National College of Ireland

Prediction of Customer Lifetime Value and Fraud Detection in BFSI using Machine Learning

MSc Research Project
Data Analytics

Vaibhav Subhash Sawant
Student ID: x19200706

School of Computing
National College of Ireland

Supervisor: Martin Alain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vaibhav Subhash Sawant
Student ID:	x19200706
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Martin Alain
Submission Due Date:	16/12/2019
Project Title:	Prediction of Customer Lifetime Value and Fraud Detection in BFSI using Machine Learning
Word Count:	7303
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Customer Lifetime Value and Fraud Detection in BFSI using Machine Learning

Vaibhav Subhash Sawant
x19200706

Abstract

Finance has a profound impact and relation to every person on the globe as it decides the way of life the individual is going lead. Banks and Financial Institutions are at the center of the financial or capitalist world. Both these entities are trying to adopt the new technology to reach a more significant number of people, contributing to higher profits. While dealing with the competition in the world market, the BSFI is experiencing different problems that make it difficult to survive. One such problem is obtaining suitable potential customers from the business perspective. While dealing with such competitive challenges, the domain is also dealing with some problems caused due to technological advancement in the business. More than 1.6 billion financial losses were accounted for in recent years due to various frauds in the domain. This makes it a severe problem to be addressed on priority.This research closely studies the two major problems faced by the BFSI domain, i.e., Fraud detection and calculating the Customer Lifetime Value, and responds with the application of the computing field of Machine learning. Much data available publicly for research purposes was collected, One from the Banking and the other from the Insurance domain. The nature of the data, the research question was studied conscientiously, and based on these parameters, the machine learning models, namely Support vector machine, Extreme Gradient boosting, Bernoulli Naïve Bayes, Decision tree, and Random Forest models were used for the Detection of Frauds in the transaction and calculating the CLV.

1 Introduction

1.1 Background

Economics has a substantial impact on the life of the human race. The impact can range from an individual level like the cost of milk in a coffee in the morning to GDP rate contribution to the development of some country. Economics can be divided into various sectors with a high level of influence. One such sector is Banking, finance, and Insurance. It is one of such sectors which is related to ordinary people as well as the world economics too. Banks are institutions that try to support the countries or world economics by balancing the debits and credits. Money lending business provides access to essential and luxury commodities for common people. On the other hand, introducing various options for investments and savings in the market helps maintain the balance between debit and credits or even the demand and supply. Insurance is another such sector that is indirectly helping the economy be prepared and accessible in risky or unpredictable situations.

Banks and financial institutions are also trying to upgrade and implement a technology-oriented approach to reach and facilitate the customers in this technology-driven generation. Internet banking or mobile applications signify the new age approach of this domain for speedy and result-oriented services to customers and hassle-free business. While developing many such banking applications as software engineers, we encountered the business understandings and insights of the domain. During this period, we also encountered some of the problems faced by banking or financial institutions leading their business in the competitive market. Our curiosity for exploring the world of machine learning inspired us to find the applications of machine learning in providing feasible, effective, and highly accurate solutions to the problems of these domains. Machine learning, which is new-age technology, can be advantageous and functional for the Banking, financial, and Insurance domain.

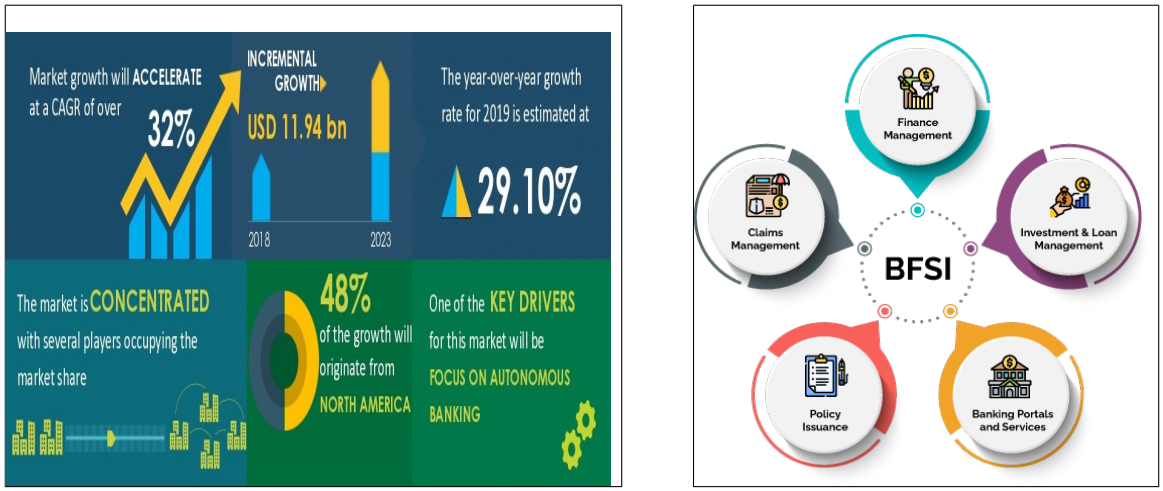


Figure 1: BFSI Industry Insights

1.2 Limitations in Literature

The BFSI domain all over the world is encountering numerous issues during their day-to-day business. Many of them are unpredictable. It can be observed that these problems are overlooked to the extent that very little research can be found for addressing these issues. This is one of the motivation of applying machine learning for this field of Economics.Chen and Han (2021); Rawte and Anuradha (2015) The research will help address some of the problems with the help of machine learning and the various related techniques and concepts related to it. It will also provide a firm base for the upcoming research where other techniques and technologies can also be used along with machine learning to provide a successful solution. Implementing new machine learning models and techniques in the research can help find different alternatives and increase the reach of research in this field.

The research intends to build a robust solution with the help of a machine learning platform for addressing multiple problems for the domain. The Idea of the research can be turned out to be a privilege for the three different predominant fields likes Banking, Finance, and Insurance. As many of the solutions will be available at the same platform, the expenditure required for setting up a different platform for a different solution like Infrastructural cost, technological licenses, human resources will be saved. Desirena et al. (2019); Chuang and Shen (2008) The developed system will be easy to integrate with the

existing technologies or platform these domains are using for the business and will also provide a scope to integrate upcoming technologies in the future.

1.3 Research Question and Objectives

The research will focus on two leading problems faced by banks and financial institutions. These days BFSI is adopting the new age technologies to provide fast and hassle-free services. This is also leading to some malpractices and frauds, causing business losses. So, one of the research questions is related to fraud detection for helping the banks to tackle this problem effectively. On the other hand, due to leading competition, these domains are experiencing difficulties in reaching the right potential customer from the business perspective and at the same time retaining the existing ones. So, the other research question will be related to this problem which can find a solution in Customer Lifetime Value.

The Research Questions are:

- How efficient machine learning algorithms are in the detection of Fraud in banking transactions?
- How effective and practical machine learning techniques are in predicting the Customer's lifetime value for BFSI?

To reach the objectives framed by the research question, we will be using one of the innovative technologies of the generation like machine learning. Machine learning is finding its applications in numerous fields ranging from E-commerce, Banking to Medical. The data required for the research will be obtained from the public datasets from the Kaggle website. Two different datasets were used in the research for addressing the research questions. Both the datasets contain data of the customers in the domain of banking and insurance that contains the data related to the demographic, financial, vehicle, or policy-related information. The link for the datasets is given below.[1] The dataset will undergo various Data processing steps like Data cleaning, preprocessing, initial visualization, etc. The data will be made ready for modeling. The machine learning models like Bernoulli Naive Bayes, Support Vector Machine, Bagging and Boosting Decision Tree, Lasso Regression, Random Forest Regression, Extreme Gradient Boosting will be used. The modeling performance will be assessed using evaluation matrices like Confusion matrix, F1 score, precision, recall, etc. Hyper tuning will be performed for fine-tuning the results, and outcomes will be presented.

1.4 Document Structure

The research will be carried out to find the solution to the problems of the BFSI domain, particularly for Fraud detection and calculation of Customer Lifetime Value. The research document comprises seven sections that describe the details about the different aspects involved in the research. Section 2 will be Related work that will help understand the background of machine learning in BFSI and describe research done in the

[1]
Fraud Detection : `https://tinyurl.com/2p99n3mn`
Customer Lifetime Value : `https://tinyurl.com/2p8ejvtc`

same. Section 3 is regarding the Methodology that was used to address the two different research questions. Section 4 will be of Design Specification that gives information about the architecture or framework design and its application in the research. Section 5 will depict the whole process carried out to implement the technical solution for the research and insights. Section 6 will take care of the different number of experiments to improve the performance of the various models used in the research to reach the required business goals. Section 7 will be the final section of the research document that will conclude the topic with the relevant results and describe the future work that can be carried out with the research as the baseline.

2 Related Work

As we are born in the age of computing, we feel privileged to experience the comfort and ease in walks of life mainly due to the various application of computing and technology. On the other hand, we face the dark sides of technology as there are risks, theft, distraction, or problems. Sifa et al. (2018) As our research concentrates mainly on the Banking Finance and Insurance Sector (BFSI), we will discuss the problems in the context of these domains. The BFSI faces numerous other problems, but we mainly try to spread light on the two most crucial problems faced, i.e., Customer Lifetime Value and Fraud detection. Machine learning is an emerging branch of computing that deals with the training of the algorithms with the help of the existing data and make them ready to deal with real-life problems. Large amounts of data either generated or collected from the customer on an everyday basis can turn out to be a crucial part of solving these problems. In the case of fraud detection, it becomes essential to identify the pattern or key characteristics that point out the fraudulence.Yao et al. (2018)Mittal and Tyagi (2019) Supervised and Unsupervised models are used to train with the available data and deploy the trained model to detect the fraud further. A technique like deep learning can also detect financial transaction fraud using the images of the account statement or passbooks. Similarly, the data related to transaction and demographic data of the customer can be helpful to build a successful model that helps find the right customer to be targeted for a particular product.

2.1 Machine learning Methods Used for Fraud Detection

Machine learning is discovering its application in providing solutions to problems like fraud detection and calculating the customer lifetime value. Some of the research is found to address the same problems other than in BFSI is the evidence of machine learning and its application as there is some scarcity of research done in these domains.

Among very few research that was carried out mainly in BFSI,Vidanelage et al. (2019)Yao et al. (2018) tried to research the frauds that are causing in the transactions. The dataset containing the demographic and other details of the customer of some banks situated in Spain were used for the research purpose. Four different machine learning models were studied for the research, which comes under the unsupervised and deep learning categories. Some of those models are namely Multinomial and Gaussian Naive Bayes (MNB GNB), K-Nearest Neighbor (KNN), and Multilayer Perceptron (MLP). As more people are obsessed with cell phones in these tech worlds, some of the searchers studied the fraudulence that is taking place in mobile transactions.Roy and George (2017)

Delecourt and Guo (2019) For this study, especially the different type of attacks was studied to know their intensity of impact on the frauds. White and black box testing where some of them. This also helped in feature selection before the modeling process. Random Forest (RF), Logistic Model (LR) are some of the models which are being used to carry out the research. The data gathered or generated can be massive and can be structured or unstructured. Handling such data is another challenge.Shivanna et al. (2020) Gyamfi and Abdulai (2018)tried to emphasize the problem. They used Big Data as a new-age implementation as it has numerous advantages. It can handle massive data of any nature. In these, the mainly used Machine learning technique were Back Propagation Networks.

We have witnessed the fact that there is little research done in the domain of BFSI, the efforts taken by some of the researchers for the implementation of machine learning showed up some initial success for the domain. The research done byYao et al. (2018)Vidanelage et al. (2019) implemented four different types of modeling techniques for the detection of frauds in the transactions. When it comes to discovering the best modeling, accuracy and processing time play an important role. The best thing about this research was that all four models showed more than 90 percent accuracy in detecting fraud and MLP with the highest accuracy. However, the only factor against MLP was the processing time taken, as it took more than 5 minutes. Shivanna et al. (2020) Gyamfi and Abdulai (2018) Implemented their research with the help of two different architectures, namely Set-1F-To-1N and Set-1F-To-4N. Back Propagation Networks are those model which were implemented for the research. After the application of models on training and testing data, the accuracy and processing time was calculated. Based on these factors, the Support vector machine outperformed Back Propagation Networks.

2.2 Machine learning Methods Used for Calculation of CLV

To deal with the competitive business world, maintaining a good relationship with customers and acquiring new ones became tough for every business industry. As a solution for targeting the right customer for business, the customer lifetime value proves beneficial.To overcome the limitations found in the traditional methods of CLV calculation and introduce the new technique, the researcher thoroughly studied the traditional method at first and tried to find the area that could be beneficial in implementing the machine for the calculation of CLV.Hao (2009) Rathi and Ravi (2017) For this purpose, the Microsoft Access data was used. After the comprehensive study of every aspect of both the techniques, the Support Vector Machines, Multilayer Perceptron (MLP), Wavelet and Artificial Neural Networks (WNN ANN), Additive Regression was implemented in the calculation of CLV. AboElHamd et al. (2020) Khajvand et al. (2011) One of the researchers tried to approach the problem of the CLV with some theoretical approach with two different combinations of features. In the first approach, the Recency, Frequency, Monetary (RFM) was taken into consideration. In the next one, all the features like Recency, Frequency, Monetary, and Count Item (RFMC) were present. Customer segmentation was the base of the research, which is one of the critical aspects of CLV, which was achieved with the help of the K-means Clustering technique of ML. Vanderveld et al. (2016) Win and Bo (2020) One such research which was addressing a similar problem but in the E-commerce industry tries to calculate CLV to maintain the customer relationship. The methods like Random Forest and Ada boost were implemented. Based on initial res-

ults, the hyperparameter tuning of Random Forest was performed to meet the best results.

As mentioned above some attempts were also made to calculate the CLV using different machine learning models.Hao (2009) Rathi and Ravi (2017) Multilayer Perceptron (MLP) is one of such models that shows a good amount of accuracy in the calculation of CLV with a smaller number of error counts. On the other hand, CART techniques have also shown equally commendable performance for the CLV. The CART technique is based on the splitting of nodes technique. AboElHamd et al. (2020) Khajvand et al. (2011) K-means clustering was used in one of such research for CLV calculation. In this case, the value of K will impact the overall result of the model. The k value was considered as 4 for the experiment.Vanderveld et al. (2016) Win and Bo (2020) Another research with two models like Random Forest and Ada boost evidenced that the Random Forest had better accuracy when the hyperparameter tuning was performed as compared to the default tuning parameters. The evaluation values of Recall, F1 score, and Precision showed a rise in percentage, which is evidence of the model's good performance.

2.3 Limitations of the Literature in Fraud Detection and CLV

After going through the various research in this domain, the first and foremost observation is that very few numbers of research are found. For example, in calculating CLV, we can find some research, but they are mainly related to E-commerce and Gaming. As the data used for research will be of different sizes and nature than BFSI, this can be considered a primary limitation. It can also be observed from some of the research that the quantity of data used for training the model was the limited performance of the models can be doubtful. Even some of the models require more time for processing that is one of the factors of concern. Finding the best alternatives for the models is one of the needs of this domain. No one was tried to implement more than one problem on a single platform. The success of this research form a base for such a platform. The research will also develop a base platform to implement more solutions to the problems of BFSI and integrate it on the same platform. It will also open the possibilities for other technologies or techniques to be implemented. Research can have a significant contribution in providing a solution for the BSFI.

3 Methodology

In this section, the details of the methodology used in the research will be explored. It will also explain more details about each step that is present in the process of successfully implementing the methodology. The section will also try to describe the technical aspects of the steps of the methodology.

List of essential steps in the methodology:

- Factors involved in Data selection
- Data importing and pre-processing
- Model Selection Process and Data preparation for modeling

- Evaluation Method Selection

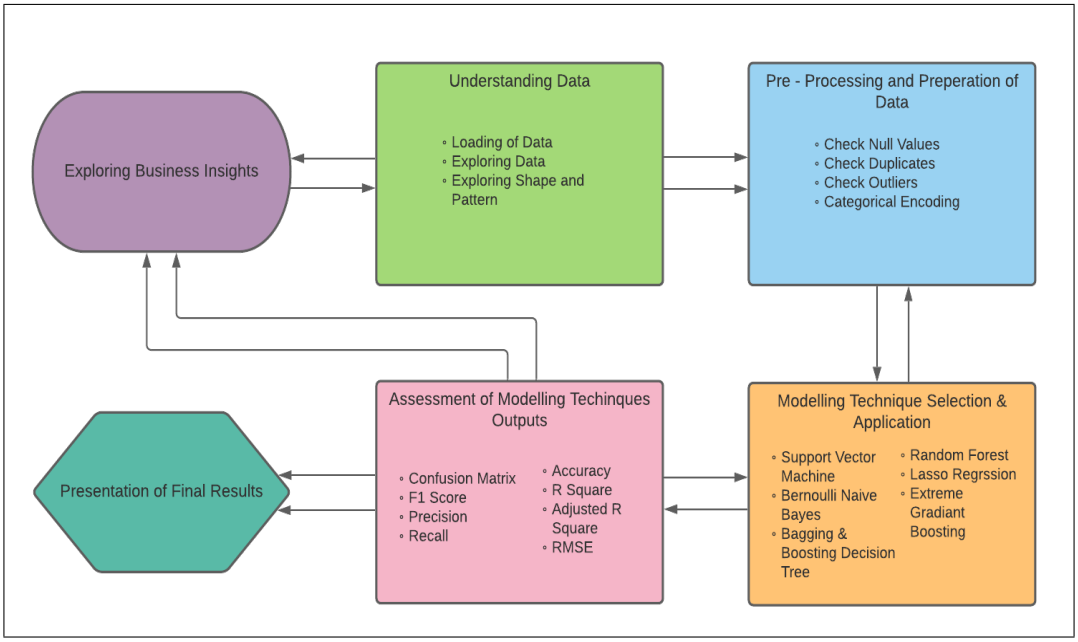


Figure 2: Research Methodology

3.1 Exploration of Business Insights

Before providing the solution to any real-world problem, it is essential to understand every aspect of that domain clearly. Every domain is different from the other, and getting the inside knowledge can help understand the problem and develop the solution according to it. According to this research, we need to understand the influential factors in fraud detection and customer lifetime value calculation. For the research, we tried to know about the nature of the business in BFSI and other factors related to it like the nature of data they use in their day-to-day business, the volume of data, sources of the data, etc. We also analyzed the various problems these domains are facing in business and found that the fraud detection and retention of customers are prominent ones. This helped us understand the business requirement and frame the research question or problem definition based on the problems.

3.2 Understanding the Details about the Data

In this phase, the first step is to gather the relevant data as per the requirement of the process followed for the research. Understanding the business insights in the above step will help decide the influential factors in addressing the research question and gathering the data according to it. The data gathered should be in a suitable form as it should get imported easily using the programming language used in the research. The data gathered can in any form like .csv, .xlsx, XML, JSON etc. In the research, the data used was made available in .csv files. After importing data next step is to understand the data and its insights—this is one of the steps in any machine learning project. Dataset used in fraud detection contains the clients' demographic, educational, and income-related data

with the dependent variable risk flag. For calculation of CLV, the dataset consists of demographic, educational, financial, vehicle, and policy details data were used in the experiment. In order to know more about the data, visualization was done in the form of graphs, plots, pie charts of different variables in the dataset with the help of different python visualization libraries, which can also help understand the pattern in the dataset.

3.3 Preparation of Data as per the Requirements

Data preparation involves various tasks executed on the raw data to convert into a desirable form. As the research will be following the process based on machine learning, it becomes essential to convert the raw data into suitable and acceptable forms by machine learning methods. Raw data contains various abnormalities and other factors that can impact the outcome of the machine learning techniques.

After loading data into the data frame, the data was further analyzed to understand the data structure, the numerical values, or the categories present in each column. The statistical values of each column were studied, and the dataset's summary was understood. Null values check was carried out on both datasets to find the number of null values present. In the same way, the duplicate record was also found for both datasets. One of the essential steps, finding the outliers points in the datasets that can impact the model's performance, was carried out. The columns like Employment Status, Vehicle Class, etc., had some outliers to be treated. The treatment for the null values, duplicated records, and outliers was performed. The next step was categorical encoding to convert the categorical variable into numerical form. It was observed that some of the columns required to be considered in the modeling process were found to be in categorical form. These columns are married, house ownership, car ownership, state, coverage, vehicle class, vehicle size, policy type was converted into the numerical form using categorical encoding. The technique used for the same was label encoding. Some of the columns in the dataset like Id, profession, city, which were significantly less correlated to the dataset, were removed.

At the end of this stage, we need to make two datasets, one for training and testing the model. We have divided the dataset into an 80:20 ratio. For training, 80 percent of the dataset was allocated. The remaining 20 percent will be used for testing the results of the models.

3.4 Model Selection and Application

This phase is known as the application phase, which is carried out in two steps the selection of the modeling techniques to be implemented and the application of the same. At first, the nature of the data and the research question is studied, and based on this, the target or the dependent variable is decided. The nature of this target variable will direct the process of modeling technique selection. In the case of our research, both the regression and classification types of target variables are present. For fraud detection, the target variable is risk flag which is of a classification nature, and the other dataset contains customer lifetime value as the target variable whose nature is regression type. This is one of the important factors in the model selection process.

In the case of fraud detection, the model selected for the application is Bernoulli Naive Bayes, Support Vector Machine, Bagging, and Boosting Decision Tree. One of the reasons for selecting the model is that that all of them are suitable for classification analysis and give the best results for classification problems. Other factors like linear function and working of SVM and Bagging and Boosting Decision Tree are ideal for dealing with extensive data as the BSFI work with massive data, making it the best choice of models for Fraud detection.

Whereas, for calculation of CLV models like Lasso Regression, Random Forest Regression, Extreme Gradient Boosting will be used. The automatic feature selection of Lasso regression, the capacity of Extreme Gradient boosting to handle null values, overfitting problems, and patterns in the dataset makes it suitable for the research. In addition to this, all three models are regression algorithms that efficiently predict numerical values. Here one of the points to be noted is that all the models used in research will be implemented for the first time to address this research question and not used in the previous research. In addition to this, most of the data in both datasets are numeric. Using the categorical encoding, even the categorical data is converted into numerical ones, making the application of all the six models suitable for these datasets. The model was first implemented without any hyper tuning as baseline models and then implemented with some hyperparameter tuning for improvement in the outcomes.

3.5 Modelling Output Assessment

Output assessment comments on the performance of the model. In this phase, the first step is to select the correct evaluation methods depending on the nature of the dependent variable of the problem statement. It is essential to select the correct evaluation method as we cannot depend only on model accuracy as it leads to inappropriate or wrong information about the model's performance in some cases. As the research consist of dependent variable of both regression and classification nature, we have selected the evaluation matrices suitable for both. For assessment of result of fraud detection, we have implemented confusion matrix, recall, precision, F1 score. On the other hand, R square, Adjusted R square, RMSE are applied to assess the outcome of calculation of CLV. Accuracy score is common assessment parameter for both types of models. The evaluation matrices are selected on the fact that they are suitable for that particularly for that type of problems and commonly used for the same. If the modeling results are as per business requirements, then the model is ready for deployment.

3.6 Presentation of Results

This is the concluding phrase that summarizes the overall output and success of the research. The various outcomes and data are studied, and business required data will be presented more visually. The results obtained in the evaluation phase will be presented in tabular form with the values of different valuation metrics values after the execution of each model. Finally, the ROC curve was plotted to show the relationship between specificity and sensitivity.

4 Design Specification

The design specification is one of the initial steps of product management, which describes the project's structure, provides detailed information about the various process to be carried out leading to the project, and has clear ideas about the objectives. In our research, we have designed a three-layer architecture to be followed to complete the project. Each layer has its importance, with more than one sub-process is involved in each of them.

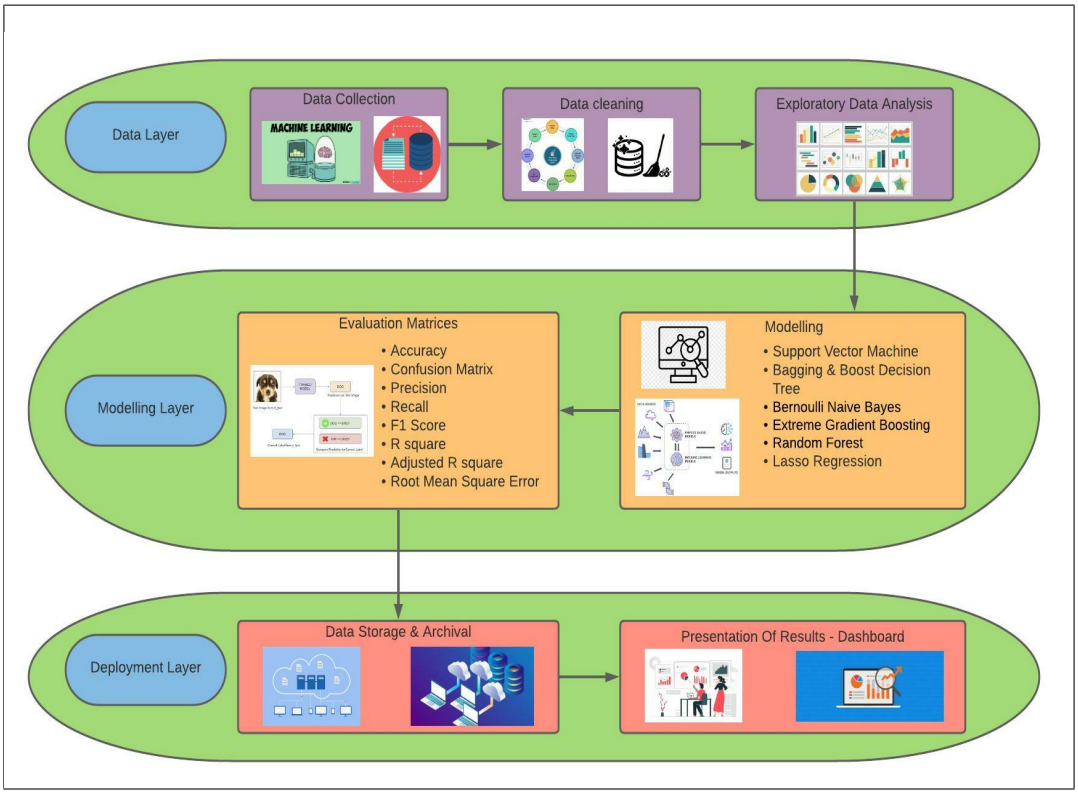


Figure 3: Research Design Specification

4.1 Data Transformation Layer

As the name suggests, the layer deals with the data, which is one of the essential parts of the whole research process. The layer is designed so that all the processes related to data will be covered in the phase. It consists of three significant processes collection of data for the research, treatment of the raw data collected, and analyze the data visually to know more about the data.

4.2 Modelling Layer

The layer mainly deals with the implementation part of the process. Two processes are carried out in the layer modeling evaluation. The modeling phase again consists of a two-part selection of the suitable model and then Implementation on the training dataset. The evaluation matrices are selected as per the nature of the research question and results obtained from it using it for each model.

4.2.1 Modeling Techniques

- **Bernoulli Naive Bayes**
  It is a machine learning algorithm widely used for the problem mainly of classification in nature.Vidanelage et al. (2019) The base of the model is the Bayes theorem that classifies the problems based on conditional probability. The nature of input data should be binary for the model.

- **Support Vector Machine**
  It is one of the models that show outstanding performance in any problem statement. It is suitable for both regression and classification problems. The theory of the model is based on identifying the group of data points and separating it. Among different types of SVM, we will be implementing SVM - Liner in our research.

- **Bagging and Boosting Decision Tree**
  These models come under the supervised category of machine learning algorithms that are well suitable for classification problems. It is based on dividing the data points node to the least possible data point. As our research is based on bank data, which is enormous in amount, these models can handle extensive data.

- **Lasso Regression**
  This is a regularization technique-based model of machine learning. It was witnessed that lasso regression has outperformed many regression models in performance. The L1 regularization-based model is known for its automatic feature selection quality, making it the research's best choice.

- **Random Forest Regressor**

  Similar to SVM, Random Forest is also suitable for classification and regression problems. Donkers et al. (2007) It works on the principle of decision there where the number of the decision tree is used in the prediction process, and at the end, the mean value of the final decision is set as out of the model. The model gives highly accurate results on the enormous data. It efficiently deals with the overfitting problems of the dataset.

- **Extreme Gradient Boosting**
  One of the machine learning models uses the theory of tree structure to identify regression values—this is one of the kinds of robust models that deals number of abnormalities present in the dataset.Chen (2018) The model is capable of handling null values and the patterns in the dataset. It consists of a cross-validation technique.

4.2.2 Evaluation Techniques

- **Accuracy**
  One of the primary evaluation techniques tells the number of correct predictions out of total predictions. The value is represented in percentage form. However, one cannot rely on the accuracy solely for the performance of the model. Wan (2021)

- **Confusion Matrix**
  It is one of the standards and widely used evaluation matrices for classification problems. Deng et al. (2021) It shows the comparison between actual and predicted

classes. It is in the form of a 2x2 (NxN) matrix. It also serves as a base for the calculation of other parameters of evaluation.

- **Precision**
  It is the calculation of true positive to total predicted positive values. The range is between 0 (Low) to 1 (High), which can be shown in percentages.

- **Recall**
  It is the calculation of true positive to total actual positive values. The range is between 0 (Low) to 1(High), which can be shown in percentages.

- **F1 Score**
  It is one of the best evaluation criteria that is proportional to both precision and recall.Bhowmik et al. (2021) It can also be called the harmonic mean of recall and precision. It is considered as one of the essential factors in modeling results. The range is between 0 (Low) to 1(High).

- **R square and Adjusted R square**
  It is an error calculation-based evaluation method that shows how fit the model is for the prediction. The model with high R values shows the best fitting and accuracy in performance. R square is based on a similar technique used with a more significant number of independent variables present.

- **Root Mean Square Error**
  It is the calculation of the standard deviation of error of the model. Ideally, the RMSE must be low for better performance of the model.

4.3 Deployment Layer

This layer deals with the process after the evaluation phase of the research. The layer consists of the two processes: data storage or archival and presentation of results.

4.3.1 Data Storage and Archival

Data is a critical element in any industry and its success. It plays a vital role in research, planning, knowing the trends in the market, and other activities. So, it becomes essential to store the data or archive it. The data used in the modeling process is stored in a suitable database platform like a cloud platform to be archived and accessed in the future whenever the business needs it.The data storage and archival is done by storing the data used in the process to a remote cloud platform, i.e., in this research, AWS RDS Service was used as a cloud platform to store data.

4.3.2 Dashboard

The data we have is in the raw form; it can be more informative if we present it in a visualized way. Python has several libraries that can present data in different visualizations like graphs, plots, charts, etc. Power BI is one of such tools that help to present data interactively and understandably. The insights of the data and the result thus achieved by the various models can be presented in visualization. In this research, we have taken the help of the ROC curve to present the result graphically for the sensitivity and specificity

information. The metrics like accuracy, precision, recall, F1 score, and confusion matrix were also shown. A function was developed to calculate all these values and display them.

4.4 Class Imbalanced Problem

While going through the exploratory data analysis in fraud detection, the target variable risk flag has highly imbalanced data. As the imbalanced data can be impactful and can create a bias in the modeling results, the imbalanced class treatment became essential to balance the dataset. We have implemented the under-sampling process to remove the unbalanced class from the data. As the dataset used has many records, under-sampling was suitable for the imbalanced class treatment.

4.5 Data Preparation

4.5.1 Categorical Encoding

This is the critical step of the machine learning model development process. The model used in the machine learning techniques accepts only numerical inputs. This shows the need to convert the categorical columns into numerical form before providing the input to the model. Among various available techniques for categorical encoding, we have choice label encoding for the implementation.

4.5.2 Splitting of Dataset into test and train sets

The data has gone through all the processes needed before modeling and is ready to be applied with a suitable model; the final step before modeling is to break the dataset into test and train sets. Importing the Sklearn library with the help of the (train test split) function, the dataset was divided into test and train sets in an 80:20 ratio.

5 Implementation

5.1 Programming language

The programming language implemented in this research is python. Python is an easy to be used and high-level programming language. The syntax of python is easy to understand and readable. It is effortless to integrate with other platforms and technologies with the help of python. It has a vast library collection, right from visualization to modeling.

5.2 Data Selection

Data used in the research was collected from the open-source website Kaggle. As we are approaching two research questions in this research, we have selected two datasets from different domains. One of the datasets was selected from the banking domain for fraud detection that contains 12 features related to the demographic data. Some of the data also comments on the customers' financial stability. It contains a massive amount of data of almost 2,52,000 rows in total. The other dataset is related to some insurance companies that deal with auto insurance. Dataset consists of 24 features; some of these are related to demographic data of the customer, Educational and financial status, details of the vehicles, policy types, and details related to it. The dataset available on the website

Kaggle is free to use. The website is specially designed for research, and its dataset can be utilized for the same. All the ethical concerns and data privacy norms have been adhered to in the research. No permission is required to use the dataset. The dataset used in the research is in .csv format.

5.3 Exploratory Data Analysis

It is one of the critical steps that help us understand the data better and show the critical points inside it. Several steps are carried out in the EDA process to study the data closely and make it perfect and ready for further processing. At first, the CSV file is loaded into a data frame with the help of the NumPy library package of python.

5.3.1 Outliers

Outliers are the unusual or unwanted data points present in the data set that can turn out to be highly impactful on the results of the models. There are several techniques to detect the outliers in the dataset. We have implemented the boxplot to detect the outliers in the dataset.

5.3.2 Data Visualization

This is plotting different dataset features against each other and understanding its insights in interactive and visualized form. Using different python visualization libraries and with the help of scatter plot, bar graph, pie plot, etc., some of each dataset's features were visualized.

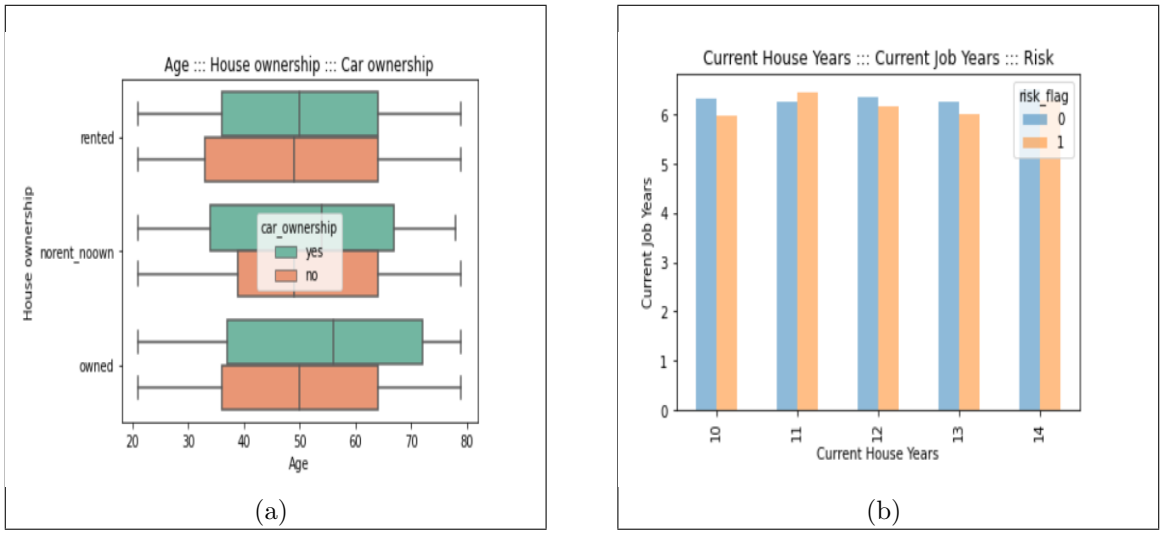
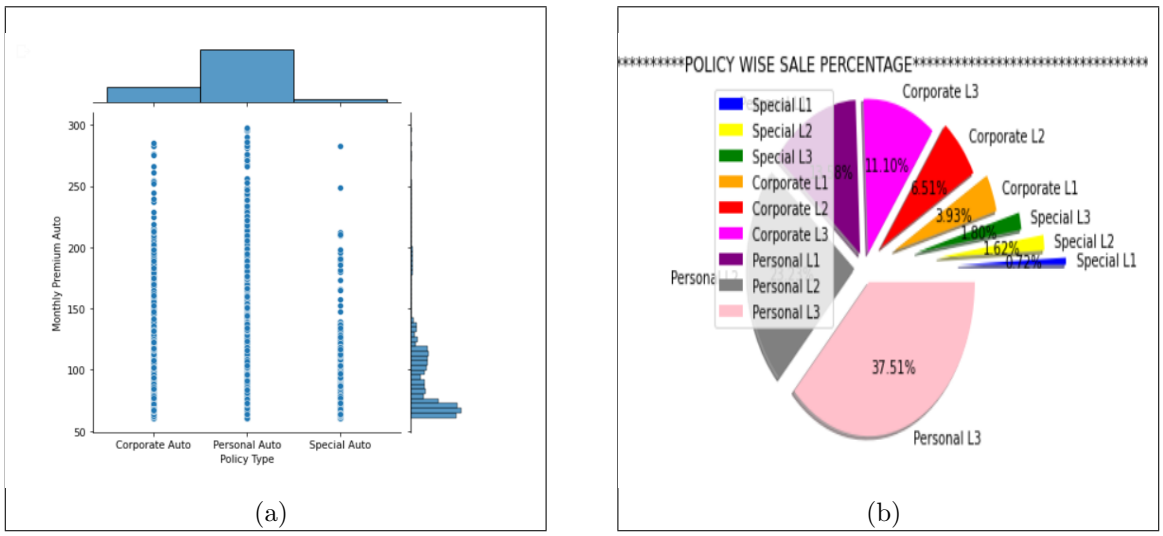


Figure 4

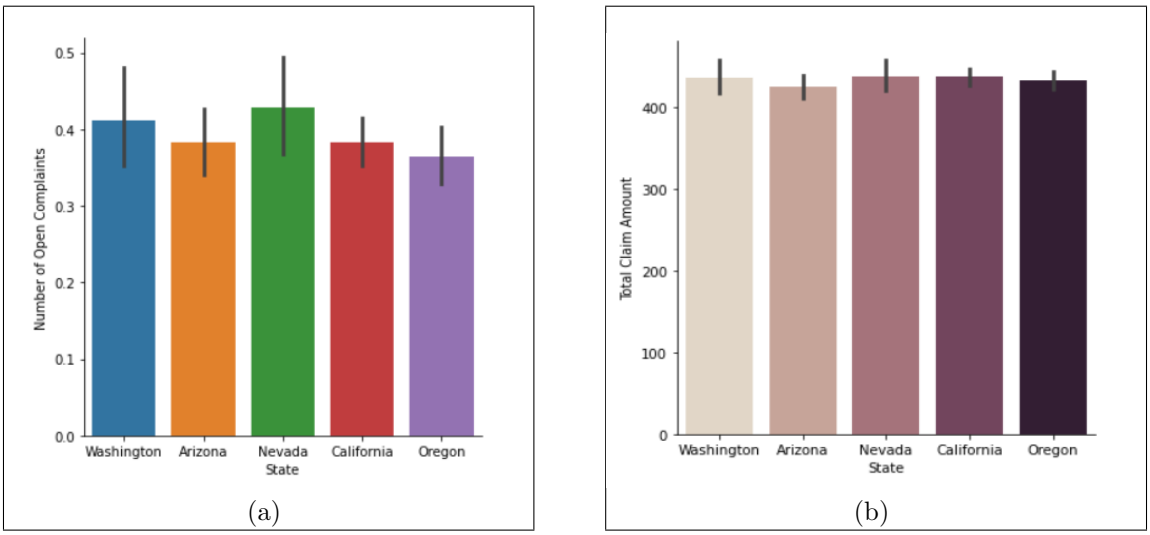
The figures 4, 5, 6 illustrate the data information in both Datasets. Figure 4(a) is the boxplot showing the comparison between the ownership of the cars and houses concerning the customer's age. The customer in the age group 35 to 65 has their cars and house. Figure 4(b) shows the risk of fraud based on experience, years in the job, and the number of years in the current house. People who have experience of 5 to 6 years and current residence years in the range of 10 to 14 years have the risk of fraud. Figure 5(a) describes

the types of policies and the premium amount. Figure 5(b) is the pie chart showing the different levels in policy types and percentage of sales. Personal L3 has the highest sales. Figures 6(a) and (b) show the state-wise total claim amount and number of complaints lodged.



(a) (b)

Figure 5



(a) (b)

Figure 6

5.4 Data Cleaning

5.4.1 Null Value Treatment

The check for null values is carried out for both the dataset with the help of IsNull() and sum() function of the python, and it was evidenced that there are no null or NA values present in either of the datasets. Both the datasets are clean from null values.

5.4.2 Duplicate Treatment

The check for duplicate values is also carried out using the duplicated() function. Only a few duplicated values were present in the dataset, which was calculated to be 2-4 percent of the total records in the datasets. As negligible numbers of duplicates are present in the dataset, deleting this record will not impact the integrity of the data or the overall results. So, these duplicated records were deleted from the dataset.

5.4.3 Outliers Treatment

After checking for outliers in the dataset, the dataset for fraud detection is free from outliers. However, another dataset for calculation of CLV contains outliers in columns like Employment Status, Vehicle Class, etc. The number of records affected with outliers was in considerable amount. The treatment of the outliers was done using replacing the values of these records with median values.

5.5 Hyperparameter Tuning

The hyperparameter tuning process is carried out when there is a scope of improvement in the model's performance. Tuning the input parameters of the model can improve the model performance significantly. In the research, the models have shown satisfactory accuracy, and values of the other evaluation matrices are too at an acceptable level. So, we found the need to hyper-tune the parameter for improvement in the performance of the models. Below is a list of hyper-tuned parameters of the models. This range of values is provided as input to these parameters. The best-fit hyperparameters value with the best performance was found using the GridSearchCV, RandomizedSearchCV of Sklearn library. Using this best combination of hyperparameter values, high performance was achieved.

List of hyperparameters used in the research:

- **Bernoulli Naive Bayes** - alpha, binarize, class prior, fit prior
- **Support Vector Machine** - C, gamma, kernel
- **Bagging and Boosting Decision** Tree - n estimators, n splits, n repeats, random state
- **Lasso Regression** - n iter, alpha
- **Random Forest Regression** - n estimators, random state
- **Extreme Gradient Boosting** - colsample bytree, learning rate, max depth, n estimators

6 Evaluation

A more detailed interpretation of the model results will be carried out in this section. The model implementation was carried out in stages for both datasets. Results of each of these stages must be analyzed and studied before reaching any conclusion over the research.

6.1 Fraud Detection without Tuning of Hyper Parameters

Modelling techniques	Accuracy	F1 Score	Precision	Recall
Support Vector Machine	0.51	0.56	0.51	0.63
Bernoulli Naive Bayes	0.52	0.57	0.52	0.63
Bagging Decision Tree	0.86	0.85	0.87	0.84
Boosting Decision Tree	0.56	0.54	0.56	0.52

After initial preprocessing of the data and carrying out each step of the machine learning process, all four models were applied to the prepared data as the base model. The results of the performance of these models will be analyzed in this section. We can see that the Bagging decision tree performed exceptionally well then all other models with an accuracy of 86 percent with an F1 score, precision, and recall in the same range. Other models like Support vector machine, Bernoulli naïve Bayes, and Boosting Decision Tree performed with accuracy in the range of 50 to 56 percent, which cannot be accepted as a good score.

6.2 Fraud Detection with Tuning of Hyper Parameters

Accuracy	Loss	F1 Score	Precision	Recall	Specificity	Sensitivity
Support Vector Machine	0.84	0.83	0.90	0.77		
Bernoulli Naive Bayes	0.50	0.57	0.52	0.63		
Bagging Decision Tree	0.85	0.88	0.87	0.84		
Boosting Decision Tree	0.83	0.84	0.82	0.85		

After applying baseline models with hyper tunned parameters, we have evidenced gradual improvement in the performance of the models. However, the accuracy of the Bagging Decision tree dropped by 1 percent. The improvement was also seen in the recall, precision, and F1 scores, along with the accuracy score. The F1, Precision and recall Score of all the models close to 1 indicates efficient modeling performance.

6.3 Calculation of CLV without Tuning of Hyper Parameters

Modelling techniques	Accuracy	R Square	Adjusted R square	RMSE
Extreme Gradient Boosting	0.61	0.6	0.6	1364.81
Random Forest	0.88	0.93	0.93	756.65
Lasso Regression	0.27	0.254	0.257	1878.05

Similarly, in this dataset, the models like Extreme gradient boosting, Random Forest regression, and Lasso regression were implemented with default parameters as baseline models. Random forest regression has shown efficiency in calculating CLV with 88 percent accuracy and good R and adjusted R square scores of 0.93. the result was followed by the Extreme gradient boosting, showing an accuracy of 61 percent. It was found that the performance of the Lasso regression was inferior, with an accuracy rate of 27 percent.

6.4 Calculation of CLV with Tuning of Hyper Parameters

Modelling techniques	Accuracy	R Square	Adjusted R square	RMSE
Extreme Gradient Boosting	0.89	0.98	0.98	717.64
Random Forest	0.88	0.93	0.935	756.62
Lasso Regression	0.27	0.258	0.256	1877.26

Even after performing the hyperparameter tuning of the models, the Random Forest regression performed similarly with accuracy and other results in the same range of 0.88. However, Extreme gradient boosting has shown commendable improvement in the performance from 61 percent to 89 percent accuracy rate. This result was also supported by the R square and adjusted R square value of 0.98, which showed significant improvement. No change was found in the lasso regression performance in this stage.

6.5 Discussion

The research was performed with two datasets to address the two-research question related to the BSFI domain. Looking at the nature of the research question, several models were selected to address each research question. On the first dataset set, four different models were applied to detect the fraud; we can say that the hyperparameter tuning has helped the models to improve the performance to reach the expectation level of the results. It can be seen from the tables present above that, Bagging Decision tree has outperformed other models showing extremely excellent results in both with and without hyperparameters stage, which was followed by the Support vector machine in the hyperparameter stage. In the case of the second research question, the practical results were evidenced in the case of a single model, i.e., Extreme gradient boosting. The lasso regression was unable to reach even satisfactory results in both the stages of model implementation. Comparing the overall results of all the three models in both cases, the Random Forest regression has given promising results in both stages.

While undergoing the research, some critical challenges are faced while dealing with the data processing task. One of the challenges was dealing with the class imbalanced challenge. As we had a dataset with massive data, we decided to go with the under-sampling process to overcome this problem and remove the imbalanced class in the dataset. This was one of the critical points in shaping the research and its success. Talking about the overall results of both the research question, all the models have shown significant achievement in the performance and addressing the research question.

7 Conclusion and Future Work

The research was implemented to develop a primary base to address the different problems faced by the Banking, Financial Services, and Insurance industries while their everyday business. The technological approach used to carry out the research followed the process of machine learning and its application. Out of some of the problems faced by BFSI, fraud detection, and customer lifetime value calculation were selected as a problem statement. Data used in the research has passed through several steps of the process studied and analyzed carefully before application as modeling input. The problem of an unbalanced dataset was handled using the under-sampling technique, which was one of the major

obstacles of the research. The outcomes of the models, which can be seen in the result tables of the evaluation section, confirm the models' excellent performance in addressing the research question. Support vector machine and bagging-boosting decision tree have shown good accuracy in fraud detection.

In contrast, random forest and extreme gradient boosting are those models that successfully calculate the customer lifetime value. After the initial stage of the application of models with default parameters, Hyperparameter tuning was also used to improve the performance of the models. This step of hyper-parameter tuning has shown commendable success and improvement in the performance of the machine learning models. As a concluding note for the research, it can be stated that the research successfully achieved its objective with a Support vector machine, bagging-boosting decision tree, random forest, and extreme gradient boosting are some of the models that outperformed and helped the research reach the goals.

In the future work the research will provide a base for developing a solution to other problems and integrating it into the same system. Machine learning can be used along with other technologies to develop a solution that may find better applications. Integration with a dashboard like Tableau and Power BI can be possible for the real-time reporting and presentation of data in an interactive way.

8 Acknowledgement

Throughout the master's and research journey, I have received lots of support and encouragement from family, friends, and mentors. This section of the report I would like to dedicate for them.

I want to thank my research project mentor, Prof. Martin Alain, for supporting and guiding me throughout the research. I am also thankful to my batchmates Mr.Piyush Kishore Dhande and Ms.Vrushali Atul Narkar-Surve, for their constant guidance. On these occasions, I cannot forget to thank my friends Ms.Manali Suhas Sawant (NY, USA), Mr.Aniket Ramchandra Baikar (INDIA), Ms.Ankita Akshay Yadav-Badiwale (INDIA), who always supported and encouraged me for the masters and research. At last, I would like to pay my gratitude to my family for believing in me and encouraging me to pursue a master's in data analytics.

References

AboElHamd, E., Shamma, H. M. and Saleh, M. (2020). Maximizing customer lifetime value using dynamic programming: Theoretical and practical implications, *Academy of Marketing Studies Journal* **24**.

Bhowmik, M., Sai Siri Chandana, T. and Rudra, B. (2021). Comparative study of machine learning algorithms for fraud detection in blockchain, pp. 539–541.

Chen, S. (2018). Estimating customer lifetime value using machine learning techniques, *Data Mining* .

Chen, Y. and Han, X. (2021). Catboost for fraud detection in financial transactions, pp. 176–179.

Chuang, H.-M. and Shen, C.-C. (2008). A study on the applications of data mining techniques to enhance customer lifetime value — based on the department store industry, **1**: 168–173.

Delecourt, S. and Guo, L. (2019). Building a robust mobile payment fraud detection system with adversarial examples, pp. 103–106.

Deng, W., Huang, Z., Zhang, J. and Xu, J. (2021). A data mining based system for transaction fraud detection, pp. 542–545.

Desirena, G., Diaz, A., Desirena, J., Moreno, I. and Garcia, D. (2019). Maximizing customer lifetime value using stacked neural networks: An insurance industry application, pp. 541–544.

Donkers, B., Verhoef, P. and Jong, M. (2007). Modeling clv: A test of competing models in the insurance industry, *Quantitative Marketing and Economics* **5**: 163–190.

Gyamfi, N. K. and Abdulai, J.-D. (2018). Bank fraud detection using support vector machine, pp. 37–41.

Hao, S. (2009). Appraisal of the customer lifetime value of commercial banks based on unascertained measurement, **2**: 399–402.

Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S. (2011). Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study, *Procedia Computer Science* **3**: 57–63. World Conference on Information Technology.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050910003868*

Mittal, S. and Tyagi, S. (2019). Performance evaluation of machine learning algorithms for credit card fraud detection, pp. 320–324.

Rathi, T. and Ravi, V. (2017). Customer lifetime value measurement using machine learning techniques.

Rawte, V. and Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques, pp. 1–5.

Roy, R. and George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques, pp. 1–6.

Shivanna, A., Ray, S., Alshouiliy, K. and Agrawal, D. P. (2020). Detection of fraudulence in credit card transactions using machine learning on azure ml, pp. 0268–0273.

Sifa, R., Runge, J., Bauckhage, C. and Klapper, D. (2018). Customer lifetime value prediction in non-contractual freemium settings: Chasing high-value users using deep neural networks and smote.

Vanderveld, A., Pandey, A., Han, A. and Parekh, R. (2016). An engagement-based customer lifetime value system for e-commerce, p. 293–302.
URL: *https://doi.org/10.1145/2939672.2939693*

Vidanelage, H. M. M. H., Tasnavijitvong, T., Suwimonsatein, P. and Meesad, P. (2019). Study on machine learning techniques with conventional tools for payment fraud detection, pp. 1–5.

Wan, F. (2021). Xgboost based supply chain fraud detection model, pp. 355–358.

Win, T. T. and Bo, K. S. (2020). Predicting customer class using customer lifetime value with random forest algorithm, pp. 236–241.

Yao, J., Zhang, J. and Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods, pp. 57–61.