Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)
IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

An Approach for Prediction of Loan Approval using Machine Learning Algorithm

Mohammad Ahmad Sheikh
School of Comuting Science & Engineering
Galgotias University
Greater Noida . India
ahmadsheikhjnp@gmail.com

Amit Kumar Goel
Professor, School of Comuting Science & Engineering
Galgotias University
Greater Noida, India
amit.goel@galgotiasuniversity.edu.in

Tapas Kumar
Professor, School of Comuting Science & Engineering
Galgotias University
Greater Noida, India
tapas.kumar@galgotiasuniversity.edu.in

***Abstract*— In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits.A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results.Model is marginally better because it includes variables (personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc.) other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.**

Keywords—loan, outlier, Prediction, component,Overfitting, Transform

I. INTRODUCTION

This paper has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this research is to predict the safety of loan [1][3]. To predict loan safety, the logistic regression algorithm is used. First the data is cleaned so as to avoid the missing values in the data set. To train our model data set of 1500 cases and 10 numerical and 8 categorical attributes has been taken. To credit a loan to customer various parameters like CIBIL Score (Credit History), Business Value, Assets of Customer etc has been considered. List of parameters as shown below:

Qualification	Categorical
In Service / Business Owner	Categorical
Individual income of Applicant	Qualitative
Individual income of Co-Applicant (if Any)	Qualitative
Amount of Loan required	Qualitative
Term for which loan Required	Qualitative
Credit History of Applicant	Qualitative
Area of Property	Categorical

II. LITERATURE SURVEY

Logistic Regression is a popular and very useful algorithm of machine learning for classification problems. The advantage of logistic regression is that it is a predictive analysis. It is used for description of data and use to explain relationship between a single binary variable and single or multiple nominal, ordinal and ration level variables which are independent in nature.

The model development for the prediction is taken in account using the sigmoid function in logistic regression as the outcome is targeted binary either 0 or 1 [11][15]. The dataset of bank customers has been divided into training and test data sets.. The train dataset contains approximately 600+ rows and 13+ columns whereas the test dataset contains 300+ rows and 12+ columns, the test dataset does not contain the target variable. Both the datasets are having missing values in their rows, and the mean, median or mode is used to fill the missing values but not removing the rows completely because the datasets are already small. Using the Feature Engineering techniques, the project is further proceeded and move towards the exploratory data analysis, where the dependent and independent variable is studied through statistics concepts such normal distribution, Probability density function etc. Study of the univariate, bivariate and multivariate analysis will give the view of the inside dependent and independent variable[13]14]. The model is focusing on to target those customers who are

eligible for loans and therefore the logistic regression is enabled using the sigmoid function as it divided the probability into binary output. Therefore the Prediction model can be developed.

III. Problem Statement

Banks, Housing Finance Companies and some NBFC deal in various types of loans like housing loan, personal loan, business loan etc in all over the part of countries. These companies have existence in Rural, Semi Urban and Urban areas. After applying loan by customer these companies validates the eligibility of customers to get the loan or not. This paper provides a solution to automate this process by employing machine learning algorithm. So the customer will fill an online loan application form. This form consist details like Sex, Marital Status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others. To automate this process by using machine learning algorithm, First the algorithm will identify those segments of the customers who are eligible to get loan amounts so bank can focus on these customers [4][7].

Loan prediction is a very common real-life problem that every finance company faces in their lending operations. If the loan approval process is automated, it can save a lot of man hours and improve the speed of service to the customers. The increase in customer satisfaction and savings in operational costs are significant[9]. However, the benefits can only be reaped if the bank has a robust model to accurately predict which customer's loan it should approve and which to reject, in order to minimize the risk of loan default

IV. Proposed Model

Prediction of granting the loan to the customers by the bank is the proposed model. Classification is the target for developing the model and hence using Logistic Regression with sigmoid function is used for developing the model. Preprocessing is the major area of the model where it consumes more time and then Exploratory Data Analysis which is followed by Feature Engineering and then Model Selection. Feeding thetwo separate datasets to the model, and then preceding the model.

Logistic regression is a type of statistical machine learning technique/algorithm which is used to classify the data by considering outcome variables on extreme ends and tries to make a logarithmic line that distinguishes between them. By this way prediction can be made through Logistic Regression.

A. Data Collection

Data has been collected from the Kaggle one of the most data source providers for the learning purpose and hence the data is collected from the Kaggle, which had two data sets one for the training and another testing[12]. The training dataset is used to train the model in which datasets is further divided into two parts such as 80:20 or 70:30 the major datasets is used for the train the model and the minor dataset is used for the test the model and hence the accuracy of our developed model is calculated.

B. Pre Processing

Data mining technique has been used in Pre-Processing for transforming raw data which is collect using online form into useful and efficient formats.There is a need to convert it in useful format because it may have some irrelevant,missing information and noisy data. To deal with this problem data cleaning technique has been used.

Before data mining the data reduction techniques is used to deal with huge volume of data. So data analysis will become easier and it intends to get accurate results. So data storage capacity increase and cost to analysis of data reduces.

The size of data can be reduced by encodingmechanisms. So itmay be lossy or lossless. If the original data is obtained after reconstruction from compressed data, such reductions are called lossless reduction else it is called lossy reduction. Wavelet transforms and PCA (Principal Component Analysis) methods are effective for reduction.

ID	0
Sex	13
Married	3
No_Dependents	15
Qualification	0
In Service / Self_Employed	32
Annual_Income_Applicant	0
Annual_income_Coapplicant	0
Amount_Loan	22
Term	14
Credit_History _ Applicant	50
Assets	0
Status_Loan	0

C. Feature Engineering

In feature engineering a proper input dataset which is compatible as permachine learning algorithm requirements is prepared. In our model **Pandas** and **Numpy** library has been imported to run. So the performance of machine learning model improves.

```
import pandas as pd
import numpy as np
```

D. List of Techniques:

*1) Imputation:*There is one more measure problem i.e. missing values when data is prepared for our machine learning model. There may be many reason of missing values like human errors, interruptions in flow of data, security concerns, and so on. The performance of machine learning model severely affected by missing values.

```
train['Gender'].fillna(train['Gender'].mode()[0],inplace=Tr
ue)
train['Married'].fillna(train['Married'].mode()[0],inplace=
True)
train['Dependents'].fillna(train['Dependents'].mode()[0],in
place=True)
```

ID	0
Sex	0
Married	0
No_Dependents	0
Qualification	0

In Service / Self_Employed	0
Annual_Income_Applicant	0
Annual_income_Coapplicant	0
Amount_Loan	22
Term	14
Credit_History _ Applicant	0
Assets	0
Status_Loan	0

2) Handling Outliers: To detect the outliers the data is demonstrated visually and afterwards handled the outliers. When the ouliers decisions visulaized are of high precision and accurate. Percentiles is another mathematical method to detect outliers. In this method,it assumes a certain percentage of value from top or taken it from bottom as an outlier. The key point is here to set the percentage value once again, and this depends on the distribution of your data as mentioned earlier.

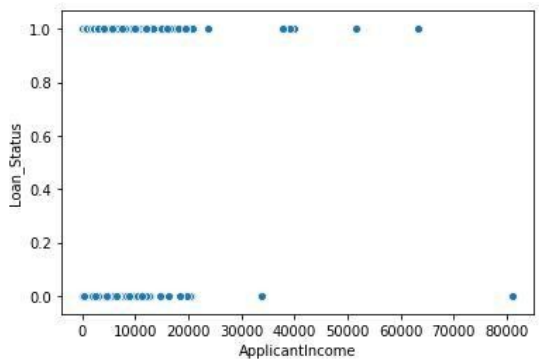
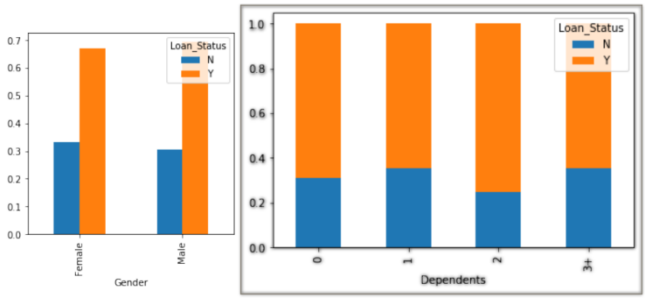


Fig. 1. Application income vs Loan Status

*3) Binning:*The key point between performance and overfitting is binning. In my opinion, for numerical values columns, except very few overfitting cases, binning might be redundant for some kind of algorithms, due to its effect on the performance of model. However, for categorical columns, the labels which have low frequencies might affected from the robustness of statistical models in negative manner. After assigning a comman category to all these less frequent values helps to keep the model robust.



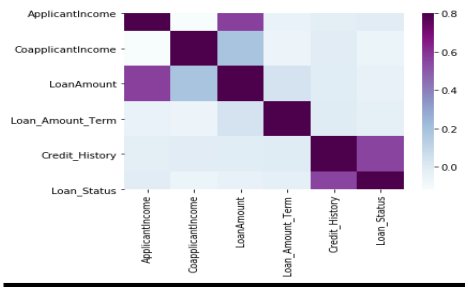
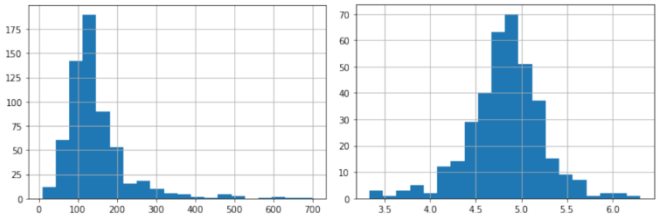


Fig. 2. Heap Map

*4) Log Transform:*Logarithm transformation (or log transform) is very common mathematical transformations technique in feature engineering. The benefit of log transformation is to handle skewed data and after transformation distribution becomes more approximate to normal. Log transformtion decreases the effect of the outliers, due to the normalization of magnitude differences and machine learning model becomes more robust.



Before Log Transform After Log Transform

*5) One Hot Encoding:*One hot encoding is commonly used encoding methods of machine learning. After using this method the values spreads in a single and multiple columns having values 0 and 1. These values shows a relation between encoded and group columns. When the categorical data by using this method has been changed then it would be difficult to understand for algorithms, to a numerical format and enables to group the categorical data without losing any of the information.

from sklearn.preprocessing import LabelEncoder
number=LabelEncoder()

G	M	D	E	SE	AI	CAI	LA	CH	LAT	PA	LAL
1	0	0	0	0	5849	0	128	360	1	2	4.85203
1	1	1	0	0	4583	1508	128	360	1	0	4.85203

G	Sex
M	Married
D	No_Dependents
E	Qualification
SE	In Service / Self_Employed
AI	Annual_Income_Applicant
CAI	Annual_income_Coapplicant

LA	Amount_Loan
CH	Credit_History _ Applicant
LAT	Loan Amount Transfer
PA	Assets
LAL	loan Amount log

V. Model Selection

The process of selecting a final machine learning model from among a group of candidate machine learning models for aparticular training dataset of Loan customer is called model selection.

There are different types of model like logistic regression, SVM, KNN, etc. All these models have some merits and demerits for example predictive error gives the statistical noise in the data, the incompleteness of the sample data, and the limitations of each different model type. The chosen modelmeets the requirements and constraints of the stakeholders (Bank and Customers) project stakeholders. A model should have parameters like

- Skillful as compared to naive models.
- Skillful relative to other tested models.
- Skillful relative to the state-of-the-art.

Thus, Prediction of loan approval is a type of a classification problem and hence this model is used.

```
from sklearn.linear_model import LogisticRegression model = LogisticRegression()

model.fit(x_train, y_train)
```

VI. Model Evaluate

Model evaluation is technique which is used for the evaluating the performance of the model based on some constraints it should be kept in mind while evaluating the model that it can't underfoot or overfit the model. Various methods are present to evaluate the performance of the model such as Confusion metrics, Accuracy, Precision, Recall, F1 score etc.

1) Confusion Metrics:

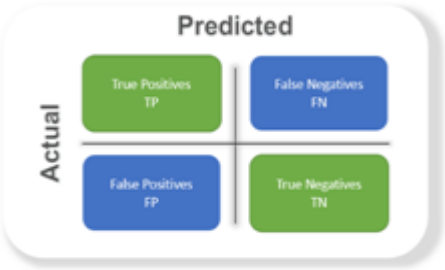


Fig. 3. Confusion Matrix

2) Accuracy:

Accuracy of the model has been measured by predefined metrics. In a balance class model shows high accuracy but in the case of unbalanced class the accuracy is very less.

$$\frac{(TP + TN)}{(TP + FP + TN + FN)}$$

3) Precision:

Percentage ratio of positive instances and total predicted positive instances gives precision value. In the below equation denominator represents the model positive prediction done from the whole given dataset. Precision value tells the perfectness of our model. In our data set good precision value has been obtained.

$$\frac{TP}{TP + FP}$$

4) Recall:

Percentage ratio of positive instances with actual total positive instances is recall value. Here denominator (TP + FN) shows the total number of positive instances which are present in whole dataset. As a result it has obtained 'how much extra right ones, the model will failed if it shows maximum right ones'.

$$\frac{TP}{TP + FN}$$

5) F1 Score:

The harmonic mean (HM) of precision and recall values is called F1 Score. Model will be best performer if it shows maximum F1 Score. Numerator shows the product of precision and recall if one goes low either precision or recall, the final F1 score goes down significantly. So a model does well in F1 score if the positive predicted (precision) having positive valueand doesn't miss out on positives and predicts them negative (recall).

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

VII. Conclusion

The process of prediction starts from cleaning and processing of data, imputation of missing values, experimental analysis of data set and then model building to evaluation of model and testing on test data. On Data set, the best case accuracy obtained on the original data set is 0.811. The following conclusions are reached after analysis that thoseapplicants whose credit score was worstwill fail to get loan approval, due to a higher probability of not paying back the loan amount. Most of the time, those applicants who have high income and demands for lower amount of loan are more likely to get approved which makes sense, more likely to pay back their loans. Some other characteristic like gender and marital status seems not to be taken into consideration by the company.

References

[1] Toby Segaran, "Programming Collective Intelligence: Building Smart Web 2.0 Applications." O'Reilly Media.

[2] Drew Conway and John Myles White," Machine Learning for Hackers: Case Studies and Algorithms to Get you Started," O'Reilly Media.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer ,Kindle

[4] PhilHyo Jin Do ,Ho-Jin Choi, "Sentiment analysis of real-life situations using loca- tion, people and time as contextual features," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[5] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[6] Bing Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," Cambridge University Press, ISBN:978-1-107-01789-4.

[7] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," Procedia computer science, 111:376–381, 2017.CrossRef.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, "An Efficient Indexing Algorithm for CBIR,"IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel ,"Face Detection and Recognition System using Digital Image Processing" , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher.

[10] Amit Kumar Goel, Kalpana Batra, Poonam Phogat," Manage big data using optical networks", Journal of Statistics and Management Systems "Volume 23, 2020, Issue 2, Taylors & Francis.

[11] Raj, J. S., & Ananthi, J. V., "Recurrent neural networks and nonlinear prediction in support vector machine" Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.

[12] Aakanksha Saha, Tamara Denning, VivekSrikumar, Sneha Kumar Kasera. "Secrets inSource Code: Reducing False Positives usingMachine Learning", 2020 InternationalConference on Communication Systems &Networks (COMSNETS), 2020.

[13] X.Frencis Jensy, V.P.Sumathi,Janani Shiva Shri, "An exploratory Data Analysis for Loan Prediction based on nature of clients", International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018.

[14] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,"Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques. Volume 5 Issue 2, Mar-Apr 2019

[15] Nikhil Madane, Siddharth Nanda,"Loan Prediction using Decision tree", Journal of the Gujrat Research History,Volume 21 Issue 14s, December 2019.

.