BT4012 Group Project Report

Fraud Detection in the Financial Services

Group 17

Pichappan Valliammai (A0192263E)
Pong Jia Min, Joan (A0188458J)

Problem Statement
As state-of-art technological capabilities and big data have grown exponentially in recent years, fraudulent activities have been rapidly increasing too. Such activities have resulted in huge losses across various industries - particularly the financial sector. These include credit card fraud and wrongful healthcare insurance claims. In 2018 alone, losses due to payment fraud globally amounted to nearly $28 billion, with the figure projected to exceed $40 billion by 2030. Additionally, fraud in the financial sector has broader ramifications, as such fraud helps fund organised crime, international narcotics trafficking, and even terrorist financing. Hence, this calls for immediate action to curb such fraudulent activities.

Both the papers we analysed pertain to fraud detection in the financial services sector. The paper 'Real-time Credit Card Fraud Detection Using Machine Learning' by A. Thennakoon et al., focuses on reducing financial loss by using machine learning models to detect credit card fraud in real time while the paper 'Mediclaim Fraud Detection and Management Using Predictive Analytics' by M.R. Sumalatha and M. Prabha, focuses on detecting fraudulent medical claims.

Main Results
The similarity in both papers lies in their usage of predictive modelling. The Mediclaim paper made use of Logistic Regression and Multi-Criteria Decision Analysis while the credit card paper primarily used Logistic Regression, Naïve Bayes and Support Vector Machine, which are supervised machine learning models.

Main Results from the Paper on Mediclaim Fraud
The paper's main contribution is using predictive analytics to accurately detect the probability of a medical claim being fraudulent. The authors rightly defined Mediclaim fraud as deliberate dishonesty or falsification of an illegal settlement, and addressed the complications of manual inspection and rule-based approach by proposing predictive modelling (*Figure 1*).

<u>Models Used</u>
With logistic regression, the predictor variables used include years insured, disease diagnosis, number of medical claims filed yearly and previous cases of rejected claims. After processing the raw data, these variables were fitted into the logistic regression model. Pearson Correlation was used to remove variables which exhibited a high correlation (>0.95) while Z-Score and Wald Statistics, set at a threshold of 0.5, were used to eliminate insignificant variables and ensure the model does not consider irrelevant patterns.

For Multi-Criteria Decision Analysis (MCDA), pairwise comparison matrices for every attribute of insurance company, hospital and pharmacy were constructed. The local weights of each combination of attributes were calculated pairwise and normalised, after which, the global weight was determined by multiplying these local weights and the options of the claim being fraudulent or not. The global weight was used as the final decision on whether the claim was fraudulent.

<u>Results</u>
Both the models have a high accuracy of 83.35% and 81.68% respectively. They also have high True Positive Rate and True Negative Rate and low False Positive Rate and False Negative Rate, based on their classification matrices and ROC curves (*Figures 2 and 3*). This indicates that the models are good at accurately distinguishing a legitimate claim from a fraudulent one despite the deceivingly similar patterns. The proposed models perform comparatively better than the existing support vector machine model, which only had an accuracy of 67.8% (*Figure 4*).

Main Results from the Paper on Credit Card Fraud

In this paper, a novel credit card fraud detection system to detect four different types of fraudulent transactions using data-driven machine learning models has been proposed (*Figure 5*). The authors also addressed the problems identified by past researchers in credit card fraud detection such as handling class imbalance, processing of large datasets with categorical and numerical data and choosing optimal machine learning models for identifying fraud. Moreover, they performed an extra step by integrating the detection algorithm in a real-time setting using a GUI, API and Data warehouse to notify the end user over the GUI the second a fraudulent transaction takes place (*Figure 6*).

Data Pre-processing and Handling Class Imbalance

Data pre-processing in this paper consisted of data cleaning, data integration and data transformation. Principal Component Analysis was used to reduce the dimensionality in the dataset.

The authors noted that the dataset was highly imbalanced, with 917781 legitimate cases but only 200 fraudulent cases. To deal with this, resampling techniques were used - Synthetic Minority Oversampling Technique (SMOTE), where the minority class was oversampled by producing synthetic examples; and undersampling techniques Condensed Nearest Neighbour (CNN) and Random Under-Sampling (RUS).

Models Used

Machine learning models used in this paper were mainly Support Vector Machine (SVM), Naive Bayes and Logistic Regression. They were selected through literature, experimenting and parameter tuning. The raw data was divided into four sets according to fraud patterns - i) risky Merchant Category Code; ii) risky ISO response code transactions; iii) unknown web addresses; iv) large amounts of value. After pre-processing the data, the data was fitted into the models, and accuracy was used to measure model performance.

Results

Logistic Regression (oversampled) and Support Vector Machine (undersampled) are best in terms of accuracy at detecting transactions from risky merchants and risky ISO response code. For classifying transactions with amounts >$100 as fraud or not, Support Vector Machine (both oversampled and undersampled) performs the best. For the case of unknown web addresses, Naive Bayes (undersampled) performed the best while Naive Bayes (oversampled) had the lowest accuracy. Overall, there seems to be a major impact of resampling techniques for obtaining a comparatively higher performance from the classifier.

Analysis

Consequences of the results

Both papers have tapped on the predictive machine learning algorithms to improve analytical capabilities in detecting fraud in financial organisations effectively. The main consequence of these papers would be that they have managed to propose models to significantly reduce financial loss that occurs due to phoney transactions or claims. Essentially, they have come up with a framework to detect risky behaviour which helps companies to make better and informed decisions, to reduce losses. There are several challenges in the fraud detection space that these two papers have overcome as well. They have also handled the processing of large datasets for the purpose of fraud detection. Another would be the issue of unbalanced class sizes with legitimate transactions far outnumbering fraudulent ones. This was handled through efficient resampling techniques. Next would be the choice of the predictive system. It is important to choose the optimal methods to reduce misclassifications as falsely identifying a legitimate transaction as fraud and vice-versa also increases the cost of the process.

Now, we will be analysing and interpreting the papers with respect to their assumptions, the models' limitations and possible future direction. We have also proposed an alternative approach using Random Forest.

Assumptions

The credit card paper has attempted to reduce the disparity between classes by undersampling the majority class in a training set. This might not be ideal as they have assumed that redundancy in data justifies the removal. However, they have failed to note that the entire dataset might be relevant for classification purposes. We feel that a suitable alternative could be to use class weights parameters instead. While the credit card paper addressed the skewed dataset issue, the Mediclaim paper did not.

Secondly, both papers have only focused on certain patterns of fraud cases. Site cloning, false merchants online, counterfeit cards, skimming, phishing could also have been considered. Instead they could have made use of unsupervised learning models to detect suspicious patterns instead of attempting to only identify certain fraud cases.

Thirdly, they have assumed that fraud patterns will remain the same over the anticipated time period. By doing so, they failed to address one of the biggest challenges of fraud analytics - concept drift, where fraud patterns keep evolving. The proposed supervised learning models lack adaptability and need to be constantly retrained. It could have been better if reinforcement learning models are used instead.

Limitations

Some general limitations would be the usage of accuracy which is not well suited for fraud detection where significant class imbalance exists in the data. Both the papers could have set aside a baseline method to compare and done parameter tuning on the models. We will now focus mainly on some of the drawbacks of the models that the papers have used to detect fraudulent activity in the financial sector.

Limitations of Logistic Regression

A major limitation of logistic regression is that it assumes linearity between the dependent variable and independent variables. This is not always the case in fraud analytics. Moreover, no multicollinearity should exist between the independent variables. Although this was taken care of in the credit card paper, the Mediclaim paper failed to address it. Also, the result of classification was not categorical (fraudulent/legitimate). Instead, it was an estimated probability of each observation belonging to a given class and a threshold had to be set manually. Maximum likelihood estimation used in logistic models is well-known to suffer from small-sample bias. This is a disadvantage here due to the small size of the fraudulent class.

Limitations of Multivariate Decision Analysis (MCDA)

We researched and found that MCDA is unsuitable where multiple attributes are being considered [1]. Furthermore, most financial fraud solutions are premised on sets of predefined rules and these are not enough to trap recent sophisticated fraud means [2]. Small increments in the quantity of evaluation criteria and alternatives result in large increases in input and thus takes a substantial amount of manual work to train. Hence, MCDA tends to be time-consuming and more costly than other methods.

In the Mediclaim paper, several attributes across the different medical sectors were into play. The nature of the claim was calculated using what seems like a rule-based approach. This is not effective in trapping suspicious patterns.

Limitations of Support Vector Machine (SVM)

In general, the SVM algorithm requires large quantities of computational time and memory for large data sets and does not perform very well when the data set has overlapping target classes. Fraud detection models involve a large number of constantly evolving data points from various sources where the fraudulent patterns mimic the legitimate ones. Hence, SVM will find it difficult to analyse and distinguish the two classes. The strength of SVMs comes from the kernel representation, however, identifying the right kernel function to map features to a high-dimensional feature space for the classification task in the financial sector is tricky.

Our Approach

Although supervised models like regression analysis and SVM provide a significant lift over domain expert-driven rules, non-linear techniques such as neural networks and random forest are essential to reduce false positives and increase quality of fraud detection while decreasing the underlying cost of prevention. From 'Comparative Approach to Predictive Analytics with Machine Learning for Fraud Detection of Real Time Financial Data (ICONC3)', the most widely used predictive models for fraud detection in financial services are:

1) Neural networks: They have the ability to model complex patterns and provide scalable nonlinear mappings. The most prevalent technique for supervised credit card fraud detection is artificial neural networks (ANN). Moreover, other papers have stated that deep learning algorithms like long short-term memory, recurrent neural network and gated recurrent unit with a larger network size perform exceptionally well. Given the benefits of neural networks in fraud detection, we propose using neural networks as an alternative method.

2) Ensemble methods - Random Forest (RF): Recently, the use of ensemble methods has found to perform the best in credit card fraud [3]. Random forests work especially well when there are many input features to learn suspicious patterns from, which is often the case in network-related classification problems such as this. They carry natural variable selection ability and have been noted to perform well with high dimensional non-linear data. Hence RF can handle the non-linear problem without any resampling. They are computationally efficient with only two adjustable parameters and can perform well even with the default values. RF overcomes limitations of the models proposed in the papers that we have read and trumps other models when it comes to fraud detection.

It can also be observed from 'APATE: A novel approach for automated credit card transaction fraud detection using network-based extension' that the Random Forest model performs the best (*Figure 7)*.

Our Experiment with Random Forest in the Financial Sector involving Cryptocurrency

Firstly, we obtained the open-source data on the Ethereum network. The data contained features such as transaction amount, from and to addresses, entity and category. Feature engineering was done based on domain knowledge and experimentation, with the new numeric features computed using the recency-frequency-monetary variables concept. Our new features include: i) Incoming and Outgoing Transactions; ii) Incoming and Outgoing High-valued transactions; iii) Balance to Transaction Amount Ratio; iv) Value Coming in and Going Out (in USD); v) Interactions with Addresses created on the same date.

We had to take into account the graph nature and time-series factor of the data. After these numeric features were created, they were all normalised using standard scalar. Using a computed feature risk score, we processed the labels for the dataset too. We then proceeded to train the Random Forest using the Sklearn package in Python. We were

attempting to understand the performance of the model - predicted hack cases vs actual hack cases. To ensure that our model has a small error margin, we retrieved a balanced dataset of hack and non-hack labels, covering different stratas. We managed to iteratively improve our scoring system with each experimentation. In each attempt, we tested on 20% of the train data and on test sets with extreme data points (*Figure 8*).

We made use of precision and recall to measure performance, as they are better metrics to use when handling imbalanced datasets. The model seemed to have a high recall regardless of the strata (*Figure 9*). This reflects the fact that the model has a consistently low level of False Negative Rate and does an excellent job at distinguishing risky from non-risky patterns. After this, we proceeded to use Shapely from the SHAP Python package to visualise various transactions and identify the typical characteristics of anomalous patterns (*Figure 10*). General fraudulent patterns include a combination of single digit transactions and large value of Ethereum in USD flowing in and out, large balance to transaction ratio and many interactions with addresses created the same day. Shapely also allowed us to understand the reason for a high False Positive Rate. It seemed like the model was able to pick up suspicious patterns that we had missed when we were labelling the data using a rules-based approach, hence predicting the legitimate data as fraudulent. This proves how efficient the Random Forest model is in fraud detection, in a financial service context.

Future Direction

The future direction of these models is that it can be generalised and implemented as a framework on a real-time basis across all services in this financial services sector including Banking, Insurance, Auditing, Stock Brokerage, etc. Anomaly detection can be extended to detect unusual and critical fraud patterns in the data as opposed to just detecting single fraudulent data points. Another development could be contribution analysis or model explainability of the fraud detection model which could provide the users with the context in which the event has occurred.

Moreover, fraud detection systems need to deal with changing fraud patterns over time. By using advanced algorithms like reinforcement learning, financial organisations can customise services and simplify management. Train and test data can be set up such that models are tested for their predictive ability in subsequent time periods to examine whether fraud patterns remain in effect over time. Also, the cost of prevention of illegal activities versus cost of deploying these detection models also needs to be considered. Hence, a good future application would be a cost-saving architecture which can create ingenious derived attributes to help classify transactions accurately without the need to manually derive them.

Conclusion

In our review paper, novel approaches such as machine learning models and predictive analytics used to detect the probability of an activity in the financial sector being fraudulent was being studied. We focused on two cases in the financial sector: credit card fraud and Mediclaim insurance fraud. After which, we conducted our very own experiment using Random Forest, a popular robust ensemble model for detecting fraud in the Ethereum network. Overall, we have managed to understand, explore and analyse different types of machine learning models, their limitations and possible future directions when it comes to detecting fraud.

Appendix

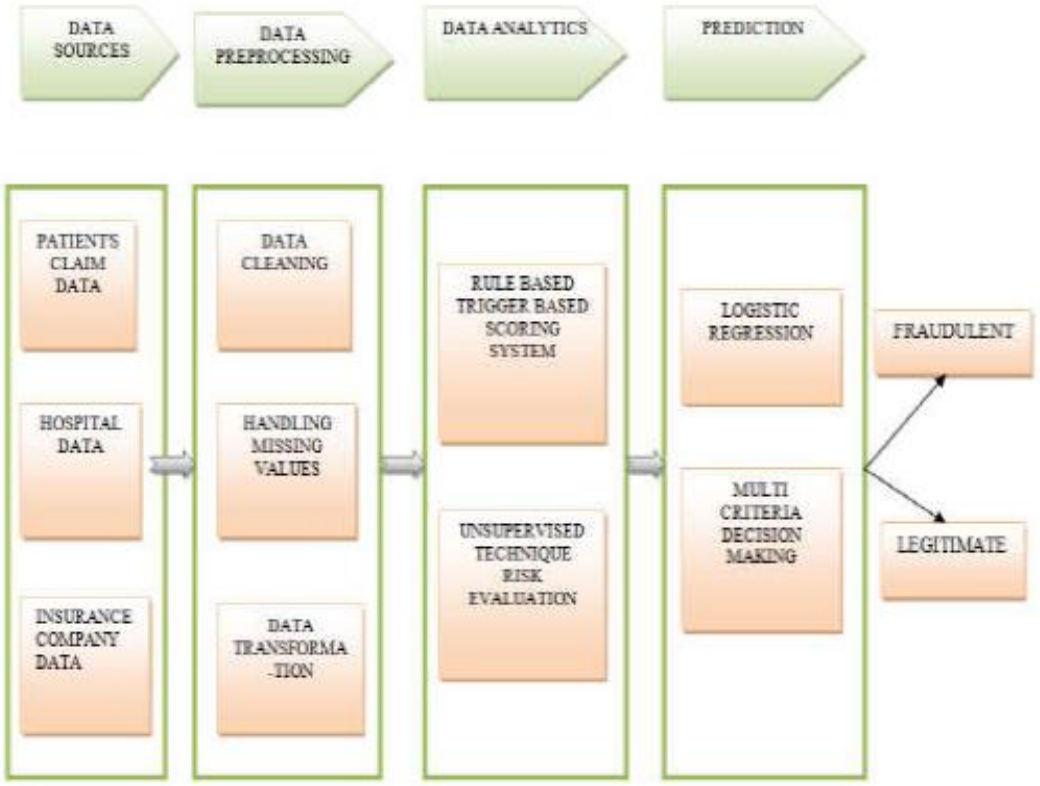


Fig 1 Mediclaim fraud detection framework

Figure 1: Mediclaim Fraud Detection Framework

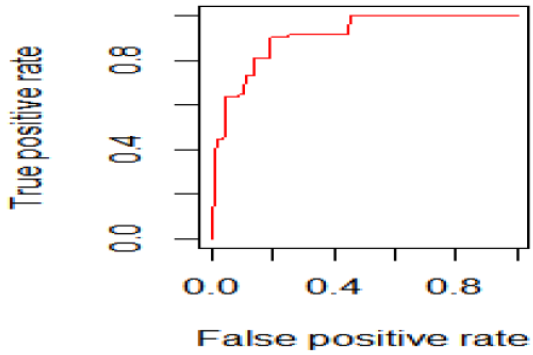


Figure 2: ROC Curve of Logistic Regression

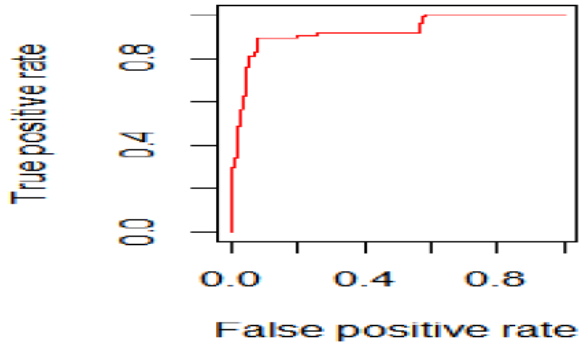


Fig 5 ROC curve of multi criteria decision analysis

Figure 3: ROC Curve of Multi-Criteria Decision Analysis

Logistic regression	Multi criteria decision analysis	Support vector machine
83.35%	81.68%	67.8%

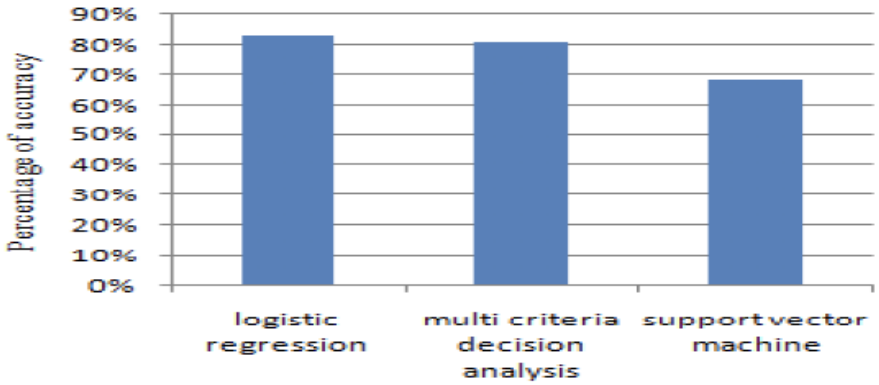


Figure 4: Comparison of Accuracy between proposed models and baseline model (SVM)

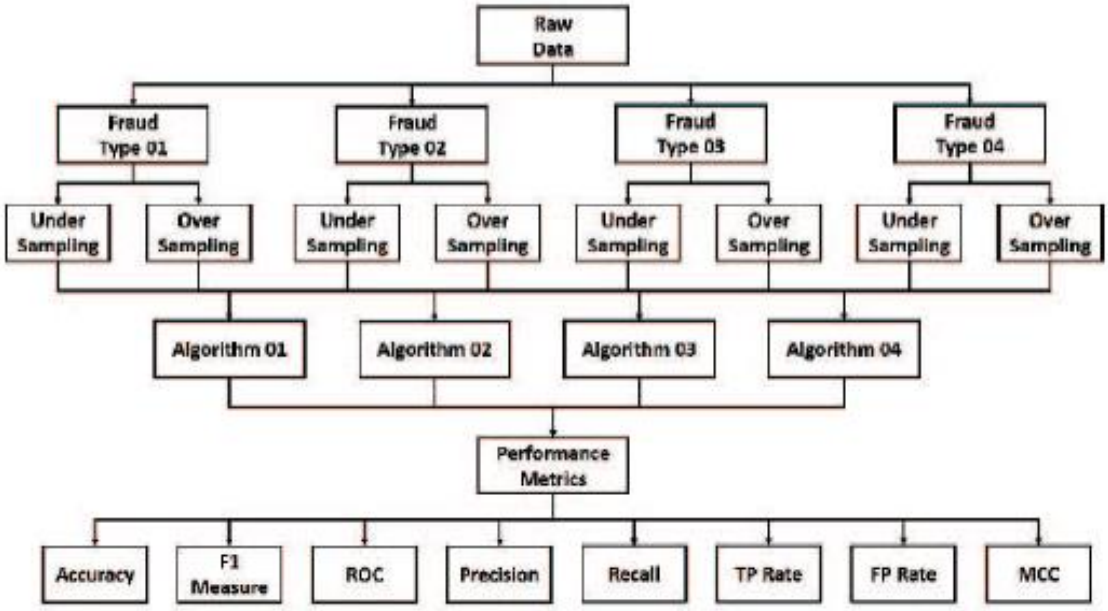


Fig. 2. Model Selection

Figure 5: Model Selection in the Credit Card Paper

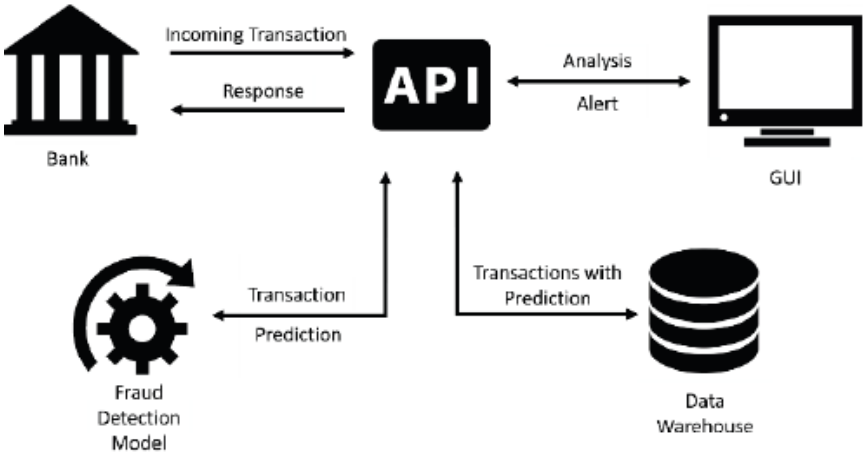
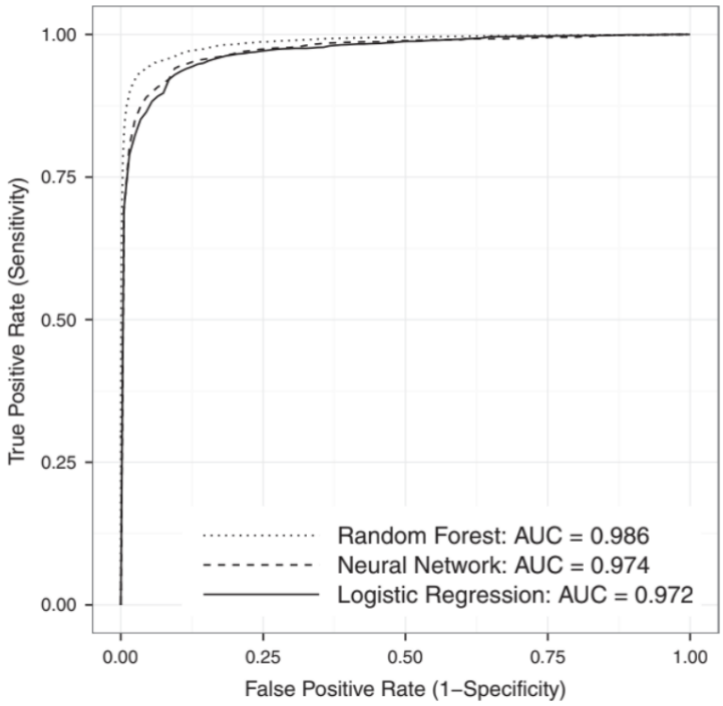


Fig. 7. System Diagram

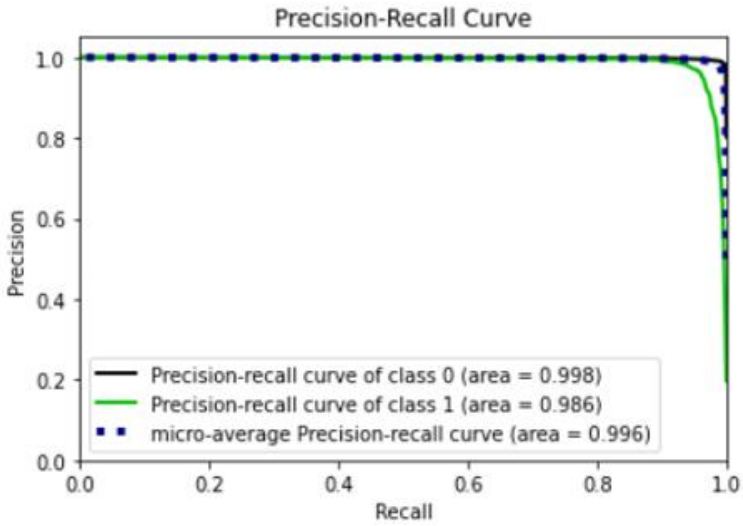
Figure 6: Proposed architecture for real-time detection of card-not-present fraud



Comparison of models.

Model	AUC	Accuracy
Logistic regression	0.972	95.92%
Neural networks	0.974	93.84%
Random forests	0.986	98.77%

Figure 7: Comparison of methods used in automated credit-card fraud detection




```
Hack_Pred # : 4152, Hack_Truth #: 4334, Total_label #: 21953


Precision of predicting risky transactions:0.98
Recall of predicting risky transactions:0.939
```

Figure 8: Performance of Random Forest Model while detecting risky entities

	Risk	Strata	Total Label	Hack Truth	No of Risky Txns Pred at 0.4	Precision 0.4	Recall 0.4	No of Risky Txns Pred at 0.5	Precision 0.5	Recall 0.5
0	Hack_risk	0	30000.0	10.0	378.0	0.026455	1.000000	347.0	0.028818	1.000000
1	Hack_risk	20	27129.0	20.0	118.0	0.169492	1.000000	80.0	0.250000	1.000000
2	Hack_risk	40	20390.0	16.0	434.0	0.029954	0.812500	338.0	0.038462	0.812500
3	Hack_risk	60	7104.0	20.0	631.0	0.031696	1.000000	584.0	0.034247	1.000000
4	Hack_risk	80-100	6925.0	3335.0	3609.0	0.875866	0.947826	3532.0	0.885334	0.937631
5	Scam_risk	0	30000.0	1.0	370.0	0.002703	1.000000	342.0	0.002924	1.000000
6	Scam_risk	20	25403.0	19.0	116.0	0.112069	0.684211	81.0	0.160494	0.684211
7	Scam_risk	40	18539.0	44.0	710.0	0.056338	0.909091	565.0	0.069027	0.886364
8	Scam_risk	60	8413.0	47.0	1542.0	0.028534	0.936170	1373.0	0.030590	0.893617
9	Scam_risk	80-100	23275.0	18749.0	20115.0	0.929704	0.997440	19908.0	0.938919	0.996960

Figure 9: Performance of model when tested on different stratas - High recall in all cases

	Columns	SHAP values	Actual values
3	incoming_large_txns	-0.108880	2.000
1	incoming_sdc_addr	0.106496	0.000
7	value_in	-0.086193	43507.799
8	value_out	-0.075792	43507.700
5	outgoing_txns	-0.065950	5.000
2	outgoing_large_txns	-0.056360	2.000
6	incoming_txns	-0.047422	5.000
4	hbr_txns	-0.040949	5.000
0	outgoing_sdc_addr	0.035741	0.000

```
entity           None
category     Untagged
hack_risk           0
scam_risk           0
Name: 4497, dtype: object
```

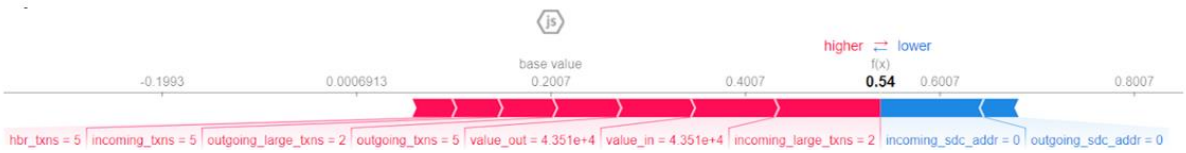


Figure 10: Using Shapely to interpret the predicted probas; this entity has been identified as risky though labelled as non-risky. The RF model identified a suspicious pattern that our rules-based approach did manage to capture.

References

[1] Chang, W. & Chang, J., (2012), An effective early fraud detection method for online auctions

[2] Viaene, S., Dedene, G. Dedene, & Derring R. A., (2005). Auto claim Fraud detection using Bayesian learning neural networks

[3] Desai, V.S., Crook, J.N., & Overstreet, J., (1996). A comparison of neural networks and linear scoring models in the credit union environment