International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 4, July-Aug-2020, ISSN (Online): 2395-566X

Fraud Detection: In Airline Loyalty Program Using Machine Learning

Cherian P Bright, Prof. Kuttyamma A J
Department of Information Technology
Rajagiri School of Engineering and Technology Ernakulam, Kerala, India
kuttyamma_aj@rajagiritech.edu.in, cherianp.bright@gmail.com,

***Abstract*-almost all the airlines within the world have currently designed frequent-flyer programs that enable airline customers to be a neighbourhood of the program to accrue points or miles which can then be redeemed for aviation or the other rewards designed by the airline. Points earned under Frequent Flyer Programs could also be supported the category of fare, distance flew thereon airline or its partners, or the quantity paid. There are other ways to earn points. For instance, in recent years, more points are earned by using co-branded Loyalty and debit cards than by aviation. Differently to earn points is by pocket money at partner shops, car hire companies, hotels, or other co-partner businesses. Points are often redeemed for air and non-air travel, other services, or for increased benefits, like class upgrades, airport lounge access, means access, or priority bookings. Loyalty points are like currency! The increase of loyalty programs amongst airlines over the years has offered more options to monetize air miles' balances built up by their customers. But this has spawned an increase of fraudsters too. During this age of digital transformation, airlines have a fundamental duty in fixing flexible and well-layered defines systems which will keep off the thrate of cyber criminals.The matter of monetary Fraud has increased abruptly nowadays. Losses thanks to the fraud transactions are costing billions of dollars per annum to the Airline Industry. To tackle this problem, decision systems that use efficient fraud detection algorithms should be invented. With the support of recent technologies, these systems can manage to research the knowledge and to make a prediction feature model. However, the invention of those systems isn't a simple but quite challenging task thanks to the large amount of various and unbalanced data. Moreover, it's tough to make a decision on which machine learning algorithm should be implemented for various business use cases.**

***Keywords*- airport lounge access , airline customers , Airline Industry , cyber criminals** etc.

I.INTRODUCTION

The main objective of machine learning is to create algorithms and models in order that it can receive input file and use statistics for prediction or categorization of an output value within a suitable range. Analytical models provide the power to automatically get deep insights, recognize unknown patterns, and make better performing predictive models from data sets, all without requiring explicit programming or manual intervention.

Machine learning technology is employed in most banking and finance industries because the right application of technology can give outstanding results and significant improvement in replacing the legacy system and developed enterprise. Machine learning technology helped the economic sector in taking a company's deciding, improving customer experience, increasing the general efficiency. Within the Aviation sector, machine learning are often applied to detect fraudulent Loyalty transactions or to predict the pricing.

II. RELATED WORKS

Various methods have been designed for generating Fraud Predictions. This section includes some of the important works done in the area of Fraud Detection using machine learning"Loyalty Fraud Detection Using Random Forest Algorithm", In this paper, they mainly specialise in Loyalty fraud detection within the world. Here the Loyalty fraud detection is predicated on fraudulent transactions. Generally, Loyalty fraud activities can happen in both online and offline. But in current times online fraud transaction activities within the Loyalty domain are increasing day by day. So to seek out online fraud transactions, various methods are utilized in the prevailing system. During this system, they used the Random Forest Algorithm (RFA) for locating the fraudulent transactions and therefore the accuracy of these transactions. This algorithm is predicated on a supervised learning algorithm where it uses decision trees for classification of the dataset. After the classification of the dataset, a confusion matrix is obtained. The performance of the Random Forest Algorithm is evaluated supported

the confusion matrix. The results obtained from processing the dataset gives an accuracy of about 90%.

"Scalable Machine Learning Techniques for Highly Imbalanced Loyalty Fraud Detection: A Comparative Study", Rafiq Ahmed Mohammed, Kok-Wai Wong, Mohd Fairuz Shiratuddin, Xuequn Wang: In this paper, different machine learning algorithms were experimented and compared to see their suitability as Scalable Algorithm when interacting with huge data sets which have class imbalance problem. They used two datasets and mainly applied Random Forest, Balanced Bagging Ensemble, and Gaussian Naive Bayes algorithm. The category imbalance is extremely common within the Fraud scenario as only 2% of all the transactions are going to be Fraud and rest are going to be legitimate transactions, hence training the model got to be done only after handling the category imbalance problem. it's found that a lot of algorithms worked well in small to medium-sized datasets but struggled to take care of similar accuracy because the size of the info sets become huge.

"A Comparative Study of Machine Learning Techniques for Loyalty Fraud Detection Based on Time Variance", Shantanu Rajora; Dong-Lin Li; Chandan Jha; Neha Bharill; Om Prakash Patel; Sudhanshu Joshi; Deepak Puthal; Mukesh Prasad: In this paper, proposes a comparative performance of ten different machine learning algorithms, done on a Loyalty fraud detection application. The machine learning methods are classified into two groups namely classification algorithms and ensemble learning groups. Each set of 5 different algorithms is taken. The 'Time' variable is introduced within the data set and accuracy and performance of the algorithms are studied with and without the 'Time' variable. Out of the ten, two algorithms of the ensemble learning group are found to figure better when the used dataset doesn't include the 'Time' feature. In Classification algorithms group, three classifiers shows better predictive accuracies when all attributes including Time variable are included within the used dataset. Most of the machine learning models have approximate similar scores between these datasets

"Fraud Detection in Online Transactions Using Supervised Learning Techniques", AkshiKumar,Garima Gupta, This paper aimed to automate prediction of fraud in a web Loyalty transaction dataset. Supervised learning techniques, namely logistic regression, nearest neighbours, linear SVM, RBF SVM, decision trees, random forest, and naïve Bayes were implemented. The performance of every technique was analysed and compared. Logistic regression proved to be the simplest among the others. As a future scope of this research, the utilization of neural networks also as other supervised, unsupervised, and reinforcement learning techniques are often explored.

"Comparing ML algorithms on Financial Fraud Detection", Chung Min Tae, Phan Duy Hung, In this paper, The problem of monetary Fraud has reached an alarming scale nowadays. Fraud transactions are causing losses to the industry. Hence it's vital to seek out decision systems that use efficient fraud detection algorithms. With the support of recent technologies, these systems can manage to research the knowledge and to make a prediction feature model. However, it's tough to seek out a system which handles such scenarios thanks to the large amount of various and unbalanced data. Therefore, our research is conducted to answer the question: which is that the best suited algorithm for the dataset during this research, especially when handling an outsized amount of uncleansed data.

III. PROPOSED SYSTEM

Machine Learning Engineering is a stream of Artificial Intelligence (AI) that permits the ML models to find out automatically and improve the accuracy with experience as we progress in time with none manual intervention or explicit program. Machine Learning focuses on computer programs which will get data and use them to coach themselves.

The training starts with data patterns or instructions to seem for data patterns and take better decisions within the future supported the info sets we offer. The most aim is to urge the models to find out automatically with none explicit programming or intervention and alter accordingly.

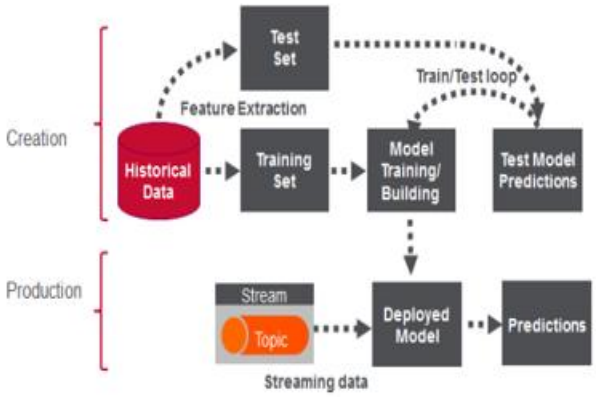

Fig. 1 High-Level Architecture

1. Implementation Modules

Module 1: Exploratory Data Analysis

This is the first step we do for any Machine Learning Implementation where the collected data sets which usually get stored in the database is analyzed. We have to perform descriptive analysis about the dataset.

Module 2: Data Cleaning

In the next step, after understanding and analyzing the dataset unwanted and redundant data sets are removed and a data cleansing activity is performed. In this process, all the duplicate values and null values that are present in the dataset will be removed and a new dataset will be obtained. The dataset in machine learning world is usually called as data frame.

Module 3: Preprocessing of the dataset

In this module, the cleaned data frame will be preprocessed where the dataset will be divided based on certain criteria and subject matters used for particular use case.

Module 4: Dataset Partition

In this module first, the dataset will be divided into two partitions as a trained dataset and testing dataset. After the data partitions, the XG Boost Algorithm is applied. After applying the XG Boost Algorithm finally a confusion matrix is obtained.

Module 5: Evaluation

In the last stage which usually called as Evaluation state the resultant data obtained in the form of confusion matrix which can be evaluated by using graphical representation usually known as elbow chart which gives better accuracy.

Most classifiers aren't designed for unbalanced data. When the unbalanced case happens, the quality machine learning algorithms would tend to classify all instances because the majority class observations to maximize overall accuracy. This is able to cause the minority class having poor accuracy, which shows a coffee recall rate. The performance of a typical algorithm is additionally affected thanks to unbalanced data and thanks to noisy and unwanted data. Therefore, an under-sampling approach is employed to get rid of the info from the bulk class until their amount is as same as (i.e. roughly equal to) the minority class.

Predictive accuracy is the method to get the performance of the machine learning algorithms, but it is not efficient in all the cases especially when we have to deal with unbalanced datasets. The performance measures calculation based on the confusion matrix shown in fig. 2 are as follows:

• Accuracy = (T P + T N)/ (T P + F P + T N + F N)
X-axis represents % F P = FP/ (T N+F P)
Y-axis represents % T P = TP/ (T P + F N)
• Recall = T P / (T P + F N)
• Precision = T P / (T P + F P)
• F1 Score= 2 PR / (P + R)

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Fig 2 Confusion Matrix

A confusion matrix may be a method of predicting the accuracy of algorithm mainly utilized in classification problem. The amount of correct and incorrect predictions are summarized with count values and grouped by each class. This is often the key to the confusion matrix. The confusion matrix shows how your classification model is confused when it makes predictions. It gives information on not only into the errors being made by a classifier but also the kinds of errors that are being made.

2. Definition of the Terms:

- Positive (P): Observation is positive (for example: is an apple).
- Negative (N): Observation is not positive (for example: is not an apple).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

3. Application Development

The proposed model is developed using an Analytical platform where the codes are written in combination with PySpark and Python. The required data for training and testing the model is taken from the data warehouse of an Airline Loyalty Program. The data is extracted in CSV file formats from the DWH using an ETL tool named Pentaho. The CSV files are then pushed into the Analytical models which then get loaded into the Hadoop database. In the Fraud detection model, a data frame with required data is created using PySpark code and then the data frame is then converted into Python Data model to be compatible with Python codes. The model is then trained based on this data frame and the algorithm used for Fraud Detection.

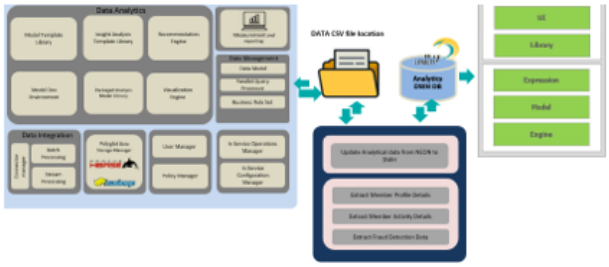

Fig 3Application Development.

In the Proposed system, we have used the XG Boost algorithm to create the Fraud Detection trained model. This trained model on passing the new set of data will provide the Fraud Predictions. Hence, the complete technology stack which is used in this model is an ETL tool, an Analytical platform that can be used for PySpark and Python coding and to load the data into the Hadoop database or onto HDFS systems. Then a scheduler to schedule for daily predictions.

IV. EVALUATION RESULTS

To evaluate the performance of the model we use the confusion matrix and based on the confusion matrix we can derive the Precision and Accuracy and compare the precision and accuracy of each Algorithm to determine which Algorithm is most suitable to predict Fraud transactions under given set of data.

Table 1 Algorithm Evaluation Report For Fraud.

	Precision	Recall	F1-score	Support
Random Forest Classifier	0.78	0.59	0.67	257
Balanced Random Forest Classifier	0.59	0.88	0.71	257
XG Boost Classifier	0.87	0.77	0.82	257

The Algorithm evaluation of existing methods has been used for comparison. Besides that, we also evaluate the efficiency of our model. Table I shows the comparison table of the Precision score of different models. It shows that our model generates better Precision than the existing models. The initial training can be done using the 3-year dataset and then the Predictions can be made in daily transactions. This model is used to predict Fraud on the daily Redemptions happening in the particular Airline Loyalty Program. The result in table I shows that the XG Boost classifier algorithm has better precision and result as compared to other algorithms hence this is used to implement our model.

V. CONCLUSION

To sum up, supported Accuracy and F1 score, derived an equivalent conclusion that the XG Boost algorithm is ranked because the best technique to detect Loyalty fraud of the mentioned dataset with Accuracy of 0.87 and F1 score of 0.82. The XG Boost algorithm is followed by the Balanced Random Forest algorithm and therefore the Random Forest algorithm. The end in that reference shows that three supervised machine learning techniques are gaining better performance as compared with other considered algorithms, namely Random Forests, AdaBoost, and Neural Networks. As are often seen, our work and investigated reference all agree about the performance evaluation of the XG Boost Algorithm.

In this work, the dataset that we take into consideration is already labelled; however, obtaining a labelled dataset during a real situation is extremely difficult. For that reason, we'll consider unsupervised techniques in future research in order that our results are often more easily applied within the world. This paper can also provide a regard to any field in Data Analytics, for instance, Pattern Recognition, etc.The Anomalies/Fraud detection model is one among the important models in AI space which is applied in most the monetary-based systems.

Random Forest Algorithm is taken into account to be the foremost accurate for Fraud Detection problem Trees get partitioned at every node by the foremost important variable of that node. Then the bulk vote decides whether a specific transaction is Fraud or not.

Real-time Loyalty fraud detection may be a challenging issue thanks to highly imbalanced massive data. This research paper is predicated on experiments that compared several popular ML techniques then selected one algorithm to implement our model. During this paper, we aren't performing on Real-time data rather we are becoming the info from the OLAP system which might have last 24 hours transactions from the OLTP system. Thus, during this system, we've a batch process that gets the info from the OLAP system and loads into the Hadoop layer for further analytical processing and daily predictions.

REFERENCES

[1] Loyalty Fraud Detection Using Random Forest Algorithm", M. Suresh Kumar; V. Soundarya; S. Kavitha; E.S. Keerthika; E. Aswini, IEEE, Feb 2019, DOI:10.1109/ICCCT2.2019.8824930

[2] Scalable Machine Learning Techniques for Highly Imbalanced Loyalty Fraud Detection: A Comparative Study", Rafiq Ahmed Mohammed, Kok-Wai Wong, Mohd Fairuz Shiratuddin, Xuequn Wang, Springer (July-2018)

[3 "A Comparative Study of Machine Learning Techniques for Loyalty Fraud Detection Based on Time Variance", Shantanu Rajora; Dong-Lin Li; Chandan Jha; Neha Bharill; Om Prakash Patel; Sudhanshu Joshi; Deepak Puthal; Mukesh Prasad, 2018 IEEE; 10.1109/SSCI.2018.8628930

[4] Fraud Detection in Online Transactions Using Supervised Learning Techniques", AkshiKumar, Garima Gupta Springer, Nov 2018

[5] Comparing ML algorithms on Financial Fraud Detection", Chung Min Tae, Phan Duy Hung