Credit Card Fraud Detection Using Historical Transaction Data

1. Problem Statement

With the growth of e-commerce websites, people and financial companies rely on online services to carry out their transactions that have led to an exponential increase in the credit card frauds [1]. Fraudulent credit card transactions lead to a loss of huge amount of money. The design of an effective fraud detection system is necessary in order to reduce the losses incurred by the customers and financial companies [2]. Research has been done on many models and methods to prevent and detect credit card frauds. Some credit card fraud transaction datasets contain the problem of imbalance in datasets. A good fraud detection system should be able to identify the fraud transaction accurately and should make the detection possible in real-time transactions. Fraud detection can be divided into two groups: anomaly detection and misuse detection. Anomaly detection systems bring normal transaction to be trained and use techniques to determine novel frauds. Conversely, a misuse fraud detection system uses the labeled transaction as normal or fraud transaction to be trained in the database history. So, this misuse detection system entails a system of supervised learning and anomaly detection system a system of unsupervised learning [2]. Fraudsters masquerade the normal behavior of customers and the fraud patterns are changing rapidly so the fraud detection system needs to constantly learn and update. Credit card frauds can be broadly classified into three categories, that is, traditional card related frauds (application, stolen, account takeover, fake and counterfeit), merchant related frauds (merchant collusion and triangulation) and Internet frauds (site cloning, credit card generators and false merchant sites) [1].

2. Background

Timely information on fraudulent activities is strategic to the banking industry as banks have huge databases with variety. Valuable business information can be extracted from these data stores. Credit card frauds can be broadly classified into three categories, that is, traditional card related frauds (application, stolen, account takeover, fake and counterfeit), merchant related frauds (merchant collusion and triangulation) and Internet frauds (site cloning, credit card generators and false merchant sites)

3. Methodology

Basically, there are five basic steps for the data mining process which defines the problem. 1) preparing data 2) exploring the data 3) development of the model 4) exploration and validation of the models 5) deployment and updation in the models. In this project, Neural network is used as the data mining technique and it utilized above mentioned steps for accurate and reliable result. Moreover, Neural network was used as it has the capability of adaption and generalization. Moreover, H2O [3] is also a good option for the experiment purpose. H2O flow is a notebook style open source interface for H2O. It is an interactive web-based environment that allows persons to combine text, plot, mathematics, executable code in a single document, very similar to iPython notebooks.

3.1 Working with H2O:

3.1.1 Start of H2O cluster: We can work with data in two ways on H2O. Either we can use commands in R Studio or we can use H2O flow interface. After starting H2O cluster the dataset named df is loaded on H2O by the name h2odf.

The input data loaded on to H2O is then split into training, testing and validation dataset. 60% of the data forms the training data, 20% forms the validation data and rest 20% forms the testing data. For the training of the data A deep neural network model is used. During the training deep learning model uses the following parameters:

1. Hidden - It specifies the number of hidden layers and number of neurons in each layer in the deep learning architecture.

2. Epochs - It represents the number of iterations to be done on the data set.

3. Activation - It represents the type of activation function to use. The major activation functions in H2O are Tanh, Rectifier, and Maxout.

4. Variable importance: It gives the importance of variables listed from greatest importance, to least importance.

A pictorial representation of fraud detection model is depicted in following figure.

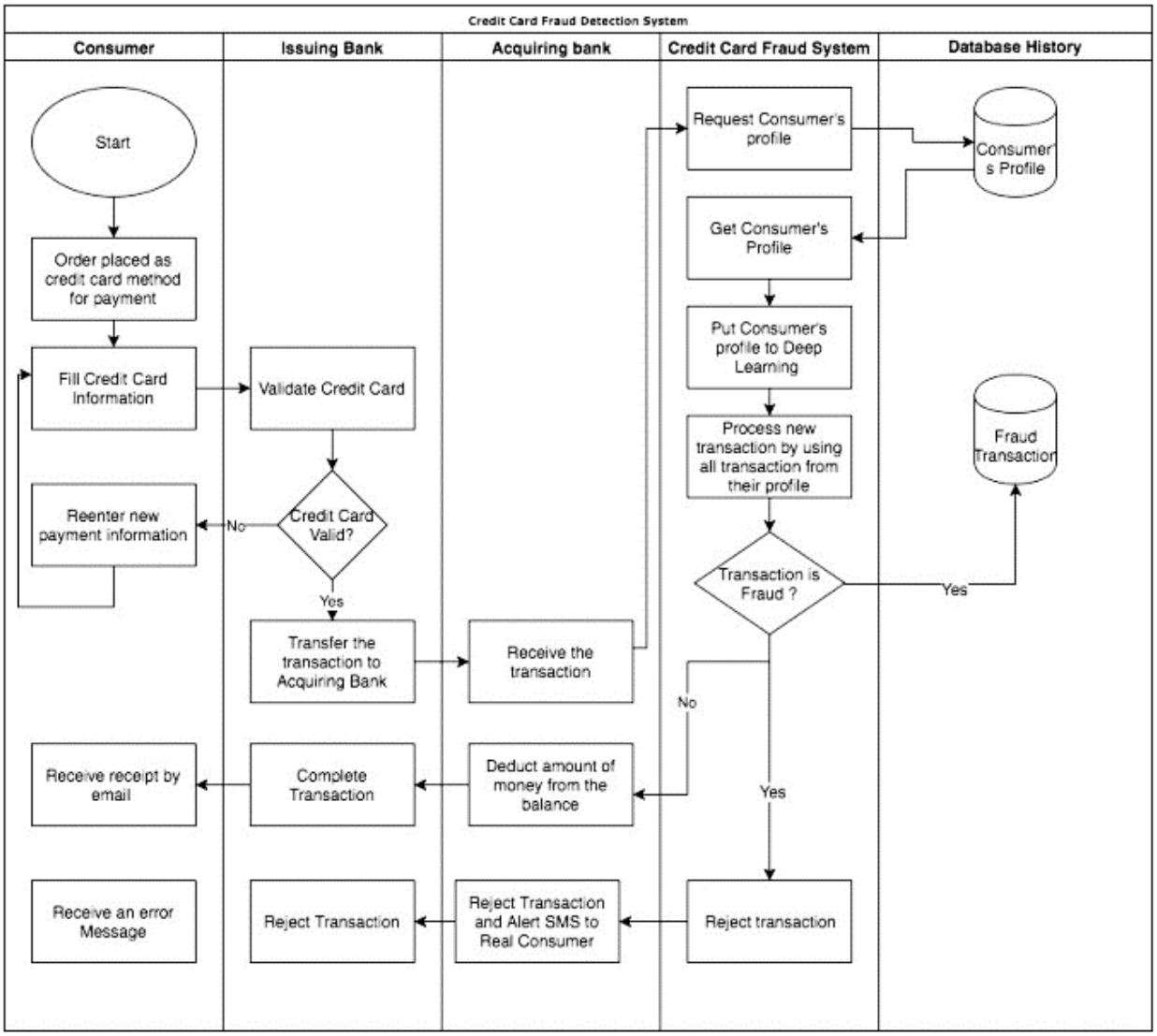

Figure 1: Fraud detection model [4]

4. Experimental Design

For the experimental purpose, various datasets are available in web. The weblink of some of such data sets as given below:

Dataset:

- *https://www.kaggle.com/datasets?sortBy=relevance&group=featured&search=credit+card+fraud+detection*
- *https://data.world/raghu543/credit-card-fraud-data*
- *https://data.world/vlad/credit-card-fraud-detection*

Evaluation Measures:

The performance of the model is evaluated by running the specific command: **h2o.varimp_plot (model_dl_1)**

We obtain a table giving the importance of the different attributes to classify a transaction as genuine or legitimate.

MSE: Mean Squared Error
RMSE: Root Mean Squared Error
MAE: Mean Absolute Errors
RMSLE: Root Mean Squared Log Error

5. Software & Hardware Requirements:

Python based Computer Vision and Deep Learning libraries will be exploited for the development and experimentation of the project. Tools such as Anaconda Python, and libraries such as OpenCV, Tensorflow, and Keras will be utilized for this process. Training will be conducted on NVIDIA GPUs for training the end-to-end version of CNN based object detection model.

6. References

[1] Smith, Michael. The Federal Cyber Role: How Federal Cybersecurity Policy has Affected the Public and Private Sector. Diss. Utica College, 2017.

[2] Seyedhossein, Leila, and Mahmoud Reza Hashemi. "Mining information from credit card time series for timelier fraud detection." Telecommunications (IST), 2010 5th International Symposium on. IEEE, 2010.

[3] https://www.h2o.ai/h2o-old/h2o-flow/

[4] Pumsirirat, Apapan, and Liu Yan. "Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 9.1 (2018): 18-25.