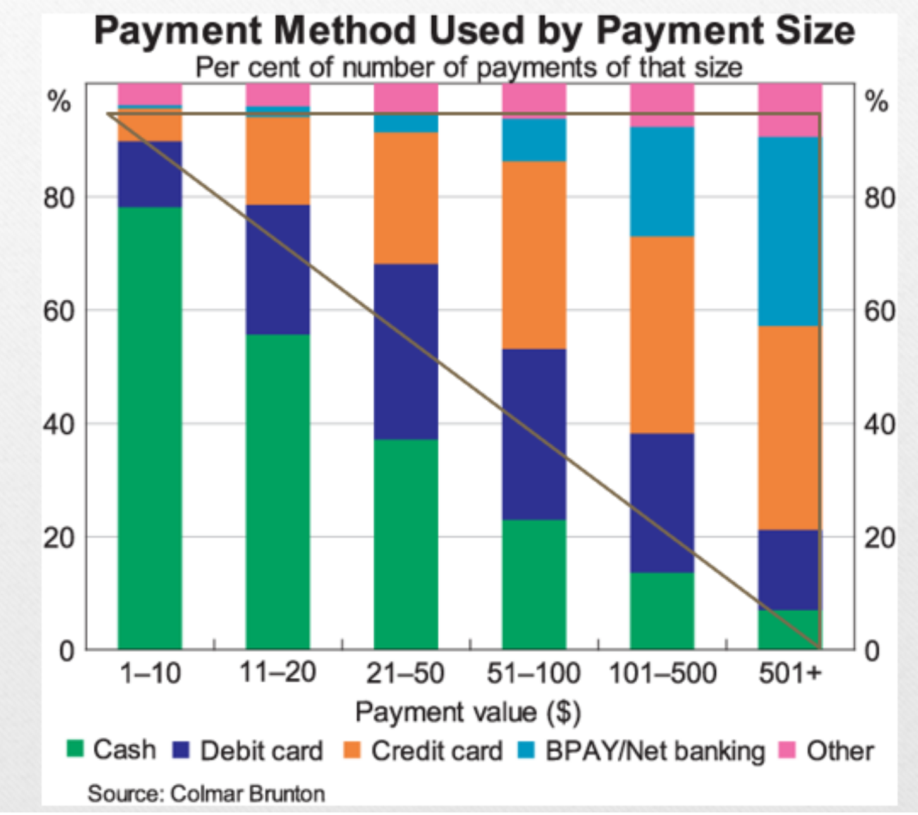


Figure 1a: Fraud trends



Source: Colmar Brunton

Figure 1b: Modes of payment

Relevance:

- Digitalization in the financial sector has made it more vulnerable towards frauds
- Increase in number of mobile and online transactions
- Extensive amount of academic literature

Research Questions:

- Which modeling technique works best for fraud detection when down-sampling the training data?
- What are the potential business savings implications from using these models to detect and prevent fraudulent activity?

Literature Review

Credit card and mobile payment fraud detection has drawn a lot of research interest and a number of techniques, with special emphasis on techniques such as logistic regression for interpretability, and random forest and XGBoost for prediction accuracy. Various methods for resampling data having unbalanced data sets has also been suggested.

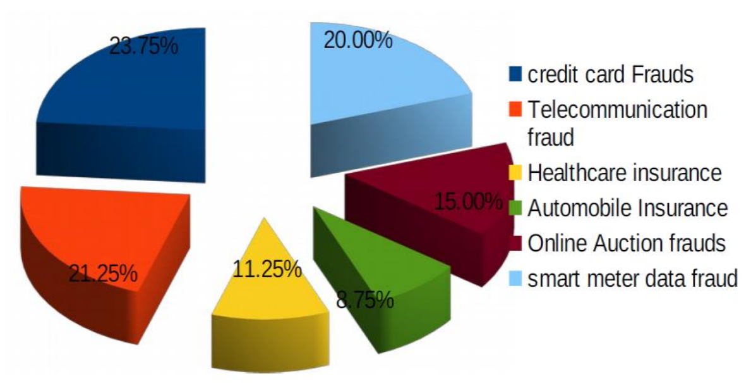


Figure 1: Number of papers reviewed in each fraud area for a decade

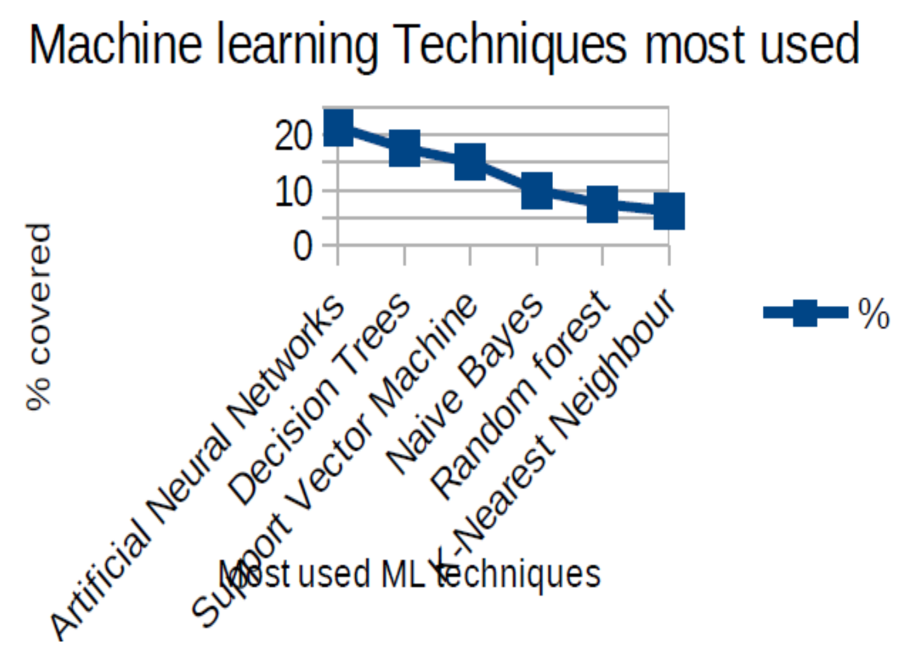


Figure 2: Literature review summary by method used

This study is novel in that we predict fraud in mobile transactions and use down sampling along with various popular models previously investigated.

Methodology

To predict fraudulent transactions, several methods have been used that include recognizing customer spending behavior, tracking network data, using advanced modeling techniques, using biotechnology-based methods etc. In this experiment we first do a down sampling to create an approximately balanced 50:50 split of fraud and non-fraud transactions.

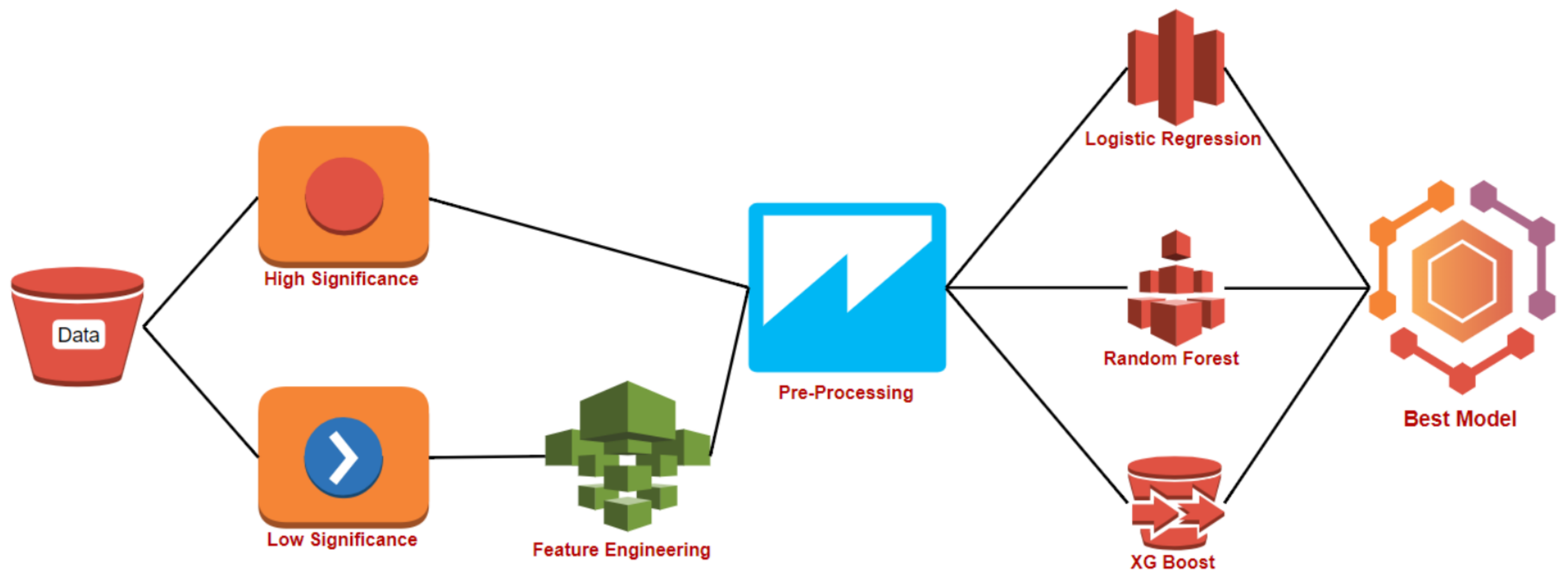
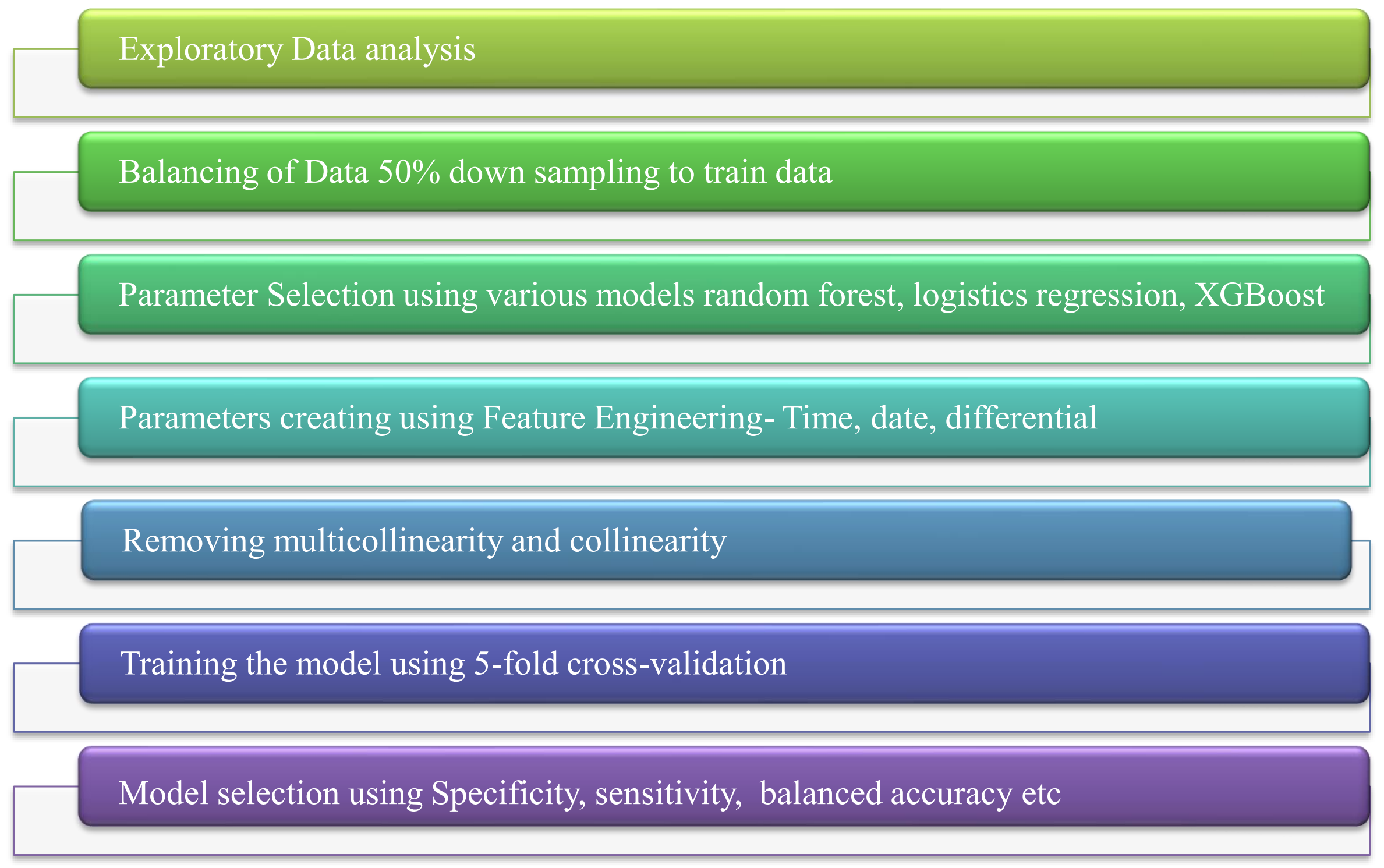


Figure 3: outlines our study design, starting from data collection, data cleaning, data pre-processing, feature creation and selection, model/approach selection and model assessment/performance measures.



Exploratory Data analysis

Balancing of Data 50% down sampling to train data

Parameter Selection using various models random forest, logistics regression, XGBoost

Parameters creating using Feature Engineering- Time, date, differential

Removing multicollinearity and collinearity

Training the model using 5-fold cross-validation

Model selection using Specificity, sensitivity, balanced accuracy etc

Figure 4: Flowchart of steps in the process

Results

We observe that the best results are obtained use the XGBoost and random forest models.

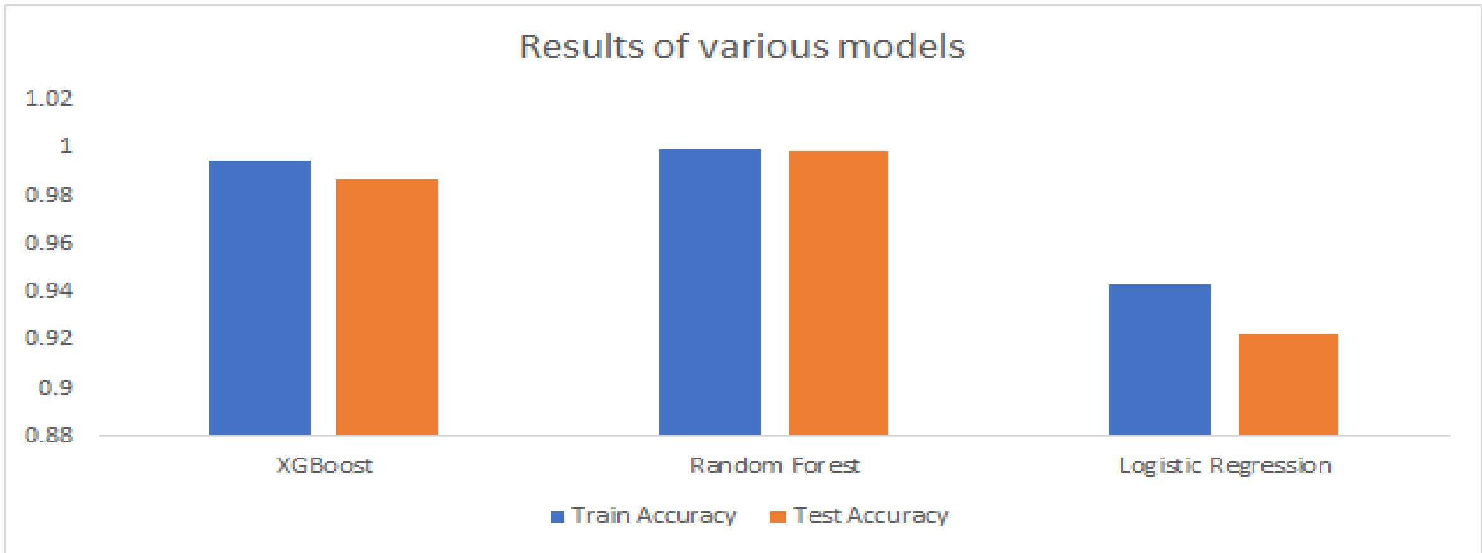


Figure 5: Model Evaluation

It can be seen from the results summary that none of the models overfit the training data based on a 10% difference in train vs. test set statistics. Based on the specificity results it can be said that 99.72% of the fraud transactions can be correctly identified. Using this model for decision-support would translate to an estimated saving of **$134,769** for a typical bank branch based on estimated daily transaction of **$300,000** at a branch..

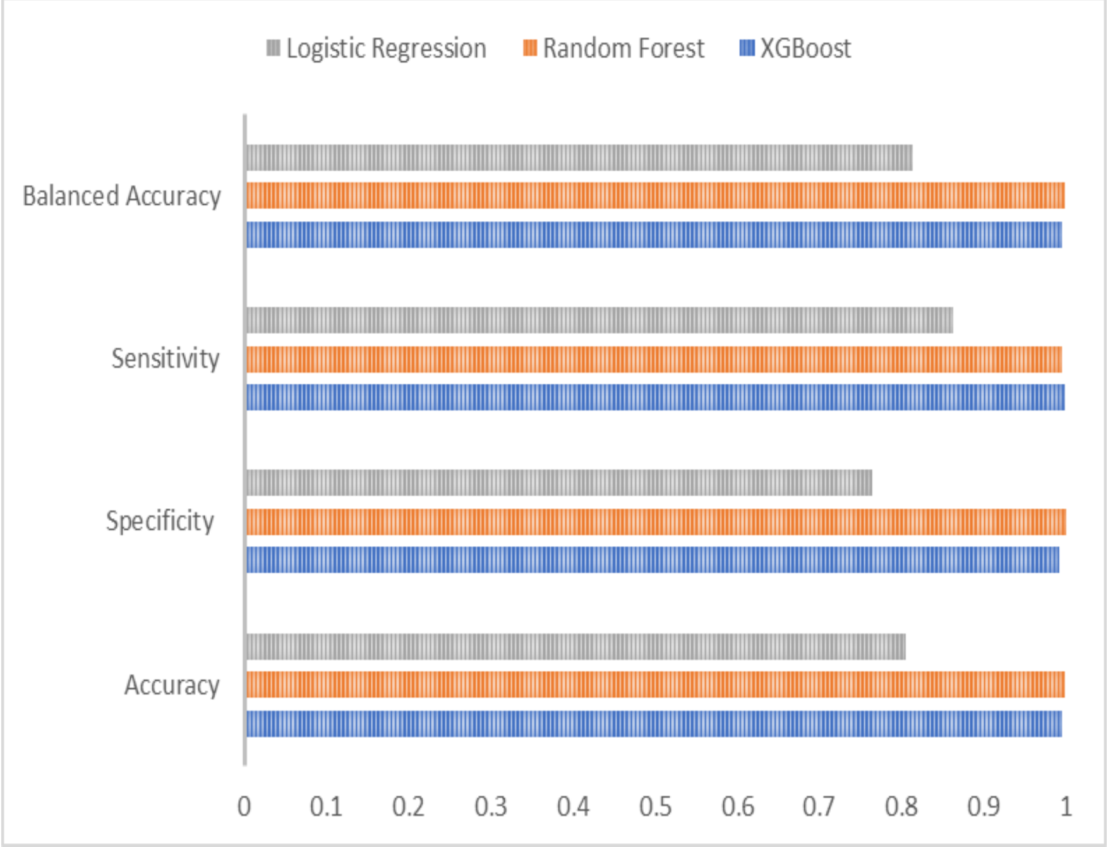


Figure 6: Model Performance

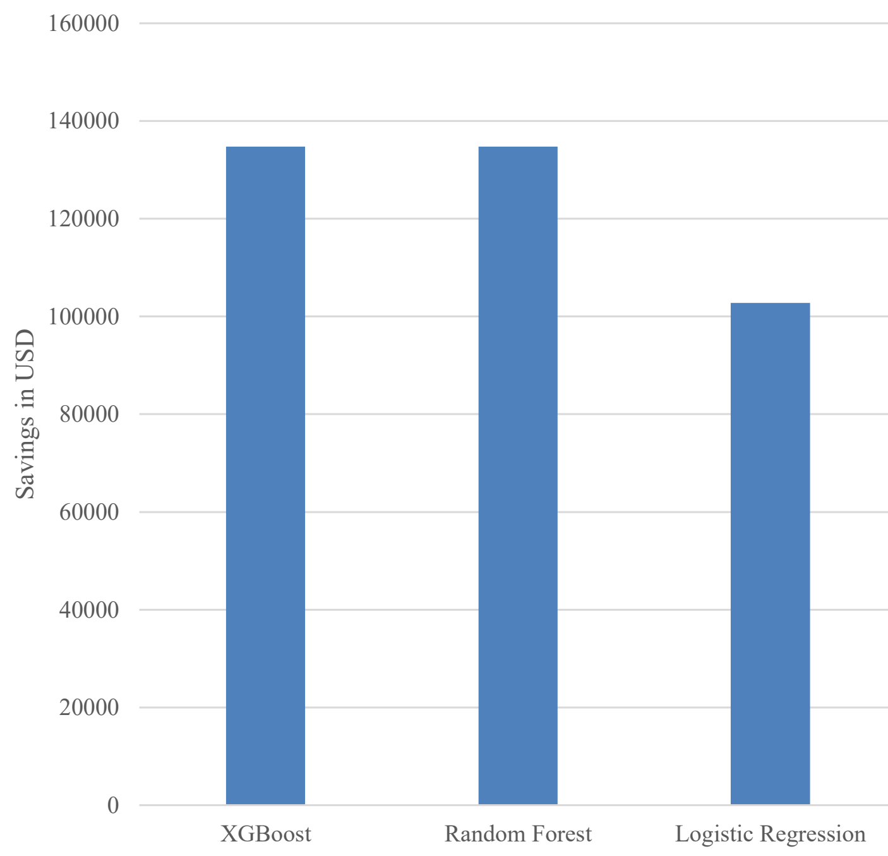


Figure 7: Comparative Savings

Conclusions

From the above research and experimentation it can be concluded that:

- **Random forest performed the best amongst the three models investigated.**
- **Logistic Regression performed the worst amongst the three.**
- **Random forest is most useful model with data spread over a very high range.**

A typical branch has potential to save approximately $**134,769** on average per year through the implementation of our model. The cost of deploying such an application has become inexpensive with advances in technology. The saving in fraud get multiplied if the model is maintained at central place and multiple branch's connect with it.

Acknowledgements

We thank Professor Lanham for allowing us to work on this project in his MGMT 571 Data Mining course.