ISSN No:-2456-2165

Using Machine Learning Methods to Forecast Credit Card Approvals

Mohammad Haseeb Dar[1], Neerendra Kumar[2]
[1,2] Department of Computer Science and IT,
Central University of Jammu, India

Abstract:- As we all know, Every day, commercial banks get a large number of credit card applications. Many of them are turned down for a variety of reasons, including large loan amounts, insufficient income, or too many queries on a person's credit record. Manually assessing these programs is tedious, time-consuming, and error-prone. Fortunately, machine learning can automate this operation, and almost every commercial bank does it nowadays. In this paper, we have used machine learning techniques to create a prediction system for automated credit card approvals, much like actual banks do.

Keywords:- *Machine Learning, Credit Cards, Logistic Regression.*

I. INTRODUCTION

Everything has fully altered as a digital attribute in modern times. Cashless transaction activity is one of those digitalized domains. This is fairly popular nowadays, and more people are willing to do so because it lowers the risk of physically misplacing cash. As a result, several financial institutions offer cashless payment options to their customers, such as debit and credit cards. A credit card is one of the most popular solutions. Most people use credit cards to complete transaction activities since it is a convenient way to make payments. Many financial organizations, including national and commercial banks, rely on client information such as basic information, living standards, wages, yearly and monthly returns, and current sources of income. All of this information is reviewed to consider an application. This thorough examination and analysis can save the institution from incurring significant technical and non-technical losses. As we see a great expansion in this business sector, appropriate analysis is necessary to eliminate any potential risk associated with unethical consumers. Banks must include precise verification when giving credit cards to applicants. Although decision-making varies per bank, the consumer's credit score is the most common consideration examined by financial institutions. As we witness an increase in the enormous growth margin of the financial institution's credit business due to more consumers applying for credit cards, there is a need to entirely automate the process to speed up the acceptance decision by banks. This assists the bank in enhancing business while saving time and requiring less labor, which is a significant financial saving. The model must categorize credit card applicants into two groups: "No Risk Present," which means the bank can lend money with the assurance that the customer will repay and the bank will incur no risk or loss, and "Risk Present," which implies the bank should not grant any credit because there is a high possibility of consumer fraud and financial loss for banks. This classification is done by considering various factors of the consumer like age, salary, the number of years he/she is been working, yearly income, assets, source of income, credit score, repay history, and existing loan dues. These entire mechanisms are not only applicable to a single consumer but also to businesses whether large scale or small scale. In the past, there were various methods introduced to examine the loan history of the consumer and to improve the precision of credit scores [1] and [2]. These are data mining models that can be classified as data that depends on statistical distribution and data that does not necessarily depend on data distribution. The logistic regression and linear regression analysis models are the best examples of models that rely on data distribution. Analysis of the linear regression model [3] is used in generating credit scores but this analysis is not favourable since data considered for approving and declining are completely different. Logistic regression supports the data unlike the linear model for parametric tests. Other decision support tools like decision trees [3] and vector machine support [4] In machine learning, non-parametric tests are used. In recent data mining research, hybrid approach-based solutions have produced the best outcomes. The neural network approach is thought to be a better approach for boosting the accuracy of credit score prediction.[5]. This study provides a model that predicts whether a financial institution will issue or deny a credit card to an applicant. Even if each bank's judgment is distinct and based on organization-designed regulations, certain common qualities are evaluated and those features are taken into account when the algorithm is implemented. The data was obtained from the University of California repository, which is a publicly accessible machine learning repository. Machine learning pre-processing techniques and data transformation techniques such as scaling, handling of missing values by predefined methods 7 known as mean imputation and label encoding for numeric and non-numeric data, dividing data into test and train sets, applying classifiers, and concluding the paper obtained results are further examined using metrics such as confusion matrix to examine the accuracy of the result are implemented.

ISSN No:-2456-2165

A. Machine Learning

It is the most rapidly evolving application of artificial intelligence. It discusses how a computer or system can carry out a task based on a set of text or data without the need for human intervention. This reduces the workload for the people and supports more automation which can save time and money in the long run. We see an enormous amount of data is being tackled by the system in very few seconds and by incorporating machine learning techniques into this fast-paced system can help in building a model. Modelling can analyse and predict patterns in high-quality and dense multiplex data. It gives insight into risk and profits to any business model because machine learning is all about linking and associating data links and predicting relations among them which can be used while taking decisions for the future of the business model. Machine learning algorithms perform their task by recognizing the patterns of the data There are two methods for spotting patterns: supervised and unsupervised learning. In supervised learning, labelled data is processed by a machine-learning algorithm to generate the desired trained model. After generating the model, unknown data is used on the model to generate the result. The well-known machine learning algorithms that use the concept of supervised learning are Random Forest, Linear Regression, Decision Trees, K-nearest Neighbours, Logistic Regression, and Naïve Bayes. Since the output is obtained based on the conclusion drawn from the labelled data, this mechanism can be used to solve real-world problems. 8 In unsupervised Learning, unlabeled data is processed by a machine-learning algorithm to generate the trained model. Since data is unknown, the trained model will identify patterns in the data. It aids in categorisation by recognising distinguishing characteristics. Clustering and association are mechanisms opted in unsupervised learning. Large scale industries that need to handle high volumes of data more efficiently like the Financial Service industry, Health Care, Retail, Government, and Transportation uses this technology to process high amount of data to obtain predictions.

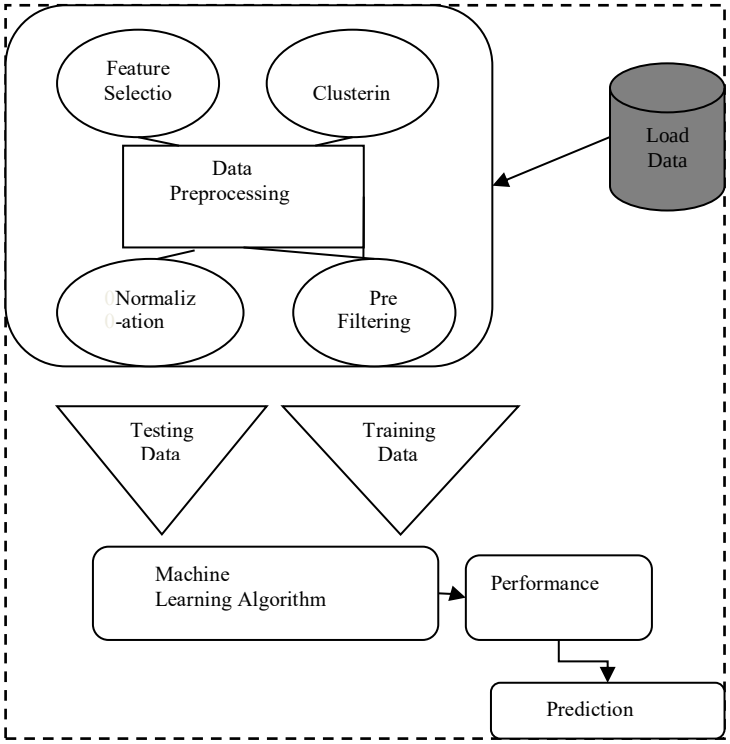

Fig.1. Working on basic data mining.

B. Predictive analytics

Predictive analytics involves a series of steps to predict the future which is based on the current as well as historical data. The predictive analytics approach is illustrated in Figure 2 below.[6]

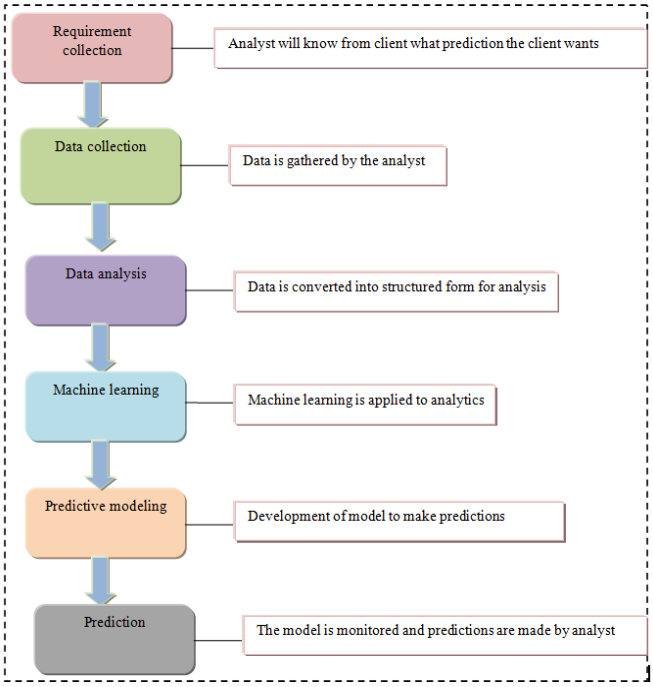

Fig.2. A generic model of predictive analytics[7]

ISSN No:-2456-2165

II. LOGISTIC REGRESSION MODEL

For credit card analysis, logistic regression (LR) is one of the most widely used statistical approaches [8] [9] [10]. It forecasts the probability of a single outcome in two different states (i.e. a dichotomy). The outcome is dependent on the circumstances and the use of one or more indications (numerical and categorical). It looks for the optimal fit parameter, according to [11] the likelihood of a binary response depending on one or more criteria features. For each credit card, independent variables were used. It gives you a probability that you can use to categorize things the acceptance or rejection of the application [8]. If the likelihood is high, It is approved if the value is greater than the threshold value. Otherwise, it isn't worth it if turned down. The client is passed as an input to the LR function outputs and characteristics of the likelihood.

$$p = \frac{\exp(\beta 0+\beta 1.x1+\cdots+\beta n.xn)}{1+exp(\beta 0+\beta 1.x1+\cdots+\beta n.xn)} \quad \text{.....(1)}$$

wherein the preceding The chance of default is denoted by p.

The explanatory factor I is xi I am the explanatory factor's regression coefficient. The number of explanatory variables is denoted by n. It is known whether each of the existing data points. The client has either accepted or declined (i.e. p=1 or p=0). The goal of this is to discover the coefficients 0, 1, 2,...,n in such a way that the model's default probability matches the detected default probability

III. LITERATURE SURVEY

Zopounidis et al. [12] used a multi-criteria analysis approach that included multi-group discrimination. An iterative binary segmentation approach underpins this method. The authors used a case study to demonstrate the method's efficacy. A credit card industry application was deployed. First, they go through a two-stage procedure. They separated the credit card applications that were accepted from the rest, while in the second, they separated the applications that needed more examination from the ones that didn't. Those who have been s rejected. Nash and Sinkey et al. [13] argued that the credit card market has gotten a lot of attention because of the enormous profits made on credit cards and the high premiums paid on credit card receivables resale. They attempted to calculate risk-reputation profile for bank credit-card offerings, as well as to investigate the impact of intangible assets in establishing creditworthiness. Credit card receivables resale premiums. Lucas et al. [14] Managing credit cards is a difficult task. One element that contributes to this complication is that credit cardholders use their cards for a variety of non-credit purposes. Reasons include payment ease, financial smoothing, and regular bill payment Emergencies, and impulsive spending In addition, handling credit card consumers involves income, chum, and bad debt issuers, as well as the explosion of in recent years competition, has made it increasingly difficult to recruit and keep lucrative employees. Customers who pose little danger. Nelson et al. [15] report, that in the United States, 24 percent of total credit card purchases in 2020 were overdue. Although the ratio has decreased from 28 percent in 2019, it still amounts to 866 billion dollars in total. Surprisingly, this figure exceeds the GDP of numerous nations, including Saudi Arabia, Turkey, Switzerland, Poland, Thailand, Sweden, and Belgium. Mehrabi, N et al. [16] focused that, supervised learning algorithms are trained to be clever by employing information gained from previous data. As a result, the machines are extremely likely to be influenced by past data and learning algorithms. C. Luo et al. [6] maximize prediction accuracy, it is utilized as a classification and regression technique. B. Mallick et al. [7] Argued that the expansion of the internet has resulted in a considerable increase in credit card usage. It is one of the most popular means of payment. Nowadays Credit card fraud is increasing at an alarming rate as the global economy grows. V. N. Vapnik et al. [17] reviewing hundreds or thousands of fraudulent and non-fraudulent credit card activity records, and SVM may learn t distinguish fraudulent credit card activity. SVM was was invented first. G. Rushin et al. [8] said one of the most often used statistical approaches for credit card analysis is logistic regression (LR). Christopher, et al. [18]focused on developing new inputs by aggregating common transactional characteristics while comparing SVM to several alternative paradigms in tackling credit card fraud. Siddhartha, et al. [19] evaluated numerous classic fraud detection models, he discovered that logistic regression was the most accurate. the conventional techniques SVM have gotten a lot of press recently. Due to its outstanding effectiveness as a classifier in recent studies Unlike SVMs are based on structural models and are ANNs that reduce empirical risk. Minimization of risk. To change the data, they utilize a nonlinear mapping. Data is entered into a multidimensional feature space. Hassibi, et al. [20] identified fraudulent transactions, authors utilized a feed-forward ANN, which proved to be successful. SVM based on structural risk minimization, unlike ANNs that minimize empirical risk. They turn the input data into a multidimensional feature space via a nonlinear mapping. TU, Yi, et al. [21] categorize a series of card transactions as fraudulent or not, the scientists employed a convolution neural network, a feed-forward neural network inspired by the animal visual brain. Seeja, K. R, et al. [22] suggested a methodology for detecting credit card fraud based on frequent item mining. Ghosh, Sushmita, et al. [23] Proposes one of the neural network-based fraud detection systems was trained on a large sample of labeled credit card account transactions and evaluated on holdout data set that comprised of all account activity over a following two months, according to the authors. When compared to previous approaches, the method produced good results, with hig fraud account detection and a factor of 20 fewer false positives.

ISSN No:-2456-2165

TABLE 1: Comprehensive overview of the review of recent papers

S. No.	Referencing	Author	Year	Method	Dataset	Accuracy
01	[12]	Zeponidis. C et al.	2001	Multi-criteria decision, binary segmentation	Five real-world applications	ACC= 97.96%
02.	[13]	Lucas.An et al.	2001	Predictive model	Public	_
03.	[15]	Nash. R et al.	1996	Risk return profiles, hidden assets hypothesis	Data from reports of condition and reports of income	ROA= 88.65%
04.	[6]	Luo. C et al.	2016	Restricted Boltzman machines	CDS dataset	ACC=87.75%
05.	[17]	Huang. P et al.	2011	Auto-correlation patterns for feature extraction	Multi-sensor, multi-modal realistic dataset collected in Arizona	ACC= 86.092%
06.	[8]	Rushin. G et al.	2017	Logistic regression, Gradient boosted tree	Dataset produced by bank containing 8 million account level transactions	AUC=0.773
07.	[18]	Whitrow. C et al.	2009	Random forest, SVM	Public	ROC= 0.090
08.	[21]	Bhattach.S et al.	2010	Random forest, SVM	Dataset obtained from international credit card operation	Precision=0.778 AUC=0.942 ACC= 94.7%
09	[20]	Hassibi. K et al.	2018	Neural network	Falcon consortium data	Account false positive rate(AFPR)= 0.89
10.	[21]	Fu. K et al.	2016	CNN-based fraud detection framework	Real credit card transaction data from a commercial bank	Entropy=0.35
11.	[22]	Seeja. K et al.	2014	Class imbalance problem	UCSD data-mining contest 2009 dataset	Mathews correlation coefficient(MCC)=0.070
12.	[23]	Ghosh.s et al.	1994	Neural network	Millions of banks' credit card dataset	_
13.	[24]	Mittal.S et al.	2019	Naïve bayies, SVM, KNN	Public	Sensitivity=0.82 Specificity= 0.97

IV. DATASETS

Datasets have a significant role in the performance prediction of credit card approvals. Plentiful of databases are available for evaluating the performance of where generalization depends on the discrepancy in samples of a dataset[15]. Based on the comprehensive review, the below table gives an overview of the current existing initiatives for evaluation. The detail of the datasets is reviewed as in Table II.

Table II. A comparative summary of available datasets

S.no.	Dataset	Samples	Reference
01.	5 financial applications	Credit card= 150 Country risk=143 Credit risk=60 Cooperate acquisition=60 Business failure prediction=80	[12] [25]
02.	CDS dataset	$12 trillion	[6]
03.	Multi-sensor, multi-modal realistic dataset	26 scenarios	[17]
04.	Dataset from two banks Bank(A) Bank(b)	175 million transactions 1.1 million transactions	[18]
05.	Dataset U	100.00 transactions	[21]
06.	Falcon consortium data	1,000,000 transactions/day	[20]
07.	Real credit card transaction dataset	260 million transactions	[8]
08.	UCSD dataset	40918 transactions	[22]
09.	Mellon bank dataset	50 data-files	[23]
10.	Public dataset	284,807 transactions	[24]

V. RESEARCH CHALLENGES/GAPS

i) Showing less accuracy: Most of the existing techniques show less accuracy and take a lot of time for execution. So, there is a requirement of applying suitable Ensemble learning techniques to show the highest accuracy and will take less amount of time.

ii) Lack of dataset: There is less information regarding the availability of datasets in the existing literature.

iii) Feature extraction: Limited amount of work has been done on the existing. The challenge here is to train the machine so that we can easily classify or predict credit card approvals. The proposed research aim is to create an effective model to overcome the accuracy and execution time.

VI. CONCLUSIONS

To predict future outcomes in several scenarios, logistic regression is widely used. Machine Learning Algorithms are best suited and powerful techniques that can be used for prediction problems**.** In this study, a focused analysis of the contemporary trends in Machine learning-based mechanisms has been presented. Various classification algorithms can be used to build a model and compare the rates or levels of accuracy to improve the model for better use. These features can improve the model to be more effective and can help the institutes to make better decisions so that they can avoid experiencing fraud and loss. Currently, factors considered are regular details related to gender, age of the consumer, his/her credit reports and worthiness, yearly income, and the number of years he/she has been working. Further, to improve this work, various other factors or conditions can be considered like their history related to any offense and their assets which can be both physical and liquid cash.

REFERENCES

[1]. P. Dennis and G. Garsva, "Selection of Support Vector Machines based classifiers for credit risk domain," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3194–3204, 2015, DOI: 10.1016/j.eswa.2014.12.001.

[2]. M. Ala'Raj and M. F. Abbod, "Classifiers consensus system approach for credit scoring," *Knowledge-Based Syst.*, vol. 104, pp. 89–105, 2016, DOI: 10.1016/j.knosys.2016.04.013.

[3]. W. Sun, Y. Song, H. Zhao, and Z. Jin, "A Face Spoofing Detection Method Based on Domain Adaptation and Lossless Size Adaptation," *IEEE Access*, vol. 8, pp. 66553–66563, 2020, DOI: 10.1109/ACCESS.2020.2985453.

[4]. S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015, DOI: 10.1016/j.ejor.2015.05.030.

[5]. G. Vinodhini and R. M. Chandrasekaran, "A comparative performance evaluation of a neural network-based approach for sentiment classification of online reviews," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 28, no. 1, pp. 2–12, 2016, DOI: 10.1016/j.jksuci.2014.03.024.

[6]. C. Luo, D. Wu, and D. Wu, "A deep learning approach for credit scoring using credit default swaps," *Eng. Appl. Artif. Intell.*, vol. 65, no. September, pp. 465–470, 2017, doi: 10.1016/j.engappai.2016.12.002.

[7]. Sourabh and B. Arora, "A Review of Credit Card Fraud Detection Techniques," vol. 45, no. 1, pp. 485–496, 2022, DOI: 10.1007/978-981-16-8248-3_40.

[8]. G. Rushin, C. Stancil, M. Sun, S. Adams, and P. Beling, "Horse race analysis in credit card fraud - Deep learning, logistic regression, and Gradient Boosted Tree," *2017 Syst. Inf. Eng. Des. Symp. SIEDS 2017*, pp. 117–121, 2017, DOI:

10.1109/SIEDS.2017.7937700.
[9]. L. Deng, X. He, and J. Gao, "Deep stacking networks for information retrieval," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 3153–3157, 2013, DOI: 10.1109/ICASSP.2013.6638239.
[10]. Y. Shen, X. He, and J. Gao, "Shgd14," pp. 373–374, 2014.
[11]. S. Arora and M. P. S. B. V. Mittal, "A robust framework for spoofing detection in faces using deep learning," *Vis. Comput.*, 2021, DOI: 10.1007/s00371-021-02123-4.
[12]. C. Zopounidis and M. Doumpos, "Multi-group discrimination using multi-criteria analysis: Illustrations from the field of finance," *Eur. J. Oper. Res.*, vol. 139, no. 2, pp. 371–389, 2002, DOI: 10.1016/S0377-2217(01)00360-5.
[13]. R. C. Nash and J. F. Sinkey, "On competition, risk, and hidden assets in the market for bank credit cards," *J. Bank. Finance.*, vol. 21, no. 1, pp. 89–112, 1997, DOI: 10.1016/S0378-4266(96)00030-1.
[14]. A. Lucas, "Statistical challenges in credit card issuing," *Appl. Stoch. Model. Bus. Ind.*, vol. 17, no. 1, pp. 83–92, 2001, DOI: 10.1002/asmb.433.
[15]. A. Khan and S. K. Ghosh, "Machine Assistance for Credit Card Approval? Random Wheel can Recommend and Explain," pp. 1–14, 2021, [Online]. Available: http://arxiv.org/abs/2105.06255.
[16]. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, 2021, DOI: 10.1145/3457607.
[17]. P. Huang, T. Damarla, and M. Hasegawa-johnson, "A training algorithm for optimal margin classifier," 2011.
[18]. C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data Min. Knowl. Discov.*, vol. 18, no. 1, pp. 30–55, 2009, DOI: 10.1007/s10618-008-0116-z.
[19]. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decis. Support Syst.*, vol. 50, no. 3, pp. 602–613, 2011, DOI: 10.1016/j.dss.2010.08.008.
[20]. K. Hassibi, "Chapter 9 Detecting Payment Card Fraud with Neural Networks," pp. 141–157.
[21]. K. Fu, D. Cheng, Y. Tu, and L. Z. B, "Neural Information Processing - 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16-21, 2016, Proceedings, Part III," pp. 483–490, 2016, DOI: 10.1007/978-3-319-46675-0.
[22]. K. R. Seeja and M. Zareapoor, "FraudMiner: A novel credit card fraud detection model based on frequent itemset mining," *Sci. World J.*, vol. 2014, 2014, DOI: 10.1155/2014/252797.
[23]. S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural network," *Proc. Hawaii Int. Conf. Syst. Sci.*, vol. 3, pp. 621–630, 1994, DOI: 10.1109/hicss.1994.323314.
[24]. S. Mittal and S. Tyagi, "Performance evaluation of machine learning algorithms for credit card fraud detection," *Proc. 9th Int. Conf. Cloud Comput. Data Sci. Eng. Conflux. 2019*, pp. 320–324, 2019, DOI: 10.1109/CONFLUENCE.2019.8776925.
[25]. Y. A. U. Rehman, L. M. Po, M. Liu, Z. Zou, W. Ou, and Y. Zhao, "Face liveness detection using a convolutional-features fusion of real and deep network generated face images," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 574–582, Feb. 2019, DOI: 10.1016/j.jvcir.2019.02.014.