**CREDIT CARD FRAUD DETECTION**

Project Id: 18-047

Project Proposal Report

Mr.Nuwan Kuruwitaarachchi.

T.M.G.A.B. Thennakoon

H.G.S. Premadasa

C.B.P. Lochana

M.D.S. Mihiranga

Bachelor of Science Special (Hons) in Information Technology Specialized in Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

April 2018

**CREDIT CARD FRAUD DETECTION**

Project Id: 18-047

Project Proposal Report

(Proposal documentation submitted in partial fulfilment of the requirement for the Degree of Bachelor of Science Special (honors) In Information Technology)

Bachelor of Science Special (Hons) in Information Technology Specialized in Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information and Technology

Sri Lanka

April 2018

# **Declaration**

“We declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.”

|  |  |  |  |
| --- | --- | --- | --- |
|  | STUDENT NAME | STUDENT NO. | SIGNATURE |
| 1 | T.M.G.A.B. Thennakoon  (GROUP LEADER) | IT15046512 |  |
| 2 | H.G.S. Premadasa | IT15142610 |  |
| 3 | C.B.P. Lochana | IT15111784 |  |
| 4 | M.D.S. Mihiranga | IT15004550 |  |

Supervisor: Mr. Nuwan Kuruwitaarachchi

Co-Supervisor: Mr. Lakmal Rupasinghe

The above candidates are carrying out research for the undergraduate dissertation under my supervision.

Signature of the supervisor

# **Abstract**

In this project we are considering the problem of credit card fraud in online transaction processes. With the increase of using of internet, online payments have drawn more attention by the fraudsters and they have obtained the benefit to perform undesirable activities on people's money to gain illegal profit. Most of the solutions for this issue has been done using machine learning. Machine learning techniques can be broadly classified into categories; supervised learning and unsupervised learning. Techniques such as Decision tree, ANN, Logistic regression and SVM fall under supervised learning and Clustering, Peer-group analysis, Breakpoint analysis and Association rule analysis fall under Unsupervised learning and those techniques have been used to build models for fraud detection. Although those fraud detection models are capable of detecting frauds, there is still a question how accurate those models are. In this project, we propose to select four supervised techniques that have been mostly used by the researchers, which are applicable for this particular problem. To figure out the best two techniques an analysis will be done and using the selected techniques a model will be implemented. Then by applying a set of preprocessed data the built model is trained. Then the result obtained by the model will be visualized using suitable data visualization technique. By providing the end user a web interface the data set can be uploaded and the result is visualized. Application of preprocessed data to the model and visualizing of the result is done by providing two API endpoints to those two processes.

# **Table of Contents**

[**Declaration** i](#_gjdgxs)

[**Abstract** ii](#_30j0zll)

[**Table of Contents** iii](#_1fob9te)

[**List of Figures** v](#_3znysh7)

[**List of Abbreviations** v](#_2et92p0)

[**List of Tables** vi](#_3dy6vkm)

[**1. Introduction** 1](#_1t3h5sf)

[**1.1 Background & Literature Survey** 1](#_4d34og8)

[**1.1.1 Background** 1](#_2s8eyo1)

[**1.1.2 Literature Survey** 2](#_17dp8vu)

[**1.2 Research Gap** 7](#_3rdcrjn)

[**1.3 Research Problem** 11](#_26in1rg)

[**2. Objectives** 14](#_lnxbz9)

[**2.1 Main Objectives** 14](#_35nkun2)

[**2.2 Specific Objectives** 14](#_1ksv4uv)

[**3. Methodology** 15](#_44sinio)

[**3.1 Tasks** 15](#_2jxsxqh)

[**3.1.1 Data Preprocessing** 15](#_z337ya)

[**3.1.2 Selecting the Optimal Algorithms** 16](#_3j2qqm3)

[**3.1.3 Creating the Fraud Detection Model and Fraud Detecting** 16](#_1y810tw)

[**3.1.4 API Integration** 17](#_4i7ojhp)

[**3.1.5 Data Visualization** 18](#_2xcytpi)

[**3.1.6 Implementing Web Application** 19](#_1ci93xb)

[**3.2**](#_3whwml4) **Technologies to be used** 19

[**3.2.1 Data Preprocessing** 19](#_2bn6wsx)

[**3.2.2 Supervised Learning Algorithm** 20](#_2p2csry)

[**3.2.3 Fraud Detection Model** 20](#_147n2zr)

[**3.2.4 API Integration** 20](#_3o7alnk)

[**3.2.5 Data Visualization** 21](#_23ckvvd)

[**3.2.6 Web Application** 21](#_ihv636)

[**3.2**](#_32hioqz) **System Diagram** 22

[**3.3**](#_1hmsyys) **Gantt chart** 23

[**4. Description of Personal and Facilities** 24](#_41mghml)

[**5.**](#_2grqrue) **References** 26

# **List of Figures**

|  |  |  |
| --- | --- | --- |
|  |  | Page |
| Figure 1.3 | Growth in fraud | 9 |
| Figure 2.1 | Techniques comparison | 10 |
| Figure 2.2 | Algorithm comparison | 11 |
| Figure 3.2.4 | Steps in API integration | 17 |
| Figure 3.3 | System diagram | 21 |
| Figure 3.4 | Gantt chart | 22 |

# **List of Abbreviations**

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| API | Application Programming Interface |
| t-SNE  SVM | t-distributed Stochastic Neighbor Embedding  Support Vector Machine |
| ANN | Artificial Neural Network |
| FDM | Fraud Detection Model |
| PCA | Principal component analysis |
| NB | Naïve Bayes |
| DL | Deep Learning |
| MCC | Matthews Correlation Coefficient |
| LOR | Logistic Regression |

# **List of Tables**

|  |  |  |
| --- | --- | --- |
|  |  | Page |
| 4.1 | Description of personal and facilities | 23 |

# **1. Introduction**

## **1.1 Background & Literature Survey**

### **1.1.1 Background**

Today, internet has become an important part of human’s life, one can engage in most of the day today purchasing and banking tasks online. Therefore a significant amount of people make their purchases online and it is growing day by day, despite the fact that credit card frauds are increasingly becoming a common type of theft.[1] Credit card fraud can be broadly categorized into behavioral fraud and application fraud. Behavioral frauds occur when fraudsters use the cardholders’ card information to purchase any item online. Application frauds happen when a fraudster use someone’s information in the application form to issue the credit card. When such frauds occur on credit cards they can make huge financial losses to both bank and card holder. So it is obvious that a solution has to be made to classify a transaction into fraud or a genuine transaction.[2]

Fraud detection approaches that have been using years ago, such as rule-based techniques are not enough to detect frauds because frauds by itself are changing and evolving. Machine learning is this generation's solution which replaces the rule based system and it can work on large data sets which is not easily possible for human beings. The dataset which the algorithm is to be trained on, have to be preprocessed before the model is applied to it. Basically the raw data we have is converted into a cleaner data set in order to prevent producing misleading results at the end of the fraud detection process.

With advanced machine learning, systems can learn and adapt to emerging patterns for preventing fraud. So that a plenty of machine learning techniques have been already developed. Under those learning methods widely used models were found through research.

### **1.1.2 Literature Survey**

According to the research papers we have referred, following are some machine learning techniques and models which have been already done by researchers.

**Review On Fraud Detection Methods in Credit Card Transactions[3]**

In this paper, they have bring together various methods to detect fraudulent transactions and comparison of these methods. One of these or combination of these methods can be used to detect fraudulent transactions. New features can be added and various sampling methods can be used to train the model more accurately.

**Selection of Optimal Credit Card Fraud Detection Models Using a Coefficient Sum Approach[4]**

This study discusses the problem of optimal selection of Fraud detection models. When whole set of selection criteria and FDMs are defined, an effective process coefficient sum method can be applied. This approach is also used as a decision maker for various analyses according to his or her preferences. For all selection criteria not any one Fraud detection model is optimal. So they proposed a method which is suitable for ranking the fraud detection models according to various comparison criteria taken all collectively. A comparatively simple mathematical formulation and basic matrix operation is used in coefficient sum method. It is also used for solving complex multi-attributes decision problems; including both quantitative and qualitative factors.

**Credit Card Fraud Detection Using AdaBoost and Majority Voting[5]**

This is a study on credit card fraud detection using machine learning algorithms has been presented in this paper. A number of standard models which include NB, SVM, and DL have been used in the empirical evaluation. A publicly available credit card data set has been used for evaluation using individual (standard) models and hybrid models using AdaBoost and majority voting combination methods. The MCC metric has been adopted as a performance measure, as it takes into account the true and false positive and negative predicted out- comes. The best MCC score is 0.823, achieved using majority voting. A real credit card data set from a financial institution has also been used for evaluation. The same individual and hybrid models have been employed. A perfect MCC score of 1 has been achieved using AdaBoost and majority voting methods. To further evaluate the hybrid models, noise from 10% to 30% has been added into the data samples. The majority voting method has yielded the best MCC score of 0.942 for 30% noise added to the data set. This shows that the majority voting method offers robust performance in the presence of noise.

**Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism[6]**

In this paper, they propose a novel fraud detection method. They utilize the behavioral patterns from the similar cardholders to build a recent behavioral profile of a cardholder. In this way, they propose a method to solve the adaptive capacity of the model. A feedback mechanism can make full use of the True label information from transactions in order to solve the concept drift problem. The classifier will adjust its own rating score according to a series of incoming transactions. This on-line fraud detection method can dynamically change its parameters to adapt to a cardholder’s transactions behaviors timely. Experimental results show the performance and effectiveness of our method. Compared with other two methods, all of them can achieve 80% accuracy at the detection of transactions.

**Adversarial Learning in Credit Card Fraud Detection[2]**

In this papers they conclude that modeling adversaries’ possible strategies in order to preemptively retrain our model proved to outperform a static model in ability to detect fraudulent transactions. As rounds progressed, the separation between AUC scores of the adversarial learning model and the static fraud model increased. Although the differences in AUC may seem small, the slightest change in AUC could potentially result in a substantial reduction of costs due to fraud. By understanding the weaknesses of our own model, we were able to preemptively adjust our classifier to provide better defense mechanisms for detection against fraud. The use of a GMM in determining a best strategy proved an effective way of finding optimal new transactions an adversary is likely to replicate. The use of SMOTE provided a useful tool in our ability to produce synthetic transactions of this best strategy. Overall, these two contributions provided tools able to mimic an adversary’s learning and thought processes, giving the credit card company the ability to preemptively react to the changing transaction strategies. In future research, there are many possible additions to our framework that would provide more information and realism in our models and could possibly improve our results. In order to differentiate the various possible fraud strategies, our GMM could be optimized to produce more regions of possible transaction types. In our SMOTE algorithm, we choose to introduce enough fraud for the next round to have 15% fraudulent transactions, though this number could also be optimized to the percentage of fraud that yields the most effective classifier

**Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis[7]**

This paper investigates the comparative performance of Naïve Bayes, K-nearest neighbour and Logistic regression models in binary classification of imbalanced credit card fraud data. The rationale for investigating these three techniques is due to less comparison they have attracted in past literature. However, a subsequent study to compare other single and ensemble techniques using our approach is underway. The contribution of the paper is summarized in the following:

1. three classifiers based on different machine learning techniques (Naïve Bayes, K-nearest neighbors and Logistic Regression) are trained on real life of credit card transactions

Data and their performances on credit card fraud detection evaluated and compared based on several relevant metrics.

2. The highly imbalanced dataset is sampled in a hybrid approach where the positive class is oversampled and the negative class under-sampled, achieving two sets of data distributions.

3. The performances of the three classifiers are examined on the two sets of data distributions using accuracy, sensitivity, specificity, precision, balanced classification rate and Matthews Correlation coefficient metrics.

Performance of classifiers varies across different evaluation metrics. Results from the experiment shows that the KNN shows significant performance for all metrics evaluated except for accuracy in the 10:90 data distribution. This study shows the effect of hybrid sampling on the performance of binary classification of imbalanced data. Expected future areas of research could be in examining meta-classifiers and meta- learning approaches in handling highly imbalanced credit card fraud data. Also effects of other sampling approaches can be investigated.

**Cost Sensitive Modeling of Credit Card Fraud Using Neural Network Strategy[8]**

In this study, they have developed a Cost Sensitive credit card fraud detection system which is called CSNN. The performance of this approach was compared to AFDM model on a real credit card dataset. The main contributions of this paper are detection rate maximization and cost minimization by minimizing FN.

**Signal Processing on Graphs for Improving Automatic Credit Card Fraud Detection[9]**

In this paper several surrogate methods based on signal processing on graphs has been proposed for improving automatic credit card fraud detection. The proposed methods were applied to different scenarios for training of the detectors, considering several ratios of fraud operation number to legitimate operation number, and surrogate proportions. The capabilities of the proposed methods to improve detection performance were demonstrated using real data and measured by ROC curves and KPIs commonly used in financial business.

**Cost Sensitive Credit Card Fraud Detection using Bayes Minimum Risk.[10]**

This paper has shown the importance of using the real financial costs of credit card fraud when selecting credit card fraud detection algorithms. Also, it is not enough to have a fixed difference between F P and F N but it is important to have the real F N cost of each transaction. Moreover, our evaluations confirmed that including the real cost by creating a cost sensitive system using a Bayes minimum risk classifier, gives rise to much better fraud detection results in the sense of higher savings

**Credit card fraud detection using ANN and BNN[11]**

Here, the goal is to detect the fraudulent behavior in a credit card transaction system by providing a trained learner with data. The implemented program should be able to classify new transactions. Problems with fraud detection – Overlapping data, Should be a way to handle noise, System should adapt to new frauds.

## **1.2 Research Gap**

In earlier researches they came up with following issues regarding credit card fraud detection mechanism.

1. Lack of real life data because of sensitivity of data and privacy issue.[3]
2. Problem in dealing with imbalance data or skewed distribution because number of fraudulent transactions are very less compare to legitimate transactions.[3]
3. Overlapping of data is one more problem as some of transactions look like fraudulent transaction.[3]
4. Data is not easily accessible due to security issues. If the data is available they are encrypted and unsupervised also does not contain full information to use these data cleaning are required.[12]
5. Data mining techniques take time to execute big data and the most important issue is the implantation of the technique differs for different platform and script taken.[12]
6. Selection of good tool is also required for extracting good features, for accurate results, cost effective and timely results. Weka, java, Matlab, oracle miner are good tools.[12]
7. Challenges regarding data set before using techniques.[12]
   1. Skewness of data.
   2. Search space dimensionally.
   3. Different cost of false positive and false negative.
   4. Durability of the model.
   5. Short time-to-answer.
8. Key experimental issues that are relevant to computational intelligence-based financial fraud detection.[13] Three of them are
   1. Choice of detection algorithm.
   2. Performance metrics.
   3. Feature selection
9. Difficult to prove the robustness or even the probability of success ratio of the methods.[14]

To overcome lack of real life data issue in [3] Many researchers have done research with real life data [15], [10], [16], [17] of bank with agreements. To deal with this problem, many tools are also available to generate synthetic data. We use publicly available data which have mentioned the amount of frauds and non-frauds also we hope to use the same model with a dataset which is from a government bank. The publicly available datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2 ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. [18]

To overcome problem in dealing with imbalance data, synthetic minoring oversampling methods are used to increase number of low incidence data in dataset that generate synthetic fraudulent transactions related with original data set by [3].

To overcome problem of Overlapping of data there is no suitable solution proposed in any of the model. They just check the accuracy to determine whether the prediction is fraud or not.

There are several ways to approach this classification problem taking into consideration this unbalance[19].

* Collect more data.
* Changing the performance metric:
  + Use the confusion matrix to calculate Precision, Recall.
  + F1score (weighted average of precision recall)
  + Use Kappa - which is a classification accuracy normalized by the imbalance of the classes in the data
  + ROC curves - calculates sensitivity/specificity ratio.
* Resampling the dataset
  + Essentially this is a method that will process the data to have an approximate 50-50 ratio.
  + One way to achieve this is by OVER-sampling, which is adding copies of the under-represented class (better when you have little data)
  + Another is UNDER-sampling, which deletes instances from the over-represented class (better when he have lots of data)

As the public data we have a data set which contain limited 30 features (28 anonymous + time + amount).First we proposed a method to compare what happen when using resampling and without resampling. This approach will test using a selected classifier then implement the model using some of the performance metrics mentioned above. Repeat the best resampling/not resampling method, by tuning the parameters. Finally perform classifications model using other classification algorithms. We will use traditional UNDER-sampling. The way we will under sample the dataset will be by creating a 50/50 ratio. This will be done by randomly selecting "x" amount of sample from the majority class, being "x" the total number of records with the minority class.

We come across with following for Accuracy, Precision and Recall work for a confusion matrix.

* Accuracy = (TP+TN)/total
* Precision = TP/(TP+FP)
* Recall = TP/(TP+FN)

TP - True Positive, TN - True Negative, FP - False Positive and FN- False Negative

As we know, due to the misbalancing of the data, many observations could be predicted as False Negatives, being, that we predict a normal transaction, but it is in fact a fraudulent one. Recall captures this. Obviously, trying to increase recall, tends to come with a decrease of precision. However, in our case, if we predict that a transaction is fraudulent and turns out not to be, is not a massive problem compared to the opposite.

For the selection of a suitable model we propose a comparison of most commonly used algorithms. The comparison will be done by a deep review of a few selected techniques, taking the same comparable characteristics as shown below.

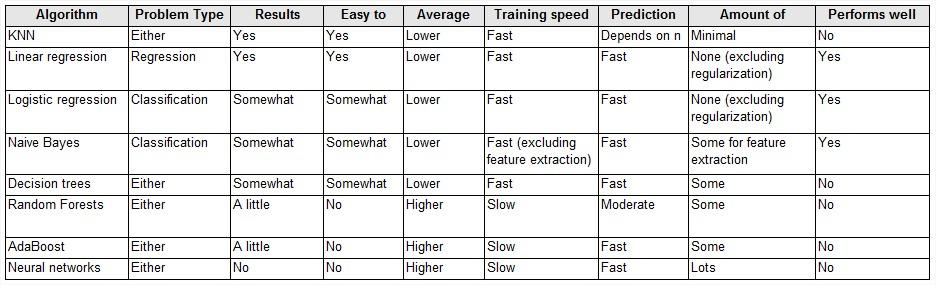


Figure 2.1

Source: [20]

Also we hope to check individual accuracy of each algorithm by the 10-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithms is evaluated in precisely the same way[21].

Output will box and whisker plot showing the spread of the accuracy scores across each cross validation fold for each algorithm.

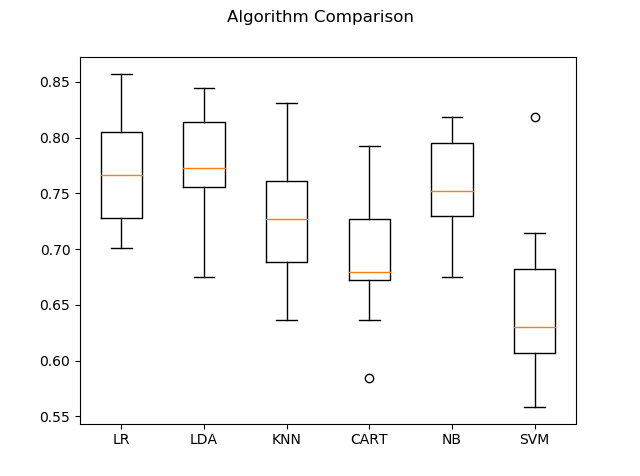


Figure 2.2

Source : [21]

After selection of the optimal techniques through review the model will be built and trained.

In our project we propose to provide a web interface for the end user to upload a suitable data set to the model. The trained model will then visualize the result of the data set to the end user. Uploaded file will be validated to get more accurate output and also the user will authenticate in each process to avoid the security issue. Also security features like data encryption which does not mentioned in early reaches will be included to minimize the security issues and to build the trust between the financial sectors with the system.

## **1.3 Research Problem**

Fraud is a wrongful or criminal deception aimed to bring ﬁnancial or personal gain. Due to the rapid growth in e-business and electronic payment systems, Fraud is rising in banking transactions associated with credit cards. [8]

So with the evolving of new technologies and payment methods the number of fraud cases have been constantly increased and new forms of fraud have been introduced. Credit card fraud occurs when a fraudster makes undesirable activities on consumers’ card information to make illegal profit. Credit card fraud affects many stakeholders including consumers as well as merchants and financial institutions. According to the Nilson Report in October 2016, more than $31 trillion were generated worldwide by online payment systems in 2015, increasing 7.3% than 2014. Worldwide losses from credit card fraud rose to $21 billion in 2015, and will possibly reach $31 billion by 2020.[22]

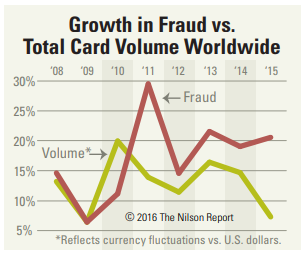


Figure 1.3

Source: [20]

The rise in credit card fraud has a big impact on the ﬁnancial industry. Loss from credit card fraud affects the merchants, where they bear all costs, including card issuer fees, charges, and administrative charges [23]. Since the retailers need to bear all the loss, the goods are priced higher and other offers and discounts are reduced by them. In addition to losing money, credit card fraud can ruin a customer’s relationship with their merchant as well. Therefore it is a must to minimize those losses and issues. Although numerous authorization techniques have been in place, credit card fraud cases have not hindered effectively. So it is imperative to build up a solution by introducing an effective mechanism to reduce fraud cases.

In avoiding losses and issues due to such fraudulent transactions, the mechanism of fraud detection can be done. According to different researches there have been a number of studies on credit card fraud detection.

# **2. Objectives**

## **2.1 Main Objectives**

1. Detecting credit card fraud.
2. Providing a web interface to upload the data set.
3. Implementing the API to provide the interaction between web interface and the model.
4. Improve the performance and accuracy of model by minimizing FP frauds.

## **2.2 Specific Objectives**

1. Analyzing four algorithms and selecting two optimal algorithms.
2. Build a suitable fraud detection model.
3. Data Preprocessing.
4. Visualizing the detected result.

# **3. Methodology**

## **3.1 Tasks**

Initially the raw data is taken which is converted into compatibility format. The next step passes the converted data set into the selected fraud detection model. The fraud detection model gives the predicted output. Then the result is visualized using a data visualization technique.

### **3.1.1 Data Preprocessing**

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Therefore preprocessed data set is used with the fraud detection model. Data preprocessing will take place in four stages.[24]

* Data clearing - Data cleaning routines works to clean the data by filling missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies.
* Data integration - In data integration part, all data will combine from multiple sources into a coherent data store.
* Data transformation - In data transformation, the data are transformed or consolidate into forms appropriate for mining.
* Data reduction - Data reduction technique is for reduce representation of the dataset without compromising the integrity of the original data.

For the data preprocessing process PCA is used as it emphasizes the variation and brings out the strong patterns in a dataset.

### **3.1.2 Selecting the Optimal Algorithms**

Four different algorithms under supervised learning will be considered. Each algorithm will be reviewed separately and common set of characteristics will be extracted from individual review and optimal algorithms will be selected.

### **3.1.3 Creating the Fraud Detection Model and Fraud Detecting**

Fraud detection model will be created using above selected algorithms. Combination of one or more algorithms will be applied. The basic steps that lead to machine learning will be applied till detecting the frauds as follows.

* **Training**

This is the process where the data is used to incrementally improve the model’s ability to predict. The training process involves initializing some random values for say A and B of our model, predict the output with those values, then compare it with the model's prediction and then adjust the values so that they match the predictions that were made previously. This process then repeats and each cycle of updating is called one training step. [25]

* **Evaluation**

Once training is complete, we have to check if it is good enough to move into this step. This is where that dataset that set aside earlier comes into play. Evaluation allows the testing of the model against data that has never been seen and used for training and is meant to be representative of how the model might perform when in the real world. [25]

* **Parameter Tuning**

Once the evaluation is over, any further improvement in the training can be possible by tuning the parameters. There were a few parameters that were implicitly assumed when the training was done. Another parameter included is the learning rate that defines how far the line is shifted during each step, based on the information from the previous training step. These values all play a role in the accuracy of the training model, and how long the training will take.

For models that are more complex, initial conditions play a significant role in the determination of the outcome of training. Differences can be seen depending on whether a model starts off training with values initialized to zeros versus some distribution of values, which then leads to the question of which distribution is to be used. Since there are many considerations at this phase of training, it’s important to define what makes a model good. These parameters are referred to as hyper parameters. The adjustment or tuning of these parameters depends on the dataset, model, and the training process. [25]

* **Prediction**

This is the point where the value of machine learning is realized. Here the model is used to predict the outcome.

### **3.1.4 API Integration**

API enables interaction between data, applications, and devices. It delivers data and facilitates connectivity between devices and programs. By creating an API for this module, there will be two endpoints. First endpoint will be the user endpoint. User can upload their dataset to the API, then detection process will start. Finally API will give the final result to the end user. End user can visualize the data in web framework by getting the output from the API endpoint.

API integration will process according to following steps.[26]

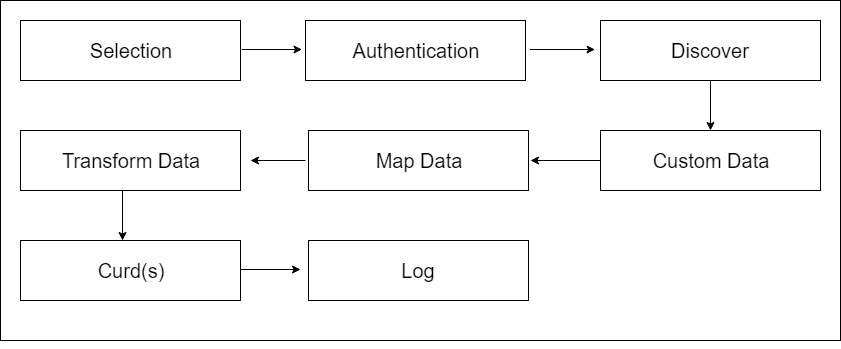
* Selection - Select the end users and the endpoints
* Authenticate - Authenticate each instance of an endpoint that users are connecting with the app. Use OAuth mechanism for authenticate.
* Discover - Discover the objects and data fields at the endpoint to provide the data structure, field names and formats that need to be mapped into the application.
* Custom Data - Endpoints provide automated discovery of custom objects.
* Map Data - Map endpoint’s standard data structure into application’s data structure.
* Transform Data - Transformations for each instance of an endpoint that the app or device connecting with.
* CURD(S) - Determine which methods the application needs to execute against each data object (Create, Retrieve, Update, Delete, Search).
* Log - Capture usage and log data to support the integration to keep operation team in the know.

Figure 3.2.4

### **3.1.5 Data Visualization**

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization. Predicted results got through the API endpoint will be displayed using a data visualization technique.

### **3.1.6 Implementing Web Application**

With the help of an API, a web interface is implemented to make the interaction between user and the fraud detection model. This will allow user to upload a raw dataset in form of a common file formats like .xls and .csv. Then the resultant output will be visualized with the help of the data visualizing technology.

## **Technologies to be used**

**3.2.1 Data Preprocessing**

There are so many data preprocessing technologies.

* **Data Preprocessing in R**

R a framework that consists of various packages that can be used for Data Preprocessing like dplyr etc.

* **Data Preprocessing in Weka**

Weka is a software that contains a collection of Machine Learning algorithms for Data Mining process. It consists of Data Preprocessing tools that are used before applying Machine Learning algorithms.

* **Data Preprocessing in RapidMiner**

RapidMiner is an open-source Predictive Analytics Platform for Data Mining process. It provides the efficient tools for performing exact Data Preprocessing process.

* **Data Preprocessing in Python**

Python is a programming language that provides various libraries that are used for Data Preprocessing. [27]

By considering above mentioned points and the product, data preprocessing in python will be the most suitable for the implementations. Use software, won’t be helpful for whole implementation. We need to achieve the task of preprocessing in programming language. Therefore the python implementation is the most effective one.

### **3.2.2 Supervised Learning Algorithm**

We can use python, JavaScript, java, R etc. to implement algorithms. But python is the most effective language for implementations. Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that can be used for both research and development and developing production systems. There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. [27]

### **3.2.3 Fraud Detection Model**

Machine learning scientists working on sentiment analysis prioritize Python (44%) and R (11%) more and JavaScript (2%) and Java (15%) less than developers working on other areas. In contrast, Java is prioritized more by those working on network security / cyber-attacks and fraud detection, the two areas where Python is the least prioritized. Network security and fraud detection algorithms are built or consumed mostly in large organizations — and especially in financial institutions — where Java is a favourite of most internal development teams. In areas that are less enterprise-focused, such as natural language processing (NLP) and sentiment analysis, developers opt for Python which offers an easier and faster way to build highly performing. [28]

### **3.2.4 API Integration**

We have decided to use python for the algorithm and model implementation. Therefore most suitable language is python to implement an API. Eve is an open source Python REST API framework designed for human beings. It allows to effortlessly build and deploy highly customizable, fully featured RESTful Web Services. [29]

### **3.2.5 Data Visualization**

Use technique called "t-SNE" that visualizes high-dimensional data by giving each data point a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. t-SNE is better than existing techniques at creating a single map that reveals structure at many different scales. This is particularly important for high-dimensional data that lie on several different, but related, low-dimensional manifolds, such as images of objects from multiple classes seen from multiple viewpoints. For visualizing the structure of very large data sets, we show how t-SNE can use random walks on neighborhood graphs to allow the implicit structure of all of the data to influence the way in which a subset of the data is displayed. We illustrate the performance of t-SNE on a wide variety of data sets and compare it with many other non-parametric visualization techniques, including Sammon mapping, Isomap, and Locally Linear Embedding. The visualizations produced by t-SNE are significantly better than those produced by the other techniques on almost all of the data sets

### **3.2.6 Web Application**

We provide a web application to the end user. By the web application user able to work with the API we provide. Develop the web application using JavaScript framework because we use JavaScript library for the visualizing part. React JS use as the framework for the front end developments. One of the main reasons why we expect a mass use of React JS is because of React JS features like server-side communication. The library of React JS empowers the programmers by giving them lifecycle “hooks” to enable the serve requests. [30]

* 1. **System Diagram**

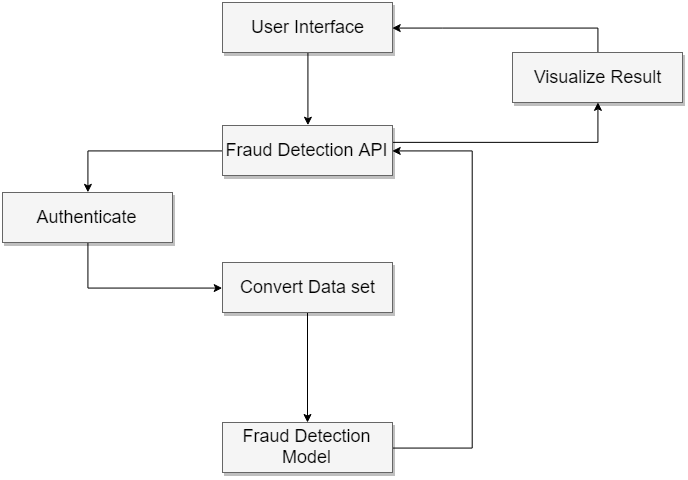


Figure 3.3

## **Gantt chart**https://lh5.googleusercontent.com/sIYF1p2nIe0kJmdSQSou8bf-ZFcdKEtyfv5rOUGL-v-Hdz3CQBlY_kOHrP-4C9m5yrMoFM6Yn4oMrxTMkhnlTpL5o-L9xcaiQQsy3IEUWV8P0-FxVx8RRF4s6ZlDqpQQE-Y0ehMV

Figure 3.4

# **4. Description of Personal and Facilities**

This section describes the workload assigned to each four members. The research problem divided into four parts with equal proportion, so all members can work with equal effort and focus. Each member’s components related with other members, so everyone should focus on entire project.

Table 1: Description of personal and facilities.

|  |  |  |
| --- | --- | --- |
| Member | Task | Description |
| IT15046512 – T.M.G.A.B.Thennakoon | * Review and implementing a supervised algorithm. * Implementing the fraud detection model. | * This section is concerned about finding facts regarding selected supervised learning algorithm and helps in selecting the best technique for the model implementation. * This section is concerned about implementing the fraud detection model from the one of the selected algorithm. |
| IT15142610 - H.G.S. Premadasa | * Review and implementing a supervised algorithm. * Data visualization. | * This section is concerned about finding facts regarding selected supervised learning algorithm and helps in selecting the best technique for the model implementation. * This section is concerned about the visualizing of the output gain from the fraud detection model through the web application interface. API endpoints will directly apply here for a better user experience. |
| IT15111784 - C.B.P. Lochana | * Review and implementing a supervised algorithm. * Data preprocessing. | * This section is concerned about finding facts regarding selected supervised learning algorithm and helps in selecting the best technique for the model implementation. * This section is concerned about preprocessing data using selected method. |
| IT15004550 - M.D.S. Mihiranga | * Review and implementing a supervised algorithm. * API and web interface implementation. | * This section is concerned about finding facts regarding selected supervised learning algorithm and helps in selecting the best technique for the model implementation. * This section is concerned about implementing API. Endpoints will be decided and developed. With the help of the API, interactions between user and the fraud detection model is handled. |

1. **References**

[1] J. Shaji and D. Panchal, “Improved fraud detection in e-commerce transactions,” *2017 2nd Int. Conf. Commun. Syst. Comput. IT Appl. CSCITA 2017 - Proc.*, pp. 121–126, 2017.

[2] M. F. Zeager, A. Sridhar, N. Fogal, S. Adams, D. E. Brown, and P. A. Beling, “Adversarial learning in credit card fraud detection,” *2017 Syst. Inf. Eng. Des. Symp.*, pp. 112–116, 2017.

[3] M. Rafało, “Real-time fraud detection in credit card transactions,” *Data Science Warsaw*. 2017.

[4] S. Arora, “Selection of Optimal Credit Card Fraud Detection Models Using a Coefficient Sum Approach,” pp. 482–487, 2017.

[5] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, “Credit card fraud detection using AdaBoost and majority voting,” *IEEE Access*, pp. 1–1, 2018.

[6] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, “Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism,” *IEEE Internet Things J.*, vol. 4662, no. c, pp. 1–1, 2018.

[7] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” *2017 Int. Conf. Comput. Netw. Informatics*, pp. 1–9, 2017.

[8] F. Ghobadi and M. Rohani, “Cost sensitive modeling of credit card fraud using neural network strategy,” *Proc. - 2016 2nd Int. Conf. Signal Process. Intell. Syst. ICSPIS 2016*, pp. 8–10, 2017.

[9] L. Vergara, A. Salazar, J. Belda, G. Safont, S. Moral, and S. Iglesias, “Signal processing on graphs for improving automatic credit card fraud detection,” *2017 Int. Carnahan Conf. Secur. Technol.*, pp. 1–6, 2017.

[10] A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten, “Cost sensitive credit card fraud detection using bayes minimum risk,” *Proc. - 2013 12th Int. Conf. Mach. Learn. Appl. ICMLA 2013*, vol. 1, pp. 333–338, 2013.

[11] K. Tuyls, S. Maes, and B. Vanschoenwinkel, “Machine Learning Techniques for Fraud Detection,” no. July, 2001.

[12] B. Sagar, P. Singh, and S. Mallika, “Online transaction fraud detection techniques: A review of data mining approaches,” *Comput. Sustain.*, pp. 3756–3761, 2016.

[13] J. West and M. Bhattacharya, “An Investigation on Experimental Issues in Financial Fraud Mining,” *Procedia Comput. Sci.*, vol. 80, pp. 1734–1744, 2016.

[14] N. Malini and M. Phil, “Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection,” *3rd Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics Anal.*, pp. 3–6, 2017.

[15] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, “Credit card fraud detection using AdaBoost and majority voting,” *IEEE Access*, vol. XX, pp. 1–1, 2018.

[16] A. Charleonnan, “Credit card fraud detection using RUS and MRN algorithms,” *2016 Manag. Innov. Technol. Int. Conf.*, p. MIT-73-MIT-76, 2016.

[17] A. El Bouchti, A. Chakroun, H. Abbar, and C. Okar, “Fraud detection in banking using deep reinforcement learning,” *2017 Seventh Int. Conf. Innov. Comput. Technol.*, no. Intech, pp. 58–63, 2017.

[18] “Credit Card Fraud Detection | Kaggle.” [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud/data. [Accessed: 05-Apr-2018].

[19] “In depth skewed data classif. (93% recall acc now) | Kaggle.” [Online]. Available: https://www.kaggle.com/joparga3/in-depth-skewed-data-classif-93-recall-acc-now. [Accessed: 05-Apr-2018].

[20] “Comparing supervised learning algorithms.” [Online]. Available: http://www.dataschool.io/comparing-supervised-learning-algorithms/. [Accessed: 05-Apr-2018].

[21] “How To Compare Machine Learning Algorithms in Python with scikit-learn - Machine Learning Mastery.” [Online]. Available: https://machinelearningmastery.com/compare-machine-learning-algorithms-python-scikit-learn/. [Accessed: 05-Apr-2018].

[22] David Robertson, “Investments &amp; Acquisitions — September 2016 Top Card Issuers in Asia–Pacific Card Fraud Losses Reach $21.84 Billion,” *Nilson Rep.*, no. 1096, 1090.

[23] Y. Sahin and E. Duman, “Detecting Credit Card Fraud by Decision Trees and Support Vector Machines,” vol. I, pp. 315–319, 2011.

[24] D. P. Methods, “Data Preprocessing Techniques for Data Mining,” *Science (80-. ).*, p. 6, 2011.

[25] “The Seven Steps Of Machine Learning.” [Online]. Available: https://www.techleer.com/articles/379-the-seven-steps-of-machine-learning/. [Accessed: 02-Apr-2018].

[26] “The 10 Step Guide to Designing an API Integration.” [Online]. Available: https://blog.cloud-elements.com/10-step-guide-designing-api-integration. [Accessed: 02-Apr-2018].

[27] “Data Preprocessing &amp; Data Wrangling in Machine Learning | Deep Learning - XenonStack Blog.” [Online]. Available: https://www.xenonstack.com/blog/data-science/data-preprocessing-data-wrangling-in-machine-learning-deep-learning. [Accessed: 02-Apr-2018].

[28] “What is the best programming language for Machine Learning?” [Online]. Available: https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7. [Accessed: 02-Apr-2018].

[29] “Python REST API Framework: Eve, the Simple Way to REST. — Eve 0.8-dev documentation.” [Online]. Available: http://python-eve.org/. [Accessed: 02-Apr-2018].

[30] “Top 10 Useful Features of React JS.” [Online]. Available: http://www.techverx.com/blog/technologies/top-10-useful-features-of-react-js/. [Accessed: 02-Apr-2018].