International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Machine learning practices in accounting and auditing

Chetanpal Singh [1,*], Rahul Thakkar [2], Rashikala Weerawarna [1] and Vimal B Patel [3]

[1] *Faculty of business, design and it, Holmesglen Institute Chadstone Campus Melbourne Australia.*
[2] *Faculty of ICT, Victorian Institute of Technology (VIT), Melbourne Australia.*
[3] *College of Agriculture, Navsari Agricultural University, Waghai, Gujarat, India.*

International Journal of Science and Research Archive, 2023, 10(01), 131–162

Publication history: Received on 24 July 2023; revised on 03 September 2023; accepted on 06 September 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.10.1.0720

Abstract

In the current technological era, Machine Learning applications are becoming popular every day. This research paper provides information about the effectiveness of the ML technique in accounting as well as the auditing process. To get proper results, different ML libraries are utilized, concluding seaborn, matplotlib, NumPy and so on. In the introduction section, the research purpose and this research objectives have been developed, through which the entire research process will be developed. "Logistic Regression Machine Learning Model" is built. Literary sources have been analyzed in the literature review section, through which it can be easy to gain different perceptions based on the research context. The effectiveness of different types of machine learning algorithms in accounting and auditing has been evaluated properly. To develop a knowledge level, it is important to increase proper attention, and this research paper will provide proper information based on ML effectiveness.

Keywords: Machine Learning; Algorithm; Accounting; Auditing; Artificial neural network

1. Introduction

Research firm Gartner Inc. suggests that to keep ahead of the competition, businesses should analyse how crucial technologies are affecting their operations. Companies risk falling behind if these technologies aren't investigated. Machine learning is an example of a strategic technology that uses artificial intelligence. Thanks to advancements in machine learning, many jobs that were once reserved for people can now be done quickly and correctly by computers.

Machine learning technologies have been developed to help systems provide the impression that they can understand, learn, predict, and adapt. Unlike conventional rule-based algorithms, they can function with minimal input from a human operator [1]. Therefore, businesses may gain increased productivity and accuracy and significant cost savings by using machine learning technologies.

The influence of machine learning can be seen in many spheres of modern life, and auditing is no exception. Auditing efficiency, error rates, risk, and cost can all be greatly improved by utilizing machine learning. There has been a recent shift toward using machine learning in auditing. Since constructing machine learning systems requires significant data science, most organizations will instead purchase pre-built machine learning applications. Although complex, machine learning technology has featured that management must understand to ensure proper implementation and alignment with intended business goals.

One department that may use some help from machine learning is accounting, which is part of the finance department. PwC acknowledges the importance of advanced themes on machine learning in undergraduate accounting programs

* Corresponding author: Chetanpal Singh

when considering the training delivered to professional accountants, contend that machine learning is a crucial branch of AI with significant potential for use in accounting, adding that it is one of the AI skills recommended by PwC.

1.1. Research Focus

Accounting users can benefit from using machine learning methods in making decisions. There are many different machine learning methods, but they may be broken down into two broad categories: predictive and explanatory strategies. The method used will be based on the circumstances. For instance, predictive techniques can make predictions based on patterns learned by the machine learning model from the data, even though the user needs to be provided with an explanation or visibility into these patterns. Complexity increases with predictive machine-learning techniques instead of explanatory machine-learning approaches.

There is a need for machine learning research in accounting since accounting decision-makers can't currently tell the difference between explanatory and predictive machine learning technologies when determining which option to utilize and rely on [2]. Promoting the use of machine learning technologies in accounting and auditing and stimulating the need for research on these topics can be accomplished by helping accounting decision makers gain a better understanding of these technologies and the difficulties that must be taken into account when adopting them.

The relevance of human comprehension and straightforward models was emphasized. Users may make a more informed choice about the best technology for their purposes if they have a firm grasp on the possibilities and drawbacks of the various options. It will also help them understand the challenges they may face and the precautions they must take before implementing these technologies.

Aim

This study aims to help readers understand how machine learning technology might be applied in accounting and auditing.

Objective

Three research questions were developed to help get us closer to our goal. Then, specific objectives were established for each research question.

- Which machine learning techniques can be integrated into the present accounting system?
- What are the potential drawbacks and advantages of using machine learning technology, and how does it work?
- Thirdly, what factors should be considered, and what procedures should be followed before applying machine learning tools?

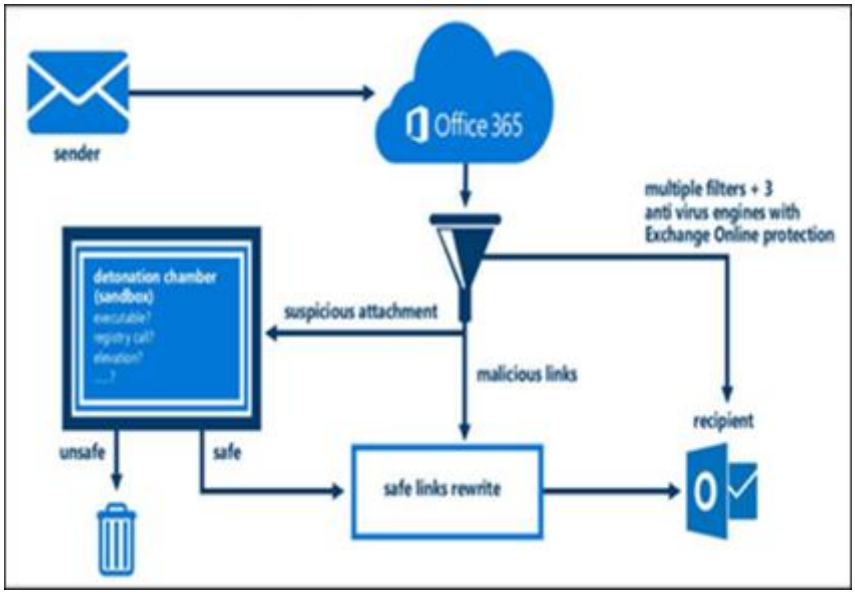
Objectives

- To explain the various accounting procedures and their constituent parts.
- Accounting process jobs that can benefit from machine learning technology must be located.
- To evaluate the potential drawbacks of machine learning methods and the advantages they may offer.
- To determine the actions to take when deploying machine learning technology to align with the accounting process's objectives.

1.2. Significance

Innovation and collaboration can help create maximum outcomes from security and other aspects. Moreover, Microsoft Office Outlook and Exchange Mail have provided a secure connection that may help increase the security of organizational system management. Office 365 and Google Workspace are widely used in collaborative business environments. Thus, proper intrusion in different organizational setups can help bring out higher productivity. Microsoft 365 has more than two hundred million users [4]. However, it does not mean that this tool is the best tool to be used in an organizational context.

Different secured email tools exist, such as Webmail, Virtu and Tanota. The company can help in getting better access through the organizational framework. Many tutorials are there and can be used to determine maximum benefits in the business context. Moreover, organizations must ensure that the users follow proper security guidelines to prevent intrusion. If there is a malware attack in the system, it will slow down, and it can also decrease productivity. Apart from that, users need to create proper analyses that may help make maximum security in the management channels. It is also necessary to train people to maximize the current situational analysis through adequate data management in different devices.



Source: [2]

Figure 1 Exchange Online ATP

1.3. Research Motivation

The auditor's opinion on whether the financial statements are prepared in all material respects by an applicable financial reporting framework is intended to increase intended users' confidence in the financial statements. There is a significant amount of skill and judgment involved in auditing. It also covers mundane but necessary actions that need to be done repeatedly and takes up too much time and which can be automated with the help of AI and ML. An audit of the notes and disclosures attached to a company's financial statement is the subject of this study. Several factors affect customer satisfaction with new information technology solutions, as identified [3]. The user's familiarity with the system, the product's perceived value, and the degree to which the user's needs and wants are met all play a role.

Understanding the benefits and drawbacks of machine learning methods in business can help users decide which methods to employ. Understanding the technology, its benefits, and uses, as well as the challenges to address and how the user can respond to the recognized dangers, is crucial for users who hope to harness the advantages of machine learning in their accounting operations.

This research will address the need for research into strategic technologies, notably machine learning, and the need to help users comprehend the technology [4]. Users will be more likely to consider using machine learning to solve problems if they have been allowed to understand the technology better.

2. Literature review

2.1. Machine Learning

Machine learning is an AI subfield that streamlines the process of creating analytical models. These models are used in machine learning for data analysis, pattern recognition, and prediction. Learning is automatic and continuous because the machines are programmed to utilize an iterative method to learn from the studied data; as the device is exposed to more and more data, it recognizes more and more robust patterns and uses the feedback to adjust its actions accordingly. Machine learning seeks to determine the combination of mathematical equations that best predicts an event instead of the probability theory and probability distributions upon which statistical analysis is founded [5]. Therefore, many issues involving classification, linear regression, and cluster analysis are amenable to machine learning.

Machine learning can be categorized into supervised and unsupervised. When trying to predict future occurrences based on historical data, such as which customers are most likely to default on their debts, supervised learning is a technique that is utilized. When there are no labels on the output variables, an unsupervised learning approach is utilized to "learn" the data patterns without being "told" what the presumed response is. Applying unsupervised

learning techniques (such as cluster analysis) on transactional data may prove useful in risk assessment by revealing previously overlooked threats. In semi-supervised learning, labelled and unlabelled final product examples are used.

When it comes to the future of AI and machine learning, artificial neural networks are crucial. Information systems can be organized into single-layer or multi-layer neural networks. Artificial neural networks, or CNNs, use nodes to form connections, just like the neurons in a real brain. Deep learning blends machine computation with neural network connectivity patterns for complicated interactions, such as those involved in medical diagnosis or location recognition.

Overfitting is a potential pitfall of sophisticated machine learning, in which the computer incorrectly interprets the data as representing patterns in the real world. This can occur, for instance, if the model is evaluated using the same data that was compiled [7]. Machines are vulnerable to "forgetting" that statistically large correlations between variables do not necessarily imply a causal relationship due to overfitting. Machine learning still relies heavily on human knowledge and judgment because of the possibility of errors in data output. Users must comprehend the nature of the inputs, the data, the machine's processing, and the result.

The supplied historical data's quality greatly affects machine learning's prediction reliability. The results may be invalid if new and unexpected occurrences are not accounted for or given the proper weight. This means that biases from humans can impact how machine learning is applied. The data sets used to train the AI, the methodologies applied to that training, and the interpretation of the results are all susceptible to bias.

2.2. Current and Potential Future Uses of Machine Learning in Accounting and Auditing

While machine learning's current capabilities are limited, it excels at routine jobs. Because of the vast amount of data involved and the complexity of the activities that need to be completed, machine learning has the potential to increase audit efficiency and quality [6]. Auditor productivity should increase due to the machine-based performance of redundant activities, allowing for more time for review and analysis and a sharper focus on high-risk areas.

2.2.1. Current applications

Companies in the auditing industry are actively investigating the potential of machine learning in audits. Argus, a machine learning technology, is used, and it can interpret legal documents like leases, derivatives contracts, and sales contracts. Key contract terms, trends, and outliers can all be pinpointed with the help of Argus's built-in algorithms. Then, auditors can zero in on the most crucial aspects of documents for analysis. It's easy to picture a machine scanning a lease agreement, picking out the main words, and figuring out whether it's a capital or operating lease. Machine learning methods, if constructed properly, can also detect anomalies and trends, such as atypical leases requiring substantial judgment (e.g., those with unusual asset retirement responsibilities). By zeroing in on the greatest risk contracts, auditors can save time and effort while still producing a high-quality report.

PricewaterhouseCoopers (PWC) also uses machine learning techniques like Halo. Journal entries containing suspicious keywords, posts from untrusted sources, or an abnormally large number of journal entries posted just below approved limitations are only some of the issues that Halo's analysis can uncover. By submitting all journal entries to testing and focusing solely on the outliers with the highest risk, auditors can boost the speed and quality of testing operations utilizing Halo.

2.2.2. Future Applications

Additional applications of machine learning in financial statement audits, notably the risk assessment process, are already being researched by CPA companies and researchers. Ting Sun and Miklos Vasarhelyi suggest that executive fraud interviews mandated by auditing standards could benefit from machine learning technology like speech recognition. When respondents use equivocal words like "sort of" or "maybe," the program will know to investigate the response further. Technology that analyses speech could also detect lengthy response pauses, another possible indicator of covert activity.

It can be difficult for a human to detect certain behavioural patterns consistently and in real-time, even though many accounting firms train their workers to conduct fraud interviews[10]. Auditors may be supplemented by voice and facial recognition technology in fraud interviews, alerting them when higher-risk responses need additional scrutiny.

CPA firms of the future may be able to spot patterns that would otherwise go unnoticed if not for machine learning technologies [8]. A restaurant, for instance, might use satellite images of parking lots, data from point-of-sale systems

detailing the number of guests, and employee schedules to show a direct connection between peak business hours, the number of guests, and employee salaries.

Using machine learning methods, auditors may benefit from this information and better comprehend the processes behind the numbers. Turns per hour, average income per turn, and outside deliveries are examples of traditional metrics with which the machine learning algorithm can uncover discrepancies; auditors would need to look into these discrepancies [9]. Customers' actions also help audit teams gain perspective as they independently try to forecast the outcomes of analyses using supervised learning methods. Risks that were previously unknown could be uncovered using unsupervised learning methods.

Non-traditional data linkages are also getting some attention from academics. Kyunghee Yoon conducted a study to determine how climate plays a role in business. In particular, it was anticipated that less business would be done due to bad weather keeping people from shopping. The results showed that while combining data from stores with similar characteristics yielded more accurate predictions, using weather variables in the peer store data enhanced predictability even further. This has crucial implications for future audits, particularly risk assessment procedures, where atypical data linkages may play an increasingly significant role.

2.3. Challenges faced by Auditors

For machine learning technologies to realize their full potential, audit firms, and regulators must overcome several obstacles. It might be challenging to collect valuable information from customers and other external sources. In most cases, auditors are restricted from accessing massive volumes of data stored by companies like Google or Facebook due to statutory and regulatory constraints. Ethical and client confidentiality concerns may also restrict auditors' access to high-quality data for use in developing training sets.

To trust the results when relevant and meaningful data is available, auditors need to learn about and exercise the internal controls over data integrity and verify the fullness and accuracy of the input data. The accuracy of machine learning will heavily depend on the security and integrity of the data utilized in training [11]. To ensure the integrity of their audits, auditors will need to consult with cyber security professionals to provide their client's information is safe from prying eyes.

If unconventional data is used more frequently, auditing documentation standards must adapt. Documentation from auditors ought to cover more than just a rundown of procedures and an explanation of how and why samples were selected to be statistically significant. With access to more data, auditors could better understand how external factors affect a company's operations. Despite the possibility of greater insight, auditors should cautiously approach machine learning conclusions. The patterns it identifies may need to be corrected or made more sense [12]. Therefore, future auditors will likely collaborate with data scientists to comprehend the algorithms, as modern auditors collaborate with IT, actuarial, and valuation specialists. For audits to be free of spurious correlations, auditors need to understand and effectively document why causal relationships are or are not present, which requires an appreciation of the output's reasonableness.

Auditors will still require knowledge of the external business environment and societal influences, as well as the specific company and industry, due to the limitations of machine pattern-finding. Facebook user accounts may be the strongest predictor of income, so they should be given more weight in the company's internal algorithm. Account integrity and the presence of "bots" may not be discernible by machines without human judgment, leading auditors to inaccurate findings [13]. To draw the right conclusion from the output, auditors must comprehend and confirm the completeness and accuracy of the input data. In addition, auditor intuition will likely remain a valuable source of knowledge because there will always be potential blind spots when analysing empirical evidence.

Human biases introduced into the system will also play a part that auditors will need to comprehend and account for. Several examples are availability, confirmation, overconfidence, and anchoring biases. Auditors should be aware of the potential of confirmation bias, which would express in overweighting or using only input data that supports pre-existing opinions, as well as the risk of availability bias, which would show in using the most easily accessible information to identify risks or draw conclusions. The auditors who rely too heavily on machine outcomes and need to investigate the appropriateness of the input data and weighting of the machine learning results may be subject to a new type of over-confidence bias that arises in machine learning. As audit customers create and improve their machine learning technologies and auditors begin to "anchor" their input data with the client's data instead of analysing alternative possibilities and contradicting evidence, the risk of anchoring biases may increase.

Given the potential for input data ambiguity and interpretable results, auditing standards for machine learning technologies may need to evolve. For instance, auditors are expected to make certain assumptions when executing analytical operations. Still, machine learning's main advantage is its capability to uncover out-of-the-ordinary connections.

2.4. Theoretical Framework

Numerous articles have been published on applying machine learning strategies to auditing and accounting processes. Comparing machine learning approaches for auditors' going concern reporting was the primary subject. The study found that artificial neural network models outperformed expert systems and multiple discriminant analyses in predicting the nature of a going concern audit report.

The concern has been laid into the feasibility of using machine learning methods to detect companies that reported false financial results. After running multiple experiments with a selection of representative learning algorithms, they concluded that the proposed stacking variant methodology outperformed the competition.

Financial statement fraud risk assessment using machine learning techniques was researched. Non-financial risk variables and a rule-based system yielded decreased error rates, as evidenced by experimental findings gained by installing a back-propagation neural network, Support vector machine, logistic regression, and the C5.0 decision tree were used in this analysis [15]. The authors concluded that the proposed method might lessen stakeholders' financial risks.

To achieve those aims, k-medians clustering, a machine learning-based peer selection method was used on companies' important financial statistics. K-medians clustering was used to analyse the data, which included 598 bankruptcies and 48,536 firm-year observations from companies that did not file for bankruptcy [14]. According to the findings, the models were improved by using a machine learning technique that took into account data from K-medians clustering of similar businesses.

Employed machine learning algorithms (a dynamic anomaly detection method) is being considered to rate the trustworthiness of each company's quarterly financial report to detect any discrepancies in the financial statements of Vietnamese-listed enterprises. The results demonstrated the model's ability to assign a creditworthiness ranking to quarterly financial reports. According to the adopted model, while most Vietnamese listed enterprises have credible financial statements, around a quarter have highly dubious claims.

The study used an actual data set taken from the ERP system of a major player in the chemical industry. The findings provided the intriguing potential for integrating tax compliance standards into IT systems, as machine learning algorithms could spot abnormalities that would have led to compliance infractions.

It is being suggested a fraud prediction model that uses raw financial data retrieved from financial statements instead of calculated financial ratios. The robustness and generalization capacity of the model was enhanced by employing the ensemble learning technique, a potent machine learning tool that integrates the predictions of multiple base estimators. The findings supported the use of an ensemble learning approach.

The work of to identify two related applications of automated text analysis on annual reports: the identification of financial fraud and the modelling of financial hardship. Hajek and Henriques use Bayesian Belief Networks on a rich feature set for fraud detection; this includes data from financial statements, yearly reports, and analyst projections. In contrast, the methods presented are based solely on the text, using linguistic credibility analysis.

Used the sentiment analysis for financial risk prediction. The information quality is evaluated by analyzing connections between risk factor disclosures and economic sentiment terms. Similarly, Matin et al. use a convolutional recurrent neural network on text segments from annual reports to evaluate the likelihood of company distress.

In view of these applications for text-based machine learning models, the continued reliance on manual analysis by auditors is puzzling [17]. This work is the first of its kind to illustrate how machine learning and natural language processing may be efficiently applied to auditing processes, thereby improving both their efficiency and their quality.

The auditing sector is not immune to the potential disruption posed by machine learning over the next few years. Since machine learning allows auditors to "avoid the trade-off between speed and quality," Deloitte's chief innovation officer, Jon Raphael, believes it will fundamentally alter the auditing process. Machine learning algorithms can allow businesses

to examine the entire population for outliers instead of just relying on representative sampling methods. When audit teams can access the whole data population, they can conduct their tests with greater precision and focus. On top of that, machine learning algorithms can "learn" from auditors' judgments on particular items and then apply that learning to other things that share those characteristics.

The use of machine learning in the auditing process is still in its early stages. Smaller CPA companies stand to gain as the feasibility of machine learning systems develops with changes to auditing requirements and professional training initiatives. For machine learning tools to realize their full potential, this article explains how they function, discusses their current and potential impact on the auditing profession, and outlines some of the issues that auditors face.

2.5. Machine Learning Algorithms Applications in Accounting and Auditing

Various factors influence the selection of an appropriate auditor. This suggests that it is feasible to develop models that can predict the type of auditor without human intervention. Data mining offers several classification techniques that can be employed to construct predictive models. Unlike fraud detection, bankruptcy prediction, or prediction of qualified opinions, data mining techniques have not yet been applied for the purpose of predicting the type of external auditor. This study uses three well-known classification methods from data mining to forecast the type of auditor [18]. The three methods used are KNN, ANN, and Support Vector Machines. These methods are compared in terms of their ability to predict the category of unknown observations. The study identifies significant factors that strongly influence the decision to select an auditor. This research has practical implications for internal and external auditors, company decision-makers, investors, and researchers. Additionally, it can assist in anticipating the most probable outcome for the selection of an auditor.

Data mining has been extensively used in the financial industry for various purposes such as bankruptcy prediction, credit card approval, loan decision, money-laundering detection, and stock analysis. However, its application in detecting financial statement fraud is limited. This research aims to predict the occurrence of financial statement fraud in financial statements as accurately as possible.

- In order to accomplish this goal, Ravisankar utilized six data mining methods, comprising Group Methodology of Data Handling (GMDH), Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Logistic Regression (LR), Genetic Programming (GP), and Probabilistic Neural Network (PNN). The data set used for this analysis was obtained from 202 Chinese financial institutions [19]. The researchers found that the Probabilistic Neural Network (PNN) performed the best without feature selection, while Multilayer Feed Forward Neural Network (MLFF) and PNN outperformed the others with feature selection and had nearly equal accuracies of Sensitivity (65.6%), Specificity (74.6%).
- Koskivaara has proposed the utilization of Neural Network (NN) based support systems for detecting material errors and managing fraud in various application areas. Koskivaara investigated the impact of several preprocessing models to enhance NN's forecast capability when auditing financial accounts [21]. The resulting Sensitivity and Specificity values were 55.6% and 75.6%, respectively.
- Sohl and Venkatachalam employed a back-propagation neural network to anticipate financial statement fraud and corresponding Sensitivity and Specificity results of 65.6% and 81.6%.
- Kirkos employed three data mining classification strategies, namely Decision Trees, Neural Networks, and Bayesian Belief Networks, and achieved Sensitivity and Specificity values of 79.9% and 85.58%, respectively.
- Zhou and Kapoor conducted a study on the effectiveness and limitations of different data processing techniques, including regression, call trees, neural networks, and theorem networks, in detecting financial statement fraud [20]. A response surface model with domain data was utilized to create a self-adjusting system that attained a fraud sensitivity of 45.6% and a specificity of 69.6%. This study aimed to determine the most effective technique for detecting financial statement fraud while also considering the limitations of each technique.
- Belinna et al. also investigated the effectiveness of the classification and regression tree (CART) technique in identifying and detecting financial statement fraud [22]. According to their findings, the technique called CART demonstrated a high level of effectiveness in distinguishing between fraudulent and non-fraudulent financial statements, with a sensitivity of 78.5% and specificity of 85.6%.
- Juszczak et al. studied the performances of different classification techniques in both supervised and semi-supervised settings for detecting financial statement fraud. They achieved a sensitivity of 85.6% and a specificity of 80.6% [23]. The goal of this study was to identify the most effective technique for detecting financial statement fraud while considering the different settings in which the technique may be used.
- In their study, Zhou and Kapoor utilized four data mining techniques: regression, decision trees, neural networks, and Bayesian networks, to assess their effectiveness and limitations in detecting financial statement

fraud [24]. They employed a self-adaptive framework that incorporated a response surface model and domain data to achieve a financial statement fraud sensitivity of 63.6% and specificity of 84.6%.

2.6. Prediction Models

In the domain of auditing, predicting audit opinions is a crucial task to ensure accurate financial reporting and regulatory compliance. To this end, various data mining techniques have been employed to develop effective prediction models that can assist auditors in making informed decisions. This section depicts three widely used data mining techniques, namely Support Vector Machines (SVM), Artificial Neural Networks (ANN), and K-Nearest Neighbor (KNN), for predicting audit opinions. Each of these techniques offers unique advantages and can be adapted to different types of data and contexts. By understanding the strengths and limitations of each approach, auditors can make better use of predictive analytics in their work and enhance the quality of their auditing outcomes.

2.6.1. Support Vector Machines

Support Vector Machines (SVM) is a popular method for classification modeling, which is widely used in data analysis for both regression and classification tasks. SVM is based on statistical learning theory, making it a reliable tool for examining data mathematically. It uses a linear optimal hyperplane to split the data into two classes with the maximum margin between the hyperplane and the nearest point. SVM maps input vectors into a high-dimensional characteristic distance through nonlinear transformation, making it effective in solving complex problems.

One of the main advantages of SVM is that it is optimal, unique, and universal. This is because the SVM solution is achieved by resolving the linearly constrained quadratic issues [25]. Another advantage is that it is based on the principle of structural risk reduction, which reduces the upper bound of the actual risk. This sets SVM apart from other classifiers that only reduce the empirical risk. Due to these advantages, SVM has been widely used in many areas, and its application and theory have been extensively studied.

SVM has proven to be successful in various financial applications, including time series prediction, insurance fraud detection, and credit rating. Previous studies have shown that SVM's performance in these areas is comparable to, or even better than, other traditional classifiers such as logistic regression and discriminate analysis.

2.6.2. Artificial Neural Networks

Artificial Neural Networks (ANN) is an algorithmic mechanism based on the human brain system capable of modeling non-linear statistical data in massive volumes. It is composed of simple handling units called neurons that work in vast parallel dispensers, as illustrated in Figure 2. Each communication is connected with a numeral value summon weight, and the output unit of ANN picks the weighted totality of the outputs from units in a former stratum [30]. The information fed from the input nodes travels through a concealed layer and reaches the output nodes. This hierarchical arrangement can be flexibly fine-tuned by altering the weights in the previous layers to improve the ranking model's performance.

One of the advantages of ANN is that it can be used to detect underlying functional relevance between inputs and outputs and can perform tasks like classification, control, pattern recognition, modeling, evaluation, and prediction [26]. ANN is particularly useful when employed to calculate accurate solutions for noisy, complex, irrelevant, or partial data. The process involves utilizing interplays among multiple variables that exhibit substantial correlation, often deemed as nonlinear, poorly related, and intricate to depict using statistical techniques.

2.6.3. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) technique is a widely-used algorithm in the field of financial and accounting studies. It works by analyzing objects with N characteristics in N distance space, where each object is considered as one spot. KNN then introduces a resemblance metric for each object and classifies new items by comparing them with existing ones, utilizing distance measurement. The algorithm calculates the distance between every item in a sampling set and determines the closest K neighbors to the unknown observation. The K cases are then used to classify the new observation by specifying it to the most popular class.

KNN is a significant classifier algorithm employed in fraud detection. The researchers assessed the effectiveness of the KNN algorithm in strengthening reviewers' views compared to methods that use logical and discriminant analysis [27]. The study used a sample of UK firms' financial statements and compared discriminate and logit analyses with KNN models. The researchers determined that KNN is more effective and efficient in terms of average classification accuracy.

The study collected two data sets that included financial data, such as financial ratios, annual statements like financial statements, income statements, cash flow statements, and equity statements, and audit opinions from all industries in British and Irish companies using the Financial Analysis Made Easy Database (FAME) software. The software contains data for about 11 million firms in the studied countries, including up to ten years of companies' data, financial information, and descriptive information about firms.

The researchers chose 49 independent variables consisting of financial and non-financial variables, and the dependent variable is audit opinion, which includes qualified (firms without credibility in the financial statement) and unqualified (companies without any fraud or misstatements). The study used two data sets, one for one year and the other for five years. An outline of the corporations analyzed in this research is provided in Table I.

2.7. Comparing Techniques Criteria

The present investigation evaluated the effectiveness of the developed models by utilizing Type I and Type II error analysis along with accuracy rate. These evaluation metrics have been widely adopted in accounting and finance research for measuring model performance [29]. The accuracy (1), Type I error (2), and Type II error (3) were calculated using the following equations:

$$\text{Average Accuracy} = TP+TN/\ (TP+FN+TN+FP) \quad (1)$$

$$\text{Type I error} = FP/\ (TN+FP) \quad (2)$$

$$\text{Type II error} = FN/\ (TP+FN) \quad (3)$$

Hence, the terms true and false denote the accuracy of the assigned classifications. True negative indicates accurately classified failed companies, whereas true positive ratio represents accurately classified healthy companies. The classification test results are presented in Table II.

3. Result and findings

3.1. Dataset Information

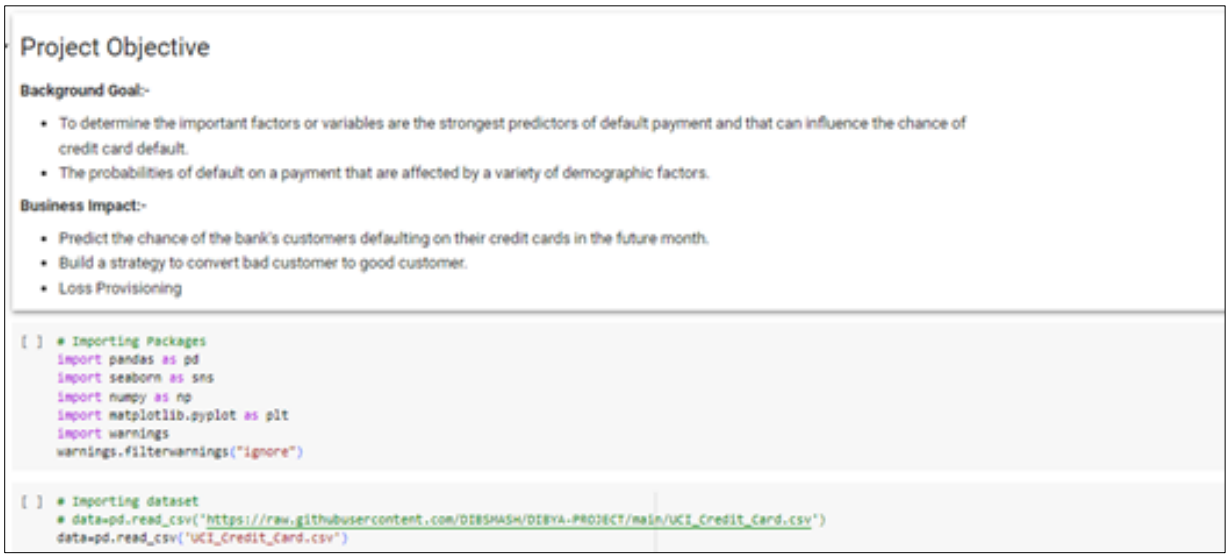


Figure 2 Project Objective and Importing datasets in python

This dataset contains information about 25 variables on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

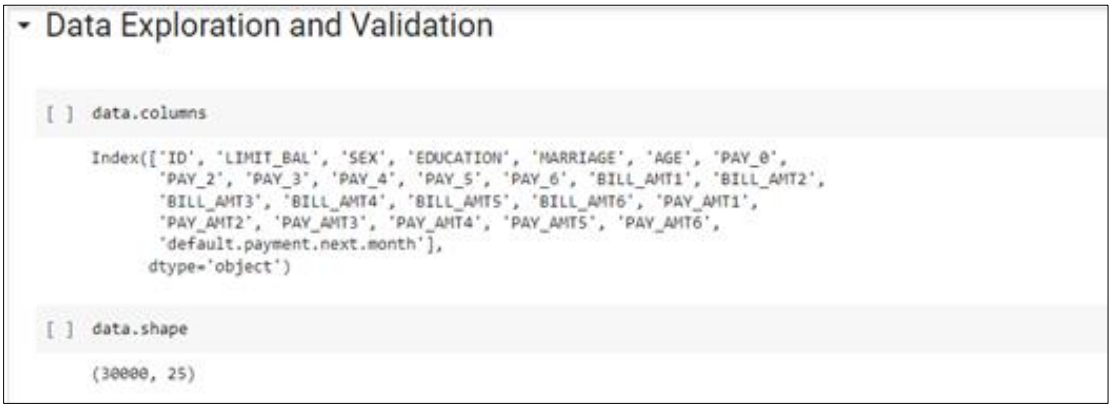
The above dataset shows the objective of the project and background goals. The key goals are to determine the essential variables that can strongly predict the default payment. It can influence the possibility of default on credit cards [28].

The above dataset shows importing of packages to read warnings regarding credit card payment utilising machine learning libraries such as pandas, seaborn, NumPy and matplotlib.

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	de
13262	13263	40000.0	2	2	1	24	2	0	0	2	...	29633.0	31548.0	30889.0	1800.0	4500.0	0.0	2400.0	0.0	3000.0	
13820	13821	60000.0	1	3	2	52	0	0	0	0	...	26305.0	23756.0	25353.0	1500.0	1500.0	2000.0	1500.0	2000.0	2000.0	
28220	28221	30000.0	2	2	1	33	3	2	2	7	...	2400.0	2400.0	2400.0	0.0	0.0	0.0	0.0	0.0	0.0	
27021	27022	100000.0	1	1	2	32	0	0	0	0	...	11700.0	11400.0	11419.0	2707.0	1363.0	1739.0	0.0	1000.0	1147.0	
24184	24185	160000.0	1	1	2	28	0	0	0	0	...	151403.0	115731.0	113636.0	6000.0	25409.0	30000.0	5000.0	5000.0	4500.0	
3427	3428	20000.0	1	2	2	48	0	0	2	0	...	10696.0	11365.0	12170.0	4500.0	0.0	1000.0	1000.0	1000.0	1000.0	
15465	15466	160000.0	1	3	2	30	0	0	0	0	...	0.0	0.0	0.0	8366.0	8368.0	0.0	0.0	0.0	0.0	
21836	21837	80000.0	2	1	2	23	0	0	0	0	...	60514.0	48965.0	29648.0	3117.0	2519.0	1838.0	1434.0	1515.0	1505.0	
19352	19353	290000.0	2	2	1	49	-1	0	0	0	...	36058.0	36965.0	38863.0	3000.0	3500.0	2500.0	2500.0	2500.0	2000.0	
21679	21680	210000.0	2	2	1	25	0	0	0	0	...	10838.0	12652.0	13060.0	1149.0	1173.0	1186.0	2000.0	1000.0	1000.0	
14969	14970	180000.0	2	1	2	31	-1	-1	-1	-1	...	3072.0	2756.0	2440.0	1359.0	316.0	3072.0	0.0	0.0	316.0	
26522	26523	140000.0	2	3	1	52	0	0	0	0	...	52676.0	51301.0	48202.0	3000.0	2200.0	2400.0	2000.0	2060.0	2150.0	
21813	21814	110000.0	2	1	2	25	0	0	0	0	...	74371.0	75126.0	73799.0	2800.0	3703.0	2788.0	2200.0	1819.0	3000.0	
21704	21705	170000.0	2	5	2	24	-2	-2	-1	-1	...	57961.0	66257.0	12106.0	0.0	1777.0	114693.0	5623.0	439.0	497.0	
1784	1785	400000.0	1	2	1	34	0	0	0	0	...	60876.0	65178.0	46933.0	6000.0	6000.0	1902.0	5000.0	0.0	5000.0	

Figure 3 Dataset showing the output variables**.**

The above dataset shows the output of Limited Balance, Education and Marriage, and Age list to analyse the payment of Alternative Minimum Tax about 15 variables.



Data Exploration and Validation

```
[ ] data.columns

Index(['ID', 'LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0',
       'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT2',
       'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1',
       'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6',
       'default.payment.next.month'],
      dtype='object')

[ ] data.shape

(30000, 25)
```

Figure 4 Data validation process in python.

Data Exploration is determined as the first step to analyse data, and it is used to visualise and explore the data to find insights from the beginning or find the patterns to dig for more information. data validation is used to check the accuracy of the result and source data quality before utilising and importing the data process. It is a form of data cleansing. The above picture shows the data exploration of the data column, data shape and data information. The data column has all the index, and the set of data shapes shows its objects. The data Info shows the use of pandas data frame and provides non-null count and data type of 23 variables.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  float64
 2   SEX                         30000 non-null  int64
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  int64
 5   AGE                         30000 non-null  int64
 6   PAY_0                       30000 non-null  int64
 7   PAY_2                       30000 non-null  int64
 8   PAY_3                       30000 non-null  int64
 9   PAY_4                       30000 non-null  int64
 10  PAY_5                       30000 non-null  int64
 11  PAY_6                       30000 non-null  int64
 12  BILL_AMT1                   30000 non-null  float64
 13  BILL_AMT2                   30000 non-null  float64
 14  BILL_AMT3                   30000 non-null  float64
 15  BILL_AMT4                   30000 non-null  float64
 16  BILL_AMT5                   30000 non-null  float64
 17  BILL_AMT6                   30000 non-null  float64
 18  PAY_AMT1                    30000 non-null  float64
 19  PAY_AMT2                    30000 non-null  float64
```

Figure 5 Classification of the data.

```
# Checking for null values in any columns
data.isnull().sum()

ID                          0
LIMIT_BAL                   0
SEX                         0
EDUCATION                   0
MARRIAGE                    0
AGE                         0
PAY_0                       0
PAY_2                       0
PAY_3                       0
PAY_4                       0
PAY_5                       0
PAY_6                       0
BILL_AMT1                   0
BILL_AMT2                   0
BILL_AMT3                   0
BILL_AMT4                   0
BILL_AMT5                   0
BILL_AMT6                   0
PAY_AMT1                    0
PAY_AMT2                    0
PAY_AMT3                    0
PAY_AMT4                    0
PAY_AMT5                    0
PAY_AMT6                    0
default.payment.next.month  0
dtype: int64
```

Figure 6 Dataset uses the application of ISnull .

The above dataset uses the application of ISnull and checks for null values in output 7.

```
[ ] data.describe()
```

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	...	30000.00000
mean	15000.500000	167484.322667	1.603733	1.853133	1.551867	35.485500	-0.016700	-0.133767	-0.166200	-0.220667	...	43262.94896
std	8660.398374	129747.661567	0.489129	0.790349	0.521970	9.217904	1.123802	1.197186	1.196868	1.169139	...	64332.85613
min	1.000000	10000.000000	1.000000	0.000000	0.000000	21.000000	-2.000000	-2.000000	-2.000000	-2.000000	...	-170000.00000
25%	7500.750000	50000.000000	1.000000	1.000000	1.000000	28.000000	-1.000000	-1.000000	-1.000000	-1.000000	...	2326.75000
50%	15000.500000	140000.000000	2.000000	2.000000	2.000000	34.000000	0.000000	0.000000	0.000000	0.000000	...	19052.00000
75%	22500.250000	240000.000000	2.000000	2.000000	2.000000	41.000000	0.000000	0.000000	0.000000	0.000000	...	54506.00000
max	30000.000000	1000000.000000	2.000000	6.000000	3.000000	79.000000	8.000000	8.000000	8.000000	8.000000	...	891586.00000

8 rows × 25 columns

Figure 7 The data output of 8 variables.

The above image has described the data output of 8 variables and depicted the data of Limited balance, sex, education, marriage and age with 4 payment statuses.

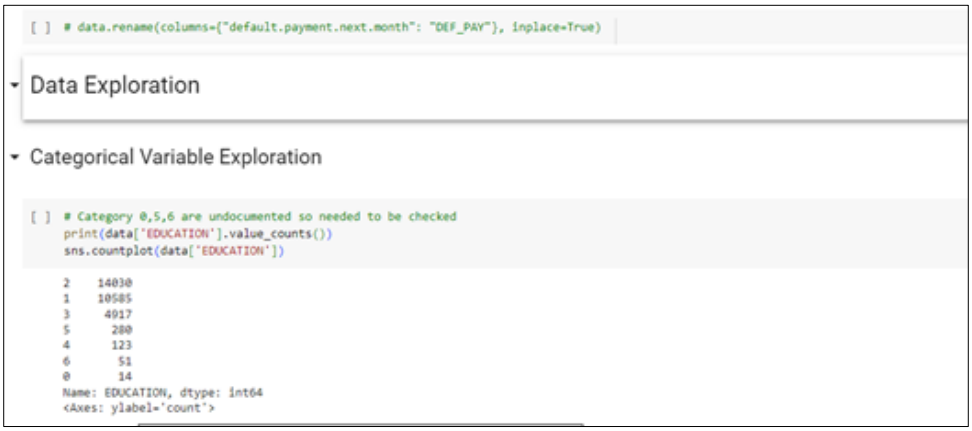


Figure 8 The data exploration in different categories.

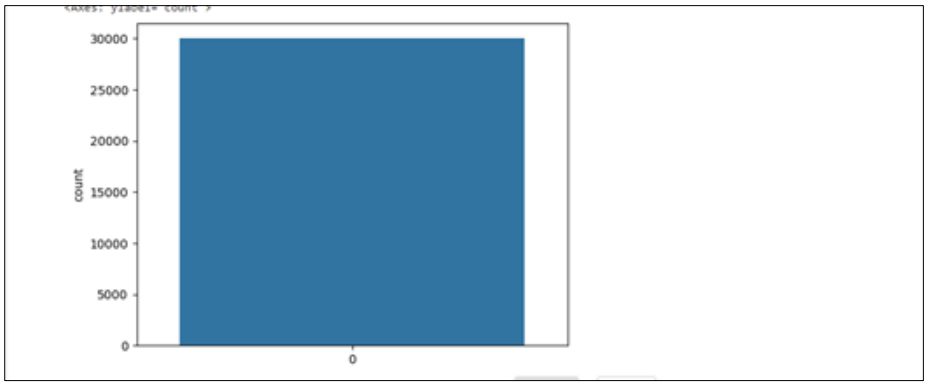


Figure 9 Visualisation of the data exploration in different categories.

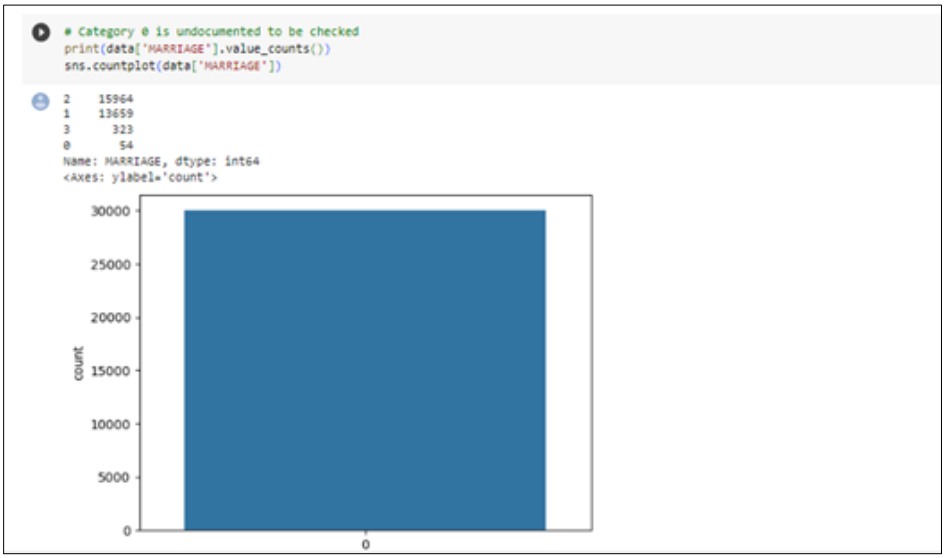


Figure 10 LaTex to conduct mathematical calculations.

The above image has used LaTex to conduct mathematical calculations in Markdown, which is a lightweight markup language. It has also depicted categorical variable exploration to check the undocumented value of 0,6,5 and provides a graph about the count.

```
[ ] # Some of the age values are more than 70 which is fine.
    plt.figure(figsize=(10,8))
    data['AGE'].value_counts()
    sns.countplot(data['AGE'])
```

Figure 11 Example of coding showing age values.

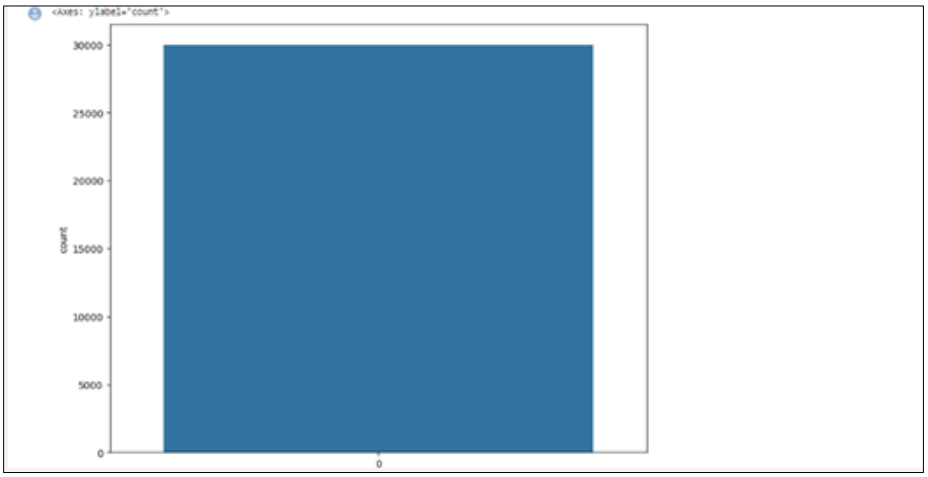


Figure 12 Depiction using matasploit library showing age values.

The above picture depicts some of the age values, which are more than 70. The objectives of figure size here are (10,8), and it has also shown the data of value count in output 12.

```
[ ] data.columns

Index(['ID', 'LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0',
       'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT2',
       'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1',
       'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6',
       'default.payment.next.month'],
      dtype='object')
```

```
[ ] # All the pay are having -2,0 category that are undocumented.
    # Given category -1 as pay duly(properly paid on time)
    # So -2,0,-1 can be treated as one category. to be checked
    print(data['PAY_0'].value_counts())
    print(data['PAY_2'].value_counts())

     0    14737
    -1     5686
     1     3688
    -2     2759
     2     2667
     3      322
     4       76
     5       26
     8       19
     6       11
     7        9
    Name: PAY_0, dtype: int64
     0    15730
    -1     6050
     2     3927
    -2     3782
     3      326
     4       99
     1       28
     5       25
     7       20
     6       12
     8        1
    Name: PAY_2, dtype: int64
```

Figure 13 Input of data columns, and the output depicts the index of default payment.

The above picture is the input of data columns, and the output depicts the index of default payment. here all the pay has (-2,0) categories that are uncounted. The -1 category is depicted as a paid duty, and -2,0 and -1 can be treated as one specific category.

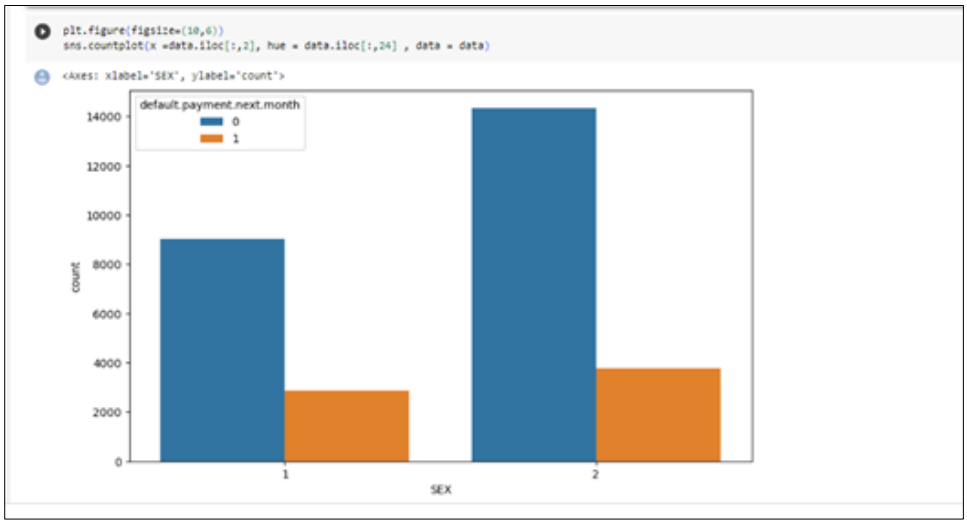


Figure 14 The library of the count plot. 'X' is the data here, and Hue is the categorical value.

The input in the above set shows a new figure of 10 inches by 6 inches, and the next line uses the library of the count plot. 'X' is the data here, and Hue is the categorical value. The x-label shows the data on sex, and the y-label shows the data on the count.

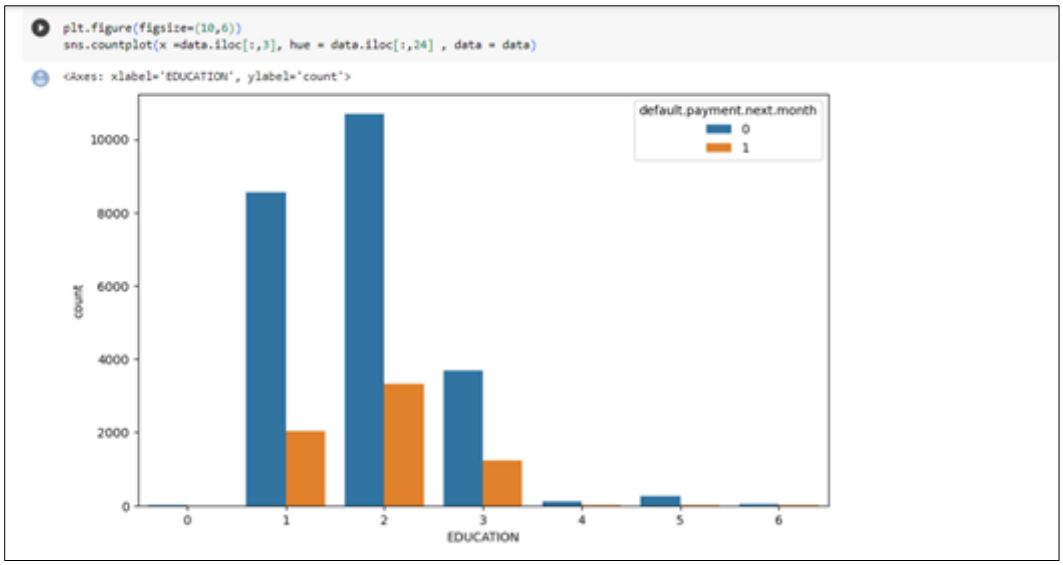


Figure 15 Depiction of the library of the count plot. 'X' is the data here, and Hue is the categorical value figure of 6 by 6 inches.

The input shows a new figure of 6 by 6 inches and provides a picture that is square-shaped. The next line used the seaborn library, the parameter 'X' is the data, and Hue is the categorical value of data. The output has an x-label axis of Education, and the y-label axis shows the count of data.

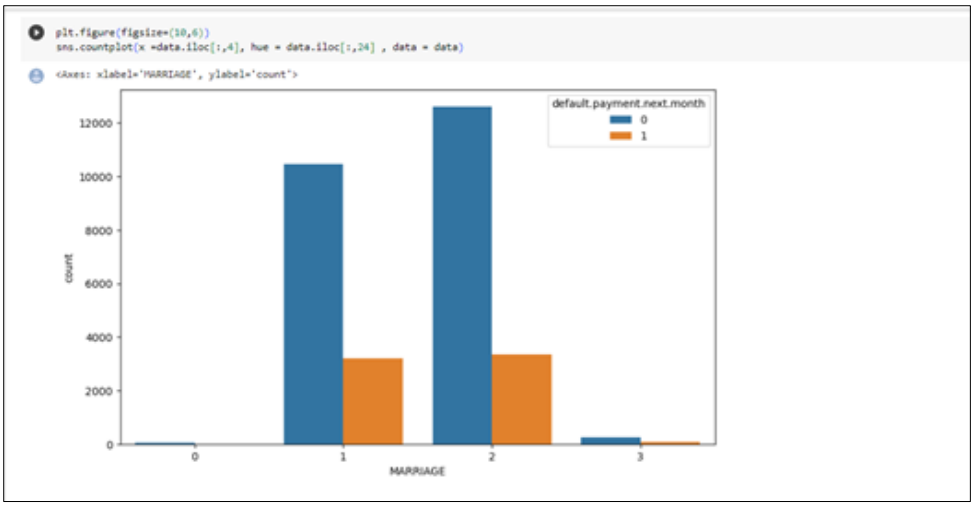


Figure 16 Depiction of the library of the count plot. 'X' is the data here, and Hue is the categorical value figure of 10 by 6 inches.

The above picture shows a figure size of 10 inches by 6 inches. The next line shows the use of count plot library where 'X' is the data and Hue is the categorical value of data. The x-label is the data of marriage and the y-label is the count of data.

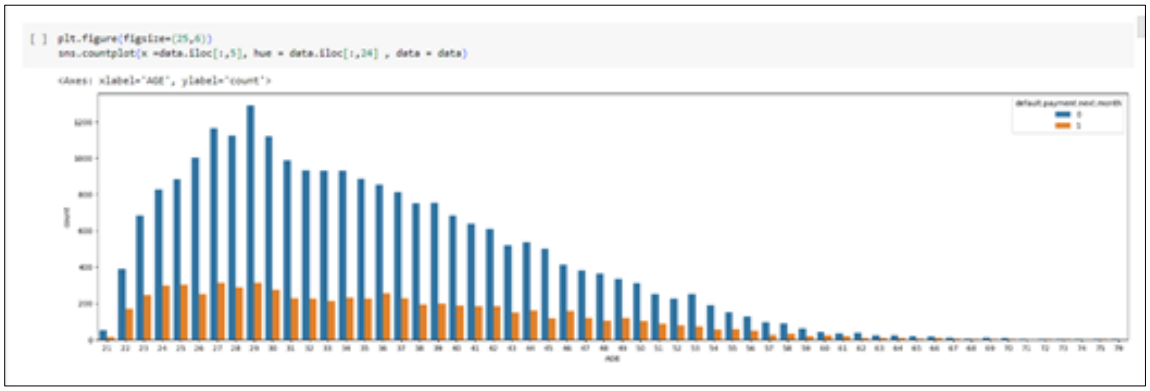


Figure 17 Depiction of the library of the count plot. 'X' is the data here, and Hue is the categorical value figure of 25 by 6 inches

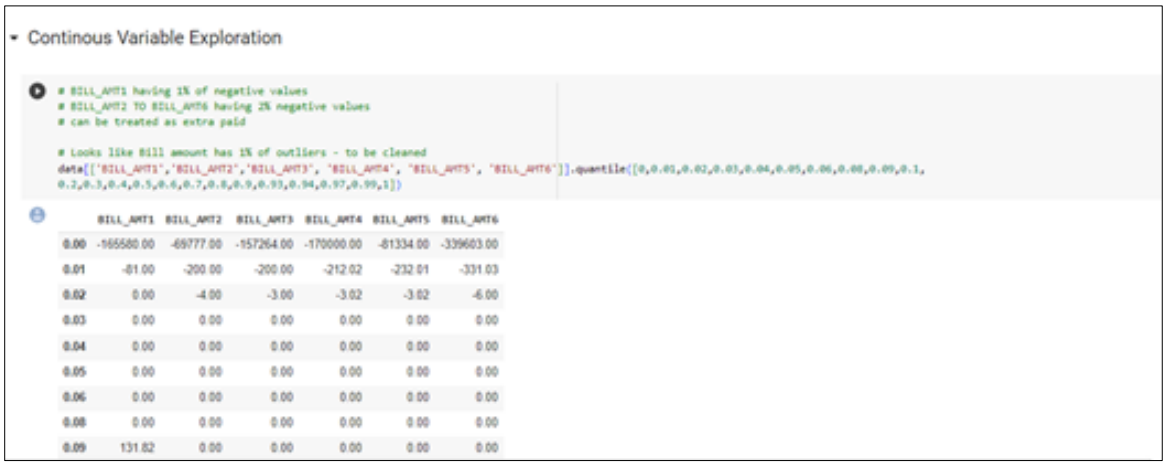


Figure 18 Coding of variable exploration**.**

0.02	0.00	-4.00	-3.00	-3.02	-3.02	-6.00
0.03	0.00	0.00	0.00	0.00	0.00	0.00
0.04	0.00	0.00	0.00	0.00	0.00	0.00
0.05	0.00	0.00	0.00	0.00	0.00	0.00
0.06	0.00	0.00	0.00	0.00	0.00	0.00
0.08	0.00	0.00	0.00	0.00	0.00	0.00
0.09	131.82	0.00	0.00	0.00	0.00	0.00
0.10	278.90	0.00	0.00	0.00	0.00	0.00
0.20	1892.80	1472.80	1187.80	968.00	763.00	476.00
0.30	6050.40	5500.00	5219.20	4643.70	3637.00	2701.70
0.40	13469.20	12799.60	12197.20	11145.00	9809.20	8770.20
0.50	22381.50	21200.00	20088.50	19052.00	18104.50	17071.00
0.60	37045.20	34773.80	31401.00	28604.40	26690.40	25508.40
0.70	52204.90	50690.00	49217.30	45456.60	40943.20	39252.20
0.80	83421.20	80292.20	76777.40	70579.00	65823.00	63150.60
0.90	142133.70	136905.50	132051.30	122418.70	115883.00	112110.40
0.93	173630.63	167485.28	160049.77	148287.35	141291.49	137836.47
0.94	187074.56	180578.40	173241.34	160944.12	151270.38	148093.84
0.97	245969.84	236662.72	228277.79	210965.51	198861.03	195114.30
0.99	350110.68	337495.28	325030.39	304997.27	285868.33	279505.06
1.00	964511.00	983931.00	1664089.00	891586.00	927171.00	961664.00

Figure 19 Result of coding variable exploration.

Input 18 shows a figure size of 25 inches by 6 inches, and it has used the count plot library. 'X' is considered as data, and Hue is the categorical value of data. In output 18, The x-label is the data of Age, and the y-label is the data of count. Input 19 shows the continuous variable exploration of 1% negative values and 25 negative values. Data of 6 bill amount is contrasted here and has cleaned 1% bill amount. Output 19 shows the data of all 6 Bill amounts from 0.00 to 1.00.

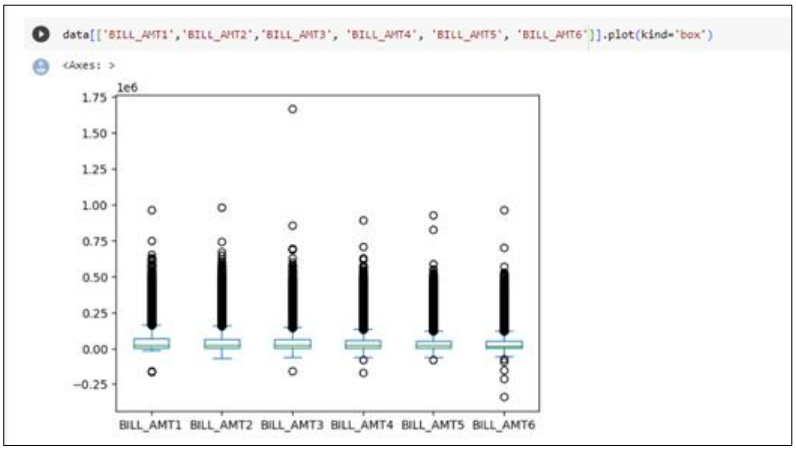
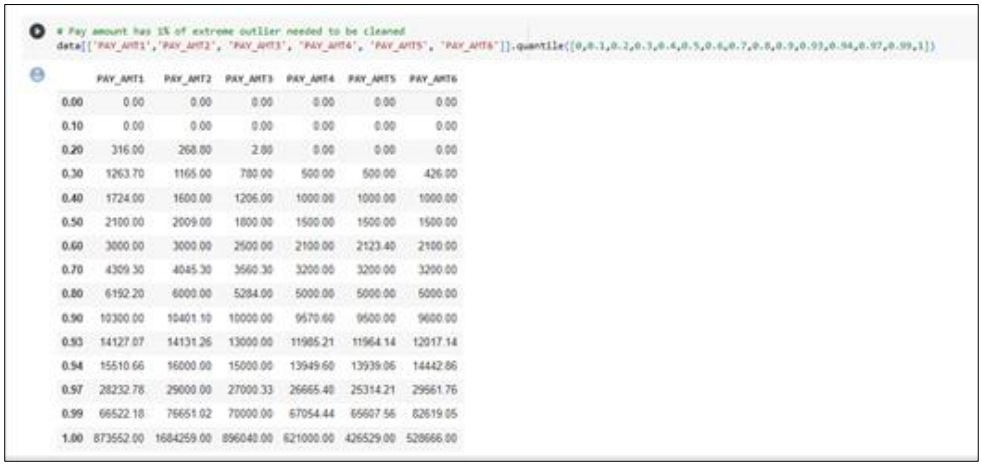


Figure 20 Output 19 shows the data of all 6 Bill amounts from 0.00 to 1.00.



	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.10	0.00	0.00	0.00	0.00	0.00	0.00
0.20	316.00	268.80	2.80	0.00	0.00	0.00
0.30	1263.70	1165.00	780.00	500.00	500.00	426.00
0.40	1724.00	1600.00	1206.00	1000.00	1000.00	1000.00
0.50	2100.00	2009.00	1800.00	1500.00	1500.00	1500.00
0.60	3000.00	3000.00	2500.00	2100.00	2123.40	2100.00
0.70	4309.30	4045.30	3660.30	3200.00	3200.00	3200.00
0.80	6192.20	6000.00	5284.00	5000.00	5000.00	5000.00
0.90	10300.00	10401.10	10000.00	9570.60	9600.00	9600.00
0.93	14127.07	14131.26	13000.00	11985.21	11964.14	12017.14
0.94	15510.66	16000.00	15000.00	13949.60	13939.06	14442.86
0.97	28232.78	29000.00	27000.33	26665.40	25314.21	29561.76
0.99	66522.18	76651.02	70000.00	67054.44	65607.56	82619.05
1.00	873552.00	1684259.00	896040.00	621000.00	426529.00	528666.00

Figure 21 Output 19 shows the data of all 6 Bill amounts from 0.00 to 1.00.

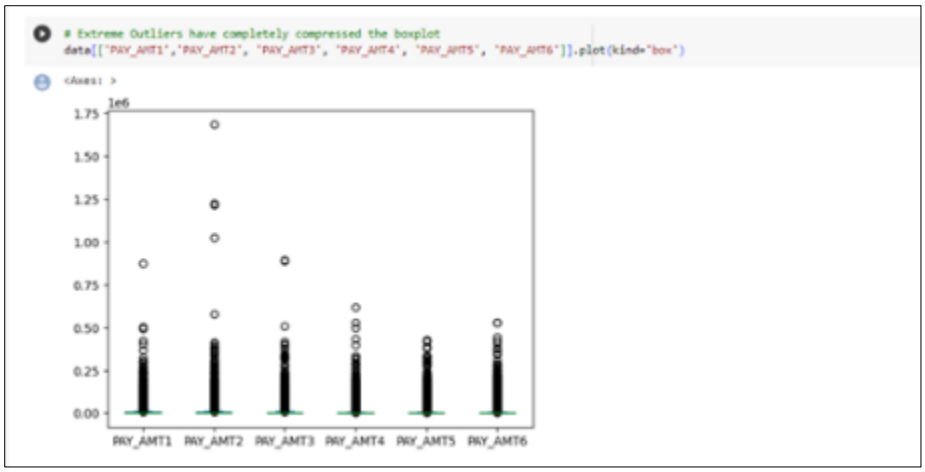


Figure 22 Output 19 shows the data of all 6 Bill amounts from 0.00 to 1.00.

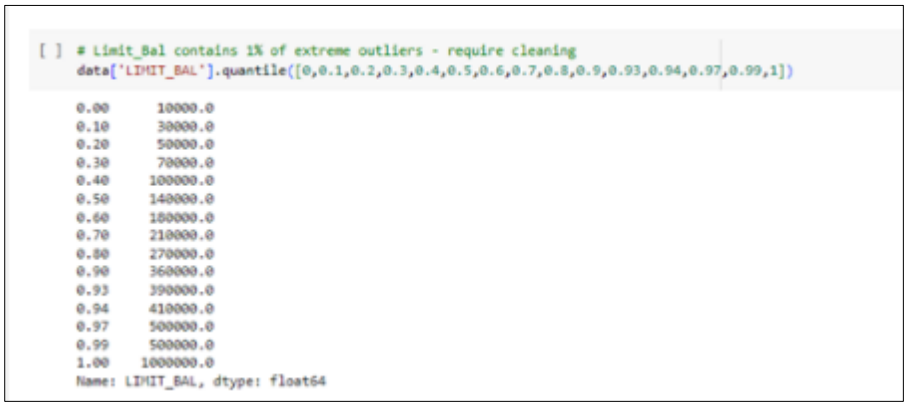


Figure 23 Uses the application of Bill AMT, an application of Python.

Input 20 shows that the figure uses the application of Bill AMT, an application of Python. It shows data of 6 Bill AMT in both axes. Input 21 states that the figure uses the application of Pay_AMT and extracts data from the quintile, and shows the data of 6 Pay_AMT in Output 21. Input 22 shows that extreme outliers have compressed the boxplot and calculated the data by utilising the application of Pay_AMT. The value of 6 Pay_AMt data is depicted in Output 22. The next image shows the input of Limited Balance that contains 1% of extreme outliers. It extracts the data from quantile, and the data type output 23 is float64.

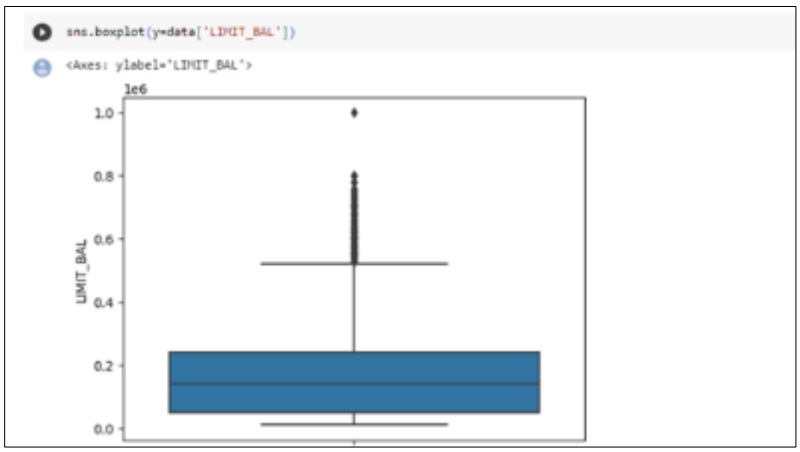


Figure 24 The input of Limited Balance that contains 1% of extreme outliers.

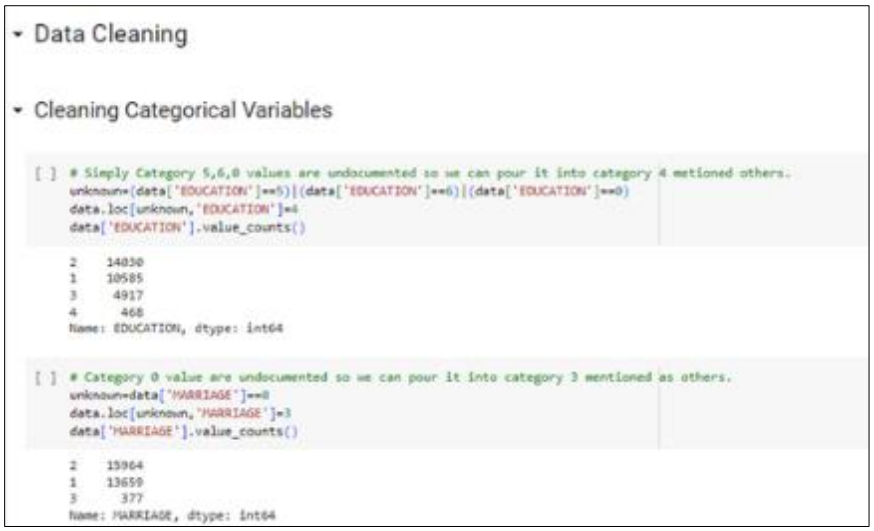


Figure 25 Data Cleaning.

Figure 26 Result of data Cleaning.

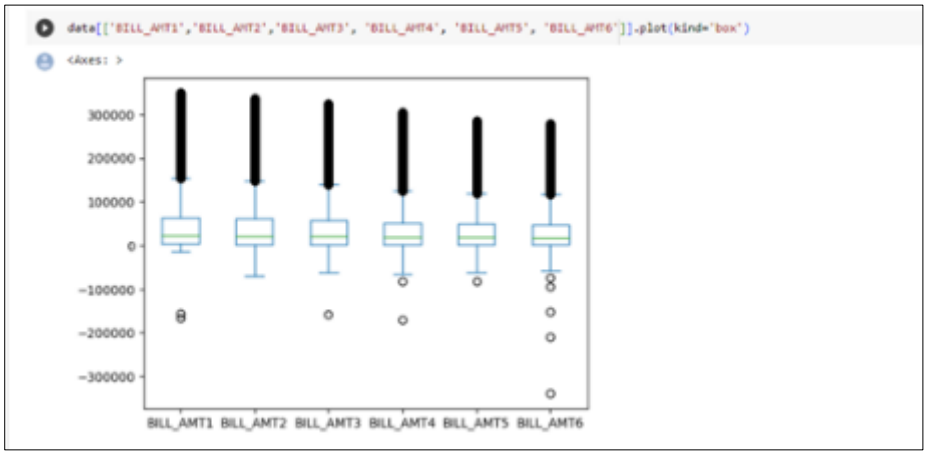


Figure 27 Result of data Cleaning.

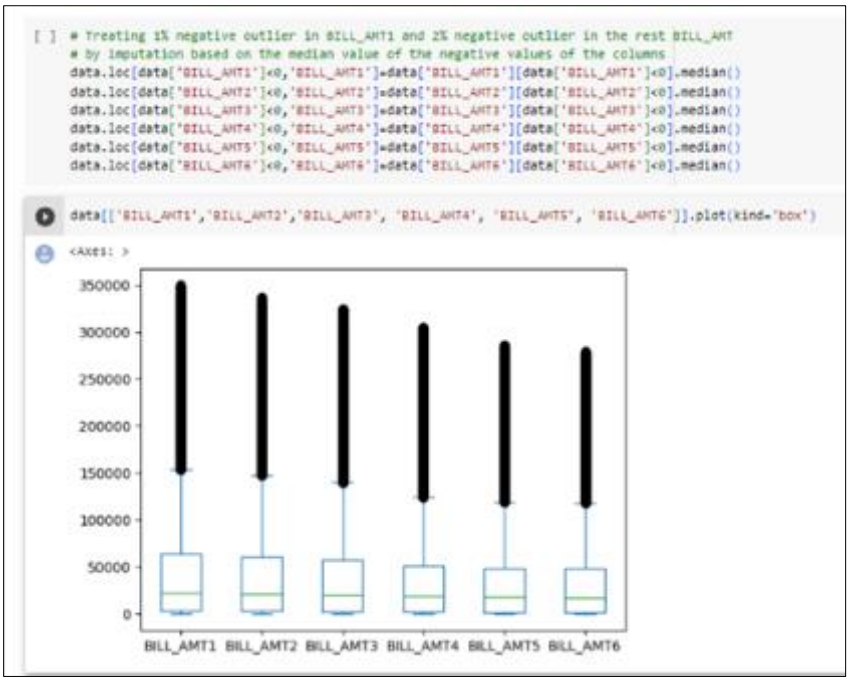


Figure 28 Result of data Cleaning

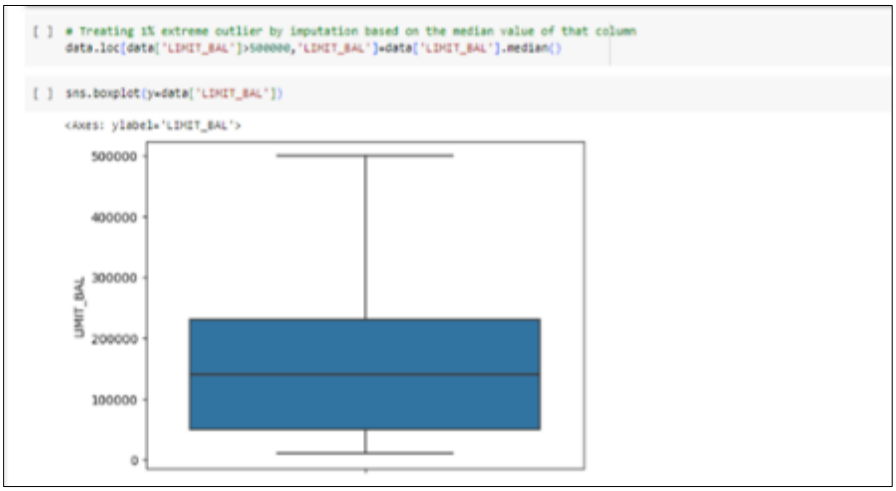


Figure 29 Result of data Cleaning.

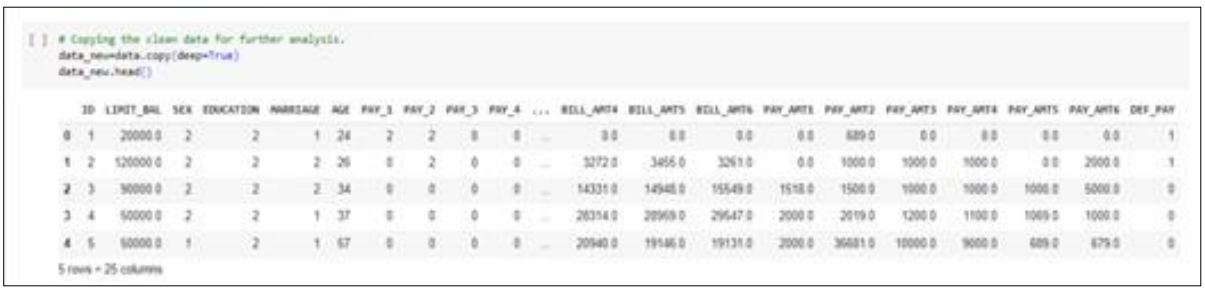


Figure 30 The above sets of data used data cleaning and to correct errors.

The above sets of data used data cleaning to correct errors and reconstruct data to make this easier, it is the process to prepare raw data to analyse and clean bad data. It also helps to organise the raw information and filling it in the null values by cleaning categorical variables and continuous variables.

Select the right model, Include Feature Engineering, and Perform Model Validation, Multiple models and performance comparison.

The figures below have selected the feature engineering model, performed model validation, multiple performance comparisons.

The problem description and objective of this project, which is to determine the default or non-default of a credit card, make it clear that this is a classification problem, and therefore we can build a Logistic Regression Machine Learning Model.

The objective and problem description of this project are to consider the default or non-default of a credit card and identify the classification problem to build a logistic regression machine learning model.

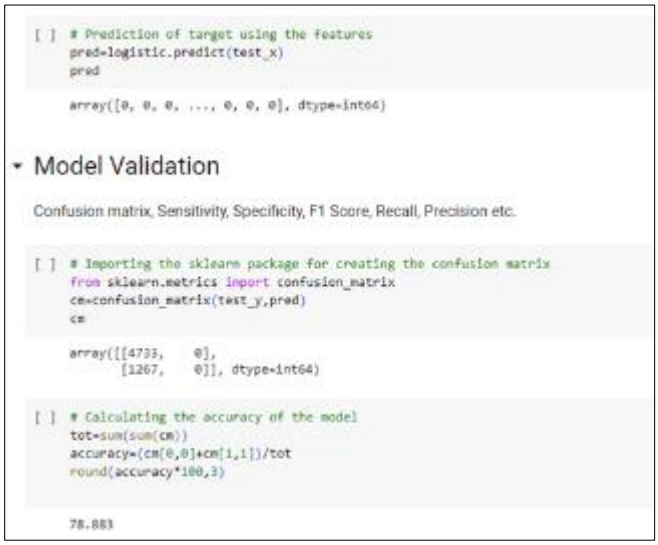


Figure 31 The above parts utilise the confusion matrix

The above parts utilise the confusion matrix, Sensitivity, F1 score, Recall and precision. The model building logistics regression has created an x array that contains features and a y array that contains the target vector. It uses SK learn a linear model to import the logistic regression.

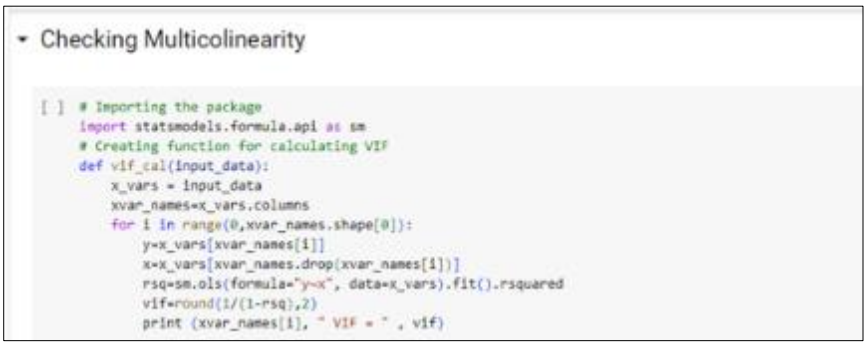
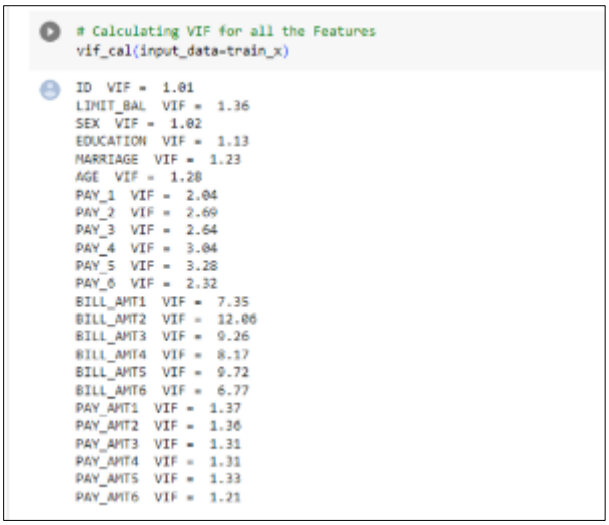


Figure 32 Use of SK learn a linear model to import the logistic regression.

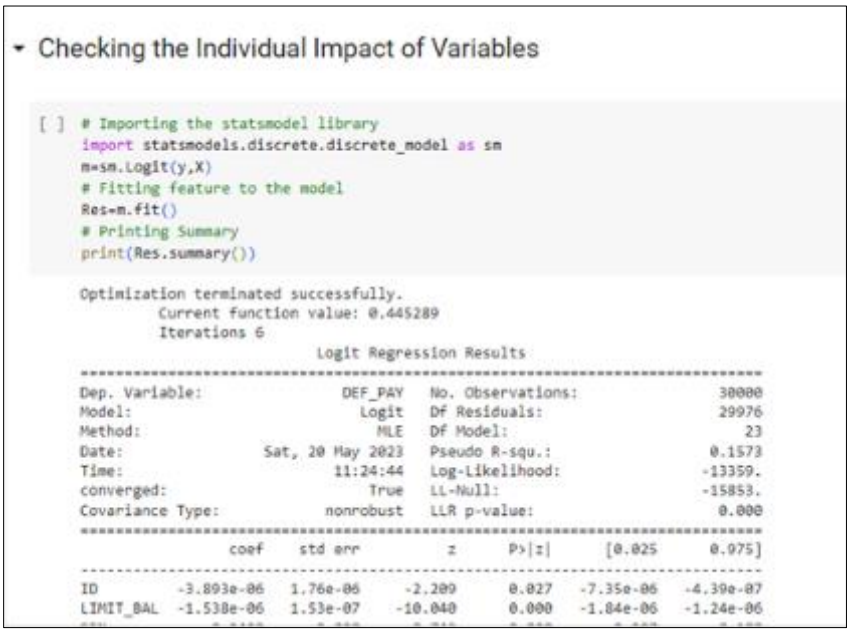


```
# Calculating VIF for all the Features
vif_cal(input_data=train_x)
```

```
ID  VIF =  1.01
LIMIT_BAL  VIF =  1.36
SEX  VIF =  1.02
EDUCATION  VIF =  1.13
MARRIAGE  VIF =  1.23
AGE  VIF =  1.28
PAY_1  VIF =  2.04
PAY_2  VIF =  2.69
PAY_3  VIF =  2.64
PAY_4  VIF =  3.04
PAY_5  VIF =  3.28
PAY_6  VIF =  2.32
BILL_AMT1  VIF =  7.35
BILL_AMT2  VIF =  12.06
BILL_AMT3  VIF =  9.26
BILL_AMT4  VIF =  8.17
BILL_AMT5  VIF =  9.72
BILL_AMT6  VIF =  6.77
PAY_AMT1  VIF =  1.37
PAY_AMT2  VIF =  1.36
PAY_AMT3  VIF =  1.31
PAY_AMT4  VIF =  1.31
PAY_AMT5  VIF =  1.33
PAY_AMT6  VIF =  1.21
```

Figure 33 The above dataset has shown VIF<5 as independent feature.

The above dataset has shown VIF<5 as independent feature to keep it and used VIF>=5 that can establish a situation that is multicollinearity.

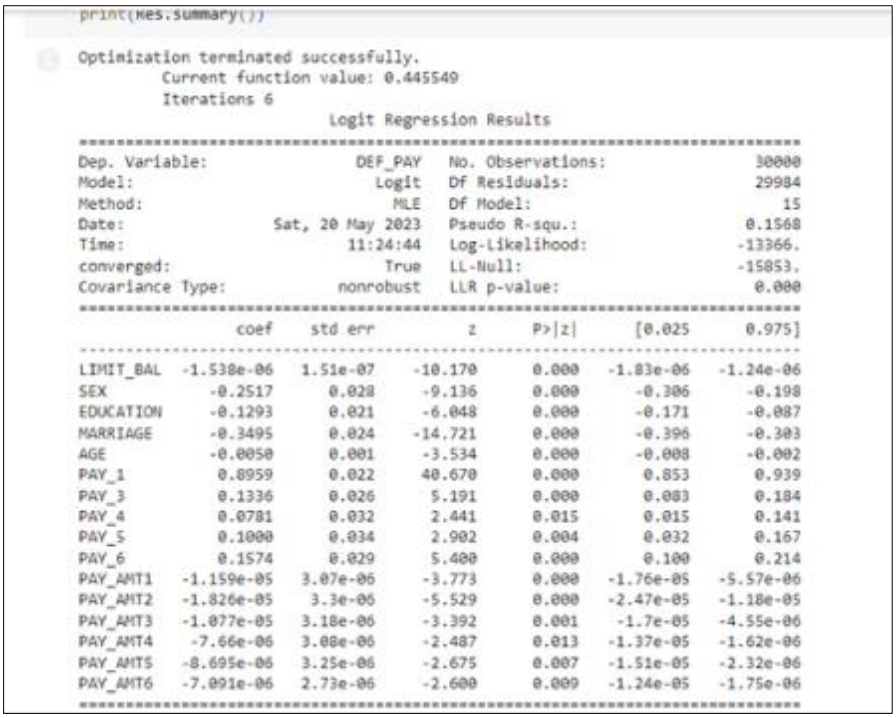


Checking the Individual Impact of Variables

```
# Importing the statsmodel library
import statsmodels.discrete.discrete_model as sm
m=sm.Logit(y,X)
# Fitting feature to the model
Res=m.fit()
# Printing Summary
print(Res.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.445289
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                DEF_PAY   No. Observations:                30000
Model:                          Logit   Df Residuals:                    29976
Method:                           MLE   Df Model:                           23
Date:                Sat, 20 May 2023   Pseudo R-squ.:                  0.1573
Time:                        11:24:44   Log-Likelihood:                -13359.
converged:                       True   LL-Null:                       -15853.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ID         -3.893e-06   1.76e-06     -2.209      0.027   -7.35e-06   -4.39e-07
LIMIT_BAL  -1.538e-06   1.53e-07    -10.040      0.000   -1.84e-06   -1.24e-06
```

Figure 34 Checking the individual impact of variables.

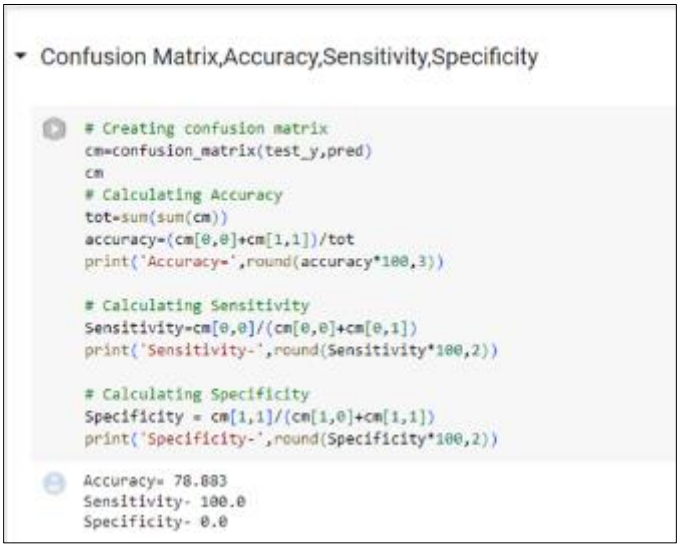


```
print(res.summary())

Optimization terminated successfully.
         Current function value: 0.445549
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                DEF_PAY   No. Observations:                30000
Model:                          Logit   Df Residuals:                    29984
Method:                           MLE   Df Model:                           15
Date:                Sat, 20 May 2023   Pseudo R-squ.:                 0.1568
Time:                        11:24:44   Log-Likelihood:                -13366.
converged:                       True   LL-Null:                       -15853.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
LIMIT_BAL  -1.538e-06   1.51e-07    -10.170      0.000   -1.83e-06   -1.24e-06
SEX           -0.2517      0.028     -9.136      0.000      -0.306      -0.198
EDUCATION     -0.1293      0.021     -6.048      0.000      -0.171      -0.087
MARRIAGE      -0.3495      0.024    -14.721      0.000      -0.396      -0.303
AGE           -0.0050      0.001     -3.534      0.000      -0.008      -0.002
PAY_1          0.8959      0.022     40.670      0.000       0.853       0.939
PAY_3          0.1336      0.026      5.191      0.000       0.083       0.184
PAY_4          0.0781      0.032      2.441      0.015       0.015       0.141
PAY_5          0.1000      0.034      2.902      0.004       0.032       0.167
PAY_6          0.1574      0.029      5.400      0.000       0.100       0.214
PAY_AMT1   -1.159e-05   3.07e-06     -3.773      0.000   -1.76e-05   -5.57e-06
PAY_AMT2   -1.826e-05    3.3e-06     -5.529      0.000   -2.47e-05   -1.18e-05
PAY_AMT3   -1.077e-05   3.18e-06     -3.392      0.001    -1.7e-05   -4.55e-06
PAY_AMT4    -7.66e-06   3.08e-06     -2.487      0.013   -1.37e-05   -1.62e-06
PAY_AMT5   -8.695e-06   3.25e-06     -2.675      0.007   -1.51e-05   -2.32e-06
PAY_AMT6   -7.091e-06   2.73e-06     -2.600      0.009   -1.24e-05   -1.75e-06
==============================================================================
```

Figure 35 The above figures show that the Id of P|z|, Bill_AMt, PAY_2, and Bill_AMT6 features are non-impactful.

The above figures show that the Id of P|z|, Bill_AMt, PAY_2, and Bill_AMT6 features are non-impactful.



Confusion Matrix,Accuracy,Sensitivity,Specificity

```
# Creating confusion matrix
cm=confusion_matrix(test_y,pred)
cm
# Calculating Accuracy
tot=sum(sum(cm))
accuracy=(cm[0,0]+cm[1,1])/tot
print('Accuracy=',round(accuracy*100,3))

# Calculating Sensitivity
Sensitivity=cm[0,0]/(cm[0,0]+cm[0,1])
print('Sensitivity-',round(Sensitivity*100,2))

# Calculating Specificity
Specificity = cm[1,1]/(cm[1,0]+cm[1,1])
print('Specificity-',round(Specificity*100,2))
```

```
Accuracy= 78.883
Sensitivity- 100.0
Specificity- 0.0
```

Figure 36 Confusion Matrix, Accuracy, sensitivity, and specificity.

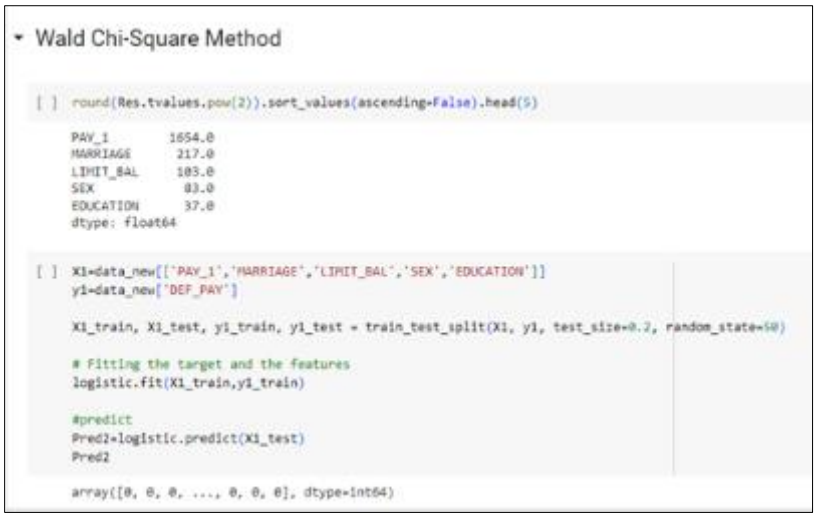


Figure 37 The use of Wald Chi-Square method to identify and choose the top 5 higher ranking features.

It has been observed that the accuracy number has not changed after dropping all the features that are unwanted. Observing the specificity and the sensitivity, it has been found that there is a huge imbalance of class.

It has used the method of Wald Chi-Square to identify and choose the top 5 higher ranking features from the observation of overall data and try to check the accuracy effect.

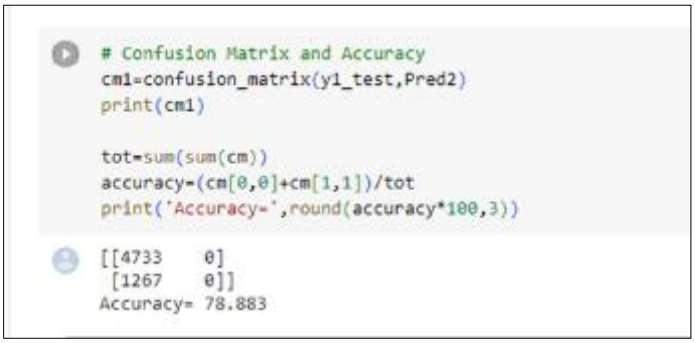


Figure 38 Checking the impact on the accuracy.In the above picture, it has been observed that there is no effect on the accuracy.

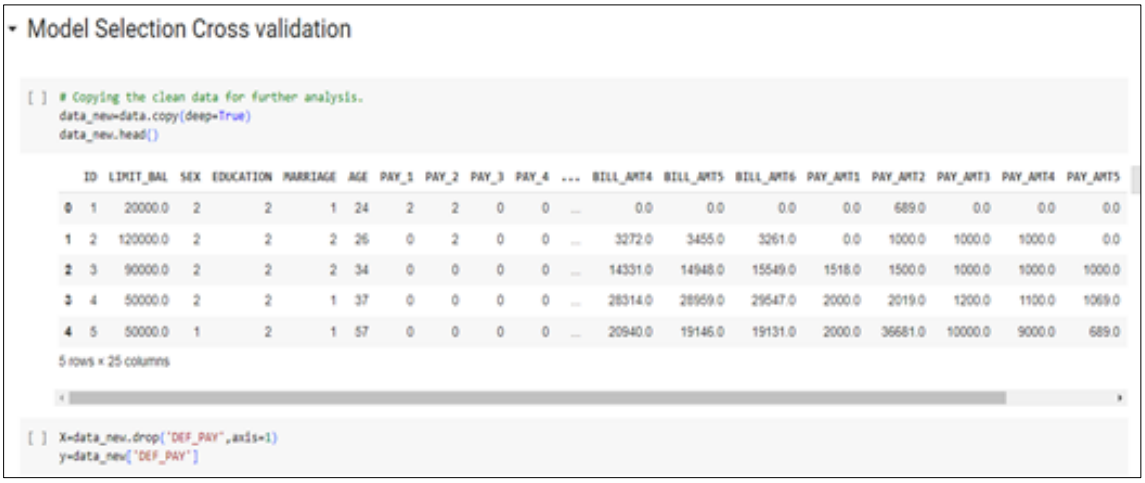
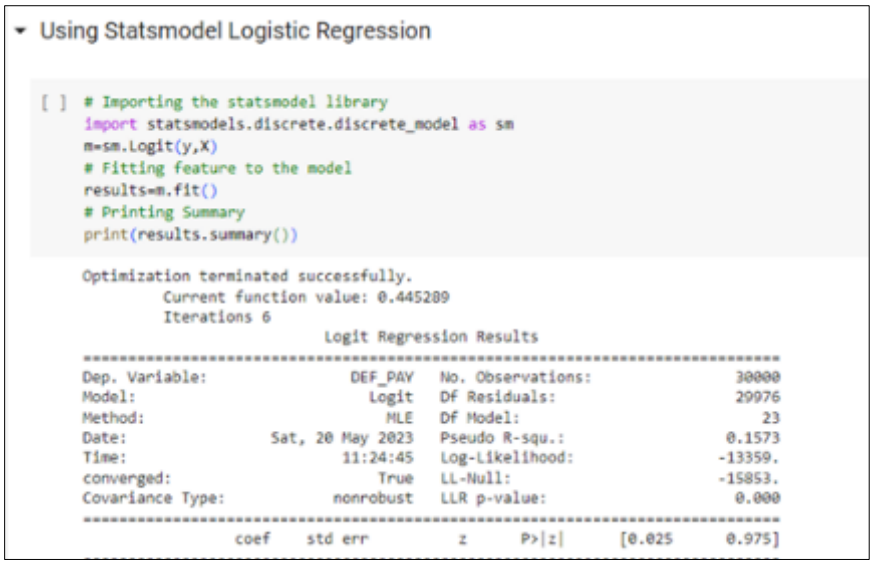


Figure 39 Model selection cross validation.

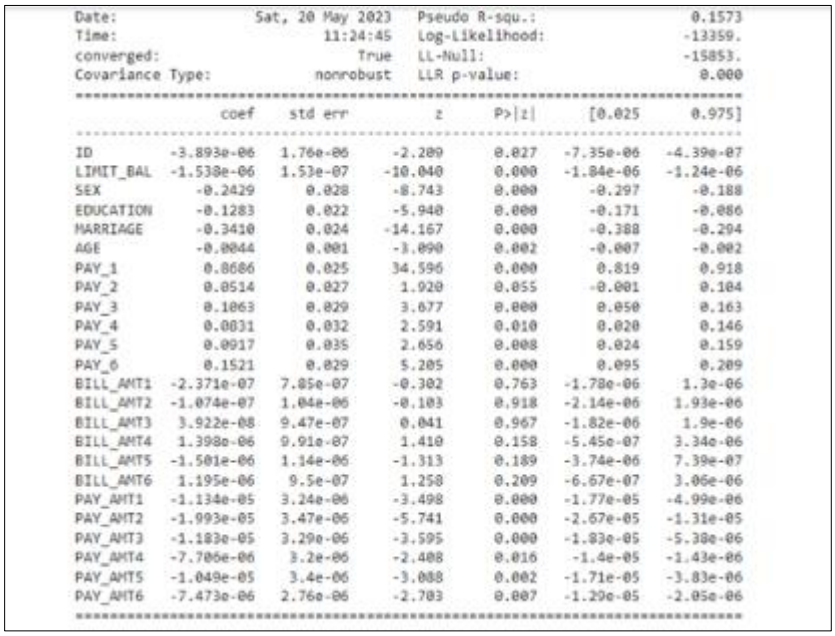


Using Statsmodel Logistic Regression

```
[ ] # Importing the statsmodel library
    import statsmodels.discrete.discrete_model as sm
    m=sm.Logit(y,X)
    # Fitting feature to the model
    results=m.fit()
    # Printing Summary
    print(results.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.445289
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                DEF_PAY   No. Observations:                30000
Model:                          Logit   Df Residuals:                    29976
Method:                           MLE   Df Model:                           23
Date:                Sat, 20 May 2023   Pseudo R-squ.:                  0.1573
Time:                        11:24:45   Log-Likelihood:                -13359.
converged:                       True   LL-Null:                       -15853.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
```

Figure 40 Use of Statsmodel Logistic regression.



```
Date:                Sat, 20 May 2023   Pseudo R-squ.:                  0.1573
Time:                        11:24:45   Log-Likelihood:                -13359.
converged:                       True   LL-Null:                       -15853.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ID         -3.893e-06   1.76e-06     -2.209      0.027   -7.35e-06   -4.39e-07
LIMIT_BAL  -1.538e-06   1.53e-07    -10.040      0.000   -1.84e-06   -1.24e-06
SEX           -0.2429      0.028     -8.743      0.000      -0.297      -0.188
EDUCATION     -0.1283      0.022     -5.940      0.000      -0.171      -0.086
MARRIAGE      -0.3410      0.024    -14.167      0.000      -0.388      -0.294
AGE           -0.0044      0.001     -3.090      0.002      -0.007      -0.002
PAY_1          0.8686      0.025     34.596      0.000       0.819       0.918
PAY_2          0.0514      0.027      1.920      0.055      -0.001       0.104
PAY_3          0.1063      0.029      3.677      0.000       0.050       0.163
PAY_4          0.0831      0.032      2.591      0.010       0.020       0.146
PAY_5          0.0917      0.035      2.656      0.008       0.024       0.159
PAY_6          0.1521      0.029      5.205      0.000       0.095       0.209
BILL_AMT1  -2.371e-07   7.85e-07     -0.302      0.763   -1.78e-06    1.3e-06
BILL_AMT2  -1.074e-07   1.04e-06     -0.103      0.918   -2.14e-06   1.93e-06
BILL_AMT3   3.922e-08   9.47e-07      0.041      0.967   -1.82e-06    1.9e-06
BILL_AMT4   1.398e-06   9.91e-07      1.410      0.158   -5.45e-07   3.34e-06
BILL_AMT5  -1.501e-06   1.14e-06     -1.313      0.189   -3.74e-06   7.39e-07
BILL_AMT6   1.195e-06    9.5e-07      1.258      0.209   -6.67e-07   3.06e-06
PAY_AMT1   -1.134e-05   3.24e-06     -3.498      0.000   -1.77e-05   -4.99e-06
PAY_AMT2   -1.993e-05   3.47e-06     -5.741      0.000   -2.67e-05   -1.31e-05
PAY_AMT3   -1.183e-05   3.29e-06     -3.595      0.000   -1.83e-05   -5.38e-06
PAY_AMT4   -7.706e-06    3.2e-06     -2.408      0.016    -1.4e-05   -1.43e-06
PAY_AMT5   -1.049e-05    3.4e-06     -3.088      0.002   -1.71e-05   -3.83e-06
PAY_AMT6   -7.473e-06   2.76e-06     -2.703      0.007   -1.29e-05   -2.05e-06
==============================================================================
```

Figure 41 Result from Statsmodel Logistic regression.

Figure 42 Confusion Matrix, Accuracy, sensitivity, and specificity.

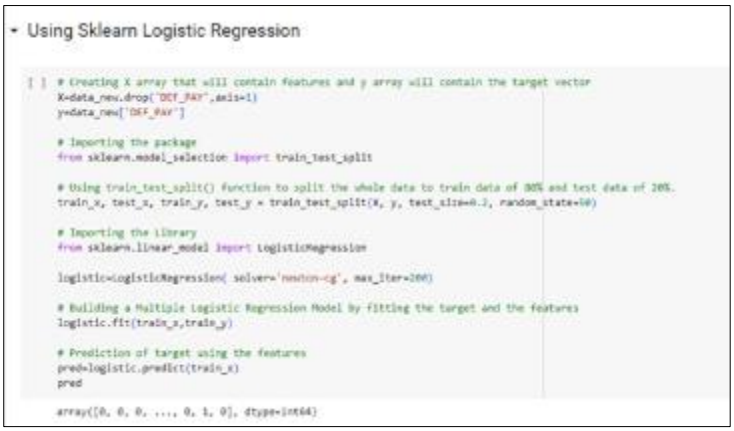


Figure 43 Use of Sklearn Logistics regression.

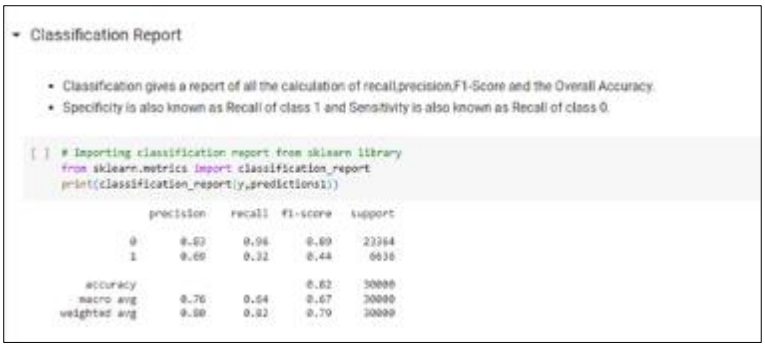


Figure 44 The classification report.

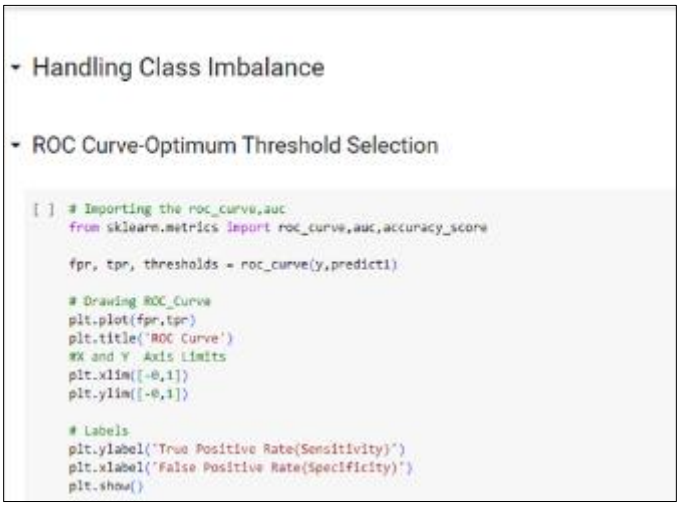
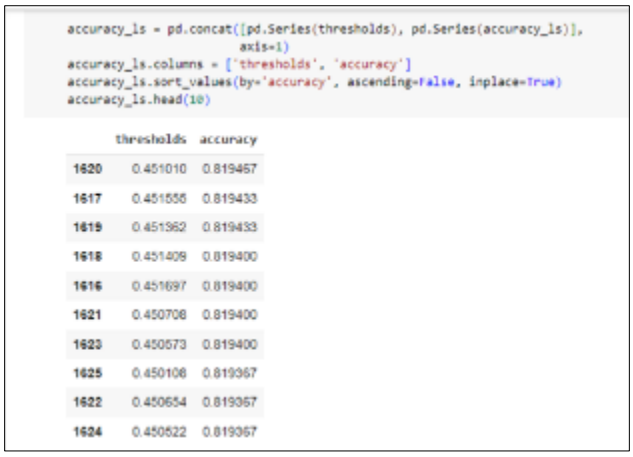


Figure 45 Handling Class imbalance.

From the data frame, all clean data are copied using the copy() and displayed in a table. Define the axis using the Logit() and use fit() to fit the model's features and show the result of Logit regression in a table format. Taking the value of Logistic regression as its threshold is 0.5, use the sklearn library for Logistic Regression where used to split the train data as 80% and the test data as 20%. the accuracy of this regression is 82%. From the above figure, the graph takes the plot() with the ROC curve value with X and Y axis.

Figure 46 Use of sklearn library for Logistic Regression.



```
accuracy_ls = pd.concat([pd.Series(thresholds), pd.Series(accuracy_ls)],
                        axis=1)
accuracy_ls.columns = ['thresholds', 'accuracy']
accuracy_ls.sort_values(by='accuracy', ascending=False, inplace=True)
accuracy_ls.head(10)
```

	thresholds	accuracy
1620	0.451010	0.819457
1617	0.451556	0.819433
1619	0.451362	0.819433
1618	0.451409	0.819400
1616	0.451697	0.819400
1621	0.450708	0.819400
1623	0.450573	0.819400
1625	0.450106	0.819367
1622	0.450654	0.819367
1624	0.450522	0.819367

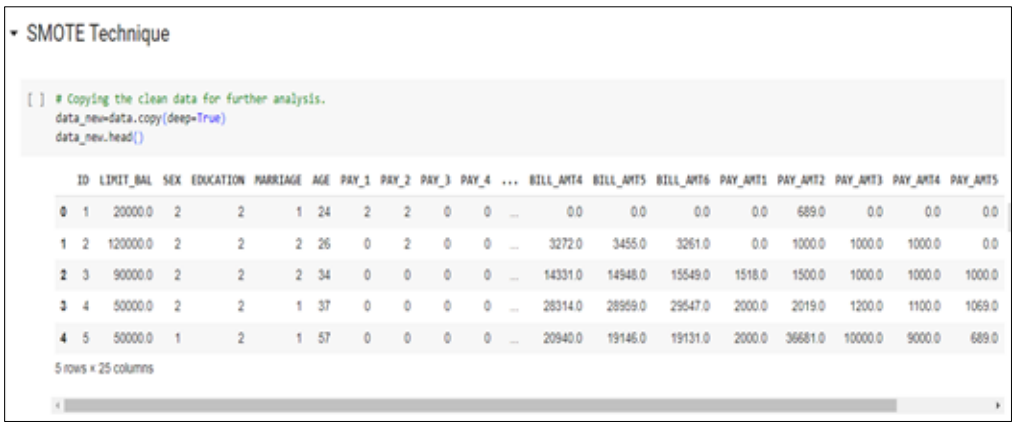
Figure 47 Using the sklearn library display the accuracy score.

Using the sklearn library display the accuracy score in a table format including the accuracy in thresholds and use concat () to concating the series of thresholds and series of accuracy.

Select the threshold as 0.45 as an optimum value and define the predictions as 0 if x is less than the thresholds else one for prediction of x.

Threshold value 0.3 and define the predictions as 0 if x is less than the thresholds else one for prediction of x.

Threshold value 0.2 and define the predictions as 0 if x is less than the thresholds else, one for prediction of x.



SMOTE Technique

```
# Copying the clean data for further analysis.
data_new=data.copy(deep=True)
data_new.head()
```

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5
0	1	20000.0	2	2	1	24	2	2	0	0	...	0.0	0.0	0.0	0.0	689.0	0.0	0.0	0.0
1	2	120000.0	2	2	2	26	0	2	0	0	...	3272.0	3455.0	3261.0	0.0	1000.0	1000.0	1000.0	0.0
2	3	90000.0	2	2	2	34	0	0	0	0	...	14331.0	14948.0	15549.0	1518.0	1500.0	1000.0	1000.0	1000.0
3	4	50000.0	2	2	1	37	0	0	0	0	...	28314.0	28959.0	29547.0	2000.0	2019.0	1200.0	1100.0	1069.0
4	5	50000.0	1	2	1	57	0	0	0	0	...	20940.0	19146.0	19131.0	2000.0	36681.0	10000.0	9000.0	689.0

5 rows × 25 columns

Figure 48 The function of the classification report has been used here to generate precision, F1 score and Recall.

The function of the classification report has been used here to generate precision, F1 score and Recall. It has used the SKlearn library and SMOTE technique. The report of classification gives data about the precision , F2 score and recall of every class and the column name support is the instance number in every class.

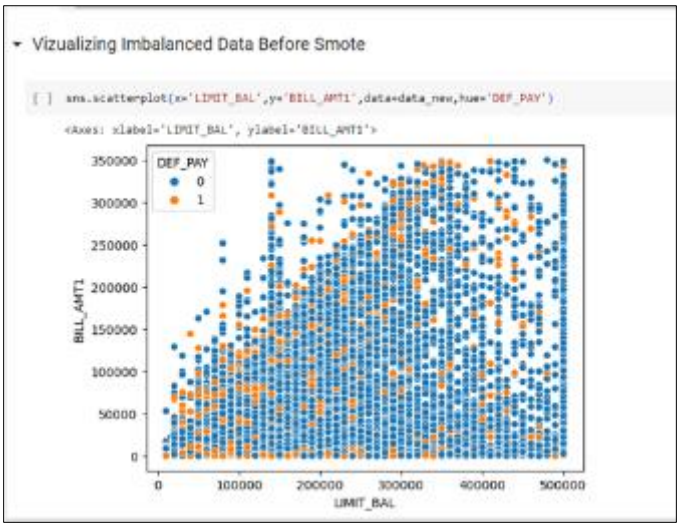
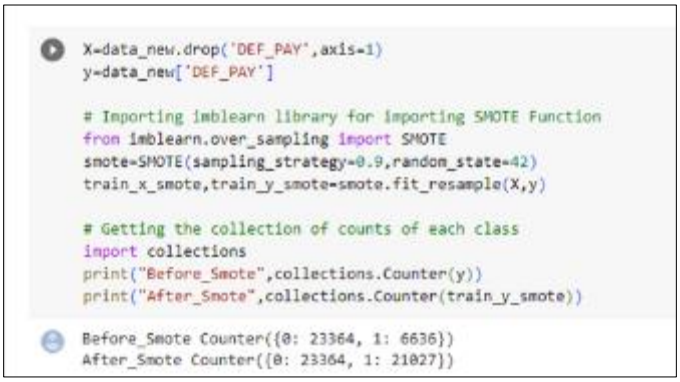


Figure 49 The X axes in the above picture show the credit limit, and the Y axes show the bill amount.

The X axes in the above picture show the credit limit, and the Y axes show the bill amount. All the points are coloured by the 'DEF_PAY' as the default status of payment. After the use of SMOTE technique, it applies the logistic regression model to balance the information by utilising the stats model library.



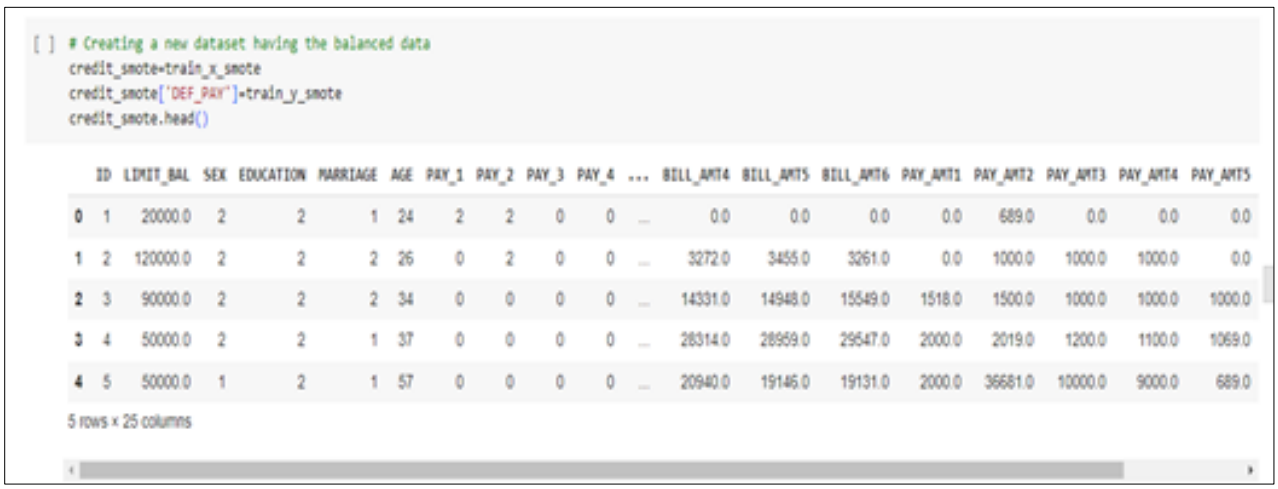
```
X=data_new.drop('DEF_PAY',axis=1)
y=data_new['DEF_PAY']

# Importing imblearn library for importing SMOTE Function
from imblearn.over_sampling import SMOTE
smote=SMOTE(sampling_strategy=0.9,random_state=42)
train_x_smote,train_y_smote=smote.fit_resample(X,y)

# Getting the collection of counts of each class
import collections
print("Before_Smote",collections.Counter(y))
print("After_Smote",collections.Counter(train_y_smote))
```

```
Before_Smote Counter({0: 23364, 1: 6636})
After_Smote Counter({0: 23364, 1: 21027})
```

Figure 50 The X axes in the coding show the credit limit, and the Y axes show the bill amount.



```
# Creating a new dataset having the balanced data
credit_smote=train_x_smote
credit_smote['DEF_PAY']=train_y_smote
credit_smote.head()
```

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5
0	1	20000.0	2	2	1	24	2	2	0	0	...	0.0	0.0	0.0	0.0	689.0	0.0	0.0	0.0
1	2	120000.0	2	2	2	26	0	2	0	0	...	3272.0	3455.0	3261.0	0.0	1000.0	1000.0	1000.0	0.0
2	3	90000.0	2	2	2	34	0	0	0	0	...	14331.0	14948.0	15549.0	1518.0	1500.0	1000.0	1000.0	1000.0
3	4	50000.0	2	2	1	37	0	0	0	0	...	28314.0	28959.0	29547.0	2000.0	2019.0	1200.0	1100.0	1069.0
4	5	50000.0	1	2	1	57	0	0	0	0	...	20940.0	19146.0	19131.0	2000.0	36681.0	10000.0	9000.0	689.0

5 rows x 25 columns

Figure 51 Creating a new dataset having the balanced data.

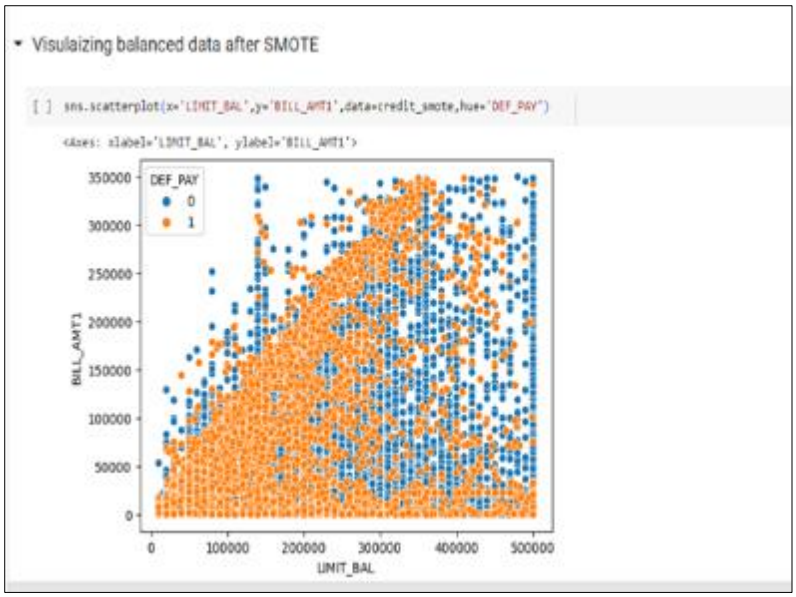


Figure 52 Visualisation of the balanced dataset.

The image attempts to assess a predictive model by using the test data and analysing its accuracy. The above datasets use the application of RandomForest so that they can predict target variables, and it uses predict method to predict the text features. The datasets create a confusion matrix to assess the performance model utilising the actual value of the target and values that are predicted. The confusion matrix is mostly dependent on the programming language and machine learning libraries. The confusion matrix shows the positive numbers and negative numbers to calculate the accuracy. The accuracy is actually the ratio of total predictions and correct predictions by calculating sensitivity and specificity.

Evaluation of various algorithms to predict the numbers of credit card defaulters has shown that the use of Random Forest is significant as it can outperform all the other algorithms to maintain a strong accuracy. It has the capability to demonstrate superior prediction by utilizing several decision trees and merging their results to conduct more accurate predictions

4. Conclusion

The purpose of an audit of financial statements is to assess evidence in an objective manner to determine the reliability of a company's annual report. Review opinions are either qualified or unqualified, depending on their level of confidence. Data mining techniques such as SVM, ANN, and KNN can be used as classification tools for auditing financial statements to improve the accuracy of the audit report, making it undetectable and free from plagiarism.

Using data mining techniques, the study used 49 variables of financial and non-financial ratios to create an accurate predictive model for assessing ratings. Despite the challenges of data collection, cleaning, and dealing with noise and missing data, experimental results show that SVM and ANN methods achieved higher average accuracy than KNN in correctly classifying companies. ANN had the lowest rate of type I error, showing a superior ability to classify healthy firms, while SVM had the lowest rate of type II error, misclassifying qualified firms into an unqualified category. Future work will explore the use of other data mining techniques, such as deep learning to improve classification and audit opinions further.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] M. Gotthardt, D. Koivulaakso, O. Paksoy, C. Saramo, M. Martikainen, and O. Lehner, 2020. Current state and challenges in the implementation of smart robotic process automation in accounting and auditing. *ACRN Journal of Finance and Risk Perspectives*.

[2] A.R. Hasan, 2021. Artificial Intelligence (AI) in accounting & auditing: A Literature review. *Open Journal of Business and Management*, *10*(1), pp.440-465.

[3] J.C. Yang, H.C. Chuang, and C.M. Kuan, 2020. Double machine learning with gradient boosting and its application to the Big N audit quality effect. *Journal of Econometrics*, *216*(1), pp.268-283.

[4] M. Schultz, and M. Tropmann-Frick, 2020. Autoencoder neural networks versus external auditors: Detecting unusual journal entries in financial statement audits.

[5] G. Dickey, S. Blanke, and L. Seaton, 2019. Machine learning in auditing. *The CPA Journal*, *89*(6), pp.16-21.

[6] C.W. Cai, M.K. Linnenluecke, M. Marrone, and A.K. Singh, 2019. Machine learning and expert judgement: analyzing emerging topics in accounting and finance research in the Asia–Pacific. *Abacus*, *55*(4), pp.709-733.

[7] S. Yoon, 2020. A study on the transformation of accounting based on new technologies: Evidence from Korea. *Sustainability*, *12*(20), p.8669.

[8] Y. Zhang, F. Xiong, Y. Xie, X. Fan, and H. Gu, 2020. The impact of artificial intelligence and blockchain on the accounting profession. *Ieee Access*, *8*, pp.110461-110477.

[9] N. Hooda, S. Bawa, and P.S. Rana, 2020. Optimizing fraudulent firm prediction using ensemble machine learning: a case study of an external audit. *Applied Artificial Intelligence*, *34*(1), pp.20-30.

[10] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, 2021, March. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

[11] F.E. Eid, H.A. Elmarakeby, Y.A. Chan, N. Fornelos, M. ElHefnawi, E.M. Van Allen, L.S. Heath, and K. Lage, 2021. Systematic auditing is essential to debiasing machine learning in biology. *Communications biology*, *4*(1), p.183.

[12] A. Barredo-Arrieta, and J. Del Ser, 2020, July. Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

[13] K. Barac, K. Plant, R. Kunz, and M. Kirstein, 2021. Generic skill profiles of future accountants and auditors–moving beyond attributes. *Higher Education, Skills and Work-Based Learning*, *11*(4), pp.908-928.

[14] M. Lokanan, V. Tran, and N.H. Vuong, 2019. Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian Journal of Accounting Research*, *4*(2), pp.181-201.

[15] Y. Zhang, F. Xiong, Y. Xie, X. Fan, and H. Gu, 2020. The impact of artificial intelligence and blockchain on the accounting profession. *Ieee Access*, *8*, pp.110461-110477.

[16] C.F. Chien, S. Dauzère-Pérès, W.T. Huh, Y.J. Jang, and J.R. Morrison, 2020. Artificial intelligence in manufacturing and logistics systems: algorithms, applications, and case studies. *International Journal of Production Research*, *58*(9), pp.2730-2731.

[17] Y. Bao, B. Ke, B. Li, Y.J. Yu, and J. Zhang, 2020. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research*, *58*(1), pp.199-235.

[18] K. Pozdniakov, E. Alonso, V. Stankovic, K. Tam, and K. Jones, 2020, June. Smart security audit: Reinforcement learning with a deep neural network approximator. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)* (pp. 1-8). IEEE.

[19] M. Gotthardt, D. Koivulaakso, O. Paksoy, C. Saramo, M. Martikainen, and O. Lehner, 2020. Current state and challenges in the implementation of smart robotic process automation in accounting and auditing. *ACRN Journal of Finance and Risk Perspectives*.

[20] S. Kruskopf, C. Lobbas, H. Meinander, K. Söderling, M. Martikainen, and O. Lehner, 2020. Digital accounting and the human factor: theory and practice. *ACRN Journal of Finance and Risk Perspectives*.

[21] D. Ucoglu, 2020. Current machine learning applications in accounting and auditing. *PressAcademia Procedia*, *12*(1), pp.1-7.

[22] C. Fieberg, M. Hesse, T. Loy, and D. Metko, 2022. Machine learning in accounting research. In *Diginomics research perspectives: The role of digitalization in business and society* (pp. 105-124). Cham: Springer International Publishing.

[23] A.R. Hasan, 2021. Artificial Intelligence (AI) in accounting & auditing: A Literature review. *Open Journal of Business and Management*, *10*(1), pp.440-465.

[24] R.A. Rahman, S. Masrom, N.B. Zakaria, and S. Halid, 2021. Auditor choice prediction model using corporate governance and ownership attributes: machine learning approach. *International Journal of Emerging Technology and Advanced Engineering*, *11*(7), pp.87-94.

[25] C.L. Jan, 2021. Detection of financial statement fraud using deep learning for sustainable development of capital markets under information asymmetry. *Sustainability*, *13*(17), p.9879.

[26] O.M. Lehner, K. Ittonen, H. Silvola, E. Ström, and A. Wührleitner, 2022. Artificial intelligence based decision-making in accounting and auditing: ethical challenges and normative thinking. *Accounting, Auditing & Accountability Journal*, *35*(9), pp.109-135.

[27] C.F. Chien, S. Dauzère-Pérès, W.T. Huh, Y.J. Jang, and J.R. Morrison, 2020. Artificial intelligence in manufacturing and logistics systems: algorithms, applications, and case studies. *International Journal of Production Research*, *58*(9), pp.2730-2731.

[28] I. Munoko, H.L. Brown-Liburd, and M. Vasarhelyi, 2020. The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics*, *167*, pp.209-234.

[29] A.K. Nawaiseh, M.F. Abbod, and T. Itagaki, 2020. Financial Statement Audit using Support Vector Machines, Artificial Neural Networks and K-Nearest Neighbor: An Empirical Study of UK and Ireland. *International Journal of Simulation--Systems, Science & Technology*, *21*(2), pp.1-6.

[30] J. Nicholls, A. Kuppa, and N.A. Le-Khac, 2021. Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape. *Ieee Access*, *9*, pp.163965-163986.