**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](mailto:djindal@purdue.edu); [sharm364@purdue.edu](mailto:sharm364@purdue.edu); [himanshu@purdue.edu](mailto:hpremcha@purdue.edu); [ndiwan@purdue.edu](mailto:ndiwan@purdue.edu); [lanhamm@purdue.edu](mailto:lanhamm@purdue.edu)

**ABSTRACT**

In this study, we are trying to analyze the pattern of fraudulent transactions by making use of different models and using the best model for predicting and hence preventing similar fraud cases in future.

In last two decades we have observed unmatched growth in e-commerce and online banking sectors, which has made our lives so easier. But nothing comes full-proof in this world.  We lose billions of dollars every year due to financial frauds in online transactions. This has urged financial institutions to continuously improve their fraud detection systems. Several studies have proposed the use of machine learning and data mining techniques could address this problem. Recent ransomware attacks showed us that we are still not prepared for unseen challenges and we need to be one step ahead to counter these challenges.

In this paper we used dataset of 6.5 million transactions within a month, simulated synthetically by an online transaction company. We performed feature engineering to develop certain variables that helped us in our prediction. We have used three different model: 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.

**Keywords:**

Fraud Transaction; Model; Fraudulent; Random Forest; Logistic Regression; XG Boost; Algorithm; Classification; Detection; Payment; Debit; Crime

**INTRODUCTION**

Money is arguably the most sought-after commodity of exchange between individuals and organizations. Some people argue that it’s not the most important thing in the world, but we can’t deny the fact that it is sole motivation for most to fulfil basic necessities. Money has evolved over years in different forms. The time value of money has changed over years, and so does its exchange form. With digitalization, we have harnessed the power of internet to transfer money to another part of world without actually visiting the banks. Today, every business organization operate their transactions online which has expanded their businesses across traditional borders. But this easiness brings in equal amount of challenges. Hackers all around the world look for loopholes in online transaction systems to steal money. This has demanded establishments and continuous improvements in fortification of information systems handling online transactions.

According to a new report from Javelin Strategy & Research, “some 15.4 million consumers were victims of identity theft or fraud last year and the total fraud costed whooping $16 billion”.

This figure is increasing every year. McAfee and the Center for Strategic and International Studies (CSIS) estimated “the likely annual cost to the global economy from cybercrime is $445 billion a year, with a range of between $375 billion and $575 billion.”

The Identity Theft Resource Center (ITRC) recently announced its [2017 Data Breach report](https://www.idtheftcenter.org/2017-data-breaches.html) and it is no surprise that breaches are up.

*“Last year there were 1,579 data breaches exposing nearly 179 million records. That represents a 44% increase in the number of breaches and a 389% increase in records exposed.”*

It is therefore paramount to enhance the fraud detection mechanisms. In most companies fraud is detected once it has been committed. Cybersecurity departments are working on measures to prevent these incidents before their occurrence. Ideally, businesses want to find ways to prevent fraud from taking place, or, if that’s not possible, to detect it before significant damage is done. For now, most of the organizations are still developing systems to prevent the occurrence of fraud in timely manner, meanwhile, they use fraud detection systems to prevent reoccurrences of previous incidents. As new fraud patterns emerge, we come across new challenges since current fraud detection systems are limited in preventing these frauds.

In our study, we have used dataset simulated synthetically by an online transaction company. We performed feature engineering to develop variables like ‘Day’, ‘Time of Day’, ‘XAST’ which is the difference between the amount transferred and the difference of account summary before and after the transaction is complete. This helped us in creating our prediction models using three different techniques: logistic regression, random forest and XG boost. Our objective is to predict future fraudulent activities with maximum accuracy using the available data. This is a classification type of model in which we are trying to predict if an ongoing transaction is fraud or not. If it is fraud, then we have recommended some measures in conclusion section. Using analytics, we aim to achieve maximum accuracy through our model.

The remainder of paper is organized as follows: A review on the literature on various research papers and studies pertaining to fraud transaction and fraud prevention are presented in the next section. In Section 3 the proposed methodology is presented, and the criteria formulation is discussed. In Section 4 various models are formulated and tested. We have compared 3 different models and since it is a classification type of problem, we have compared it based on its balanced accuracy, specificity and sensitivity. Section 5 outlines the performance of our models. Section 6 concludes the paper with a discussion of the implications of this study, future research directions, and concluding remarks.

**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, Random forest, XG Boost, data mining and distributed data mining have been suggested.

**Logistic Regression:**

In his paper, the importance of logistic regression implementations in the Turkish livestock sector and logistic regression implementations/fields, Murat KORKMAZ highlights that *“Logistic regression analysis is one of the mostly preferred regression methods that can be implemented in modelling binary dependent variables. If there are medium-significant variables after beginning with variable selection using single variable logistic regression analysis, then multivariable logistic regression method should be selected.”*

It should be determined whether these medium-significant variables will be included into the model as continuous or discrete, and interaction between variables should be examined. Statistically meaningful interaction terms are included in the model after tested by Odds rate test, thereby the model will be defined. The relationship between the logistic regression analysis and independent variables that affect the dependent variable has been reviewed, and the analyses have been reviewed as suitable for each hypothesis on the logistic regression method. The basic form of logistic regression model is Linear Regression.

**Random Forest:**

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Freund and Schapire[1996]), but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

As mentioned by Gerard Biau [1] in his research paper *Analysis of a Random Forests Model (Journal of Machine Learning Research 13 (2012) 1063-1095)*,

*“Formally, a random forest is a predictor consisting of a collection of randomized base regression trees {rn(x,Θm,Dn),m ≥ 1}, where Θ1,Θ2,... are i.i.d. outputs of a randomizing variable Θ.”*

In our model, we used the Random forest algorithm because of the multivariate nature and the extremely expanded range of data. Hence, there is a need for a robust model that keeps a check on the range of variations our data may have. Because the model traverses through randomized base regression trees, it is one of the best fit models in this analysis. We defend our statement by mentioning the fact that transaction amounts in the data ranged from a mere $20 to a whopping $100000. To meet the ends of such a largely ranged data, Random Forest is the best fit.

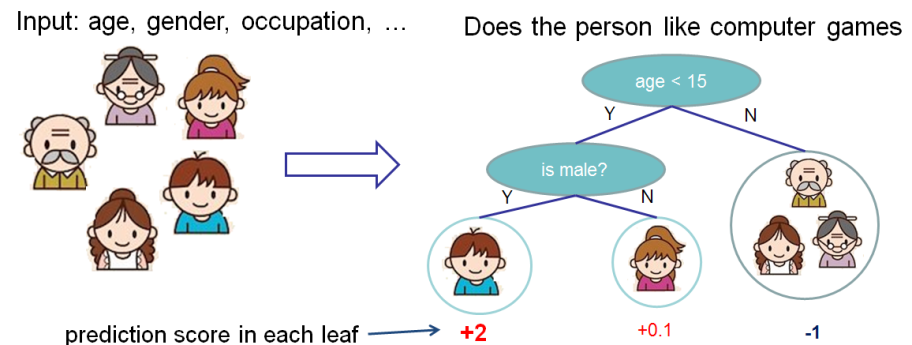
**XGBoost:**

In their research paper *XGBoost: A Scalable Tree Boosting System*, Tianqi Chen and Carlos Guestrin suggested that “[2]

*“XGBoost is a scalable machine learning system for tree boosting. The system is available as an open source package. The impact of the system has been widely recognized in a number of machine learning and data mining challenges.”*

We took the Kaggle competition, PAYSIM, a challenge to successfully predict fraudulent transactions by using machine learning techniques. On doing our analysis on a few more challenges, we found that among the 29 challenge winning solutions 3 published at Kaggle’s blog during 2015, 17 solutions used XGBoost. Among these solutions, eight solely used XGBoost to train the model, while most others combined XGBoost with random forest and Logistic regression in ensembles. The success of the system was also witnessed in KDDCup 2015, where XGBoost was used by every winning team in the top10. Moreover, the winning teams reported that the use of other machine algorithms combined with XGBoost outperform a well-configured XGBoost by only a small amount.

XGBoost works as an ensemble of decision trees. It consists of a set of classification and regression trees (CART). The figure below shows a simple example of CART to classify whether someone will like a computer game:



We classify the members of a family into different leaves and assign them the score on the corresponding leaf. A CART is a bit different from decision trees, in which the leaf only contains decision values. In CART, a real score is associated with each of the leaves, which gives us richer interpretations that go beyond classification. This also allows for a principled, unified approach to optimization.

Mathematically, we can write our model in the form:

**y^i=∑k=1Kfk(xi),fk∈Fy^i=∑k=1Kfk(xi),fk∈F**

where KK is the number of trees, ff is a function in the functional space FF, and FF is the set of all possible CARTs. The objective function to be optimized is given by:

**obj(θ)=∑inl(yi,y^i)+∑k=1KΩ(fk)**

**DATA**

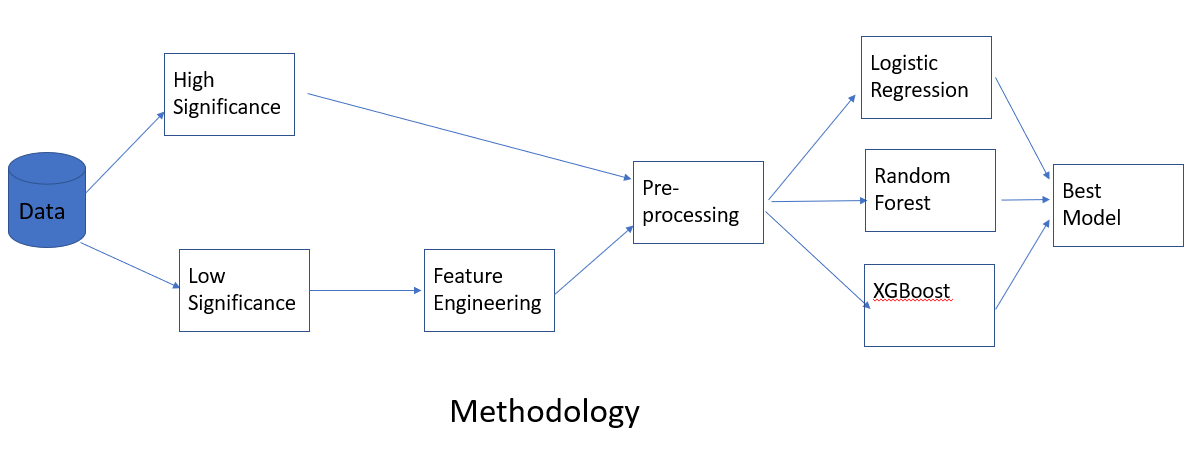
We used the data set provided for a Kaggle competition: <https://www.kaggle.com/ntnu-testimon/paysim1>. The competition had dataset which was simulated by the organizer. The data has 10 feature parameters and about 6.5 Million rows.

**Table 1:** Data used in study:

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| Step | Numeric | Hours passed from starting |
| Type | Factor | 5 Levels. CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER |
| Amount | Numeric | Amount of the transaction (Imaginary Unit) |
| NameOrig | String | customer who started the transaction |
| oldbalanceOrg | Numeric | initial balance before the transaction |
| newbalanceOrig | Numeric | customer's balance after the transaction. |
| nameDest | String | recipient ID of the transaction. |
| oldbalanceDest | Numeric | initial recipient balance before the transaction. |
| newbalanceDest | Numeric | recipient's balance after the transaction. |
| isFraud | Binary | identifies a fraudulent transaction (1) and non-fraudulent (0 |
| isFlaggedFraud | Binary | flags illegal attempts to transfer more than 200.000 in a single transaction |

## METHODOLOGY

To predict fraud transaction 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 50:50 split of fraud and non-fraud transactions. This is necessary to provide sufficient incentive for the algorithm to identify fraud transaction which comprise less than 0.1% of the original dataset. Such a transformation is always required for a highly skewed dataset. Through logistic regression we first find parameters that show significant statistical significance in predicting the transaction nature and differentiate them from those which don’t. We then used the different combinations of the non-significant features to create new features that have better prediction capabilities. We performed pre-processing on data which includes completing missing values in data, removing highly correlated features, removing linear combinations, removing features with low variance, creating factor variables. The processed data was split in 70:30 ratio for the training and the test set. We decided to choose this ratio after we tried multiple ratios and compared their performance with each other. 5-fold cross validation is used to train logistic regression, random forest and XGBoost models[3]. The AUC and sensitivity are chosen to be the statistical business performance measure because it is important for the business to identify fraudulent transaction whose entire cost is born by banks and overall performance is important because rise in false negatives will result in high operational overhead for banks.



## MODELS

We used three models for predictions which are logistic regression, random forest and XGBoost. Logistic regression is a type of linear model.

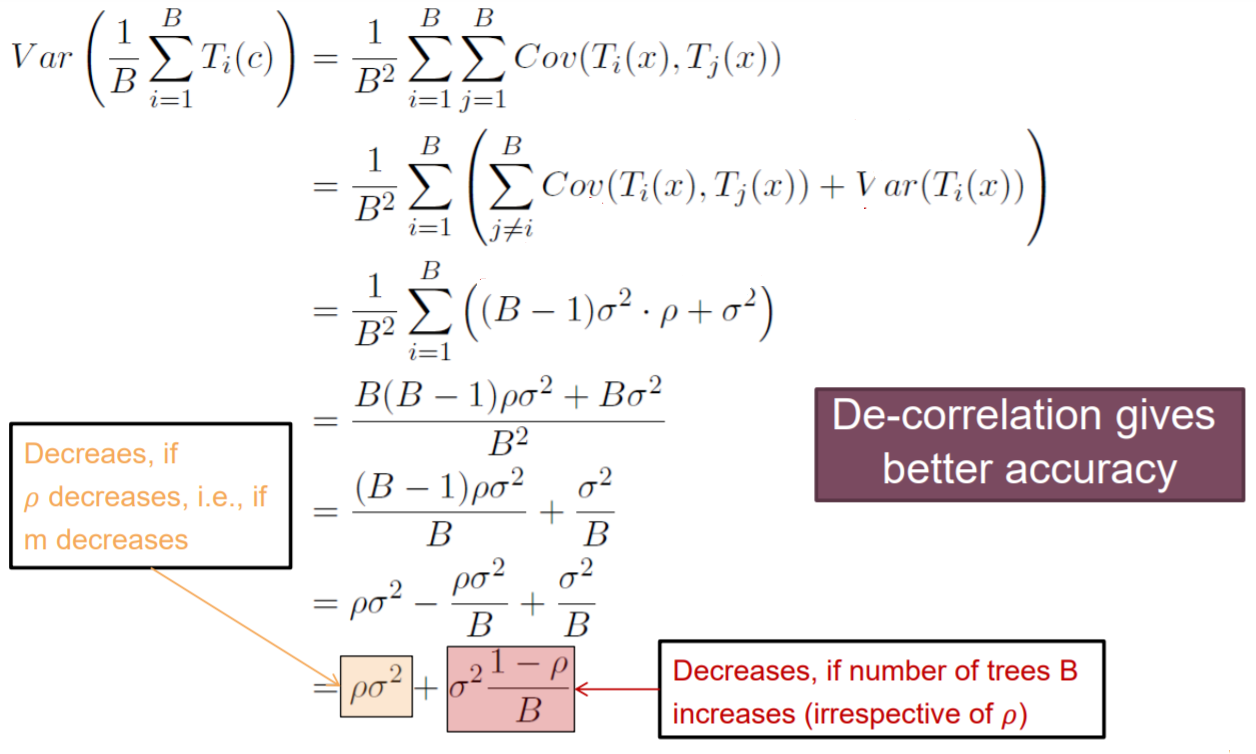
* **Logistic regression** is easy to interpret and performs well with few categorical variables but does not capture complex relationship between independent and dependent variables.
* **Random forest** is a type of tree based hierarchical classification model. The algorithm utilizes the rules of a full tree or a subtree to arrive at an outcome. The subtree is weaker than a full tree, but the combination full tree and subtree leads to better overall performance. The models created through random forest tend to have very high quality and are also fast to train. The models are very large because of large number of tree and are also difficult to understand.
* **Gradient boosting** uses even weaker tree to build its decision, this leads to even high performance from the random forest model. As a result, the models become larger than whey they are for random forest and are not easy to understand. Even small changes in the feature set or the training set results in radical changes for the model.

We chose these models because fraud detection is a supervised classification problem and previous studies such as Bruno Carneiro da Rocha et Al showed promising results by using tree-based classification [4]. We carried out logistic regression to understand the relative impact of features to the model outcome.

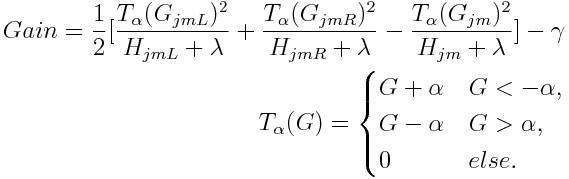
**Logistic Regression**

The logit function is generalized as

**Random Forest**



**XG Boost**



## RESULTS

**Train Data Results:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Specificity** | **Sensitivity** | **Balanced Accuracy** | **AUC** |
| XGBoost | 0.9915 | 0.9912 | 0.9997 | 0.9945 | 0.9921 |
| Random Forest | 0.9991 | 0.9982 | 0.9968 | 0.9991 | 0.9988 |
| Logistic Regression | 0.9265 | 0.8915 | 0.9722 | 0.9433 | 0.9322 |

**Test Data Results:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Specificity** | **Sensitivity** | **Balanced Accuracy** | **AUC** |
| XGBoost | 0.9903 | 0.9809 | 0.9996 | 0.9865 | 0.9877 |
| Random Forest | 0.9986 | 0.9972 | 0.9953 | 0.9986 | 0.9981 |
| Logistic Regression | 0.9225 | 0.8815 | 0.9635 | 0.9225 | 0.9244 |

**Elaboration of the above results:**

It can be seen from the results summary that none of the models overfit the data if we follow the empirical rule of 15% difference allowance in the train and test data. Logistic Regression Performed the worst for all the parameters under consideration. For Accuracy, Sensitivity and Balanced Accuracy, XGBoost performed better than random forest whereas for Specificity and AUC, random forest had a better performance. Since specificity and AUC are the important performance measures for fraud detection problem it is the chosen best model it.

Based on the specificity results it can be said that 99.72% of the fraud transactions can be identified through the model developer. This translates to an estimated saving of $219,000 for a typical bank branch based on estimated daily transaction of $300,000 at a branch. If all the branches are processed through a single model infrastructure, then the saving would be even greater. The one time and variable cost of maintain the IT infrastructure can easily be covered from these saving. Hence, it is beneficial for business affected by fraud transaction to use the developed model to increase their profits.

**CONCLUSIONS**

From the above research and experimentation done on Logistic regression, Random Forest, and XG-Boost, it can be concluded that:

* Random forest performed the best amongst the three.
* Logistic Regression performed the worst amongst the three.
* Random forest is the most useful model in predicting the dependent variable in a multivariate analysis with data spread over a very high range.

With changing technology banks face increasing risk of fraudulent transaction which result in losses for the banks. A typical branch can save approximately $219,00 in year through the implementation of our model and the cost of deploying such as application has become inexpensive with advances in technology. The saving in fraud get multiplied if the model is maintained at central place and multiple branched connect with it.

We assumed that the data generated by the simulation machine is ideal. In the real-world things are different. It can thus affect the overall performance of the model. We also assumed that the complete behavior of fraudulent transactions has been generate and no more characteristics are to be found after the model has been generated. As a result, with changes in technology and time the model can lose it efficiency of predication and will need to be retrained.

Further investigation can be done toward finding the factors that capture the fraudulent behavior of transaction such as location, date etc. These features would make the model more robust and increase functionality to new areas. Another area that can be investigated is the use of a model that trains itself on the latest data such reinforced learning. This would allow the model to keep updated with changes in the pattern of fraudulent transactions.

**REFERENCES**

[1] Gerard Biau

**Analysis of a Random Forests Model**

Journal of Machine Learning Research 13 (2012) 1063-1095

[2] Tianqi Chen and Carlos Guestrin

**XGBoost: A Scalable Tree Boosting System**

[Paul G. Allen School of Computer Science and Engineering](http://www.cs.washington.edu/)

[University of Washington](http://www.washington.edu/)

[3] P. Ravisankar, V. Ravi, G. Raghava Rao, I.Bose

**Detection of financial statement fraud and feature selection using data mining techniques**

[4] Bruno Carneiro da Rocha

**Identifying Bank Frauds Using CRISP-DM and Decision Trees**

International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010

[5] Maria R. Lepoivre, Chloé O. Avanzini, Guillaume Bignon, Loïc Legendre, and Aristide K

**CREDIT CARD FRAUD DETECTION WITH UNSUPERVISED ALGORITHMS**

Journal of Advances in Information Technology Vol. 7, No. 1, February 2016