REDUCING RISK OF BANK PERSONAL MODELLING USING LOGISTIC REGRESSION ALGORITHM AND COMPARE WITH NAÏVE BAYES ALGORITHM

Sk. L. Muskan[1], Dr. A. Mohan[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

In the banking system, banks have a variety of products to provide, but credit lines are their primary source of revenue. As a result, they will profit from the interest earned on the loans they make. Loans, or whether customers repay or default on their loans, affect a bank's profit or loss. The bank's Non-Performing Assets will be reduced by forecasting loan defaulters. As a result, further investigation into this occurrence is essential. Because precise forecasts are essential for benefit maximisation, it's crucial to analyse and compare the various methodologies. The logistic regression model is an important predictive analytics tool for detecting loan defaulters. In order to assess and forecast, data from Kaggle is acquired.

Materials and Methods: The dataset needed for the machine learning model for Bank personal loan is acquired from Google's Kaggle Website. The dataset coloumn have the columns person name ,age, gender, personal loan details, account details and reasons. We analyze, design and implement the infrastructures of the machine learning framework and machine learning application. The dataset are imported and logistic regression algorithm and naive bayes algorithm are tested. The number of groups is 2 for two algorithms the sample size is 75 per group**. Results:** The results are acquired in the form of accuracy for the inputs provided. The IBM SPSS tool is used in order to obtain the results. from the results we have obtained, the statistical significance difference was observed between the logistic regression algorithm and has an accuracy 70.56%.the naïve bayes algorithm has an accuracy 69.01%,p=0.01 which is more accurate than the value. The independent sample t-test was performed to find the mean, standard deviation, mean statistical significance between the groups.

Conclusion: In this paper, based on results we have obtained, the logistic regression Algorithm has more accuracy than Naive Bayes Algorithm.

Keywords: Logistic regression, Naive Bayes, Loan prediction, Data analysis, Machine learning model,Innovation

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and technical Science, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

1. Introduction

Banks have many products to sell in our banking system, but their main source of income is their credit lines. As a result, they are likely to profit from the interest on the loans they make. Loans, or whether customers repay or default on their loans, affect a bank's profit or loss. The bank can minimize its Non-Performing Assets by forecasting loan defaulters. Because precise predictions are crucial for maximising earnings, it's essential to look at the different methodologies and compare them. A logistic regression model is a critical approach in predictive analytics for analysing the problem of predicting loan defaulters. Kaggle data is taken in order to investigate and predict. Logistic Regression models were used to calculate the various performance measures. Model is significantly better because it includes variables (personal attributes of customers include graduation, dependents, credit score, credit amount, credit period, and so on.) other than checking account information (which indicates a customer's wealth) that should be considered when correctly calculating the probability of loan default. As a result, by evaluating the likelihood of default on a loan, the right customers to target for loan granting can be easily identified using a logistic regression approach. The model predicts that a bank should not solely approve loans to wealthy consumers rather should also consider a customer's other characteristics, which play an important role in credit decisions and predicting loan defaulters. As the demand for products and services rises, so does the amount of capital credit given, and people are more eager to take credit than ever before. As a result, computer software has replaced the human interface as more people from all over the world (Urban, Rural, and semi-urban) push for a high demand for credit.

A Machine Learning software algorithm has been developed in order to construct a robust and efficient software algorithm that classifies individuals based on 13 characteristics (Gender, Education, Number of Dependents, Marital Status, Employment, Credit Score, Loan Amount, and others) whether they would be eligible for a loan or not. Although this is the first line of command, it will undoubtedly lower the workload of all other bank employees because the process will be automated to identify client segments and those who are qualified for a loan amount, allowing them to target those clients individually. And this will indicate whether or not the loan applicant meets the eligibility criteria for loan approval based on those 13 elements. To provide a convenient, prompt, and accurate method of selecting deserving applicants for loan eligibility. Our team has extensive knowledge and research experience that has translated into high quality publications(Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022)

The problems in the existing research of machine learning models is less accuracy. There are certain algorithms with more accuracy when comparing it with existing ones. The main aim of the study is to improve the accuracy by implementing a logistic regression Algorithm.

2. Materials and Methods

The proposed work is done in the cloud computing lab, Department of computer science and Engineering, Saveetha School of engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. The number of groups is 2 for two algorithms. The sample size is 75 per group and the group in total is 130.

The dataset named 'bank marketing analysis' is downloaded from Google's Kaggle Website bank marketing about personal loans. Bank marketing analysis -comparison of algorithms /Kaggle. The data in this dataset explains about the loan analysis by a person performed in different sectors attended using different algorithms /Kaggle.com by keeping threshold 0.05% and G power 80% and confidence level is 90% and enrolment ratio as 1.In this dataset we have information about persons and analytics about total persons taking loan and performing in different sectors. The first format provides the information about personal details and in the second format provides information about person personal bank loan details . The statistical comparison of the bank marketing analysis using two sample groups was done through SSPS version 21.0 . Analysis was done for mean, standard deviation, independent- sample T-test.

Logistic regression:

Logistic regression algorithm is one of the most popular Machine learning algorithms, which comes under the supervised learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Step 1: It predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

Step2 : it is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems,

whereas **Logistic regression is used for solving the classification problems**.
Step 3: In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
Step 4: it is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
Step 5: it can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Naïve bayes:
step 1: Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
Step 2: It is mainly used in *text classification* that includes a high-dimensional training dataset.
Step 3: Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
Step 4: **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
Step 5: Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

For the logistic regression algorithm, the test size is 40% of the total dataset and the remaining 60% is used in training data sets. Accuracy of both the algorithms are tested from sample sizes 40 to 180. The dataset used for this paper machine learning based model is obtained from google's official dataset website Kaggle.

Statistical Analysis
The statistical software used for performing analysis is IBM SPSS version 21.0. IBM SPSS is a statistical software software tool used for the analysis of data. The data sets are normalized and then the data is converted into arrays. The number of variables needed and data are visualized and analyzed and the existing algorithms are obtained. Algorithms and accuracy are dependent variables and the accuracy approximation values are independent variables.

3. Results

In table 1 a file collection of people who are having loans based on their age ,job,marital status in the first format is given . The dataset is taken from Kaggle and it contains age of the person ,job of the person, marital status of the person, account balance ,whether the person is having housing or property, and loan taken by that person using machine learning. The loan and the duration time having for that loan is represented in numerical values which are taken from that dataset conducted in the first format. The statistical comparison of the loan taken and duration using two sample groups was done through SPSS version 21. Analysis was done for mean ,standard deviation, independent T- test.

In table 2 data of people who has taken loan in the second format is given. The dataset is taken from Kaggle and it contains the age of the person , job of that person, marital status of that person, and duration of that loan campaign. The age coloumn is the respective number of each row,loan taken ,age of the person and loan taken by the person is represented in numerical format which is taken from the dataset of loan campaigns conducted in second format.

From table 3,representing the accuracy for both the algorithm with five sample sizes An overview of the collection of the accuracies of both the algorithms with different sample sizes are given. For each sample size the accuracy is calculated for both of the respective algorithms filled in respective columns. It was observed that increase in sample size increases the accuracy of the algorithms. In final average accuracy is calculated for both the algorithms. At sample size 20 the average accuracy of logistic regression is 70.01%.and naive bayes algorithm is 69.09%.

From table 4, the results we have obtained ,Statistical significance difference was observed between the Logistic regression algorithm with an accuracy of 70.01% over the naïve bayes algorithm with an accuracy of 69.09%.

In fig 1, the picture displays the outcome that is formed from the set of data . The average accuracy is easily calculated using outcome that are formed from the set of given data and data points.

In fig 2, the figure represents the working of naïve bayes algorithm. This naïve bayes algorithm explains the working in step by step and procedural manner in a structured way.

In fig 3, the bar chart plotted with the accuracies of both the algorithms for different sample sizes is represented. The bar chart is plotted by taking algorithms as X-axis and accuracy as Y-axis. From the bar chart ,it can be seen that the logistic regression algorithm is more accurate than the naïve bayes algorithm.

4. Discussion

From the results of the study we can see that logistic regression algorithm has a better performance than the naïve bayes algorithm .

logistic regression algorithm has an accuracy of 70.01% whereas naïve bayes algorithm has an accuracy of 69.09%.

There are some similar findings related to the machine learning based model in bank marketing analysis. In this paper on machine learning based model ,the model has successfully created In our model prediction of whether the loan would be accepted or not, we achieved the highest accuracy from the logistic regression model. On the dataset, the best case accuracy attained is 0.785. t bank marketing campaigns focus on competitive strategies. Some of the strategies include demographic targeting, customer outreach, loyalty programs, and technology adoption. These strategies not only help the banks to reach many customers but to sell their products to the general public. By targeting a specific group of customers, banks can achieve their organizational objectives. One of the goals is an increase in the number of subscriptions to term deposits (Grzonka et al., 2016). The literature review tries to understand the findings of various researchers. The focus of the review is the factors that can increase customers' subscription to a term deposit.

Poor quality of data is the main challenge faced in implementing the machine learning model . Obviously, if your training data has lots of errors, outliers, and noise, it will make it impossible for your machine learning model to detect a proper underlying pattern. Hence, it will not perform well So put in **every ounce of effort** in cleaning up your training data. No matter how good you are in selecting and hyper tuning the model, this part plays a major role in helping us make an accurate machine learning model."Most Data Scientists spend a significant part of their time in cleaning data.

5. Conclusion

In our model prediction of whether the loan would be accepted or not, we achieved the highest accuracy from the logistic regression model. On the dataset, the best case accuracy attained is 0.785.

Our model was able to forecast whether the applicants in the dataset would be eligible for the loan when the project was completed it was also able anticipate the loan eligibility of a specific applicant by pointing out his row number. Applicants with a high income and smaller loan requests are more likely to be approved, which makes sense because they are more likely to payback their debts. Gender and marital status, for example do not appear to be considered. The loan credibility prediction system can assist companies in making the best judgement on whether to approve or deny a customer's loan credibility prediction system can assist companies in making the best judgement on whether to approve or deny a customer's loan request This will undoubtedly assist the banking industry in establishing more effective distribution routes. It is necessary to create and test new strategies that outperform the performance of common data mining models for the domain.

Declaration:

Conflict of Interests

No Conflict of Interest In Manuscript

Authors Contributions

Author L Muskan was involved in data collection ,data analysis and manuscript writing .Author s.subbiah was involved in Conceptualization ,data validation, and critical review manuscript.

Acknowledgment:

We would like to thank our management, Saveetha school of Enginnering ,Saveetha Institute of Medical and Technical Sciences for providing facilities for our research study.

Funding:

We thank the following organizations for providing financial support that enabled us ti complete the study.

1. The big event
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha school of Engineering.

6. References

Baraneedharan, P., Sethumathavan Vadivel, C. A. Anil, S. Beer Mohamed, and Saravanan Rajendran. 2022. "Advances in Preparation, Mechanism and Applications of Various Carbon Materials in Environmental Applications: A Review." *Chemosphere*. https://doi.org/10.1016/j.chemosphere.2022.134596.

Deena, Santhana Raj, A. S. Vickram, S. Manikandan, R. Subbaiya, N. Karmegam, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2022. "Enhanced Biogas Production from Food Waste and Activated Sludge Using Advanced Techniques – A Review." *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2022.127234.

Karpagam, M., R. Beaulah Jeyavathana, Sathiya Kumar Chinnappan, K. V. Kanimozhi, and M. Sambath. 2022. "A Novel Face Recognition Model for Fighting against

Human Trafficking in Surveillance Videos and Rescuing Victims." *Soft Computing*. https://doi.org/10.1007/s00500-022-06931-1.

Kumar, P. Ganesh, P. Ganesh Kumar, Rajendran Prabakaran, D. Sakthivadivel, P. Somasundaram, V. S. Vigneswaran, and Sung Chul Kim. 2022. "Ultrasonication Time Optimization for Multi-Walled Carbon Nanotube Based Therminol-55 Nanofluid: An Experimental Investigation." *Journal of Thermal Analysis and Calorimetry*. https://doi.org/10.1007/s10973-022-11298-4.

Nagarajan, Karthik, Arul Rajagopalan, S. Angalaeswari, L. Natrayan, and Wubishet Degife Mammo. 2022. "Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm." *Computational Intelligence and Neuroscience* 2022 (April): 6461690.

Nagaraju, V., B. R. Tapas Bapu, P. Bhuvaneswari, R. Anita, P. G. Kuppusamy, and S. Usha. 2022. "Role of Silicon Carbide Nanoparticle on Electromagnetic Interference Shielding Behavior of Carbon Fibre Epoxy Nanocomposites in 3-18GHz Frequency Bands." *Silicon*. https://doi.org/10.1007/s12633-022-01825-1.

Pandiyan, P., R. Sitharthan, S. Saravanan, Natarajan Prabaharan, M. Ramji Tiwari, T. Chinnadurai, T. Yuvaraj, and K. R. Devabalaji. 2022. "A Comprehensive Review of the Prospects for Rural Electrification Using Stand-Alone and Hybrid Energy Technologies." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102155.

Venu, Harish, Ibham Veza, Lokesh Selvam, Prabhu Appavu, V. Dhana Raju, Lingesan Subramani, and Jayashri N. Nair. 2022. "Analysis of Particle Size Diameter (PSD), Mass Fraction Burnt (MFB) and Particulate Number (PN) Emissions in a Diesel Engine Powered by Diesel/biodiesel/n-Amyl Alcohol Blends." *Energy*. https://doi.org/10.1016/j.energy.2022.123806.

Whangchai, Niwooti, Daovieng Yaibouathong, Pattranan Junluthin, Deepanraj Balakrishnan, Yuwalee Unpaprom, Rameshprabu Ramaraj, and Tipsukhon Pimpimol. 2022. "Effect of Biogas Sludge Meal Supplement in Feed on Growth Performance Molting Period and Production Cost of Giant Freshwater Prawn Culture." *Chemosphere* 301 (August): 134638.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Advances in the Application of Immobilized Enzyme for the Remediation of Hazardous Pollutant: A Review." *Chemosphere* 299 (July): 134390.

Tables and Figures

Table 1. Represents the file containing details about the people who has taken loan in first format which includes housing and duration of loan

s.no	Age	Job	Education	Marital	Housing	Loan	Duration	Campaign
1	59	Admin	Secondary	Married	2343	Yes	May	1042
2	56	Admin	Secondary	Married	45	Yes	May	1389
3	41	Technician	Secondary	Married	1270	No	May	571
4	55	Admin	Secondary	Single	2476	Yes	May	673
5	54	Management	Secondary	Married	184	No	May	562
6	42	Retired	Secondary	Married	830	Yes	May	1201

Table 2. represents the file containing details about the people who have taken loan in the second format which includes balance and duration of that loan.

age	balance	day	duration	campaign	pdays	previous

count	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000
mean	41.231948	1528.538524	15.658036	371.993818	2.508421	51.330407
std	11.913369	3225.413326	8.420740	347.128386	2.722077	108.758282
min	18.000000	-6847.000000	1.000000	2.000000	1.000000	-1.000000
25%	32.000000	122.000000	8.000000	138.000000	1.000000	-1.000000

Table 3 represents the accuracy of the both the algorithms [Logistic regression algorithm average accuracy is 70.01% and Naïve bayes algorithm accuracy is 69.09%]

s.no	No.of samples	Accuracy of logistic regression Algorithm	Accuracy of Naïve bayes algorithm
1	10	70	69
2	20	69	68
3	40	67	65
4	60	65	63

Table 4 Represents the statistics taken from SPSS software.[logistic regression algorithm accuracy is 70.01% and naïve bayes algorithm accuracy is 69.09%.

	Algorithm	N	Mean	STD.Deviation	STD.Error Mean
Accuracy	Logistic Regression	10	68.9420	.72229	.22841
	Naive bayes	10	68.3510	.68502	.21662

Table 5 represents results from an independent sample test. Independent Sample T-test is applied for dataset fixing confidence intervals as 95%(Logistic regression algorithm appears to perform better than Naïve bayes algorithm)

	Levens test forEquality of variances					t-test for equality of means		95% confidence interval of the Difference	
	F	sig.	t	df	sig.(2-tailed	mean difference	std.error difference	lower	upper

Ac curacy	Equal Variances Assumed	.063	.804	1.877	18	.077	.59100	.31480	-.07036	1.25236
	Equal Variances Not Assumed			1.877	17.950	.077	.59100	.31480	-.07049	1.25249

Graph

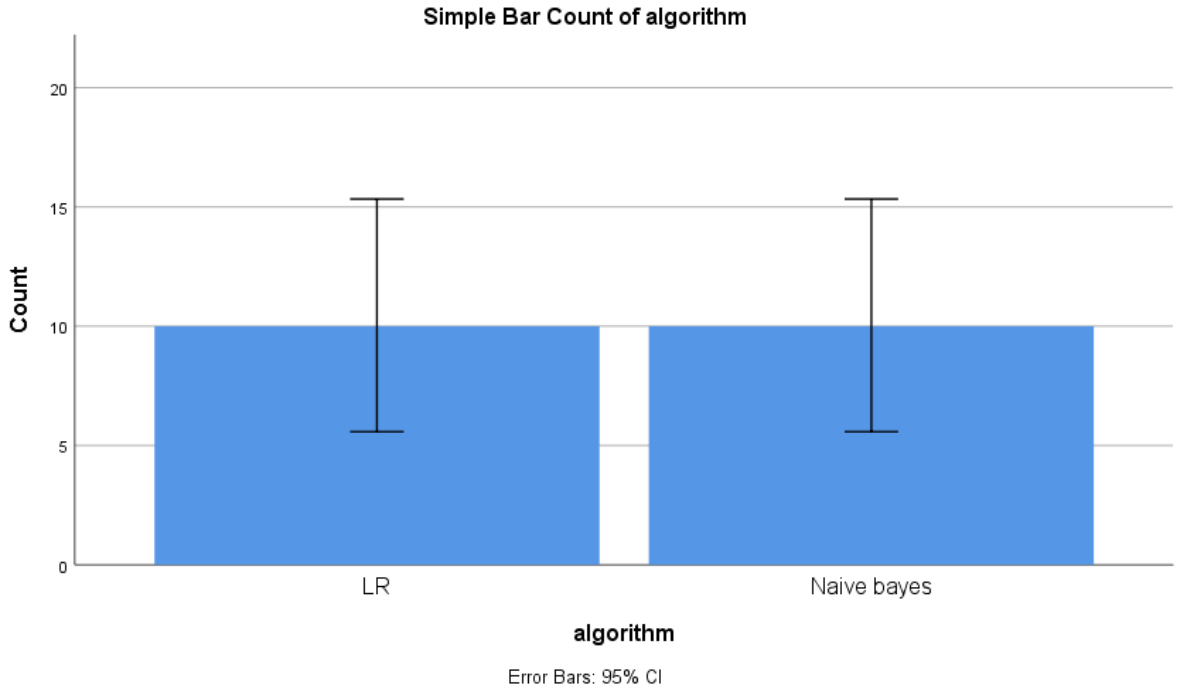

Fig 1 Barcharts represent the comparisons of mean accuracy and standard errors for Logistic regression algorithm .Logistic regression algorithm is better than real Naïve bayes algorithm in terms of mean accuracy and standard deviation .X-axis : Logistic regression vs Naïve bayes algorithm Y-axis : mean accuracy of detection += 1 SD.