Fraud Detection Using Deep Learning

Apoorv Tyagi[1], Durvesh Satish Deshmukh[2], Nitish Surana[3], Aditya Ranjan[4], M. Viswanath[5]
[1,2,3,4,5]Student, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

***Abstract*: The increment of computer technology use and the continued growth of companies have enabled most financial transactions to be performed through the electronic commerce systems, such as using the credit card system, telecommunication system, healthcare insurance system, etc. Unfortunately, these systems are used by both legitimate users and fraudsters. It is hard to predict the exact scale of the fraud because most of the time it remains undetected. Even after numerous mechanisms to stop fraud, fraudsters are continuously trying to find new ways and tricks to commit fraud. Thus, to stop these frauds, we need a powerful fraud detection system capable of identifying fraudulent transactions with high accuracy. For this paper, we will be using the Fraud Detection data hosted on Kaggle.com to find fraudulent transactions among the genuine ones. We plan on applying numerous machine learning algorithms to create different models that can evaluate the dataset predict whether the respective transaction is fraudulent or not. We will use the training dataset hosted on the site to train the model and evaluate the following model using the testing dataset provided.**

***Keywords*: Online Fraud Detection, Machine Learning, Classification, Neural Networks, Data Mining**

1. Introduction

Fraud is a large-scale problem that affects the various entities from the public sector to private sectors including government, profit, and non-profit organizations. It is hard to predict the exact scale of the fraud because most of the time it remains undetected. It is very important to detect financial frauds and save the company's or the tax payer's money. To detect these frauds, we need to understand the patterns present in the fraudulent transactions and use it to predict the future frauds.

A good fraud detection system should be able to identify the fraud transaction accurately and should make the detection possible in real-time transactions.

Traditionally, many major banks have relied on old rules-based expert systems to catch fraud, but these systems have proved all too easy to beat; the financial services industry are relying on increasing complex fraud detection algorithms. Many in the financial services industry have updated their fraud detection to include some basic machine learning algorithms including various clustering classifiers, linear approaches, and support vector machines. The most advanced companies in the financial services industry, such as PayPal, have been pioneering more advanced artificial intelligence techniques such as deep neural networks and autoencoders.

Neural networks were introduced to detect credit card frauds in the past. Now, we focus on deep learning that is a subfield of machine learning (ML). This would be a two-step process involving pattern recognition and fraud detection. By extracting key information about some pattern in which fraudulent activities take place can help in providing a probability about fraudulent transactions taking place. Neural network-based fraud detection is based totally on the human brain working principal. Neural network technology has made a computer capable of thinking.

As the human brain learns through past experience and uses its knowledge or experience in making the decision in daily life problem the same technique is applied with the credit card fraud detection technology. When a particular consumer uses its credit card for example in case of an online transaction, there is a fixed pattern of credit card use, made by the way consumer uses its credit card. So, when the credit card is being used by unauthorized user, the neural network-based fraud detection system checks for the pattern used by the fraudster and matches with the pattern of the original cardholder on which the neural network has been trained, if the pattern matches the neural network declare the transaction as "ok".

When we say that the transactions need to match the pattern to understand if its genuine or not, we mean to say that it should be identical to the pattern that is stored in the neural network. When the pattern is sent as an input to the neural network it determines how similar it is to the prescribed pattern. Based on the similarity measure it categories the respective transactions as fraudulent or not.

The similarity as we understand for a normal user on an average can have variable usage patterns, which is exactly why we have a measure to determine the transaction and classify it accordingly. Having these classifications, help the developers understand the wide variety of the user's better and improve the overall experience for them by keeping a margin for the outliers and also being capable enough to handle fraudulent transactions thereby making the system more robust and capable.

2. Modules

There are 5 major modules. They are listed below:

Data Cleaning: To detect and correct corrupt or inaccurate records from a record set, table, and to identify incomplete, incorrect, inaccurate parts of the data and then replacing, modifying, or deleting the irrelevant data.

Exploratory Data Analysis and Feature Engineering: To analyze data set to summarize their main characteristics, with some visual methods as well.

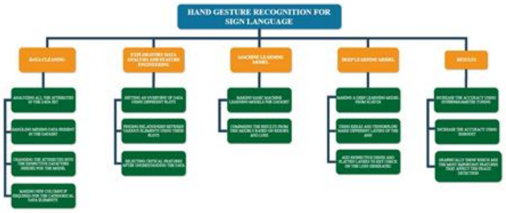
Machine Learning Model: A machine learning model can be a mathematical representation of a real-world process. To generate a machine learning model we will need to provide training data to a machine learning algorithm to learn from it.

Deep learning model: Deep learning models are built using neural networks. A neural network takes in inputs, which are then processed in hidden layers using weights that are adjusted during training. Then the model spits out a prediction. The weights are adjusted to find patterns in order to make better predictions.

Results: To verify the result of our prediction with our test dataset. Implementation of Hyperparameter tuning to attain the parameters that get the best result from the model.

3. Work break down structure

- *Data Cleaning:* It is the first, most important and most time-consuming step any Data Science or Machine learning workflow. Cleaning the data means creating a dataset that is free of errors. Filtering and modifying the data such that it is easier to explore, understand, and model. Without clean data, it will be a hard time seeing the actually important parts of your exploration.
- *Exploratory Data Analysis and Feature Engineering:* Exploratory Data Analysis is the process of performing an initial analysis of data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Feature engineering is the process of using domain knowledge of the data to create features that make machine-learning algorithms work. If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process.
- *Machine Learning Model:* With no specific machine learning assigned for a particular set of problems, we need to try to evaluate the data set on varied algorithm implementation to get a model that best represents the dataset's behavior. Classifications models such as Naïve Bayes, SVM, KNN, Decision Tree, Random Forest, etc. would be implemented to help us classify a given problem into binary classifiers or multi-class classifiers.
- *Deep Learning Model:* An artificial neural network consists of an interconnected group of artificial neurons. The principle of the neural network is motivated by the functions of the brain especially pattern recognition and associative memory. The neural network recognizes similar patterns, predicts future values or events based upon the associative memory of the patterns it was learned. It is widely applied in classification and clustering. The advantages of neural networks over other techniques are that these models are able to learn from the past and thus, improve results as time passes. They can also extract rules and predict future activity based on the current situation. By employing neural networks, effectively, banks can detect fraudulent use of a card, faster and more efficiently.
- *Results:* The model created would be tested against the test dataset, with the accuracy and loss factor the entities that determine the merit of the model used. Implementation of Hyperparameter tuning to attain the parameters that get the best result from the model. Predictive analysis helps in identifying the factors that affect the results (accuracy and loss) of the model in a positive and negative manner thus conferring insights that would affect the result.


HAND GESTURE RECOGNITION FOR SIGN LANGUAGE
DATA CLEANING
EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING
MACHINE LEARNING MODEL
DEEP LEARNING MODEL
RESULTS
ANALYZING ALL THE ATTRIBUTES IN THE DATA SET
HANDLING MISSING DATA PRESENT IN THE DATASET
CHANGING THE ATTRIBUTES WITH THE RESPECTIVE DATATYPES NEEDED FOR THE MODEL
MAKING NEW COLUMNS IF REQUIRED FOR THE CATEGORICAL DATA ELEMENTS
GETTING AN OVERVIEW OF DATA USING DIFFERENT PLOTS
FINDING RELATIONSHIP BETWEEN VARIOUS ELEMENTS USING THESE PLOTS
SELECTING CRITICAL FEATURES AFTER UNDERSTANDING THE DATA
MAKING BASIC MACHINE LEARNING MODELS FOR DATASET
COMPARING THE RESULTS FROM THE MODELS BASED ON SCORING AND LOSS
MAKING A DEEP LEARNING MODEL FROM SCRATCH
USING KERAS AND TENSORFLOW MAKE DIFFERENT LAYERS OF THE ANN
ADD RESPECTIVE DENSE AND FLATTEN LAYERS TO KEY CHECK ON THE LOSS GENERATED
INCREASE THE ACCURACY USING HYPERPARAMETER TUNING
INCREASE THE ACCURACY USING XGBOOST
GRAPHICALLY SHOW WHICH ARE THE MOST IMPORTANT FEATURES THAT AFFECT THE FRAUD DETECTION

4. Design of the model

A. Preprocessing the Data

- *Data Formatting:* The data should be in standardized record format, with the variables in the features representing each attribute in a similar way. Data also must be consistent.
- *Data Cleaning:* It includes a set of procedures for removing noise and fixing inconsistencies in data. Dealing with the missing values, features with inconsistent data types, redundant features removal and outlier detection. It also includes removing incomplete, useless and redundant data objects.
- *Data Scaling:* - Data in the dataset may have numeric attributes that span different ranges, for example, millimeters, meters, and kilometers. Scaling is about converting these attributes so that they will have the same scale, such as between 0 and 1, or 1 and 10 for the smallest and biggest value for an attribute.

B. Exploratory Data Analysis

- *Data Visualization:* A large amount of information represented in graphic form is easier to understand and analyze. Its main goal is to distil large datasets into visual graphics to allow for easy understanding of complex relationships within the data. Data visualization can provide insight that traditional descriptive statistics cannot.
- *Feature Engineering:* Preparing the proper input dataset, so that is compatible with the machine learning algorithm requirements. The features included in the dataset influence the result. Hence, features must be carefully chosen so that they are

relevant to the problem statement and result.

C. *Dataset Splitting*

- *Training set:* The training set is a major fraction of the entire dataset that is fed to the machine learning models, learning from the algorithm that is employed in the model. The parameters for the model have to learn from this set of data and make predictions.
- *Testing set:* A test set is needed for an evaluation of the trained model and its capability for generalization. It is a way of recognizing a model's ability to identify patterns in new unseen data after having been trained over a training dataset. It is important to use different subsets for training and testing to avoid model overfitting.
- *Validation set:* Validation set is used to tweak a model's Hyperparameter. Tuning the Hyperparameter makes the model recognize the patterns faster and also help in increasing the accuracy of the model.

D. *Modelling*

Model Training: Model Training is used to develop a model that learns from the data and output accurate predictions based on the patterns learned from the input data. The data after it has been preprocessed and split into subsets, the training data is fed to the model. The algorithm trains the data and outputs a model that can find the target value of new data that would be fed into the model.

E. *Algorithms*

- *Neural network:* An artificial neural network consists of an interconnected group of artificial neurons. The principle of the neural network is motivated by the functions of the brain especially pattern recognition and associative memory. The neural network recognizes similar patterns, predicts future values or events based upon the associative memory of the patterns it was learned. It is widely applied in classification and clustering. Neural networks have shown a tendency to produce the best result for only large transaction dataset. Since the Neural networks learn from examples, the more the data that is available better it is. While employing the neural networks, we would employ a technique which will drop a feature that has the least relevant effect on the outcome and observe if it increases or decreases the performance. If the performance increases then the feature is dropped, else the feature stays. This would help us in identifying the features that dictate the outcome for the model and hence would be more crucial in classification.
- *XGBOOST:* XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. It is a highly flexible and versatile tool that can work through most regression, classification and ranking problems as well as user-built objective functions. It is capable of performing the three main forms of gradient boosting and it is robust enough to support fine-tuning and addition of regularization parameters. The algorithm was developed to efficiently reduce computing time and allocate an optimal usage of memory resources. Important features of implementation include handling of missing values, Block Structure to support parallelization in tree construction and the ability to fit and boost on new data added to a trained model.

F. *Model Evaluation*

- *Accuracy:* The accuracy measure gives us an intimation of the number of target output variables that match the actual output values. It is the fraction of instances that the model was able to rightly predict. It gives a good idea about the performance of a model or the algorithm for a given problem.
- *Confusion Matrix:* A confusion matrix is a table that is used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.
- *ROC Curve:* Measuring the area under the ROC curve is also a very useful method for evaluating a model. ROC is the ratio of True Positive Rate (TP) and False Positive Rate (FP). It is a performance measurement for the classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes.

G. *Model Tuning*

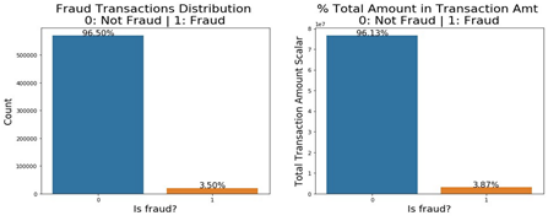
Hyperparameter Tuning: We do not know the best hyperparameters for a model for any problem, so we search for the best hyperparameter values suiting the model through the trial and error process. Thus, hyperparameter tuning is the process of finding the ideal model architecture that best suits the problem and hence increases the accuracy and speed of convergence.

Model parameters are learned during training when we optimize a loss function using something like gradient descent.

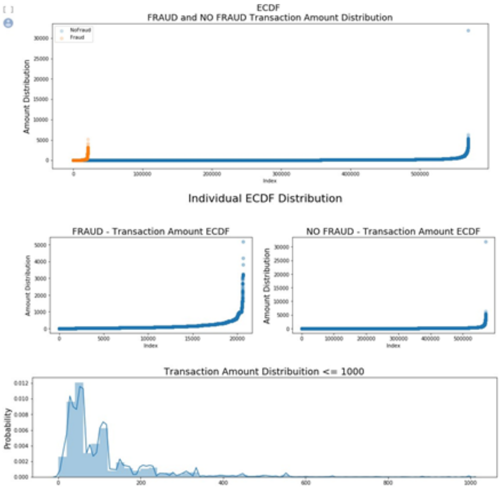
5. Data exploration

A. *How much percentage of the transactions are fraudulent and what is the total amount lost in it?*

It was inferred that around 96.5 % of the transactions in the dataset are, not fraudulent and only 3.5% of the transactions are fraudulent. In the total amount of the transaction money provided in the dataset, only 3.87% is lost in fraud.

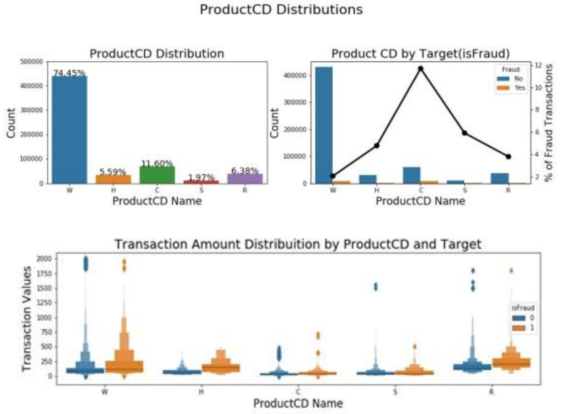


B. How is the transaction amount distributed in the dataset? What is the highest amount that was detected as a fraud?



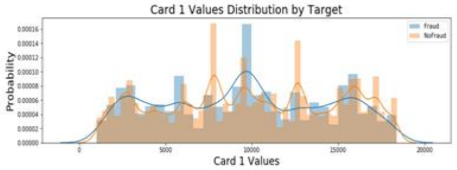
It was inferred that the highest amount that was detected as fraud is around 5000 USD. From the subsequent graphs, it was inferred that most of the fraudulent transactions take place between 1000-3000 ranges. The respective probability of fraudulent transactions under 1000 USD was found to be as represented in the last graph.

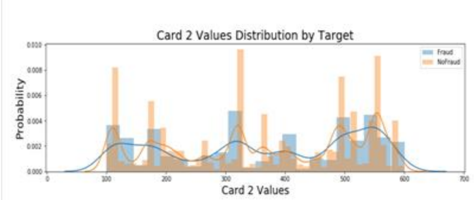
C. What is the different type of products provided and its respective distribution? How does fraudulence varies based on it and its respective fraudulent and non-fraudulent amounts?

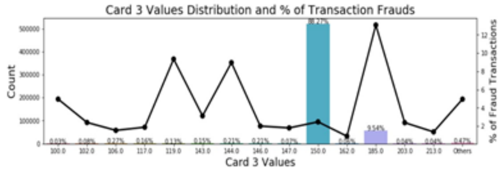


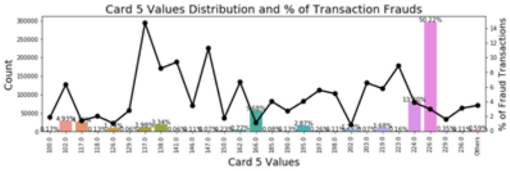
It is observed that W, C and R were the most frequent ones and by comparing all of them to the fraudulent transactions and drawing a ratio – in W, H and R the distribution of Fraud values are slightly higher than the Non-Fraud Transactions.

D. How are the card values distributed? Is there any strong correlation between different card values and fraud committed? If yes, how much fraud amount has been seen from these card values?





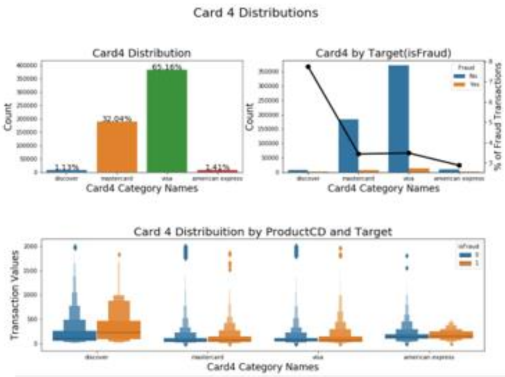


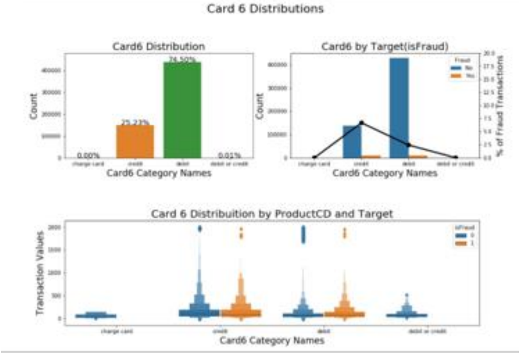


The non-categorical data has been visually plotted and shown above. The first diagram shows how the values are distributed among all the six different card values provided to us, how many unique values, how many null values and the respective data type after reducing the size of the dataset. The next two diagrams show how the data is being distributed among the different card and how much fraudulent amount is related to them respectively. In Card3, we can see that 150 and 185 are the most common values in the column. The values with highest Fraudulent Transactions are in 185 followed by 119 and then 144. In Card5, the most frequent values are 226, 224, and 166 that represents 73% of data. Also is possible to see high percentage of frauds in 137, 147, and 141 that has few entries for values.

E. How are the categorical card features distributed? Is there any strong correlation between these categorical attributes and fraudulent transactions? What is its distribution and the amount lost in the fraud?

As the first graph shows, the maximum number of transactions are made using Visa and MasterCard. Even though Discover has the least number of transactions, the maximum amount of fraudulent transactions is observed there. The following graph shows the respective distribution of the Card4 values in a boxplot, coupled with the transaction money. Their distribution is provided in the following boxplots. From the card value graph, we can see that nearly all card types are credit or debit. Fraudulent transactions were observed more in credit than in debit.

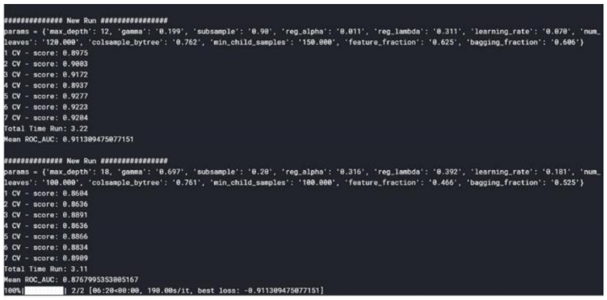




6. Results and discussion

A. XGBoost

The dataset was run on stratified K-fold iterator on the XGBoost algorithm. In order to obtain the best result, a ist of hyper parameter were listed to be tried out by the model to get the combination of best parameters to subsequently increase the accuracy of the model. The ROC AUC (Receiver operating characteristics – Area Under the Curve) was taken to be the metric system for the model's performance evaluation purpose.



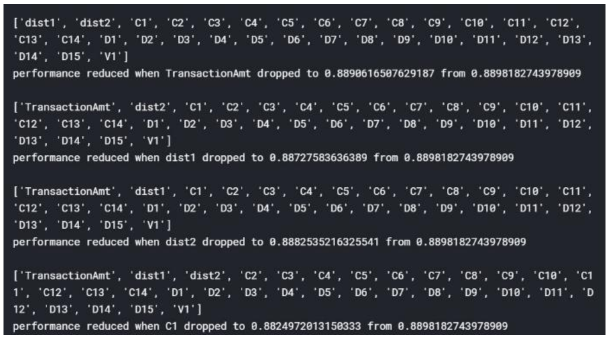
```
############## New Run ################
params = {'max_depth': 12, 'gamma': '0.199', 'subsample': '0.90', 'reg_alpha': '0.011', 'reg_lambda': '0.311', 'learning_rate': '0.070', 'num_
leaves': '120.000', 'colsample_bytree': '0.762', 'min_child_samples': '150.000', 'feature_fraction': '0.625', 'bagging_fraction': '0.606'}
1 CV - score: 0.8975
2 CV - score: 0.9083
3 CV - score: 0.9172
4 CV - score: 0.8937
5 CV - score: 0.9277
6 CV - score: 0.9223
7 CV - score: 0.9284
Total Time Run: 3.22
Mean ROC_AUC: 0.911309475077151

############## New Run ################
params = {'max_depth': 18, 'gamma': '0.697', 'subsample': '0.20', 'reg_alpha': '0.316', 'reg_lambda': '0.392', 'learning_rate': '0.181', 'num_
leaves': '100.000', 'colsample_bytree': '0.761', 'min_child_samples': '100.000', 'feature_fraction': '0.466', 'bagging_fraction': '0.525'}
1 CV - score: 0.8604
2 CV - score: 0.8636
3 CV - score: 0.8891
4 CV - score: 0.8636
5 CV - score: 0.8866
6 CV - score: 0.8834
7 CV - score: 0.8909
Total Time Run: 3.11
Mean ROC_AUC: 0.8767995353085167
100%|██████████| 2/2 [06:20<00:00, 190.00s/it, best loss: -0.911309475077151]
```

The final accuracy was found to be 91.1%.

B. Deep Learning Model

After the data preprocessing is done, different numerical and categorical entities are separated. The NA or empty columns in numerical entities are replaced by zero and then transferred to standard scaler to normalize the values before it can be sent into the model. Categorical entities are encoded using the Label Encoder provided in the sklearn library. The neural network used is fairly standard. We will use the embedding layer for categorical and the numerical will go through feed-forward dense layers. We create our embedding layers such that we have as many rows as we had categories and the dimension of the embedding is the log1p + 1 of the number of categories. Therefore, this means that categorical variables with very high cardinality will have more dimensions but not significantly more so; the information will still be compressed down to only about 13 dimensions and the smaller number of categories will be only 2-3. We will then pass the embeddings through a spatial dropout layer which will drop dimensions within the embedding across batches and then flatten and concatenate. Then we will concatenate this to the numerical features and apply batch norm and then add some more dense layers after. We then extract the features we actually want to pass to the NN.



```
['dist1', 'dist2', 'C1', 'C2', 'C3', 'C4', 'C5', 'C6', 'C7', 'C8', 'C9', 'C10', 'C11', 'C12',
'C13', 'C14', 'D1', 'D2', 'D3', 'D4', 'D5', 'D6', 'D7', 'D8', 'D9', 'D10', 'D11', 'D12', 'D13',
'D14', 'D15', 'V1']
performance reduced when TransactionAmt dropped to 0.8890616507629187 from 0.8898182743978909

['TransactionAmt', 'dist2', 'C1', 'C2', 'C3', 'C4', 'C5', 'C6', 'C7', 'C8', 'C9', 'C10', 'C11',
'C12', 'C13', 'C14', 'D1', 'D2', 'D3', 'D4', 'D5', 'D6', 'D7', 'D8', 'D9', 'D10', 'D11', 'D12',
'D13', 'D14', 'D15', 'V1']
performance reduced when dist1 dropped to 0.88727583636389 from 0.8898182743978909

['TransactionAmt', 'dist1', 'C1', 'C2', 'C3', 'C4', 'C5', 'C6', 'C7', 'C8', 'C9', 'C10', 'C11',
'C12', 'C13', 'C14', 'D1', 'D2', 'D3', 'D4', 'D5', 'D6', 'D7', 'D8', 'D9', 'D10', 'D11', 'D12',
'D13', 'D14', 'D15', 'V1']
performance reduced when dist2 dropped to 0.8882535216325541 from 0.8898182743978909

['TransactionAmt', 'dist1', 'dist2', 'C2', 'C3', 'C4', 'C5', 'C6', 'C7', 'C8', 'C9', 'C10', 'C1
1', 'C12', 'C13', 'C14', 'D1', 'D2', 'D3', 'D4', 'D5', 'D6', 'D7', 'D8', 'D9', 'D10', 'D11', 'D
12', 'D13', 'D14', 'D15', 'V1']
performance reduced when C1 dropped to 0.8824972013150333 from 0.8898182743978909
```

The final accuracy obtained is 88.98%.

7. Conclusion

The outcome of this paper is a model that is able to detect frauds with a high probability of 91.1% and 88.98%. Crucially, this paper provides the foundation for the development of a definitive neural network encompassing the usage pattern of the user via the network-based fraud detection system. This will eventually become a methodology to identify a wide variety of users and help developers increase the experience for them by keeping a margin for the outliers and being capable enough to handle fraudulent transactions thereby making the system more robust and capable.

References

[1] Yufeng Kou, Chang-Tien Lu, S. Sirwongwattana and Yo-Ping Huang, "Survey of fraud detection techniques," *IEEE International Conference on Networking, Sensing and Control, 2004*, Taipei, Taiwan, 2004, pp. 749-754 Vol.2.

[2] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research.

[3] Ghosh and Reilly, "Credit card fraud detection with a neural-network," *1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences*, Wailea, HI, USA, 1994, pp. 621-630.

[4] R. Brause, T. Langsdorf and M. Hepp, "Neural data mining for credit card fraud detection," *Proceedings 11th International Conference on Tools with Artificial Intelligence*, Chicago, IL, USA, 1999, pp. 103-106.