A Research paper on

“Credit Card Fraud Detection using Machine Learning Algorithms and Neural Network”

Name of student: Sunil Ghimire

UN ID: 1928584

BSc (Hons) Software Engineering

Supervisor: Sachin Kafle

Herald College Kathmandu

University of Wolverhampton

Bishalnagar, Kathmandu, Nepal

11th May 2020

Acknowledgement

I would like to express my deepest gratitude to Mr. Sachin Kafle, Department of Information and technology, Herald College Kathmandu for providing us his regular support, cooperation, and coordination. I would like to acknowledge with appreciation to Herald College Kathmandu for providing me with the platform to enhance my knowledge of the project and its application. Also, I would like to thank the teaching and non-teaching staff of my college whose immense effort and gratitude made the project possible. They were who provided me a friendly environment, valuable instructions, friendly behavior, and guidance and provide prompt comments and rectification necessary before finalization of the report. I would like to express my thankfulness to Info Developer Team who provided me necessary instructions when necessary.

I would like to thank my colleagues who directly or indirectly help us during the project and during making this report. It will be not possible for me, if you guys were not there so, thank you all for making this project successful.

Abstract

Table of Contents

Table of figure

Table

1. Introduction

1.1. General Introduction

When people in the old days decided to get a home, they needed to save the funds for it and build it on their own or employ someone else to do it. Without bank-like financial intermediaries' transactions like this were done in available funds. So, with the changing fields of technology and communication, financial institutions have become a hot industry among entrepreneurs which makes it easier for people with money to contact people who want to borrow money. So, financial institutions are one of the most important factors of the financial system of any country which plays a crucial role in assessing the efficiency and effectiveness of the financial system (Ghazi, 2019). A credit card is a part of a financial product that is issued by banks to make purchases on credit. In other words, a credit card also known as a charge card is defined as a plastic card having a magnetic stripe on the side of the card which contains the details of the critical cardholder and should therefore not be abused. The issuing bank assigns a different credit limit for each card where people can shop in stores, malls, and even online where using a card up to limits issued by the card provider. Also, the credit cards can be used by individuals every month for any amount up to the limit. When people use their card, the issuing company charges to the retailers on their behalf. So this the reason, people buy goods and services without having to pay directly out of their wallets (Salman & Salman, 2015).

Credit card frauds happen in different ways and have been a concern for years all over the world. There are various forms of credit card systems and programs but the illegal usage of a missing or stolen card is one of the simplest processes. Among the missing or stolen card, Account Takeover, Internet Fraud, Non-Receipt Fraud, and counterfeit credit card fraud are the major types involved in credit card fraud which not only affect the victims but also credit card companies and retailers are struck by the impact of this fraud (Barker & Sheridon, 2008). Amount of fraudulent behavior is increasing rapidly, with individuals and organizations at high risk. So, detection of fraud involves identifying a fraud that is either about to occur or after it has happened which can never be fully prevented. According to Experian, "Customers making an application want the best service but also want to be protected against fraudsters and identity theft. You have to balance protection while giving your genuine customers the best decision in the shortest possible time." (Al-Jumeily, et al., 2015)

Fraud is an illegal method of obtaining commodities and money and credit card fraud is a growing issue nowadays regarding payment cards as an illegal source of funds in transactions. The main goal of such illegal activity may be to get goods from an account without paying or obtaining an illegitimate fund. In the fraud detection system of the modern world, the investigator is not able to verify all the transactions. Also verifying all the transactions is a time consuming and costly process. So, identifying this kind of fraud is troublesome and may endanger companies and business organizations (Shirgave, et al., 2019).

1.2. Current Scenario

The existing traditional finance system is quite challenging because the traditional finance sector was focused on taking deposits for a fee from the depositors and lending this money at a price (interest) to interested borrowers. Nowadays, most account holders depend on plastic currency to withdraw money, debit card, instead of standing in lines to withdraw cash. Also, most of the depositors may not own a cheque book to avoid unnecessary restrictions that come with owning cheque books such as added fees and obligatory balance criteria. In a dynamic economy, Nepal Rastra Bank (NRM) plays a key role in directing network growth in a competitive environment, including structures such as payment systems. The first successful e-payment network device, Electronic Cheque Clearing System (ECCS), has been introduced and operational since November 2011 along with the efforts of banks and central bank that enables the clearing of the cheque of the same day irrespective of the location of the Banks and their branches (Giri, 2013).

1.3. Proposed System

Credit cards are currently one of the leading online and offline payment types and the use of credit cards has risen exponentially which means there is a high chance of misleading transactions. The making of detection is not possible, none of the standard programs monitor for credit card fraud. For example, if we want to accept the ‘ABC’ card as a real card or a fraud. There is no obligation to distinguish the actions of the card. And there is also no guarantee that the card is valid or counterfeit. Besides this there is no fixed software or system that can provide genuine, fraud log of transactions. People are more curious to figure out whether thirty people used their card, or those used by other third-party apps. However, the main problem is the avoidance of potential misuse of credit cards.

Thus, Information technology is rapidly increasing with respect to automated systems. In such systems, people utilize computer-based expert systems to analyze and handle real-life problems such as Online Payment systems. The proposed Credit Card Fraud Detection (CCFD) is a solution for fraudulent identification that tracks the time and amount of money in the everyday transactions to determine if the credit card is accepted or not where dataset for this project of credit was obtained from Kaggle which includes input variables that arise from a PCA transformation. Unfortunately, the dataset cannot include the original features and further detailed details regarding the data because of confidentiality concerns. Features V1, V2 ... V28 are the key components obtained with PCA, Time and Amount are only the features which have not been transformed with PCA. The 'Time' features include the seconds between every transaction and the first transaction inside the dataset. The 'Amount' feature is the transaction amount which can be used for the example-dependent cost-sensitive learning. The feature 'Class' is the target variable which takes value 1 as a fraud transaction and value 0 for normal transactions. CCFD uses a machine-learning algorithm like Logistic Regression and Random Forest and similarly deep learning algorithm like Autoencoder which finds a structured pattern of normal credit cards and fraudulent credit cards. At last, tkinter is used which is the part of Python standard library that provides an object-oriented interface onto TK/TCL Also, tkinter helps to develop to create a cross-platform GUI for FYP without more dependencies.

1.4. Project Scope

Machine learning and deep learning can hardly be claimed to be one of the best researches and growth opportunities. It is growing to uncharted height, innovating revolutionary technology dramatically changes computer science trends and studying across the globe. So, with time passing rapidly, any more development would be dampened with any second wasted. For to operate on these large amounts of data which need technologies that can speed the latest techniques and models. ML's and DL's techniques and models will simplify and visualize results, and reliably assess profound observations and trends from the training collection. Machine Learning and Deep Learning is a new-generation technology that allows the newest technologies to be built from various mechanisms. The proposed application considers Logistic Regression and Random Forest as machine learning algorithms where logistic regression is used to minimize the wrong prediction and random forest which is used for classifiers which is used to handle the missing values and maintain the accuracy of a large proportion of data. And deep learning algorithm i.e. Autoencoder is used to denoise the testing dataset in the prediction process and classify the sample dataset. Along with ML and DL model techniques like SMOTE and K-fold cross validation in data-preprocessing.

1.5. Aims and Objectives

1.5.1. Aims

The main aim of this report is to gain the ability to research various machine learning and deep learning algorithms along with it’s wrong mechanisms based on fraud credit cards and gain knowledge about the techniques which make complete algorithms.

1.5.2. Objectives

The objectives of this report are as follows:

- Getting information by proper research
- Understanding algorithms and its working mechanism
- Able to understand different algorithm based on CCFD
- Detect precision, recall, f1-score based on algorithm
- Understand the technique to handle imbalanced dataset
- Able to visualize the graph of dataset
- Create the report based on the project

1.6. Academic Questions

There are certain questions arose during the planning of the proposed algorithms which are listed below:

- What sort of problem this project going to solve?
- How actual is fraud detected?
- What are the challenges involved in developing an algorithm to detect fraud card?
- Are there any similar projects?

2. Report Structure

The diagram for the report structure is shown below:

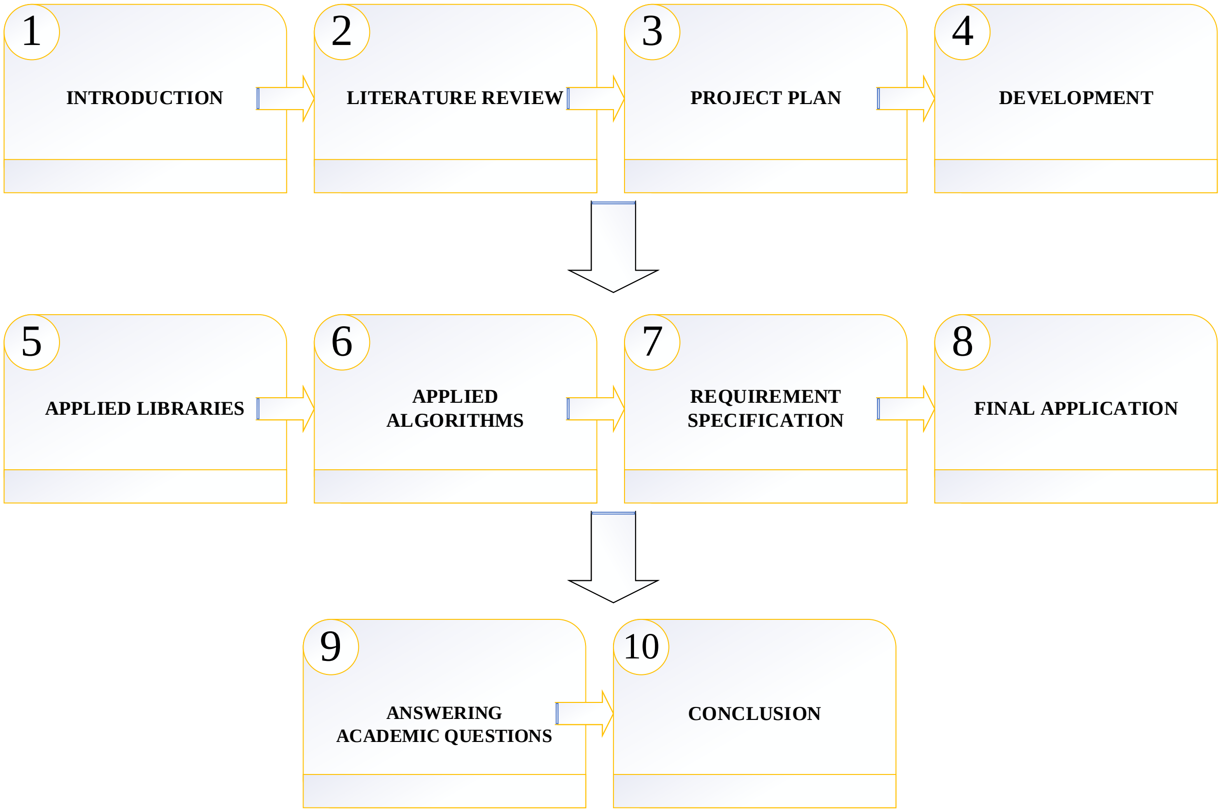


Figure 1: Structure of the report

The brief description of structure of this report is as follow:

- Introduction: This section includes general information about the topic along with aims and objectives.
- Literature review: This section includes all the background research regarding the similar systems.
- Project plan: This section includes the detail plan to complete the project
- Applied libraries: This section includes the libraries used during project development.
- Applied algorithms: This section includes the machine and deep algorithms used during development.

- Requirement specifications: This section includes functional and non-functional requirements of the project.
- Final application: This section includes the final execution of all machine learning and deep learning algorithms with confusion matrix and classification report.
- Answering academic questions: This section includes the answer to the academic questions.
- Conclusion: This section includes the conclusion to the entire project and future escalations.

3. Literature Review

There are several published research papers related to detecting fraud credit card. A paper was published on "A Comparative Analysis of Various Credit Card Fraud Detection Techniques". Credit-card fraud has cost merchants and banks trillions of dollars worldwide. Even after various strategies for preventing fraud, fraudsters are actively finding different forms and techniques to commit fraud. Thus, to stop fraud this paper suggests an effective fraud detection system that not just identifies the fraud card but also identifies it before it occurs and adopts new methods to monitor fraud cards in a specific manner also discussed numerous various possible techniques for fraud detection such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Bayesian Network, K-Nearest Neighbor (KNN), Hidden Markov Model, Fuzzy Logic Based System and Decision Tree and these techniques are used to create an accurate, precise and fast fraud detection system capable of detecting not only internet theft such as phishing and site duplication, but also credit card misuse itself, i.e. utilizing a tempered credit card to trigger a warning. The main drawbacks of this system are that they are non-guaranteed, they give best results for dataset and poor results for other types of dataset. The algorithm ANN and Naive Bayesian gives high detection rates and high accuracy. ANN considers recent incoming transactions as fraud or genuine transactions based on previous records and Naive Bayesian consists of nodes and edges where nodes represent random variables and edges between nodes represent correlations with certain random variables and their probabilistic distribution which is used to calculate the minimum and maximum probability of fraud and valid transactions. For new incoming transactions if the probability of legal transactions is less than fraud transactions it is known as fraud transactions. Some other algorithms like KNN and SVM give excellent results for small datasets. The KNN model calculates the prevalent class for every new transaction and marks the transactions as belonging to the prevalent class. The SVM model generates a hyperplane that studies activity of fraud and legitimate transactions and then classifies new transactions according to which class it belongs. For better results sampled and preprocessed data SVM and decision tree is used whereas logistic regression and fuzzy systems give better accuracy with raw unsampled data. Decision Tree used for classification and prediction contains internal nodes that represent a test on an attribute and each branch denotes the product of an outcome and each leaf node has a class label using a technique called depth first greedy approach or breadth first greedy approach and stops until all the transactions are allocated to a certain class. Logistic regression is used for clustering with the goal of estimating the values of parameter coefficients using the sigmoid function, and it examines the values of its attributes when a transaction is ongoing and tells whether the transaction should proceed. And the last proposed fuzzy logic is used for continuous data when there is absence of discrete truth value in the dataset where Fuzzification, Rule Based and Defuzzification are the three important components. Fuzzification is used for incoming transactions according to the numerical interest correlated with the transaction in the categories of large, small, or medium. Rule-based regulations are expanded based on consumer behavior and last component Defuzzification, it is not permitted to proceed if a transaction does not conform with the predefined collection of laws. This is halted automatically, and then cross-checked with the customer that approval to proceed or be terminated will be given (Jain, et al., 2019).

Algorithms used to detect fraud card along with Accuracy, Detection Rate and False Alarm Rate is shown below:

Techniques	**Accuracy**	**Detection Rate (Precision)**	**False Alarm Rate**
Support Vector Machine (SVM)	94.65%	85.45%	5.2%

Artificial Neural Networks (ANN)	99.71%	99.68%	0.12%
Bayesian Network	97.52%	97.04%	2.50%
K – Nearest Neighbor (KNN)	97.15%	96.84%	2.88%
Fuzzy Logic Based System	95.2%	86.84%	1.15%
Decision Trees	97.93%	98.52%	2.19%
Logistic Regression	94.7%	77.8%	2.9%

Table 1: Comparison of different machine learning techniques (Jain, et al., 2019)

From the above table, ANN and Bayesian Network gives higher accuracy, KNN, SVM and decision tree gives medium level of accuracy and fuzzy logic-based system and logistic regression gives low level of accuracy as compared with others.

Another paper was published on “Credit card fraud and detection techniques: A review”. The key goal of this paper is to recognize the different forms of credit card fraud and to explore the alternative methods used in fraud detection. The sub-aim is to present, evaluate and examine recent results in the identification of credit card fraud. This article defines specific terminology for credit card fraud, outlining the important facts and figures and minimization of credit card fraud faced by banks or credit card companies. Techniques to counter fraud cards are checked and details are given from European markets where fraud happens when a lender is tricked or fooled by a borrower offering him / her transactions, assuming the borrower’s credit card account would compensate for these transactions. Also, the paper describe different types of fraud such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud and proposes a different techniques for minimization of credit card frauds which can be carried out by Decision Trees, Genetic Algorithms, Clustering Techniques and neural networks. Algorithms have presented the best result to detect fraudulent credit cards. The results showed that data mining techniques can be enough to detect fraud credit cards. This paper investigated different statistical techniques used by different countries to detect fraud credit cards. However, the most used technique is neural network which is technically an online fraud detection system based on neural classifiers (Delamaire, et al., 2009).

Study	Country	Method	Details
Aleskerov (1997)	Germany	Neural Network	Card-watch
Bently (2000)	UK	Genetic Programming	Logic rules and scoring process
Brause and his team (1999)	Germany	Data mining techniques and neural network	Data mining application combined probabilistic and neuro-adaptive approach
Bolton and Hand (2002)	UK	Clustering Techniques	Peer group analysis and break point analysis
Ghosh and Reilly (1994)	USA	Neural Network	FOS (Fraud Detection System)
Dorronsoro (1997)	Spain	Neural Network	Neural Classifier
Leonard (1995)	Canada	Expert System	Rule based Expert System for fraud detection (fraud modelling)
Zaslavsky and Strinzkak (2006)	Ukraine	Neural Network	SOM, algorithm for detection of fraudulent operations in payment system
Kokkinaki (1997)	Cyprus	Decision Tree	Similarity tree based on decision tree logic.
Chan and his team (1999)	USA	Algorithms	Suspect behavioral prediction
Ezawa and Norton (1996)	USA	Bayesian networks	Telecommunication industry
Kim and Kim (2002)	Korea	Neural Classifier	Improving detection efficiency and focusing on bias of training sample as in skewed distribution. To reduce "mis-detections".

Table 2: Investigating different statistical techniques in credit card fraud (Delamaire, et al., 2009)

Another research paper was published on “Credit Card Fraud Detection System Based on Machine Learning Techniques”. The paper proposes data mining and machine learning techniques like Naive Bayesian, Support Vector Machine and Decision tree used for the evaluation terms which is the initial phase because they require fewer assumptions and deliver higher analytical accuracy also known as “Standard model” to detect the fraudulent credit card. The machine learning tree-based ensemble model is especially common with bagging and boosting where hierarchical tree model is capable of modeling non-linear relationships which is usually used for regression and classification and is likely to do well with large independent variables. So, random forest is an assembly machine learning approach that uses bagging to use several trees as classifiers and after taking majority voting across all classifiers the random forest integrates knowledge over all trees to expose variable importance. The second phase introduce the proposed algorithm (hybrid methods) based on “AdaBoost and Gradient Boosting Machine ” known as Boosting methods which is another form of ensemble method for improving the accuracy of any given learning algorithm and closely related to random forest that use model efficacy to evaluated only use of the credit card collection which is publicly available that may be concerning credit risk, customer profit, stock prices and automated trading (Kumar, et al., 2019).

The below table shows the accuracy after using different model where proposed algorithm has higher accuracy as compared with other

	KNN	**Random Tree**	**Proposed Algorithm**
Accuracy	0.9691	0.9432	0.9824
Sensitivity	0.8835	0	0.9767
Specificity	0.9711	0	0.9824

Limitations	Cannot detect the fraud at the time of transactions	No suitable for Randomness dataset	Not applied for non-linear data

Table 3: : Comparison of algorithms (Kumar, et al., 2019)

Application offered by Jumio i.e. NetVerify is a web application to detect and prevent credit card fraud by using machine learning, biometric facial recognition, and computer vision where human reviews are needed to see recognized patterns. The web application also detects some effort to modify the ID, such as cropping a portion of the picture or identity. Consumers must have valid ID card verification, identity verification and document verification where software can catch fake IDs. Companies add another feature i.e. biometric facial recognition system with eyeball tracking and catch the fraudsters by most minute facial moment. The company verified mobile transactions including KYC and more than 120 million of identities through the web where algorithms are used by computer vision to enable to recognize of trends (Jesus, 2019).

4. Scope Identification

4.1. Fact Finding Techniques

Fact finding is a method of evidence collection based and knowledge based on strategies that involve sampling of the existing documents, analysis, observation, questionnaires, interviews, prototyping, and joint requirements planning. System analyst uses effective fact-finding techniques to build and execute current programs. The compilation of necessary information is very critical for the implementation in the System Development Life Cycle because the system or software cannot be used efficiently and effectively without proper extraction from facts. Fact-Finding Techniques are applied in the early stage of the System Development life Cycle including the phase of system analysis, design, and post-implementation review. Facts used in any information system should be evaluated based on three steps (Tulasi, 2020):

a. **Data-facts**: Used to create useful information system

b. **Process-function**: To perform the objectives

c. **Interface-design**: To interact with users.

Following are the fact-finding techniques:

4.1.1. Interview

Interview is the most popular and widely used method to collect knowledge from face-to-face interviews. The aim of the interview is to define, check, explain information, inspire the end users concerned, recognize criteria, and gather ideas and information (Mathers, et al., 2000). Interview greatly serves to obtain insight on the operation of the program from experienced professionals and helps to gain practical guidance, interviewee's past mistake mistakes and their challenges in overcoming strategies. Therefore, while doing CCFD projects also have a part to contribute to the practical operating scenario of the program and to get feedback about how to be successful in the method to produce meaningful outcomes via interviews.

Questions to be asked are as follows:

4.1.1.1. For the Data Science Expert

4.1.1.1.1. How do you handle the imbalanced dataset?

4.1.1.1.2. How do you classify the feature and target class of the dataset?

4.1.1.1.3. On What basis you remove the null values from the dataset?

4.1.1.1.4. On what basis is the visualization allocated?

4.1.1.1.5. What is the curse of dimensionality and what are some ways to deal with it?

4.1.1.1.6. The dataset obtained from Kaggle contains only numerical input variables which are the result of PCA transformation. So, why is PCA needed in Machine Learning?

4.1.1.1.7. Can you classify a dataset after dimensionality reduction?

4.1.1.2. For the Machine Leaning Engineer

4.1.1.2.1. What are the types of data mining techniques that can detect the actual card and fraudulent card?

4.1.1.2.2. How does one choose which algorithm is best suitable for the dataset at hand?

4.1.1.2.3. How to apply machine learning in fraud detection?

4.1.1.2.4. What are the factors I must consider before comparing the performance of two-meta algorithms applied to a problem?

4.1.1.2.5. Why do we need a validation set and test set?

4.1.1.2.6. On what basis is k-Fold cross validation allocated?

4.1.1.2.7. What are some factors that explain the success and recent rise of machine learning and deep learning?

4.1.1.3. For Account Manager

4.1.1.3.1. How is the price of credit card determined?

4.1.1.3.2. Is there a provision of reminder notification provided to the user?

4.1.2. Observation

Observation is one of the most effective fact-find methods used to gather data and information in social research (Kawulich, 2012). The field research has been carried out in many specific observing types, three aspects of observing types are listed below:

a. Participant observation

Participant observation was an important research strategy for this project, involving a relatively unstructured and flexible combination of informal interview data and information being gathered (Kawulich, 2012). In this type of observation, particular sorts of activities like knowing about credit card, working mechanism of ATM system, working mechanism of online transaction even knowing what type of database are using by financial institution are observed and recorded by directly participates in the study of population's activities, perhaps accompanying and helping bankers member to gather information. This allows for an understanding of the study population from their own perspective activities. At the same time, it is possible to check difference between what people are saying and what they are doing. So, in participant observations, knowledge of local language is important. A knowledge and correct banking information like ATM, credit cards, type of online transaction helps greatly in understanding theme of people. System takes long time to develop. With studies of financial institution such as description of system, techniques or marketing, this may be less of a problem than for more sensitive topics such as government law, politics, though the usage of CCFD system may be part and parcel of these sensitive issues.

b. Nonparticipant observation

Nonparticipant observation also was an important research strategy for this project to study about different kind of research paper and system to detect and prevent fraud credit card. The research can be organized, at least to the point of choosing to concentrate on a particular task such as finding best machine learning and deep learning algorithms and technique used to balance the imbalanced dataset. During this type of observation, visualization of data is equally important so that one can know about best algorithm and techniques with confusion matrix. In the case of forming theories from research paper to be checked, or covering up unresolved interactions, there can be a deliberate structuring of findings. So, it is important to be careful not to enforce preconceived ideas and should stay agile and accessible to new understanding. And the main thing observe during this observation is that honesty is the best policy for both ethical and practical factors, the people who are observing CCFD should have right to know about the scope and purpose of the system.

c. Time allocation studies

An important aspect of CCFD survey is how much time is spent on tasks like collecting data, understanding data, understanding machine learning and deep learning and other related activities like handling imbalanced data, feature reduction technique like PCA. It may be difficult to get precise information because algorithm involved in detecting fraud and genuine credit is hard to understand and take more time. The biggest problem with our data is that large data size, recalling one our initial warnings on avoid collection of surplus data. Considerable time is spent in learning algorithms and train our data.

4.1.3. Research

Research is most important process of analyzing the problems which had already solved by other sources that can be either documents or human (tutorialspoint.com, 2020). The main purpose of research is to inform action, to prove a theory, and contribute to developing knowledge in a field or study.

Research in detecting fraud credit card helps to be more familiar with the terms like:

Representation: A classifier needs to be in a structured language that the computer is able to handle.

Evaluation: A evaluation function which is also known as objective function or scoring function which is used to classify good classifier from bad ones.

Optimization: The optimization technique is used to search among the classifiers in the language for the highest score one.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support Vector machines	Squared Error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naïve Bayes	Posterior Probability	Continuous optimization

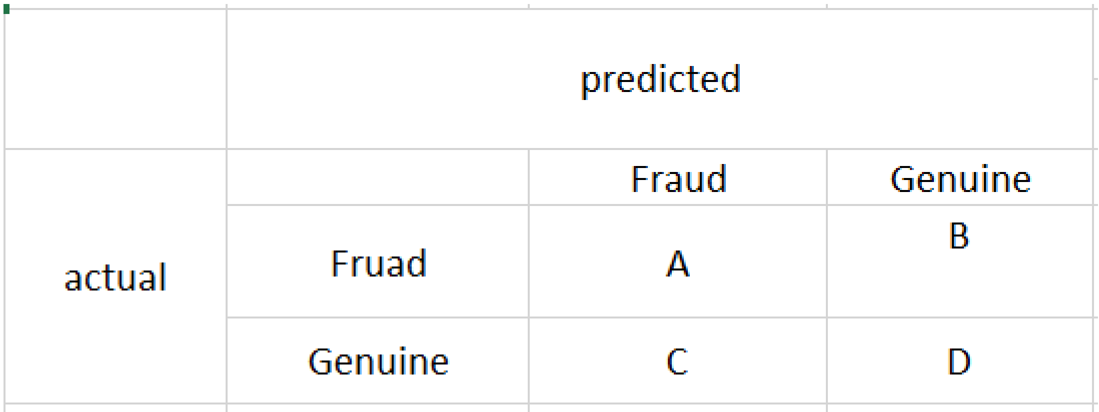
Logistic regression Decision Set of rules Propositional rules Logic Programs Neural Network Graphical models Bayesian networks Conditional random fields	Information gain K-L divergence Cost/Utility Margin	Unconstrained Gradient descent Conjugate gradient Quasi-Newton methods Constrained Linear Programming Quadratic Programming

Table 4: Component of algorithms (Domingos, 2012)

How does a confusion matrix work?

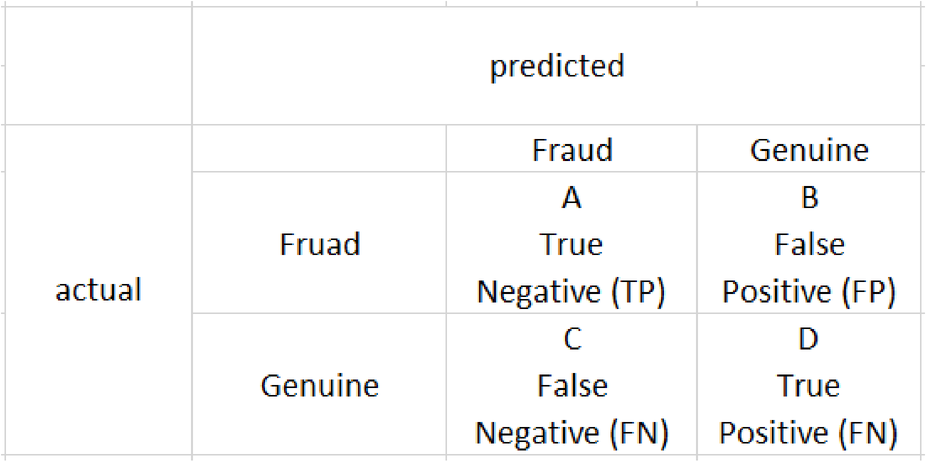
A confusion matrix is typically computed to calculate a cross tabulation of observed (true) and predicted classes (model) in any machine learning classifier such as logistic regression, decision tree, support vector machine. Naïve Bayes and may more. There are several matrices such as precision and recall that helps to find the accuracy of model and choose the best model.

Suppose a 2-class case confusion matrix with Fraud and Genuine is shown below where row represents the instances of an actual class and each column represents the instances of a predicted class.



	predicted		
actual		Fraud	Genuine
	Fruad	A	B
	Genuine	C	D

The matrix fields say the following



	predicted		
actual		Fraud	Genuine
	Fruad	A True Negative (TP)	B False Positive (FP)
	Genuine	C False Negative (FN)	D True Positive (FN)

Now, we can describe important performance measures in machine learning

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP}$$

The accuracy is not acceptable performance measure in machine learning. Let's say we have 1000 samples of data where 995 were fraud class and only 5 of them are genuine class. When we use classifier for this sample data, the accuracy would be a remarkable 99.5%, also classifier is unable to classify any positive samples (Visa, et al., 2011).

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{True Negative Rate} = \frac{FP}{TN+FP}$$

$$\text{Precision} = \frac{TP}{FP+TP}$$

4.2. Business Process Model (BPM)

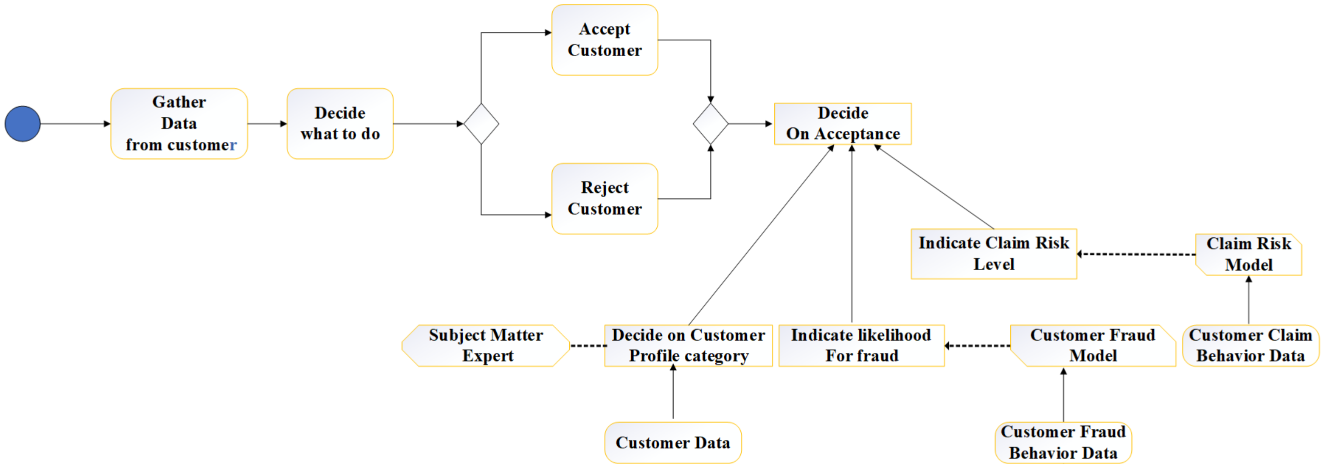

Gather Data from customer
Decide what to do
Accept Customer
Reject Customer
Decide On Acceptance
Indicate Claim Risk Level
Claim Risk Model
Subject Matter Expert
Decide on Customer Profile category
Indicate likelihood For fraud
Customer Fraud Model
Customer Claim Behavior Data
Customer Data
Customer Fraud Behavior Data

Figure 2: Business Process Model

5. Software Requirement Specification (SRS)

SRS is basically an organization understanding of a customer or potential clients system and dependencies at a particular point in time prior to any actual design or development work. SRS has been developed for future reference in case of any ambiguity and misunderstanding. SRS provides of the requirements, behaviors, constraints, and performance of the system (tutorialspoint.com, 2020).

5.1. Requirement analysis

Requirement analysis is for transformation of operational need into software description, software performance parameter, and software configuration through use of standard, iterative, process of analysis and trade-off studies for understanding what the customer wants analyzing need, assessing feasibility, negotiating a reasonable solution validating the specification and managing the requirements.

5.1.1. Purpose of SRS

The aim of SRS is to identify the criteria for the identification of fraud in credit cards also include a broad outline of our project, including user requirements, product perspective and requirement overview and general constraints. It will also include the functionality needed for this project and specific requirements for this project such as interface, functional requirements, and performance requirements.

5.1.2. Scope of SRS

The scope of this SRS document like functionality, performance, constraints, interface, and reliability which persists for the entire life cycle of the project which defines the final state of the software requirements agreed upon by the customers and designers. Finally, at the end of the project, all tasks can be tracked from the SRS to the component.

5.1.3. Overview

The software requirement specification document for the system covers the following two sections i.e. General Descriptions and Specifications Requirements where general description provides general description of the project which includes the description about the production function, user characteristics and general constraint and specification requirement describes both functional and non-functional requirement of the project.

5.1.4. General Descriptions

A fraud in credit cards has been developed to alert the customer to their credit card fraud. After the payment method, the transactions carried out are checked if the transaction out is a true transaction or a false transaction and reduce the false alarm by applying a machine learning and deep learning algorithms.

5.1.4.1. Product Function

The project is expected is expected to provide consistent outcomes and the functionality of the product to detect number of fraud transaction effectively and offering flexibility to the customer in a safe and reliable manner.

5.1.4.2. User Characteristics

Customers and administrator are classified as the user of the system.

- Customer are those who make the transaction through any means
- Administrator are those who computes on the transaction and reports about the fraud usage

5.1.4.3. General Constraints

- **Audit Functions:** There shall be no audit functions.
- **Interfaces to other application:** There shall be no interfaces.
- **Control Functions:** There shall be no control functions.
- **Hardware Limitations:** There are no hardware limitations.
- **Parallel Operations:** There are parallel operations.

5.1.5. Functional Requirements

The interaction between input and output of the system is defined by SRS functional criteria.

5.1.5.1. Technical Issues

Many software projects have failed due to incomplete or incorrect analysis, including technical problems. Technical issues are a crucial factor in designing the software program.

5.1.5.2. Risk Analysis

Project risk estimation is for expense estimation with defined precision and cost for capital investment projects. The key task is to decide how to model and visualize the complex relationships between risks, to identify and track the effect of risk, to assess the probability of risk incidence, to minimize the negative influence of risks, and to control the success of the project with risks and uncertainties.

5.1.6. Interface Requirements

The performance of the system is adequate. Mostly vendor deals for the user internet access, 60 percent is up to the client side.

5.1.6.1. Hardware Requirements

- **Processor type**: Pentium III-compatible processor or faster.
- **Processor speed: Minimum**: 1.0 GHz, Recommended: 2.0 GHz or faster
- **RAM**: 512 MB or more
- **HARD DISK**: 20GB or more
- **Monitor**: VGA or higher resolution 800x600 or higher resolution
- **Pointing device**: Microsoft Mouse or compatible pointing device
- **CD-ROM**: Actual requirements will vary based on system configuration and the applications and features chosen to install.
- PC/laptop/Server

5.1.6.2. Software Requirements

- **Operating System:** Windows XP Professional or more or Linux
- **Back End:** SQL server
- **Application Software Framework:** Python
- **Library:** Scikit Learn

5.1.7. Performance Requirements

Following are the performance requirements of the project:

- The key criterion is that no fault situation forces the project to exit suddenly
- Any error that happened in some phase should produce a clear error message.

- The response / answer should be relatively fast, the intervention participants should not be confused at any point of time regarding the activity that is taking place.
- The performance of the system is adequate.

5.1.8. Non-Functional Requirements

- Secure access of confidential data (user's details)
- 24 X 7 availability and should be efficient
- Better component design to get better performance at peak time
- Flexible service-based architecture will be highly desirable for future extension
- The system must display necessary information in case of failure preventing system breakdown.

5.2. Feasibility Study

5.2.1. Economic Feasibility

The project requires a high-end graphics processing unit (GPU) and CPU for creation however for CPU for creation however for the user to use the product a decent GPU and CPU will be enough. Hence, the system we are going to develop does not require an enormous amount of money so it will be economically feasible.

5.2.2. Operational Feasibility

The project requires the general user-friendly environment to store the dataset then predicts the frame using trained with dataset.

5.2.3. Technical Feasibility

This project requires large amount of database space to stores credit card transaction details and processing power for processing real time data to recognize fraud transaction, and genuine transactions. For the administrator to use the product a descent GPU and CPU will be enough.

6. System Design and Architecture

6.1. System Design

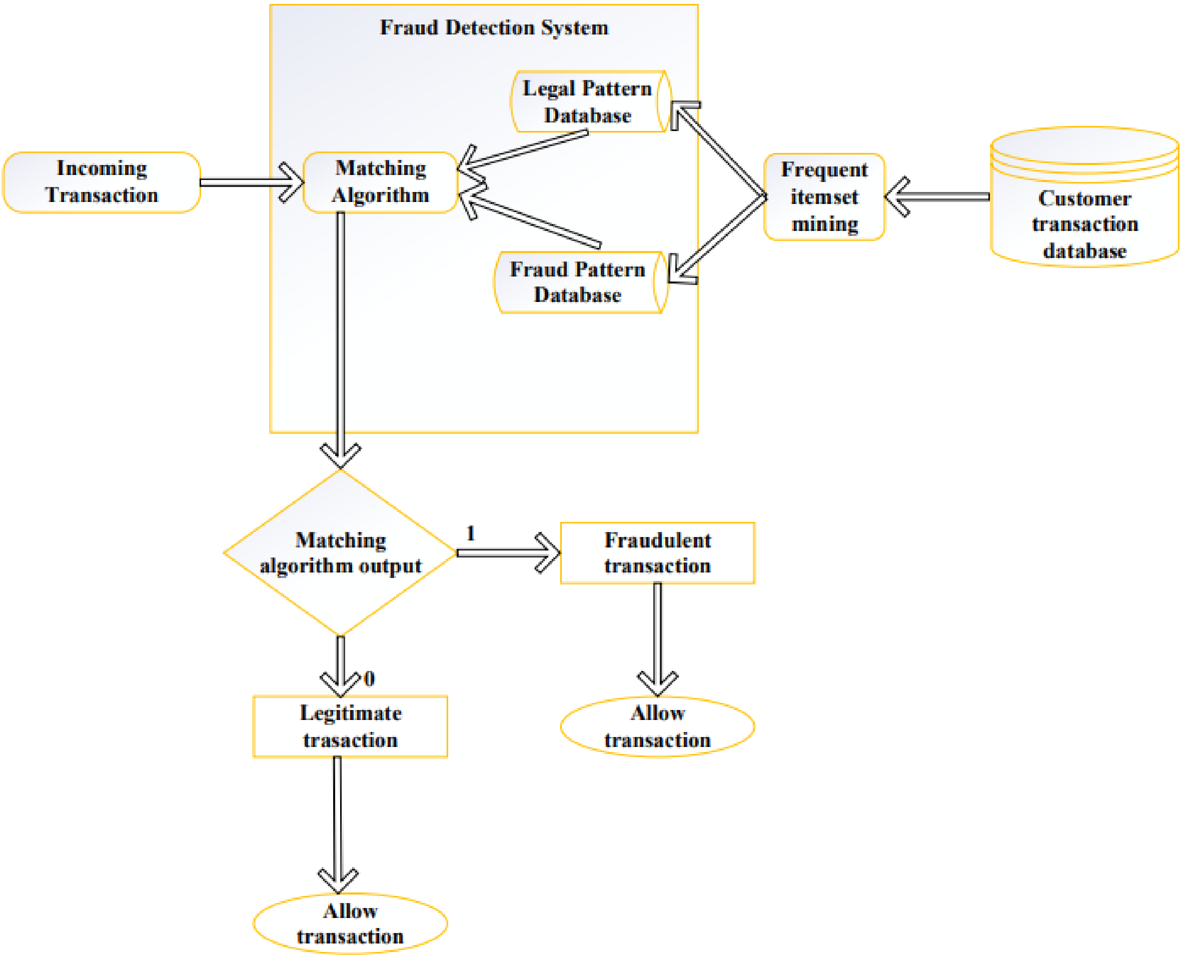


Figure 3: System Design

6.2. Use Case Diagram

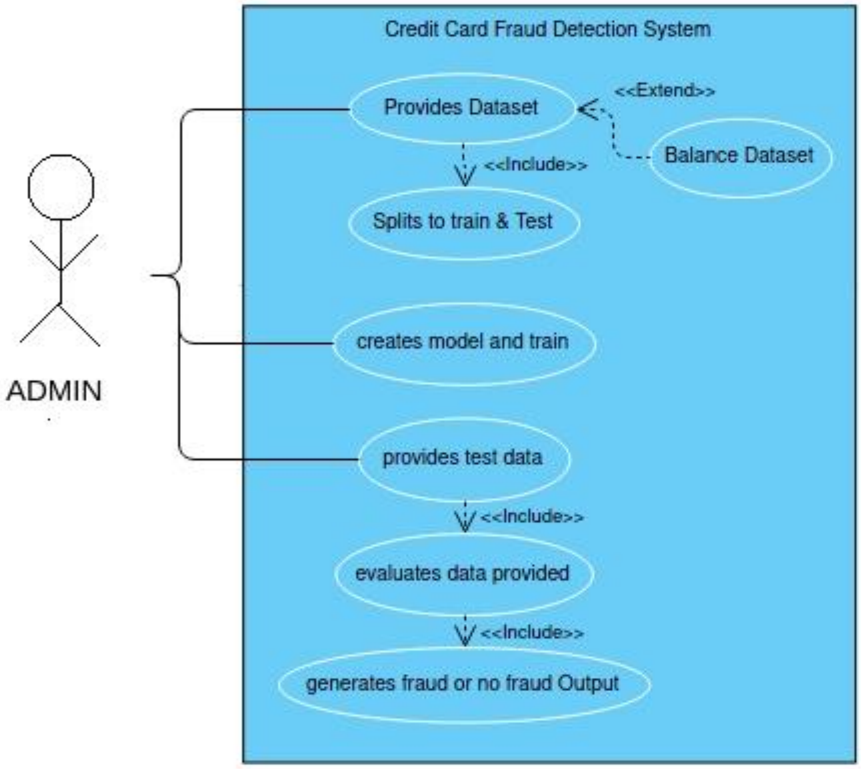


Figure 4: Use case diagram

7. Approach

7.1. Sampling Technique

Learning from class imbalanced data continues to be a popular and difficult problem in supervised learning as traditional classification algorithms are structured to handle balanced class distributions. Although there is various approach to solve this problem, methods that produce artificial data to achieve a balanced class distribution are more robust than modifications to the classification algorithm. These methods, such as over-samplers, change training results, enabling every classifier to be used on class-imbalanced datasets. Several algorithms have been suggested for this problem, but most of them are complicated to understand and appear to produce unwanted noise. The research provides a clear and efficient form of oversampling focused on SMOTE oversampling, which prevents noise production and essentially overcomes imbalances between and within classes. Empirical findings of detailed studies from credit card fraud transaction dataset indicate that over-sampled training data for the suggested approach increases classification performance where an implementation is supported in the python programming language (Last, et al., 2017).

7.1.1. SMOTE

Synthetic Minority Oversampling Technique also known as SMOTE is a technique based on machine learning for classification of data where Kaggle data for this research is trained using SMOTE technique to solve data imbalance which is mainly used to differentiate fraud transactions from the original transactions carried out by cardholders. Initially, the transaction is processed in a confluence form. Thus, the confluence data was trained by the SMOTE method to synthesize fraud transactions from non-fraud transactions (Sahayasakila.V, et al., 2019).

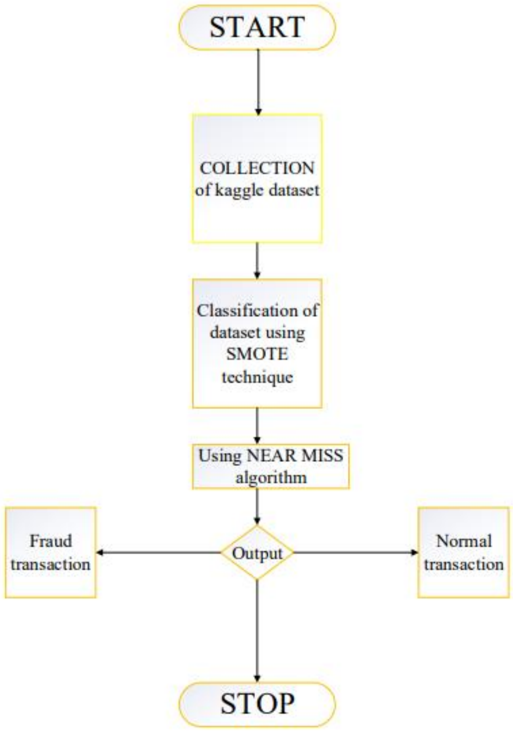

START
COLLECTION of kaggle dataset
Classification of dataset using SMOTE technique
Using NEAR MISS algorithm
Fraud transaction
Output
Normal transaction
STOP

Figure 5: Flow Chart of SMOTE Technique (Sahayasakila.V, et al., 2019)

More deep insights into how SMOTE algorithm works

Step 1: The initial step is to establish the minority class let's say set A, for each x ∈ A, the k-nearest neighbors of x are obtained by calculating the Euclidean distance between x and each other sample in set A.

Step 2: The sampling rate N is determined by the imbalanced proportion, for each x ∈ A, N (i.e. x_1, x_2, ... x_n) are randomly chosen from k-nearest neighbors, and they create the set A_1.

Step 3: Consider an example, x_k ∈ A_1(k = 1,2,3 ... N) the below formula is used to new example(sample) (geeksforgeeks.org, 2020).

$$\boldsymbol{x' = \mathrm{x} + \mathrm{rand}\ (0{,}1) * |\ \mathrm{x} - x_k\ |}$$

Where,

rand (0,1) = random number between 0 and 1.

In the above figure, synthesized transactions are re-sampled to test the consistency of the records. Synthesized fraud transactions are optimized by using near miss algorithm which aims to balance distribution of class by randomly eliminating majority class samples. If the instances of two separated classes are very similar to each other, the instances of the majority class are excluded to maximize the difference between the two class samples which helps in classification process. Also, it is used to prevent information loss of dataset.

The dataset contains high volume of majority class than minority class where majority class is the genuine transaction and minority class is the fraud transactions of the dataset.

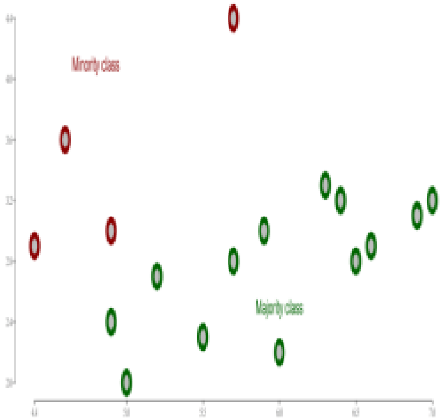


Figure 6: SMOTE Technique (Sahayasakila.V, et al., 2019)

The fundamental theory on the function of near-neighbor method is as follows:

Step 1: First, the method identifies the distance between all cases of the majority class and those of the minority class. That is where the largest class tends to be under-sampled.

Step 2: N instances of the majority class that have the lowest differences to those of the minority class are then selected.

Step 3: If there are k instances in the minority class, the closest method will result in k*n instances in the majority class.

To find n closest instances in the majority class, there are several variations in the NearMiss algorithm:

NearMiss – Version 1: The k closest instances of the minority class are smallest when it selects of the minority class of each average distance.

NearMiss – Version 2: The k farthest instances of the minority class are smallest when it selects sample of the majority class for each average distance.

NearMiss – Version 3: It is working in two steps. First, their nearest M-neighbors will be stored for every minority class example. Finally, the majority class instances are selected for which the average distance to the nearest N-neighbors is the largest (geeksforgeeks.org, 2020).

Fraud transaction from the original non-fraud transactions with the smote synthesis is shown below:

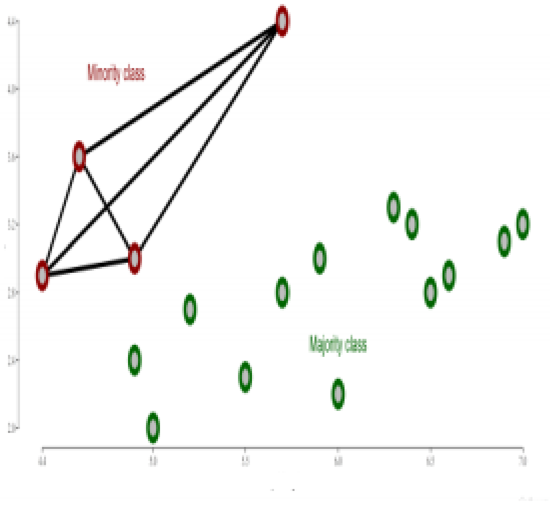


Figure 7: Synthesis of minority class (Sahayasakila.V, et al., 2019)

The dataset consists of transactions made with credit cards. This sample contains 492 scam transactions out of 284,807 transactions. As a result, the positive class (frauds) accounts for 0.172 percent of all transactions. After applying SMOTE algorithm, we get below output as our fraud transactions.

```
>>> from imblearn.over_sampling import SMOTE
>>> sm = SMOTE(random_state= 42)
>>> x_Sampled,y_Sampled = sm.fit_sample(xData,yData.values.ravel())
>>> Source_data_no_fraud_count = len(data[data.Class==0])
>>> Source_data_fraud_count = len(data[data.Class==1])
>>> print('Percentage of fraud counts in original dataset:{}%'.format
...      ((Source_data_fraud_count*100)/(Source_data_no_fraud_count
...                                    +Source_data_fraud_count)))
>>> Sampled_data_no_fraud_count = len(y_Sampled[y_Sampled==0])
>>> Sampled_data_fraud_count = len(y_Sampled[y_Sampled==1])
>>> print('Percentage of fraud counts in the new data:{}%'.format
...     ((Sampled_data_fraud_count*100)/(Sampled_data_no_fraud_count
...                                    +Sampled_data_fraud_count)))
```

Output:

```
Percentage of fraud counts in original dataset:0.1727485630620034%
Percentage of fraud counts in the new data:50.0%
```

7.2. Algorithm Used

A lot of work has been performed on the prevention of credit card fraud. The general history of credit card system and non-technical knowledge about the credit can be learnt from introduction part of this report. While selecting specific resource for literature review, most of the credit card fraud detection are build using supervised algorithms such as neural networks. The results of classifier models developed using a deep learning model and well-known logistic regression and random forest with imbalanced dataset are compared in this analysis.

7.2.1. Logistic Regression

A well-established statistical method for predicting binomial or multinomial results is logistic regression. Multinomial Logistic Regression algorithm can generate models when a set field with two or more possible values is the target field. And Binomial Logistic Regression algorithm is restricted to such models where the goal area is a flag or binary area.

For the classification of fraud identification, we use Logistic Regression. Logistic Regression is a form of probabilistic model of statistical classification which uses the logistic curve to detect fraud. The univariate logistic curve formal is shown below:

$$p = \frac{e^{(c_o + c_1 x_1)}}{1 + e^{(c_o + c_1 x_1)}} \text{ ------------ } (1)$$

The logistic curve gives a value between 0 and 1, so that it can be interpreted as a class membership probability. To execute the regression, the logistic function can be applied as shown below:

$$log_e\left(\frac{p}{1-p}\right) \text{ -------------- } (2)$$

Here, 1 – p is the probability that tuple will not be in class and p is the probability that tuple will be in class. However, the algorithm selects coefficient of C_o and C_1 that optimize probability of incoming transaction The algorithm flow of logistic regression is shown below where confusion matrix is used to illustrate the performance of the classifier (Sahin & Duman, 2011).

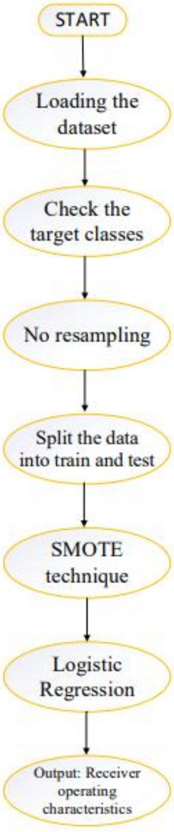


Figure 8: Algorithm flow of Logistic Regression

7.2.2. Random Forest

Random forest which is the part of supervised learning technique used to solve classification as well as regression problem whereas random forest also known as collection of decision tree classifiers. In this research we have not use decision tree as our classifier model because random forest has an advantage over the decision tree as it corrects training set with the habit of overfitting. A subset of the training set is sampled randomly such that each tree is trained and then decision tree is built, each node then splits on a feature chosen from a random subset of the compete feature set. In the random forest, training is extremely even for large sets with many features and data instances and because each tree is trained independently of the others. It has been observed that the random forest algorithm gives a reasonable approximation of the generalization error and is resistant to overfitting.

$$\underset{\substack{\downarrow \\ \text{Posterior Probability}}}{\mathrm{P}\,(\mathrm{c} \mid \mathrm{x})} = \frac{\overset{\substack{\text{Likelihood} \\ \uparrow}}{P\ (x\mid c\ )}\, P(c)}{\underset{\substack{\downarrow \\ \text{Predictor Prior Probability}}}{P(x)}} \longrightarrow \text{Class Prior Probability}$$

Random forest ranks the importance of variable in a natural way in regression and classification problem. The random forest uses the below pseudocode to perform the prediction of fraudulent transaction.

Step 1: Extract the incoming transaction test features and use the rules of every randomly created decision tree to predict the outcome and store the predicted outcome(target).

Step 2: Calculate the votes for every target output predicted

Step 3: Evaluates the highly voted projected goal as the final statistical performance from specific decision tree (Xuan, et al., 2018).

7.2.3. Autoencoder Neural network

An autoencoder is a type of artificial neural network that is used in an unsupervised way to learn effective data coding. For more concrete understanding, autoencoder is neural network which is trained to recreate whatever is fed as input as output. There is a series of hidden layers in between the input layer and the output layer. There is a layer right in the middle, containing fewer neurons than the input. The output of that layer is the result of the autoencoder's so called encoder part.

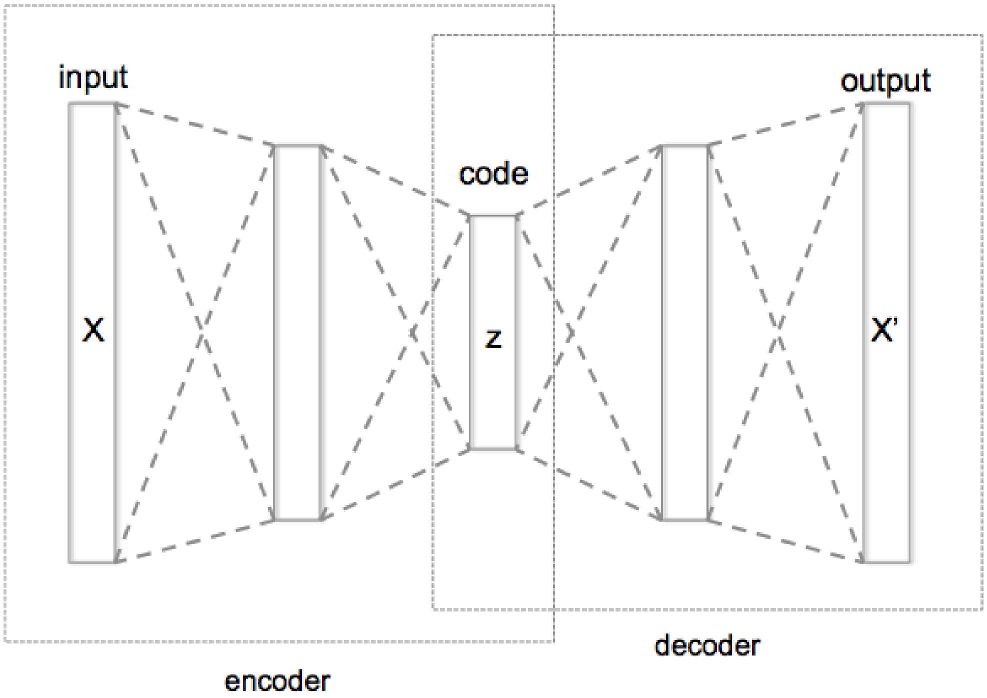


Figure 9: Architecture of autoencoder neural network

The reasonable behind utilizing autoencoders is that in a pleasant way the secrete layers map the data to a vector space. The autoencoder in this research map our input from a high dimensional space to one with fewer dimensions by using the layer with few neurons. In the above figure 9, the network structure has layer-to-layer connections, but does not have connection within each layer, where X is the input sample and X' is the output feature. Autoencoder neural network training in this project is programmed to minimize reconstruction error by using the samples provided. The cost function defined in the project for autoencoder neural network is shown below:

$$J_{A,E} = \frac{1}{m}\sum_{i=1}^{m} (\frac{1}{2} \, ||X' - X||^2) \text{-----} \quad (1)$$

where m represents number of input samples.

Gaussian noise, and salt and pepper noise are the commonly used noises. And cost function of of denoising autoencoder neural network is shown below:

$$J_{D,A,E} = \frac{1}{m}\sum_{i=1}^{m} (\frac{1}{2} \, ||X' - X||^2) \text{-----} \quad (2)$$

Where, **X' = f ((**$\sum \boldsymbol{wX} + \boldsymbol{b}$**)), w = weights** and **b = bias**

There is range of new autoencoder named as denoising autoencoder that might allow autoencoder to learn how to eliminate noise and recreate uninterrupted inputs as often possible.

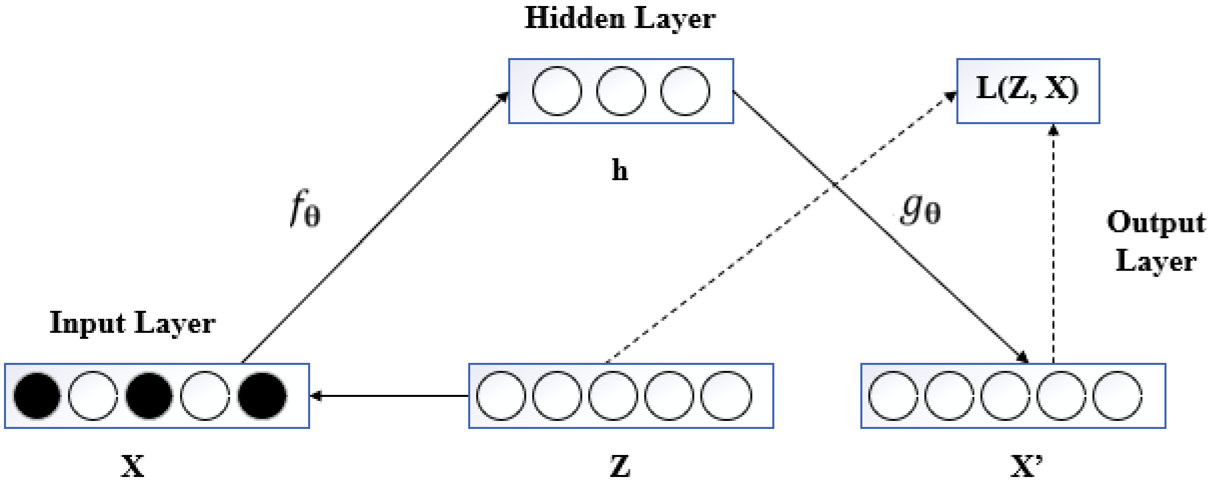


Figure 10: Denoising autoencoder neural network

The original data X and Z as shown in the figure is the is the data corrupted with noise. The output X' is the complete denoising autoencoder process. The loss function aims to reduce the difference between the output and the original data in such a way that the autoencoder can remove the effect of noise and retrieve features of distorted results. Hence, the features produced by learning distorted input with noise are more robust, which enhanced the autoencoder neural network model's ability to generalize data to input. Entropy is a measure of the content of information and could be defined as the unpredictability of an event. And the greater the probability, the lower the unpredictability, which ensures that the value of the knowledge is therefore quite high. If an occurrence eventually happens with a 100% chance, so the unpredictability and the value of information are 0. Cross entropy loss function takes advantage of entropy equation functionality whereas cross entropy is used for multiclassification problem with the SoftMax activation function. So, SoftMax activation function in our autoencoder neural network model used in out last layer of our neural network which first calculates the exponential value of each output, then normalize all the output and allows the output sum to be equal to 1. Cross-entropy loss function will calculate the efficiency of a classification model, which is shown below (Yuan, et al., 2019).

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} 1(y_i = j) log \frac{e^{\theta^T{}_j{}^{x_i}}}{\sum_{i=1}^{k}\theta^T j x_i} \text{ -----} \quad (3)$$

8. Comparative analysis

9. References

Al-Jumeily, D., Hussain, A. J., MacDermott, Á. & Tawfik, H., 2015. The Development of Fraud Detection Systems for Detection of Potentially Fraudulent Applications. 12, pp. 7-13.

Barker, K. J. & Sheridon, J. D. P., 2008. Credit card fraud: awareness and prevention. *Journal of Financial Crime,* Volume 15, pp. 398-410.

Delamaire, L., Abdou, H. A. & Pointon, J., 2009. Credit card fraud and detection techniques: A review. *Banks and Bank Systems,* Volume 4.

Domingos, P., 2012. *Communications of the ACM.* Seattle, WA 98195-2350, U.S.A.: Communications of the ACM.

geeksforgeeks.org, 2020. *ML | Handling Imbalanced Data with SMOTE and Near Miss Algorithm in Python.* [Online]
Available at: https://www.geeksforgeeks.org
[Accessed 10 05 2020].

Ghazi, R., 2019. Financial institutions and their role in the development and financing of small individual projects. 07.

Giri, P., 2013. PAYMENT AND SETTLEMENT SYSTEM DEVELOPMENT IN NEPAL: HURDLES AND WAY OUT. *Economic Journal of Development Issues,* Volume 15 and 16 , pp. 1-2.

Jain, Y., Tiwari, N. & Jain , a., 2019. A comparative analysis of various credit card fraud detection techniques. *International Journal of Recent Technology and Engineering,* Volume 7, pp. 402 - 407 .

Jesus, A. d., 2019. *Machine Learning for Credit Card Fraud – 7 Applications for Detection and Prevention.* [Online]
Available at: https://emerj.com/ai-sector-overviews/machine-learning-for-credit-card-fraud/
[Accessed 06 05 2020].

Kawulich, B., 2012. Collecting data through observation. In: C. Wagner, B. Kawulich & M. Garner, eds. s.l.:McGraw Hill, pp. 150 - 160.

Kumar, 1. S., Naidu, M. V. & Sujatha, D., 2019. Credit Card Fraud Detection System Based On Machine Learning Techniques. *IOSR Journal of Computer Engineering (IOSR-JCE),* 21(3), pp. 45-52.

Last, F., Douzas, G. & Bação, F., 2017. Oversampling for Imbalanced Learning Based on K-Means and SMOTE. 02 11.

Mathers, N., Fox, N. J. & Hunn, A., 2000. Research Approaches in Primary Car. In: *Using Interviews in a Research Project* . s.l.:Radcliffe Medical Press/Trent Focus, pp. 113 - 134.

Sahayasakila.V, Monisha, D. K., Aishwarya & S. V., 2019. Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm. *International Journal of Engineering and Advanced Technology (IJEAT)* , 8(5), pp. 190 - 192.

Sahin, Y. & Duman, E., 2011. Detecting credit card fraud by ANN and logistic regression. *nternational Symposium on INnovations in Intelligent SysTems and Applications.*

Salman, A. & Salman, A., 2015. Relationship between the Incentives Offered on Credit Card and its Usage. 01, p. 30.

Shirgave, S. K., Awati, C. J., More, R. & Patil, S. S., 2019. A Review On Credit Card Fraud Detection Using Machine Learning. *International Journal of Scientific & Technology Research,* Volume 8, pp. 1217 - 1220.

Tulasi, 2020. *Fact Finding Techniques || Fact-Finding Techniques for Requirements Discovery || Bcis Notes.* [Online]
Available at: https://bcisnotes.com/thirdsemester/system-analysis-and-design/fact-finding-techniques-fact-finding/#:~:text=Fact%2Dfinding%20techniques%20are%20a,prototyping%2C%20and%20joint%20requirements%20planning.
[Accessed 03 05 2020].

tutorialspoint.com, 2020. *https://www.tutorialspoint.com.* [Online]
Available at:
https://www.tutorialspoint.com/system_analysis_and_design/system_analysis_and_design_planning.htm#:~:text=Information%20Gathering%20Techniques,precise%20SRS%20understood%20by%20user.&text=be%20complete%2C%20Unambiguous%2C%20and%20Jargon,tactical%2C%20and%
[Accessed 03 05 2020].

tutorialspoint.com, 2020. *https://www.tutorialspoint.com.* [Online]
Available at: https://www.tutorialspoint.com/software_engineering/software_requirements.htm
[Accessed 06 05 2020].

Visa, S., Ramsay, B., Ralescu, A. & Knaap, E. v. d., 2011. Confusion Matrix-based Feature Selection.. *CEUR Workshop Proceedings,* Volume 710, pp. 120 - 127.

Xuan, S., Liu, G., Li, Z. & Zheng, L., 2018. Random forest for credit card fraud detection. *IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),* pp. 1 - 6.

Yuan, F.-N., Zhang, L., Shi, J.-T. & Xia, X., 2019. Theories and Applications of Auto-Encoder Neural Networks: A Literature Survey. *Jisuanji Xuebao/Chinese Journal of Computers,* Volume 42, pp. 203 - 230.