GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

IDENTIFYING THE POTENTIAL CUSTOMERS FOR LOANS

A Report for the Evaluation 3 of Project 2

Submitted by

PIYUSH SHUKLA

(1613105070/16SCSE105044)

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION OF CLOUD COMPUTING AND VIRTUALIZATION

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr.Ashutosh Upadhyay, Assistant Professor

APRIL / MAY- 2020

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **" IDENTIFY THE POTENTIAL CUSTOMERS FOR LOANS"** is the bonafide work of **"PIYUSH SHUKLA (1613105070)"** who carried out the project work under my supervision

SIGNATURE OF HEAD

Dr. MUNISH SHABARWAL,

PhD (Management), PhD (CS)

Professor & Dean,

School of Computing Science & Engineering

SIGNATURE OF SUPEVISOR

Dr. SANJEEV KUMAR PIPAL,

M.Tech,PhD,

Professor

School of Computing Science & Engineering

ABSTRACT

These days it is very important in the banking sector to identify which customer is potential and which are not for the loan purpose. Different people use a different approach to find this out. In this paper, we are focusing on the machine learning algorithm to find out the customers which are potential for the loan. For the problem statement, we have chosen the open-source dataset which is of Thera Bank from Kaggle. We have tried to build different models using different classification algorithms which are there is machine learning and performance is measured using different metrics like confusion matrix, log loss, etc.

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a

healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

The department wants to build a model that will help them identify the potential customers who have higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign.

TABLE OF CONTENTS

INTRODUCTION

Various classification methods and algorithms have been developed for resolving machine learning problems, including statistical models, decision and regression trees, rules, connectionist networks, probabilistic networks. Supervised classification is a principle core of what has been recently called the data mining. The applications of supervised classification in real life are very vast, like automatic speech treatment, face detection, signature recognition, customer discovery, spam detection, systems biology etc. Many decision making problems in a variety of domains such as medical science, engineering, human sciences and management science can be considered as classification problems.

This project is about a bank in which the department of the bank wants to build a model that will help them to identify the potential customers who have higher probability of purchasing the loan. This will increase the

success ratio while at the same time reduce the cost of the campaign. Using machine learning algorithms like logistic regression, KNN and Naïve byes models are build and potential customers are found out.

Distribution of the loans is the important business part of almost every bank. The main bank assets is directly came from The profit earned from the loans distributed by the banks. The prime motive in banking sector is to invest their assets in safe hand. Today many banks approves loan after a long process of verification but still there is no surety whether the chosen applicant is the deserving or not out of all the applicants. Through this system we can predict weather which applicant is capable of returning the loan amount to bank in time. And the whole process of prediction the prime applicant of features is automated by machine learning technique. Loan predictions is very helpful for employee of bank as well as for the applicant also. The aim of this paper to help them identify the potential customers who have higher probability of purchasing the loan. It can provide advantages to the bank.

Machine learning is about prediction on unseen data or testing data. In machine learning first learn to perform a task by training dataset. Then perform the same task with the testing data [1]. In supervised learning we pass both input and output data and result is already known. Supervised learning is two types classification based and regression

based. In this paper we are using classification based supervised learning [2]. Logistic Regression is a popular and powerful supervised machine learning technique used to build a model relating the independent predictors(x variable) with the response (y variables) that is categorical in nature. Where the class is known already, it can help find factors distinguishing between records in different classes in terms of the predictor variables in the dataset [3]. KNN is a simple algorithm that stores all available cases and classifies based on a similarity measures [4]. Naïve Bayes is a classification algorithm for binary (two-class) and multi-class classification problems [5].

The implementation of the model includes sex basic steps of machine learning that are:

1. Collect data
2. Choose algorithm
3. Creating object of the model
4. Train the model by training dataset
5. Making prediction of unseen data and testing data
6. Evaluation of the model.

LITERATURE REVIEW

In this survey we study about many researchers are working on the problem profiling bank customers using different technique and different data set.

In this Research paper [6] proposed that a segmentation of bank customers using machine learning technique that is clustering technique nowdays due to large amount of customers data un the baking sector, it can detect the hidden patterns in the data set to improve the banking sector of each group of customer.

In this Research paper [7] proposed that in today's world there are large amount of dte present in unsystematic way to classify the data using supervised Artificial neural network algorithm.

In this research paper [8] proposed that a use of two algorithm as a predictive model in machine learning and data mining. In this it analysis between these two machine learning algorithms is done and the result is who is credit card holder or not.

In this research paper [9] proposed that a three classification algorithm are used, which are Naïve Byaes, Random Forest, and Decision Tree. To

increasing the accuracy of their customers profiling through classification algorithms.

The previous literature survey shows that various or different machine learning algorithm were used for predicting and classification. All of them are using machine learning algorithms. with the existing level in this work, we are using the supervised technique and use it as a target for the logistic regression, KNN Algorithm, Naïve Bayes Classifier.

METHODS AND MATERIALS

In this study, Data and attributes are taken from an open source, Anaconda platform along with libraries like Pandas and Numpy are used for data manipulation and analysis. First the data is imported from source and stored in the form of pdf file, in a structured format. From the above structured format then EDA(Exploratory data analysis) is done, with the help of EDA we saw inside the datasets, how the data represented in datasets then saw the pair plot before sub so that the user would understand the distribution then with the help of SNS.DISKPLOT we obtain the distribution plot. then we saw sound plot 1 means how much use user is potential and 0 means how much user is non potential. After this we use box plot to saw inside then we use second box plot to saw

that what number of outlier inside then after this we made a correlation matrix is used to systematic way to organize data, as an input for more advance analysis. then look at the value of everything. How is the security account we saw then what number of user are potential customers and what number are non potential customers after this we saw different accounts after this all the procedure we split our date sets into 70/30 ratio 70% of data is for training data and 30% of data is for testing data. we scale the data using standard scalar.

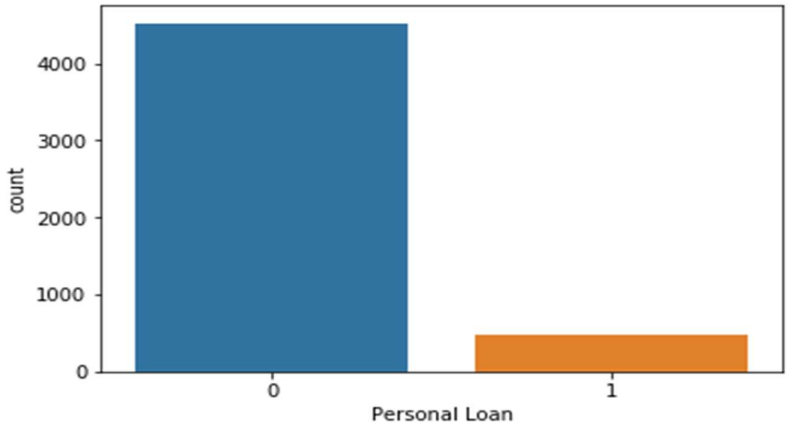

Fig [1] Personal loans

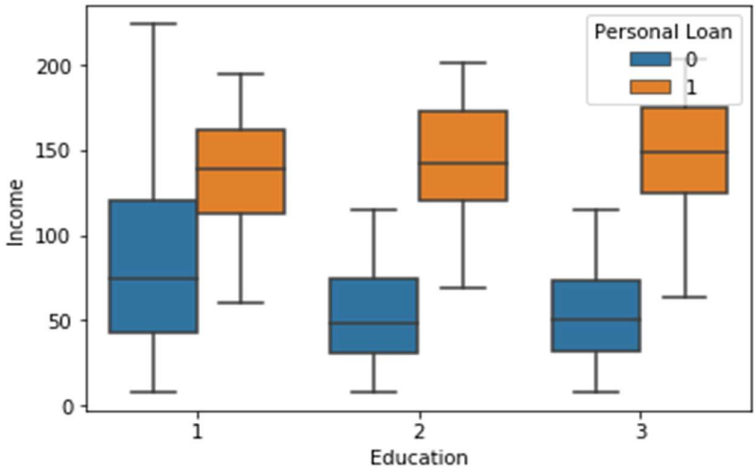

Fig [2] education loans

We have used different models of machine learning and combined them to all to obtain better predictive performance could be obtained from any of the model. Whose accuracy is high it is use for prediction.

Method that can be followed to achieve the goal is given below:

1) Pre-processing: Pre-processing of data is the mandatory step which is followed in machine learning before applying any machine learning algorithm. Exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

2) Visualization: Second step is the visualization of the data the data is analysed visually using pair plot Fig [3]. Pair plot allows us to see both

distribution of single variables and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis

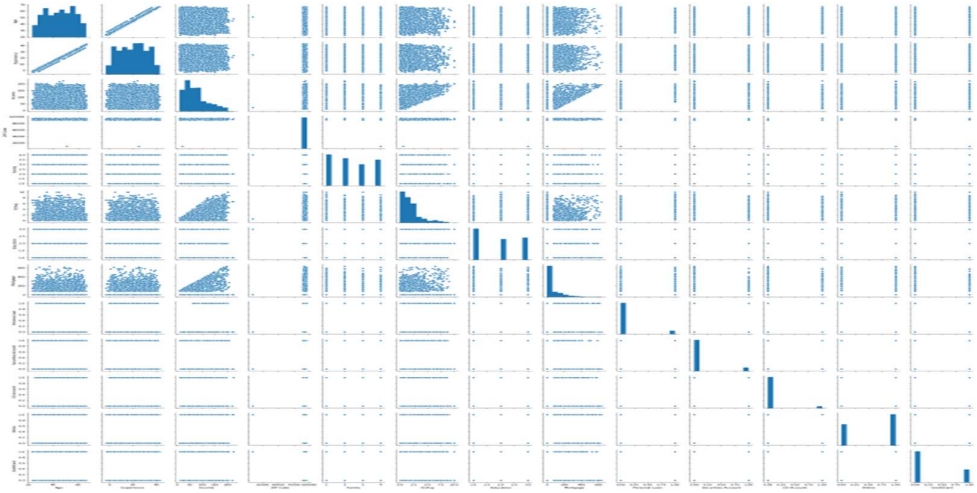

Fig [3] Pair plot

The three models are used in this study are discussed below:

1) **Logistic regression:** In this we use logistic regression with default parameters with L2 penalty. Logistic Regression is a popular and powerful machine learning method used to create a model with respect to the independent predictors (x) with the response variable (y) being naturally distributed. Where the class is known already, it can help find factors distinguishing between records in different classes in terms of the predictor variables in the dataset. Where the outcome variable has only two classes. Binomial logistic regression is used and multinational logistic regression is used when we have more than two classes.

Structural regression is a mathematical method and provides a detailed summary of the calculations depending on the significance of the predictor variables and how each predictor variable contributes to the availability of the Y-variance categories. These unique attributes make this algorithm highly relevant to the Banking and Finance domain to provide a detailed and quantitative description of the forecast variables.

```
Log_reg = LogisticRegression()
log_reg.fit(x_train,y_train)
y_pred_logregression = log_reg.prdict(x_test)
```

2) **KNN algorithm:** The k-nearest neighbor (KNN) is a simple, easy-to-use machine-controlled learning algorithm that can be used to solve both partitioning and imaging problems. The KNN algorithm assumes that the same objects exist in close proximity. In other words, the same things are closer to each other. KNN incorporates a sense of similarity (sometimes called distance, proximity, or proximity) to a specific statistic we can learn in our childhood - calculating the distance between points on a graph. In this KNN use different neighbor first two different neighbor then it increase the number of neighbor max a range of different neighbor it use range 1 to 20 for the best accuracy.

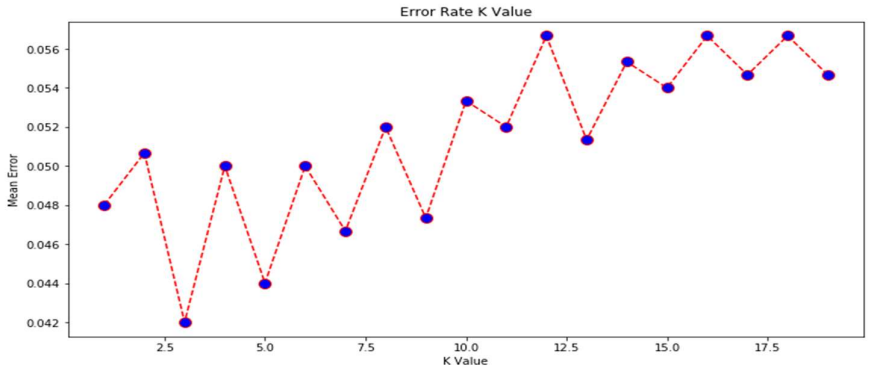

Fig (4)

In above figure we can see that the lowest value of this error we can see that lowest value of this error is when the value equals 5 so we implementing KNN Algorithm for K=5.

```
Knn=KNeighboursClassifier(n_neighbour=5)
Knn.fit(X_train,y_train)
pred_KNN = Knn.predict(X_test)
confusion_matrix_Knn=confusion_matrix(y_test,pred_KNN)
print(confusion_matrix_Knn)
```

3) **Naïve byes classifier:** Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or

categorical input values In this we use multinominal Naïve byes classifier with no additional parameter best default parameters.

```
Pred_GNB=clf_GNB.predict(X_test)
confusion_matrix_GNB=confusion_matrix(y_test,pred_GNB)
print(confusion_matrix_GNB)
```

RESULTS

Logistic regression: 0.9513333333333334

KNN Algorithm: 0.95333

Naïve Bayes classifier: 0.8846666666666667

Implementing three different models and looking at the confusion matrix. We can say tha KNN predicted rightly, more number of true positives and negatives and the sum of false positive and negative is lower for the KNN model. KNN has a high score compared to all the others models which is 95.5333% .

CONCLUSION

This study designed for analysis and identifying the potential customers for loans and verify tha who is potential customers or non portential customer.

1. Age seems to be distributed normally with no outliers.

2. Experience seems to be distributed normally as well with no outliers.

3. Income seems to be right skewed and contains a lot of outliers.

4. Family also does not have any outliers with median being 2.

5. CCavg column contains a lot of outliers is right skewed.

6. Mortgage column again is right skewed and has a lot of outliers. 7. Out of the total customers, 4478 users do not have security account whereas 522 customers do have the security account.

8. Out of the total customers, 4698 users do not have CD account whereas 522 customers do have the CD account.

9. Out of the total customers, 2984 use internet banking facilities whereas 2016 do not.

10. Out of the total customers, 3530 do not use credit card whereas 1470 use credit card.

REFERENCES

[1] I.Louridas,P, and Ebort, C, (2016).Machine learning software,33(5),110-115, doi:10.1109/me.2016.114

[2]Osisanwo F.Y.*1,Akinsola J.E.T.*2,Awedele O.*3,Hinmikaiye J.O.*4,Olakanmi O.*5,Akinjobi J.**6 *Department of computer science, Babcock University, Hishan Remo, Ogun State, Nigeria.

[3]CHAO-YING JOANNE PENG , KUK LIDA LEE, GARY M. INGERSOLL Indiana University-Bloomington

[4] Gongde Gue 1,Hui Wang1, David Bell2,Yaxin Bi2, and Kieran Greer1. School of computing and mathermatics,University of Ulster Newtownabbey, BT37 0QB,Northern Ireland,UK1. School of Computer Science, Queen,s University Belfast Belfast, BT71NN,UK2

[5] Pouria kaviani1,Mrs.Sunita Dhotre 2. 1.M.Tech student, Department of Computer Engineering,Bharati Vidyapeeth University, College of Engineering, pune. 2.Associate Professor,Department of Computer Engineering, Bharati Vidyapeeth University, Collegeof Engineering,pine.

[6]M.Sharahi,M.Aligholi,"Classify the data of bank customers using data mining and clustering techniques(case study:Sepah bank branches Tehran)",J.Appl.Environ,Biol.

[7] P. S. Patil, N. V. Dharwadkar, "Analysis of banking data using machine learning", Proc. Int. Conf. IoT Social Mobile Analytics Cloud (I-SMAC), pp. 876-881, Feb. 2017.

[8] N. H. Niloy, M. A. I. Navid, "Naïve Bayesian classifier and classification trees for the predictive accuracy of probability of default credit card clients", Amer. J. Data Mining Knowl. Discovery, vol. 3, no. 1, pp. 1, 2018.

[9] S. Palaniappan, A. Mustapha, C. F. M. Foozy, R. Atan, "Customer profiling using classification approach for bank telemarketing", Int. J. Inform. Vis., vol. 1, no. 2, pp. 214-217, 2017.