WHITE PAPER

INTEGRATED ML AND BIG DATA APPROACH FOR AIRLINE CHARGEBACK FRAUD DETECTION

Abstract

Airline ticket booking using credit cards has become an important, convenient and widespread mode of transactions, occupying more than 60% of total bookings. However, the chargeback fraud associated with such credit card transactions are also growing significantly. This vulnerability, many times, goes undetected until the fraudulent transaction is investigated and confirmed. By the time it reaches the confirmation stage, Airlines would have spent a lot of time on investigation with a substantial risk of losing significant amount of money, as it is not a recoverable loss. In addition, fraudsters quickly adapt to other methods, which cannot be easily identifiable unless a comprehensive investigation happens. Given the continuously changing nature of fraud transactions, Machine Learning Models could be of significant use in predicting fraudulent transactions. Many Machine Learning and Deep Learning approaches have been developed to proactively detect various credit card fraudulent transactions. However, a Machine Learning approach with a single technique may fail to capture the multidimensional aspect of fraudulent transactions. Hence, a comprehensive integrated approach, which captures the deviations in trend, pattern and anomalies at a time is very much required. Integration of business rules on top of the blended Machine Learning solution will be even more robust. This paper discusses the how the Chargeback fraudulent transactions can be predicted live or well in advance using integrated Machine Learning approach in an optimally designed Big Data platform.

Infosys®
Navigate your next

Introduction

Credit cards play a major role in online purchase of any goods and services and has become the most preferred payment mode. When consumers make purchases with credit cards, the merchants typically receive funds for those card transactions a few days after the transaction date. However, in few transaction cases, the payments are not necessarily guaranteed for the merchants due to chargebacks. The chargebacks have grown significantly these days, bringing turbulence in the Airline industry, leading to huge loss in revenue and bad customer experience. The chargebacks are difficult to handle, as they are much more complex and hidden. Detection, confirmation and prevention of such chargebacks is a lengthier procedure. In addition, there is a need for timely detection of frauds to prevent such transactions even before the utilization of the service. It would be even more complex if the fraud is detected or confirmed once the service is utilized. Hence the early detection, investigation, confirmation and prevention of fraudulent transactions is crucial.

The existing approaches, many times are partially helping the Airlines because of their limitations. If the historical trend and pattern is captured by supervised learning methods, unsupervised learning methods try to capture anomalies in the transactions. While the historical trend is less likely to repeat as the fraudsters quite often change their way of doing such fraudulent transactions, in many cases anomalies may not be fraudulent transactions. As per our empirical analysis, genuine transactions may also be classified as anomalies because of their unique nature. After all, both might be missing dynamic inputs from business, as it cannot be included in the models. As a result, there is a need for an integrated approach to capture the trend, pattern, anomalies at a time, which can be further integrated with the business inputs or rules.

As a result, we have approached this problem using a three-layered solution. In this paper, we have proposed an integrated solution based on integrations of supervised, unsupervised methods and business rules. Supervised models focus on capturing the historical trend, unsupervised models capture the anomalies. Finally, the predictions of both will be integrated with the business rules. The experimental results using synthetic dataset demonstrates that proposed method takes less investigation cost and predict potential frauds better than the single technique-based approach. The proposed solution helps Airlines to mitigate the risk at maximum even before it occurs and reduces the false positives and negatives. Thus, it also helps to eliminate unnecessary chargebacks before they initiated and reduce the revenue leakages due to missed opportunity.

Problem Statement

This paper focuses on the credit cards chargeback frauds in the Airline industry. The chargeback process is a complex procedure and involves many parties; everything from the initial customer compliant all the way through arbitration and long-term effects on merchants. Cardholders may dispute their payment card transactions for different reasons including friendly frauds.

There are several key players involved in the process, including 1. Cardholder, 2. Merchant, 3. Issuer, 4. Acquirer, 5. Card Association. Below is the brief explanation of the various steps followed:

1. The process starts with the cardholder raising a dispute challenging a transaction with the issuing bank
2. If the bank determines the claim is invalid, the chargeback will simply be declined. If there is an indication that an error occurred, the case will proceed through the chargeback lifecycle.
3. A refund will be issued to the cardholder on condition basis and the merchant's account will be debited the original transaction value, along with any applicable fees
4. A numeric reason code will be assigned by the issuing bank for the chargeback, and then electronically transmits all the chargeback data to the merchant
5. The merchant has the option to accept or dispute the chargeback. Merchant does his own investigation and confirm back the back about his acceptance or decline.

The back-and-forth cycle continue until the dispute gets resolved and, in many cases, merchant will end up in losing the money.

The research suggests that majority of these chargebacks can be prevented, if merchant can proactively predict such kind of transactions well in advance. Therefore, relying on static technology for fraud detection costs heavily to any Airline. A quick and efficient fraud detection system is a need of an hour for all the Airlines.

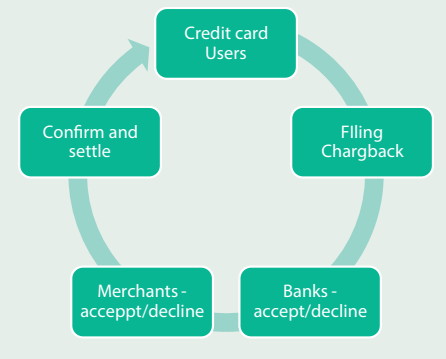

Figure 1. Fraud Chargeback Procedure

Business Case

Many Airlines across the globe have been manually reviewing the suspected fraudulent transactions and spending a lot of time on the investigation, confirmation and prevention. Many times, fraudulent transactions go unobserved and unattended resulting huge revenue loss. We have considered such a case here for our study and proved an integrated approach a best solution to resolve this issue effectively using synthetic data. We have created a hypothetical business case, which would be truly representative of actual scenario. We would like to name this hypothetical Airline as ABC Airline, and henceforth this hypothetical name will be used throughout this paper.

ABC Airline, one of the largest passenger carriers, based out of Europe, operates 300 flights and 60 destinations every day. It receives around 6K booking requests everyday and accepts the payment through various mode. Among them, online credit card payment is an important mode where it accounts for approximately 60% of the total transactions. Customers book tickets using credit cards issued by many banks/ issuers. The payment happens through a payment gateway associated with many banks/issuers on a regular basis. The payment window spreads between 60 to 90 days after any transaction. However, in the meantime, Airline receives chargeback files from Banks triggered by the respective customers, withholding the payments for their purchases. Airline does the usual manual investigation of the suspected frauds and either accepts or rejects the chargeback files. If it is accepted, it is a huge revenue loss to the business; if it does not accept also, the Airlines will not get money. This could have been avoided if the airline is having more accurate fraud detection solution. The current solution that Airline is having is set of business rules. They suspect a very high number of false positives and false negatives, which requires lot of time to investigate and lead huge revenue leakage due to missed opportunity. In this regard, Airline is having clear objective of reducing false positives and false negatives with comprehensive solution. In addition, Airline wants to reduce the revenue leakages and missing opportunities through optimized and near real time fraud detection solution.

We had set the hypothesis after verifying the data that frauds are being occurred by usual pattern and can be captured through supervised learning models. In addition, we could see few anomalies, which may be the potential fraud transactions. At the same time, integrated the business rules with both the results. Thus, we have created a comprehensive and a multi-layer approach to safeguard revenue of the Airline through a sophisticated fraud detection system.

Literature Review

Credit card chargeback fraud detection has drawn a lot of research interest among fraud researchers. They have been adopting a number of techniques since beginning:

J. S. Mishra et al. [5] in his paper "A Novel Approach for Credit Card Fraud Detection Targeting the Indian Market", explains how a fraud can be reported instantly while the fraudulent transaction is in process. Andrei Sorin SABAU [6], in his paper, "Survey of clustering based financial fraud detection research", explains the clustering techniques that was used in fraud detection over the last ten years. Fig. 2 shows various data mining techniques and types of financial fraud.

Anamika and Mayuri, et al. [2] compares the performance of various techniques for fraud techniques and concludes that the self-organizing maps and hierarchical clustering provided good classification accuracy.

M. A. Vasarhelyi et al. [6] describes "Application of Anomaly Detection Techniques to Identify Fraudulent Refunds". This paper describes classification based, clustering-based anomaly detection techniques and their applications. As an illustration, the paper applies K-Means, a clustering-based algorithm, to a refund transactions dataset from a telecommunication company, with the intent of identifying fraudulent refunds.

Sharmik Sural et al. [10] in their paper "Credit card fraud detection using hidden Markov model", explains how the sequence of operations in card transaction processing is modeled using a Hidden Markov Model (HMM) and how it can be leveraged for the detection of frauds. At first, an HMM is trained with the normal behavior of a cardholder. it is considered to be fraudulent transaction, if an incoming credit card transaction is not accepted by the trained HMM model with required probability. At the same time, it is taken care in such a way that genuine transactions are not rejected. Detailed experimental results are presented to prove the effectiveness of the approach and comparing it with other techniques available in the literature.

Benson Raj et al. [11] in their paper "Analysis of Credit Card Fraud Detection Methods", evaluates various techniques used in credit card fraud detection mechanisms based on certain design criteria. The paper concludes that the hybridized algorithm named BLAH-FDS identifies and detects fraudulent transactions using sequence alignment tool. The processing speed of BLAST-SSAHA enable on-line detection of credit card fraud in a very faster process. The ANN and BNN are used to detect cellular phone fraud, Network Intrusion. Many techniques of credit card chargeback fraud detection discussed in this survey paper have its own strengths and weaknesses. At the end, the author conveys the importance of need of developing a hybrid approach for identifying fraudulent credit card transactions.

D. Khataria, S. Muzamil, et al [3] in their paper "A Comparative Study of Airline Recommendation System Using Sentiment Analysis on Customer Tweets" highlights the importance of recognizing user's personalized historic behaviors. It uses the different tweets made by them for the analysis. This paper compares various techniques used to develop Airline recommendation system. The based motive of this paper is to build analytical solution based on the sentiment analysis of customer reviews.

Khyati and Jyoti et al. [3], in their paper "A Review of Fraud Detection Technique: Credit Card" illustrates how the integrated data mining techniques approach helps for high fraud coverage along with high and low false alarm rate. In this paper, 13 classification techniques were used to build fraud detection system. This work demonstrates the advantages of applying the data mining techniques including ANN and LR, BN techniques for the purpose of reducing the bank's financial risks.

Proposed Approach

We propose three-layer approach to the above business case, which includes – identification of trend, anomalies detection and adding business inputs through rules. In addition, as a continuous approach and learning, the output from this solution will be imputed further in every next calibration of the model. The entire solution framework is illustrated below in Figure 2.

The solution starts with the standard Data Science approach – data extraction, data preparation, variable selection based on both the domain and statistical knowledge, correlation and association analysis and visualization. In the next level, the data is split into different samples – training, testing and out of time validation. The supervised and unsupervised models are trained based with many iterative procedures. The appropriate model is selected based on the different criteria – model performance, consistency across sample and out of time validation. The suspected frauds from both the Supervised and Unsupervised models will be integrated with the business rule. We will discuss this approach in detail in the next section.

The first layer of the detections happens through Supervised Leaning, which captures the historical trends. This specifically uses recent past data and tries to detect potential fraud based on historical pattern. The second layer tries to detect anomalies, which are unusual transactions and have the high potential to become fraudulent transactions.

At third level, business inputs will be added in terms the business rules to detect the suspected frauds. The solution is then deployed in the production environment to predict the potential fraud; every new transaction will be passed in the solution on a regular interval.

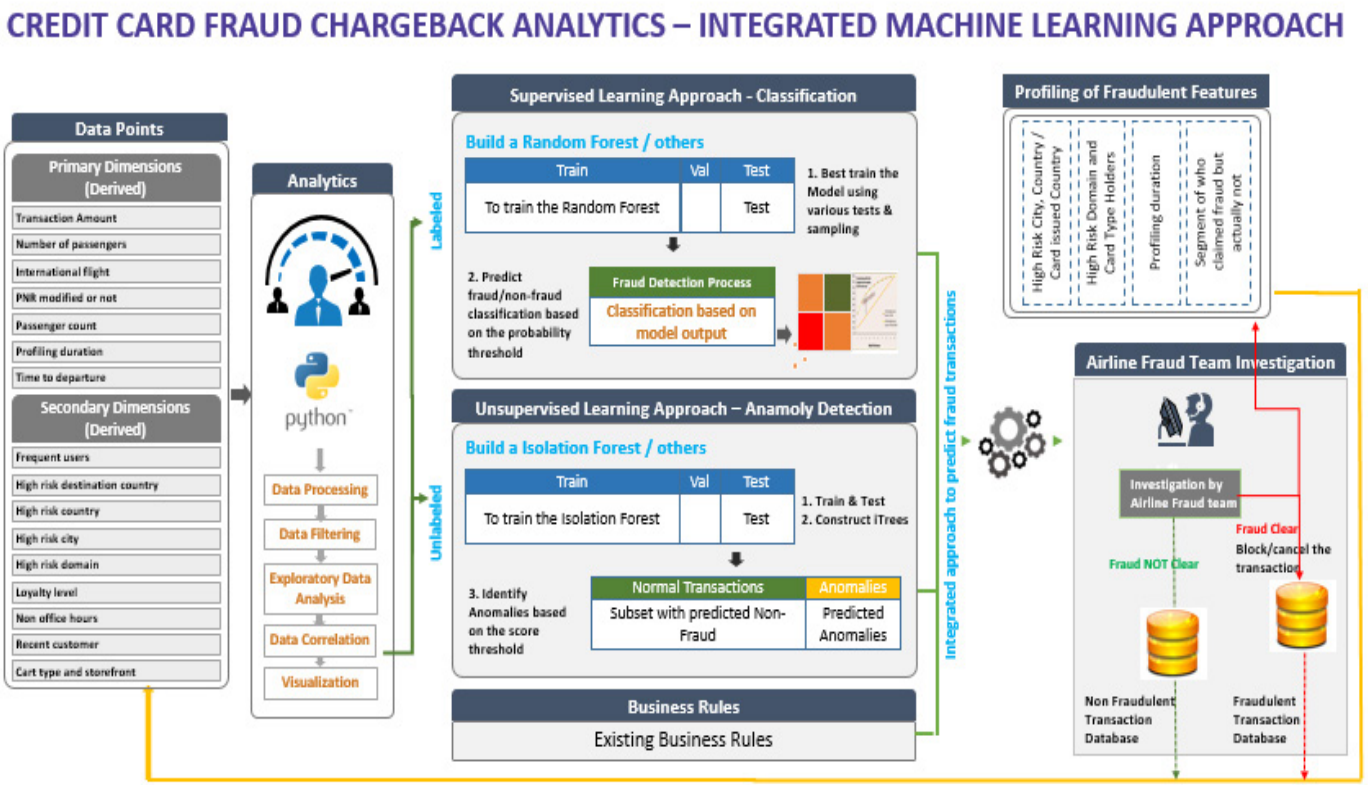

Figure 2. Solution Approach

We would recommend the fraud investigation team of the Airline to refresh the solution every 2 hours. This is to proactively identify the fraudulent transactions well in advance before the service is utilized. If the solution detects the potential fraud, those will be observed by fraud investigation team to investigate and confirm. The fraud investigation team works on the procedure and finalizes the list of fraud and non-fraud transactions. If the transaction is confirmed as fraud, those transactions will be blocked and prevented before the services is used. At the end, both fraud and genuine transaction features will be further added to the database in terms of profiling and behavioral pattern. This helps the solution to be most updated in terms of the recent pattern and business inputs. Thus, 3 layers detection system is adopted to identify the fraudulent transactions with different dimensions. The solution ultimately provides the list of suspected fraudulent transactions with all the required details for the instant investigation like PNR number, passenger name, phone number, email ID, origin and destination city, origin and destination country etc. The empirical analysis of this approach using synthetic data is explained in the next section.

Business Case

Input Data

We have 12 months of hypothetical transaction level data (sample example: Jan to Dec 2018) with 27 attributes, which include 19 derived attributes, 8 original attributes and 1 response attribute. There were approximately 2 million transactions. For this study, we have considered only web-based transactions, as call center-based transaction would differ slightly. All the web-based transactions using credit cards during this period are considered for the analysis. The synthetic data (for both dependent and independent variables) is created based on the Joint Probability distribution of a similar real time transaction data.

Scope

This study considers all those countries, cities where the Airline operates including all types of web channels. All types of passengers are included in the analysis. All type of credit card and all type of issuers and payment services (VISA, MAESTRO) are included in the analysis. The recent one-year data as mentioned in the section 5.1 is included, as the fraudsters will keep adopting the new patterns quite frequently. In addition, they tend to avoid those methods, which were followed historically.

Data Preparation

The data preparation has been done based on standard Data Science approach - the missing value treatment is done based on variable types using standard criteria, error in the data entry is fixed, duplicate values are removed after reconciliation and outliers are treated as per the client suggestions.

Feature Engineering and Variable Selection

The variables are selected based on both domain knowledge, business input and the statistical importance. We have selected 13 direct attributes and 13 derived attributes from the existing data.

			VARIABLES SELECTION - ORIGINAL AND DERIVED				
#	VARIABLES	O/D	DESCRIPTION	Type	FRAUD (%) - 1	FRAUD (%) - 0	F SCORE
1	Highriskorig_country	D	Two paratmers - 1. Fraud proportion > 0.8, 2. Fraud relative proportion > 0.0008 (based on origin country)	Binary	1.11	0.32	3,416.17
2	HighriskCountry	D	Two paratmers - 1. Fraud proportion > 0.8, 2. Fraud relative proportion > 0.0008 (based on bill Country)	Binary	1.02	0.24	2,975.21
3	highrisk_loyaltyLevel	D	Loyalty vs. others. Two paratmers - 1. Fraud proportion > 0.8, 2. Fraud relative proportion > 0.0008 (based on origin country)	Binary	0.49	0.08	2,218.40
4	frequent_users	D	Customer has flown more than 4 times during analysis period	Binary	0.09	0.51	1,823.83
5	Highrisk_dest_country	D	Two paratmers - 1. Fraud proportion > 0.8, 2. Fraud relative proportion > 0.0008 (based on destination country)	Binary	1.17	0.34	1,728.15
6	highrisk_storefront	D	Copa Storefront of Webpage. E.g: CMUS = United States	Binary	0.90	0.25	1,652.40
7	Highriskdomain	D	Two paratmers - 1. Fraud proportion > 0.8, 2. Fraud relative proportion > 0.0008 (based on bill domain)	Binary	0.55	0.19	1,562.71
8	highrisk_amountCurrency	D	Two paratmers - 1. Fraud proportion > 0.5, 2. Fraud relative proportion > 0.0008 (based on local currency)	Binary	0.90	0.25	1,517.70
9	Dep_Arr_City	D	Indicates whether Departure and Arrival City are same	Binary	0.29	0.69	1,213.23
10	HighriskCity	D	Two paratmers - 1. Fraud proportion > 0.8, 2. Fraud relative proportion > 0.0008 (based on bill City)	Binary	1.21	0.29	1,174.30
11	highrisk_tcScheme	D	Brand of Credit Card (Visa, Mastercard, Diners, American Express)	Binary	0.52	0.30	465.71
12	sameIP_NA	D	Indicates either IP Address or True IP Address is blank	Binary	0.82	0.35	210.89
13	internationalFlight	O	Indicates whether it is international flight of not	Binary	0.38	0.08	183.78
14	modifiedPNR	O	is there any change of date or route of reservation?	Binary	0.19	0.37	139.19
15	bill_ip_country	D	Indicates if bill country and IP country are same or not	Binary	0.32	0.52	126.53
16	bill_bin_country	D	Indicates if bill country and bin country are same	Binary	0.34	0.50	126.53
17	Recent_Customer	D	If Customer has flown within 30 days of his last booking	Binary	0.29	0.40	72.25
18	highrisk_passType1	D	Type of passenges - Adult, Children or Infant	Binary	0.37	0.41	64.01
19	non_office_hours	D	whether the transaction is done during office hours or non-office hours	Binary	0.39	0.32	55.63
20	sameIP_TRUE	D	Indicates whether IP Address and True IP Address are same	Binary	0.36	0.39	50.43
21	cardholderPAX	O	Is cardholder passenger (PAX)?	Binary	0.37	0.37	0.08
22	amountUSD	O	Transaction Amount in USD	Numeric			
23	NoLegs	D	Number of landings before reaching the final desitnation derived from 'complete route' colum	Numeric			
24	passCount	O	Total passengers in reservation	Numeric			
25	profilingDuration	O	Device information of record if it is a Web transaction	Numeric			
26	ttd	O	Departure time - Time of transaction. It is in hours	Numeric			
27	feedbackreason	O	Dependent variable with confirmed Fraud and non-Fraud classification	Binary			

Table 1. Model Variables

The different categories of variables are selected for the modeling purpose – 1. Transactional, 2. Geographical, 3. Behavioral, 4. Billing, 5. Time related, 6. Domain related attributes etc. The selected attributes are screened for the complete understanding of distribution, frequency and parameters involved. Based on the understanding, the feature engineering process was initiated. We decided to create binary attributes for categorical variables based on criteria explained by assigning 1 and 0. This is done based on the number of iterative procedures, 1 is assigned for those who come under high-risk category and 0 is assigned for those who come under low-risk category.

Fraud Definition and Dependent Variable

The fraud is defined as those hypothetical transactions, which were claimed by the issuers and confirmed by the Airline after the investigation. The confirmed frauds are those Airline had agreed with the card issues and accepted not to claim the money for those transactions. The dependent variable is categorized as 1 for confirmed fraud and 0 for non-confirmed frauds. Feature engineering and dependent attribute creation is explained below in the table. The column 'FRAUD (%) -1' and 'FRAUD (%) – 0' shows the fraud rate in the high-risk category and low risk category respectively.

Exploratory Data Analysis

The Exploratory Data Analysis has been done for all the explanatory variables and dependent variable. Firstly, dependent variable fraud has been plotted to sees the pattern by months. Below graph shows the distribution of the fraud across the study period. As we can see below, the fraudulent transactions are randomly distributed over a period. The highest fraud rate has been recorded in the month of Jun. As we are focusing on the short window of analysis period, we not focusing more on the seasonality factors. Here, the question arises whether fraudulent transactions are correlated with the seasonality. Anyways, we leave that to another research topic.

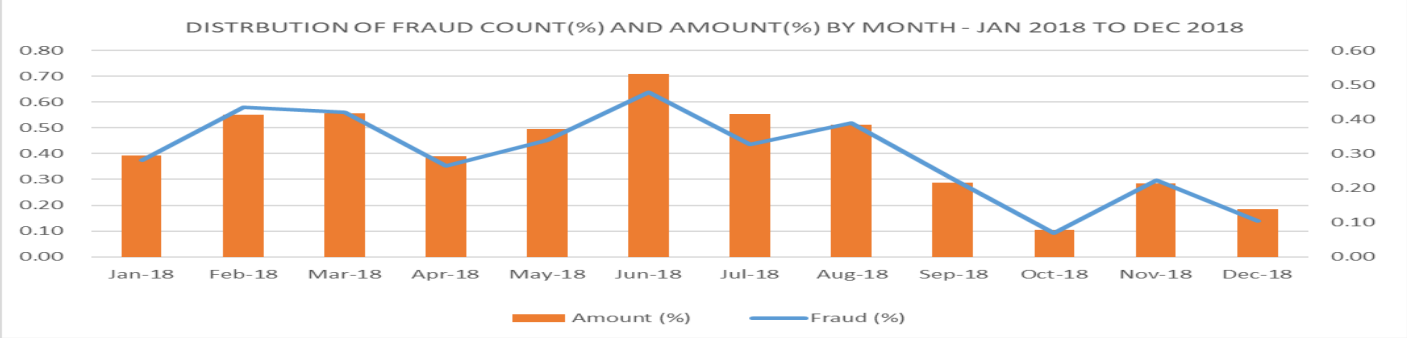

Figure 3. Distribution of Fraud Count (%) and Amount by Month

Above graph shows, the distribution of fraudulent transaction and associated revenue leakage my month. The average fraud rate for the entire one year is 0.30%. The average monthly fraud count is 250 and daily fraud is on an average 8 to 10 approximately. As a results of this, average monthly revenue loss is 0.42 % and the average revenue loss is $298,598 pm and daily loss is approximately $10K. On the other hand, we have plotted, analyzed and explored the explanatory variables also.

Sampling

The next level is to decide the sampling structure. We have splitted the data in two – training data and testing data. The entire data set for the period of Jan to Nov 2018 is splitted randomly by 70% and 30% respectively. The Dec 2018 data is kept aside for the out of time validation. This is because, the model needs to be validated for the recent past data for its performance measure.

With this and considering the above tiny figures of fraudulent transactions, we decided to increase the fraudulent transactions by oversampling method. Among, many we explored, we observed Synthetic Minority Oversampling Technique (SMOTE) was very effective in improving the model performance. The fraudulent transaction percentage was increased to optimal 5% from 0.3% after many iterations of model performance measure.

Model Development

In this section, we will describe the models that we have developed in order the capture the fraudulent transactions at maximum and in order to reduce the false positives and false negatives.

Supervised Models: At first, we started with supervised model as we had prepared the labeled data. The labels were assigned explained in the feature engineering section. Below are the supervised models we tried with – 1. Logistic Regression, 2. Decision Tree, 3. K Nearest Neighbors, 4. Random Forest, 5. Gradient Boosting, 6. Multilayer Perceptron. Among them, Random Forest was the best predictor, which had the good performance measures – Precision – 0.88, Recall – 0.82 and F1 score – 0.84.

Actual/Predicted	Non-Fraud (P)	Fraud (P)
Non-Fraud (A)	7,56,000	121
Fraud (A)	1,200	3,000

Table 2. Confusion Metrics – Supervised Learning

We are not explaining more on the technique here as our focus is on the integrated approach rather how a particular technique work. There are already many studies and papers explaining particular technique.

Unsupervised Models: In the next level, we started with an Unsupervised Learning to capture the unusual transactions. The hypothesis here is unusual transactions sometimes may turn out to be fraudulent transactions. We tried with many unsupervised models and finalized Isolation Forest based on the performance measures.

Actual/Predicted	Non-Fraud (P)	Fraud (P)
Non-Fraud (A)	6,95,000	650
Fraud (A)	1750	450

Table 3. Confusion Metrics – Unsupervised Learning

The idea here is whatever the Supervised model could not capture because of their uniqueness; Unsupervised models could capture those anomalies. In addition, the anomalies are having less tendency to repeat over a period. They need to be captured with Unsupervised models only.

Business Rules

At the end, the already implemented dynamic business rules will join the solution adding the remained fraudulent transactions captured using the business sense. These are the results of key business inputs from the business. Few examples of business rules are 1. Transactions originating from a high-risk zip-code, 2. Transactions of fraud prone air routes etc.

Actual/Predicted	Non-Fraud (P)	Fraud (P)
Non-Fraud (A)	6,50,000	7500
Fraud (A)	4733	75

Table 4. Confusion Metrics – Business Rules

Integration and Implementation

It is time now to integrate both the model results and the output of business rules to develop a more robust and comprehensive solution. The integration happens through the predictions each approach creates in turn the suspected frauds and non-fraud transactions. This is a recommended logical integration based on multi-layer approach in the Big Data platform using PySpark. The integration output has been illustrated below in the table.

Actual/ Predicted	Non-Fraud (P)	Fraud (P)
Non-Fraud (A)	5,251	35
Fraud (A)	3	9

Table 5. Confusion Metrics – Supervised Models

Actual/ Predicted	Non-Fraud (P)	Fraud (P)
Non-Fraud (A)	5,150	135
Fraud (A)	8	5

Table 6. Confusion Metrics – Unsupervised Models

Actual/ Predicted	Non-Fraud (P)	Fraud (P)
Non-Fraud (A)	4,732	552
Fraud (A)	10	4

Table 7. Confusion Metrics – Business Rules

At first, both Supervised and Unsupervised models are executed in the hypothetical production environment for the single day transactions. This is in order to identify the potential fraudulent transactions out of the total transaction during the day. Here, we are considering one day time period, i.e., all the reservations happened from morning to evening. However, our recommendation is to make the solution real time or run it for at least 2 hours in advance.

In the next stage, the fraudulent transaction suspected by the Supervised learning method will be given high priority as the fraud positives and fraud negatives are less and the predicted frauds are closure to the actual fraud. In the next stage, suspected fraudulent transactions by Unsupervised learning along with the another set of suspected fraudulent transactions by business rules will be taken for the investigations. This is because the false positives and false negatives will get reduced by integration of the results of both these approaches. The suspected fraudulent transactions identified in the level will be mapped with those identified by the Supervised learning. The incremental difference will be added to the cumulative list.

In this section, we will describe the models that we have developed in order the capture the fraudulent transactions at maximum and in order to reduce the false positives and false negatives. Proceeding further, in the next level, the remaining anomalies will be screened and investigated for adding another round of incremental frauds. Finally, the output of business rules will be considered for further investigation. The business rules are very important here as few dynamic factors and scenarios cannot be easily imputed in the models. Also, these scenarios vary quite frequently from time to time. For example, fraudulent transactions may vary with respect to some particular event or from those countries/cities where the economic scenario would have been changed to worse. Such factors can be easily captured by the business rules rather than machine learning models.

We would also like to introduce the criticality level for the suspected fraudulent transactions. The criticality levels are assigned as per the models' accuracy and rate of false positives and false negatives. Assigning criticality level is requested in order to give the priority for the investigation team. Because this is really important that how soon we identify most of the fraudulent transactions and prevent them well before they are utilized.

The solution, thus, tries to capture the suspected fraudulent transactions to the maximum by integrating the multidimensional approach. As we can see in the table, the cumulative suspected frauds are incremental as we progress with the 3 stages. Though, the incremental rate looks to be lesser as we progress to the next stage, the optimal stage will be reached at the end of the process. This also prevents many fraudulent transactions from escaping from the coverage and lead to many revenue leakages. Also, the time the investigation team needs to spend to perform the investigation will come down significantly as this solution narrow down the number of false positive and false negatives. Ultimately, as part of continuous learning experience, the features extracted from all the above techniques will be keep adding in the calibration process. Thus, those transactions with older patterns captured by Supervised method, new patterns will be captured by Unsupervised method and frauds under dynamic circumstance will be captured by business rules.

INTEGRATED SOLUTION - MODEL DEPLOYMENT - VALUE ADDITION											
#	OUTPUT	SUPERVISED	UNSUPERVISED	BUSINESS RULES	SUSPECTED	IDENTIFIED	INTERSECTION	NET ADDITION	CUMULATIVE	ACTUAL FRAUD	CRITICAL LEVELS
A	Trend	1	0	0	44	12	12	12	12	22	1
B	Anamolies	0	1	0	119	7	2 (A ∩ B)	5	17	22	2
C	Business Inputs	0	0	1	452	4	2 (A ∩ B ∩ C)	2	19	22	3

Table 8. Integration Approach

The above table explains the integration approach in detail and logically. In the columns, we have Supervised, Unsupervised and Business rules which explain the trend, anomalies and business inputs (rows) respectively. The value 1 represents indication of potential fraud by the particular technique and 0 represents the predicted genuine transactions. In the columns, we have Supervised, Unsupervised and Business rules which explain the trend, anomalies and business inputs (rows) respectively.

As a first step, in the integration, we will take all the suspected 44 fraudulent transactions by Supervised learning and suggest the investigation team to take those on high priority for the investigation. By doing so, they will be able to identify 12 actual fraudulent transactions. With this, a total of 55% of fraudulent transactions are covered. In the second step, even though the Unsupervised technique identified 140 potential transactions, the solution automatically deducts those transactions, which are already suspected by the Supervised technique. Hence, there will only 119 transactions recommended for the next level of investigation (assuming 21 are overlapping between Supervised and Unsupervised techniques). In this stage, there will be net addition of 5 fraudulent transactions to the total identified list leasing to 77% coverage of total fraudulent transactions. In the third stage, we suggest the investigation team to consider suspected frauds by business rules. Again, we assume 104 transactions are overlapping with A and B and the remaining non overlapping numbers are 452. At the end of the third stage, the integrated approach would be able to capture total 19 genuine fraudulent transactions covering total 86% of total.

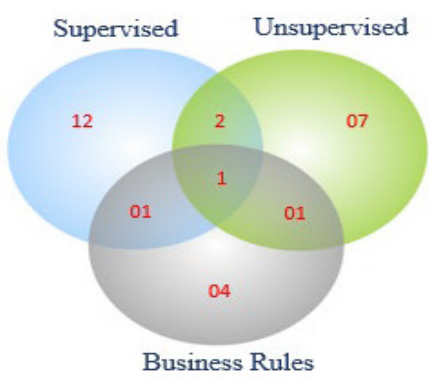

Figure 4. Integration Logic

The above figure explains how the fraudulent transactions are being captured by each technique and how they are being added incrementally. At the same time, few fraudulent transactions may be overlapped also. For example, out of the 12 suspected transactions identified by Supervised model, 2 transactions were also identified by the Unsupervised model and then1 transaction was identified by Business rules. Also, 1 transaction was identified by all 3 approaches. This was shown in the intersection sections of the figure above. Finally, the cumulative total of 19 which is the total number of fraudulent transactions captured by the integrated solution maximizes the fraud detection cases.

Below is an example of out of time validation we have done for a single day snapshot. However, we have done such validations for number of days. The overall approach showed the consistency and comprehensive coverage.

Approaches	Fraud Coverage	Actual Fraud	Percentage	Average Ticket Price	Total Potential Fraud Loss	Individual Approach Savings	Potential Loss
Supervised	12	22	55	1,000	22,000	12,000	10,000
Unsupervised	7	22	32	1,000	22,000	7,000	15,000
Business	4	22	18	1,000	22,000	4,000	18,000
Integrated	19	22	86	1,000	22,000	**19,000**	**3,000**

Table 9. Quantified Financial Benefits of Integrated Approach overs Others

The above table shows the financial benefits of following integrated approach over individual approach. For example, had we followed only single approach, the potential savings would only be $12,000 or $7,000 or $4000 respectively for Supervised, Unsupervised and Business approaches. However, by following an integrated approach we will minimize the loss to the extent of $3,000 ($22,000 – $19,000).

Conclusion

This study aims to detect potential fraudulent transactions using multidimensional and multilayer approach. It tries to integrate most of the possible scenarios of fraudulent transactions occurring during Airline booking, making the solution most comprehensive and robust. It tries to avoid missing out of such transactions from both historical perspectives, from anomalies perspective and ultimately from business sense perspective.

The empirical analysis proves that that the integrated approach is efficient and is able to capture 36% incremental fraudulent transactions in comparison with the any single approach followed generally. More importantly, it reduces false positives and false negatives significantly. Thus, Airline will be able to capture relatively more fraudulent flyers well in advance before they fly with reduced effort and time resulting in increased revenue. We have also observed that integration of text analytics is having a significant value addition to the Airline fraud detection solution, which we will bring up in the next paper.

About the Authors

Dr. Vinay M R
Senior Data Scientist

Lakshmi N Subramanian
Associate Principal-Data Science

Suresh Mani
Senior Lead Analyst

About the Mentor

Amitabh Manu
Delivery Manager

References

[1] Anamika G, Mayuri K, "Analyzing the Performance of Various Fraud Detection Technique", International Journal of Security and Its Applications, pp.21-36, vol.12, No. 5, (2018)

[2] D. Khataria, S. Muzamil, "A Comparative Study of Airline Recommendation System Using Sentiment Analysis on Customer Tweets", International Journal of Advance Science and Technology, vol 111 (2018), pp. 107-114

[3] Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, "A review of Fraud Detection Techniques: Credit Card", International Journal of Computer Applications (0975 – 8887), Volume 45– No.1, May 2012

[4] J.S. Mishra, S Panda, "A Novel Approach for Credit Card Fraud Detection Targeting the Indian Market", IJCSI International Journal of Computer Science Issue, Vol. 10, Issue 3, No. 2, May 2013

[5] M. A. Vasarhelyi, H. Issa, "Application of Anomaly Detection Techniques to Identify Fraudulent Refunds", 2011

[6] Mohd A.Z. Khan, J.D. Pathan and A.H.E. Ahmed, "Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering", IJARCCE, vol. 3, Pages 5458-5461, February 2014.

[7] Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). "A comprehensive survey of data mining-based fraud detection research", Artificial Intelligence Review, 1–14. Retrieved November 26, 2010, from http://arxiv.org/pdf/1009.6119

[8] Sharmik Sural, "Credit Card Fraud Detection Using Hidden Markov Model", IEEE Transactions on Dependable and Secure Computing", Vol. 5, No. 1, Jan to Mar 2008

[9] S. Benson Edwin Raj, A. Annie Portia, "Analysis on Credit Card Fraud Detection Methods", International Conference on Computer, Communication and Electrical Technology – ICCCET2011, 18th & 19th March, 2011

[10] V. Dheepa, Dr. R. Dhanapal, "Analysis of Credit Card Fraud Detection Methods" IJRTE, vol 2, No. 3, November 2009.

For more information, contact askus@infosys.com

Infosys®
Navigate your next