

MACHINE LEARNING TECHNIQUES FOR QUALITY DECISION MAKING IN BANKING SECTOR

Fauziya N. Shaikh Student, NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai Fauziyashaikh605@gmail.com	Simran P. Shinde Student, NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai imranshinde35@gmail.com	Dr. Jayalekshmi K.R Associate Professor, M.C.A. Dept. SIMS, Navi Mumbai jlekshmi76@gmail.com
--	---	---

MACHINE LEARNING TECHNIQUES FOR QUALITY DECISION MAKING IN BANKING SECTOR

Fauziya Shaikh¹, Simran Pradeep Shinde², Dr. Jayalekshmi K.R.³

¹NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai-400706

²NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai--400706

³NCRD's Sterling Institute of Management Studies, Neru, INavi Mumbai-400706

ABSTRACT

An increasing number of customers are applying for personal loans for their purchases. Personal loans help the consumer to meet any shortfall they experience in buying a needful item. fulfill their needs consumer apply for loans. As many Loan Application increases, day by day Bank needs to detect and analyze which customer is eligible for a personal loan. The term banking can be defined as receiving and protection money that is deposited by the individual or the entities. The primary objective of the bank is to provide their wealth in safer hands. In recent times, banks approve the loan after verifying and validating the documents provided by the customer. Yet there is no guarantee whether the applicant is deserving or not. This paper classifies customers based on certain criteria. In this paper, the main focus is to identify and analyze the risk of giving a loan in the banking sector. The Machine Learning techniques are used to analyze risk giving loan, it helps in summarize into a piece of valuable information. This will improve the quality of banking system thus improving customer retention. Applying it on a dataset of the customer and predicting the risk percentage for an individual to give loan. These techniques facilitate useful data analysis and can help to get

better outcomes into the processes behind the data. The banks and many investment companies are pioneers in taking advantage of Machine Learning. The main objective is to predict whether assigning the loan to a particular person will be safe or not.

Keywords: Machine learning, Classification, Prediction , Support Vector Machine , Decision-tree

INTRODUCTION

In today's world, as the price of products are increasing, the no. People opting for a loan are increasing. There are a variety of reasons for customers to take a loan for which they approach banks like home loan, personal loan, etc. Its bank's responsibility to check all the credentials and documents related to the scenario and provide people with a loan as per their requirement. Sometimes it happens such that persons opting for a loan do not have any valid documents or are not able to satisfy other requirements needed to lend loan, such as unbalanced income and expenditure ratio, etc. In such scenarios bank needs to take a final decision according to the customer's requirements. Granting a loan is the core business part of the banking sector. Revenue is generated by lending money at rates higher than the cost of money lent to prevent possible losses and bankruptcy, effective machine learning techniques can be used. Loan prediction has been performed in the literature using various techniques including machine learning, data mining, and hybrid techniques. These techniques support companies to identify, predict and retain churning customers, help in decision making and CRM. The decision trees are the most commonly recognized methods used for prediction of problems associated with the loan prediction.

OBJECTIVE

Customer Loan prediction for its eligibility based on secondary data and applying techniques such as Logistic Regression, SVM and decision tree classifier which provides results. The most obvious reason to monitor a consumer profile is that banks want to avoid or prevent loan losses. To find desired output consumer monitoring is, therefore, necessary to detect which loans are likely to become risky, and which loans might default and lead to financial loss. The main task is to identify the main attributes, those attributes result in the diversion of results. As Machine Learning tools such as Loan approval prediction has been performed in the literature using various techniques including machine learning, data mining. These

techniques support the bank to identify, predict and retain details customers, helps in decision making. The classification used in its prediction of problems associated with the customer.

LITERATURE REVIEW

Kumar Arun, Garg Ishan, Kaur Sanmeet (2016), "Loan Approval Prediction based on Machine Learning Approach" published in National Conference on Recent Trends in Computer Science and Information Technology(NCRTCSIT) in their paper few attributes used with parameter setting for ML model (Decision Tree, RR, SVM, AdaBoost and NN)and predict that numbers cases of computer glitches, errors in content and most important weight of features are fixed in automated prediction system.

X.Francis Jency, V.P.Sumathi, Janani Shiva Sri(2018), "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients" published in International Journal of Recent Technology and Engineering (IJRTE) in their paper they classify and analyse the nature of the loan applicants. From the graphs, many conclusions have been made and information was inferred such as short-term loan was preferred by the majority of the loan applicants and the clients majorly apply loan for debt consolidation.

E. Chandra Blessie, R. Rekhathey(2019), "Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process " published in International Journal of Recent Technology and Engineering (IJRTE) use LR, SVM, DT technology to predict the loan model accuracy. By properly analyzing positive qualities and constraints, it can be concluded with confidence that the Naïve Bayes model is extremely efficient and gives a better result when compared to other models.

RESEARCH METHODOLOGY

The analysis is done on secondary data . Here we used the Jupyter Notebook with python code. Jupyter Notebook is a JSON document, as a result of versioned schema, and containing an ordered list of input/output cells which can contain code, text, graph, plots and rich media, and various library usually ending with the ". ipynb" extension. The main components are the kernels and a dashboard. A kernel is a program that runs and examines the user's code. The

Jupyter Notebook App provides a kernel for Python code, but there are also different kernels available for other programming languages.

We import the various in build library to perform the preprocessing and classification of data. We have applied the PCA algorithm to reduce the data into dimensions to visualize the classification of data using different classification techniques. After splitting the training data set and test data set now it supplied to a machine learning model, based on this data set the model is trained. Every new consumer details filled at the time of loan request application form acts as a test data set. After the operation performed on the testing data set, the model predicts whether the new consumer is a fit person for approval of the loan or not based upon the basis of the analysis it con

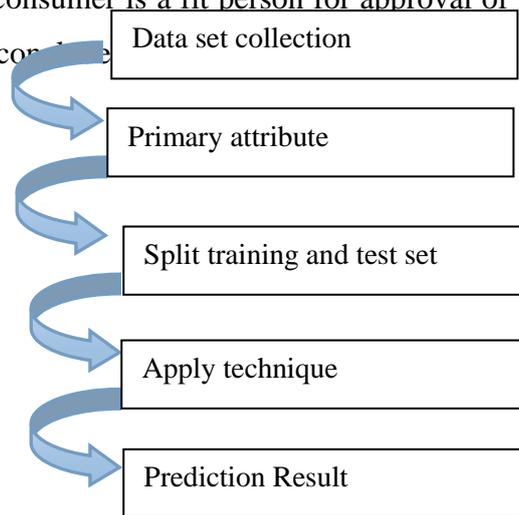


Figure-1

In the study python pandas, Matplotlib, Jupyter notebook etc are used for the model creation. Source of secondary data <https://www.kaggle.com/>

DATA PREPROCESSING

For achieving better results from the applied model in Machine Learning the format of the data has to be in a proper manner. Data set is often incomplete and inconsistent, and /or lacking in certain behaviour or trends, and is likely to contain many errors. In python, we can solve such issue using data pre-processing method. On this data set, we have used the

IsNull() method and filled the missing value with the majority and mean of the numeric values.

- **Code to check missing values:**

```
train.apply(lambda x: sum(x.isnull()),axis=0)
```

- **Filling the missing values with majority:**

```
train.Gender = train.Gender.fillna('Male')
```

```
train.Married = train.Married.fillna('Yes')
```

```
train.Dependents = train.Dependents.fillna('0')
```

```
train.Self_Employed = train.Self_Employed.fillna('No')
```

```
train.LoanAmount = train.LoanAmount.fillna(train.LoanAmount.mean())
```

```
train.Loan_Amount_Term = train.Loan_Amount_Term.fillna(360.0)
```

```
train.Credit_History = train.Credit_History.fillna(1.0)
```

- **Convert the coded form of categorical features into numeric values:**

```
labelencoder_train = LabelEncoder()
```

- **Devide the dataset into the training dataset and the testing datasets**

```
F_train, F_test, s_train,s_test = train_test_split(F, s, test_size = 1/3, random_state = 0)
```

Different models are considered for predicting the Loan eligibility. The loan eligibility prediction models are created by using Logistic regression, Decision Tree and Support Vector Machine(SVM).. The data set of 615 records are split into training and test set with tests set consisting of 25% of the data and training set with 75% data. Here we applied the the PCA for dimentionality reduction.

- **PCA**

```
pca = PCA(n_components = 2)
```

```
F_train = pca.fit_transform(F_train)
```

```
F_test = pca.fit_transform(F_test)
```

explained_variance = pca.explained_variance_ratio_

The data set after preprocessing consists of the following attributes.

Variable	Description	Type
Loan_ID	Unique Loan ID	object
Gender	Male/ Female	object
Married	Applicant married	object
Dependents	Number of dependents	object
Education	Graduate/Not Graduate	object
Self_Employed	Self Employed (Y/N)	object
ApplicantIncome	Applicant income	Int64
CoapplicantIncome	Coapplicant income	Float6
LoanAmount	Loan amount in	Float6
Loan_Amount_Term	Term of loan in months	Float6
Credit_History	credit history(Y/N)	Float6
Property_Area	Urban/ Semi Urban/	object
Own House	Self/Rent	object
Loan_Status	Loan Approved(Y/N)	object

Table-1

- **Logistic regression**

Logistic regression is a Machine Learning classification algorithm used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary(0 and 1) variable that contains data coded. Logistic regression model predicts $P(S=1)$ as a function of F . The results are saved in the variable “s_test” and later the accuracy score is measured and printed.

Code:

```
from sklearn.linear_model import LogisticRegression

#Model

LogisticRegression()
```

```
# Predicting the Test set results  
s_pred = classifier.predict(F_test)  
  
#Print accuracy  
  
print('The accuracy of Logistic Regression is: ', metrics.accuracy_score(s_pred, s_test))
```

- **Decision Tree**

Decision Tree is one of the most best, popular and powerful algorithm. Decision-tree algorithm is the type of the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. In python, sklearn is a machine learning package which include a lot of algorithms. Here, we are using some of its modules like train_test_split, DecisionTreeClassifier and accuracy_score.

Code:

```
#Model  
  
DecisionTreeClassifier()  
  
# Predicting the Test set results  
s_pred = classifier.predict(F_test)  
  
#Print accuracy  
  
print('The accuracy of Decision Tree Classifier is: ', metrics.accuracy_score(s_pred, s_test))
```

- **Support Vector Machine**

Support Vector Machine (SVM) is a supervised machine learning algorithm used to perform classification, regression and even outlier detection. The linear SVM classifier performed by drawing a straight line between two classes.

Code:

```
from sklearn.svm import SVC  
  
#Model  
  
SVC()  
  
# Predicting the Test set results
```

```
s_pred = classifier.predict(F_test)
#Print accuracy
print('The accuracy of SVM is: ', metrics.accuracy_score(s_pred, s_test))
```

RESULTS & VISUALIZATION



Figure-2

Models are analyzed for performance of prediction. The prediction model is evaluated by accuracy. Accuracy of the model indicates the instances that are correctly classified. Accuracy level in Jupyter :Notebook for the given set are:

The accuracy of Logistic Regression is: 57.56 %

The accuracy of SVM is: 70.73 %

The accuracy of Decision Tree Classifier is: 42.43 %

Accuracy depends on how data is collected, and judged on basis of comparison of several parameters Hence on the basis the SVM is the best model for prediction.

CONCLUSION

This study will help the bank to check the loan eligibility criteria of the customer based on customer detail provided while filing a loan request form. To use this process, they have given a problem to identify the customer's attributes, those are eligible for loan amount so that they can specifically target these customers. Here we used the secondary data set by adding the few different attributes in them. To address this problem the predictive models

that are created and they proved that we can give loan to the new consumer-based one primary attribute. This will help the bank employee to detect that the user is able to repayment the loan. Here we use the PCA and SVM to get the more accuracy level, which helps the bank to minimize or prevent future losses. Reference research paper helps us to choose the right attribute to provide an accurate and reliable result.

FUTURE ENHANCEMENT

As the result are predicted on a small amount of dataset, the result diversion may vary in a large amount of data. Also, the key attributed used are primary attribute so in this research paper feature selection is not performed hence the number of attributes increases the feature may change accordingly which will affect the deviation of accuracy level.

REFERENCES:

[1] Kumar Arun, Garg Ishan, Kaur Sanmeet, “Loan Approval Prediction based on Machine Learning Approach” e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 18-21

[2] An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients X.Francis Jency, V.P.Sumathi, Janani Shiva Sri [Nov 2018]

[3] “Loan Sanctioning Prediction System”by Aditi Kacheria, Nidhi Shivakumar, Shreya Sawkar, Archana Gupta [Sept 2016]

[4] E. Chandra Blessie, R. Rekha[Nov 2019],“Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process “

[5]Web References:

I. <https://www.kaggle.com/>

II. <https://github.com/>

III. <https://datahack.analyticsvidhya.com/>