Loan Prediction Using Machine Learning

*

Kunal Shah
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, USA
kshah88@stevens.edu

***Abstract*—In today's world, taking Home loans from financial institutions has become a very common phenomenon. Everyday large number of applications for loans are sent to financial institutions and some of the applicants may not be not reliable. So finding out to whom the loan can be granted is a day-to-day process that is performed by financial institutions.**

A financial institution manually validates the home loan application for eligibility. This manual process is cumbersome for them, due to hundreds of applications coming every day. So the aim of this project is to automate the loan eligibility process to identify potential customers who are eligible for loan. This is done by making predictive model using different machine learning techniques.

I. Introduction

Importance of loans in our day-to-day life has increased to a great extent. People are becoming more and more dependent on acquiring loans, be it education loan, housing loan, car loan, business loans etc. from the financial institutions like banks and credit unions. However, it is no longer surprising to see that some people are not able to properly estimate the amount of loan that they can afford. The consequences of such scenarios are late payments or missing payments, defaulting or in the worst-case scenario not being able to pay back those bulk amount to the banks.

Dream Housing Finance Company is dealing with all home loans and has presence across all urban, semi urban and rural areas. Assessing the risk involved in a loan application is one of the key concerns of this finance company for survival in the highly competitive market and for profitability. This company receives number of loan applications from their customers and other people on daily basis. All these applicants may not be reliable thus everyone cannot be approved. Company use their own risk assessment techniques in order to analyze the loan application and to make decisions.

Company wants to automate this loan eligibility process. Fortunately, new technology is making that increasingly possible. In this project, Machine learning algorithms will be used to study the loan-applications data and extract patterns, which would help in classifying the person who is eligible for loan from people who are not eligible for loan, thereby helping the company to specifically target those customers who are eligible. Company has provided data about the customers and different machine learning algorithms would be applied on this dataset to extract patterns and to obtain results with desired accuracy.

The remaining parts of this paper are organized as follows: In Section 2, I briefly discuss the related work to Loan Prediction Problem. Then, Section 3 describes about the each variable in the dataset ,Section 4 is about python packages used for working with data, Section 5 shows loan prediction methodology, Section 6 is about Exploratory Data Analysis, Section 7 discuss strategy used to preprocess data, Section 8 discuss feature engineering and then feature removal process, Section 9 and 10 is about machine learning models used for the project and the comparison of each models in terms of accuracy , their advantages and disadvantages. Finally, section 11 and 12 discusses interesting directions for future research and summarizes the paper respectively.

II. Related Work

The model proposed in "M. Sudhakar, and C.V.K. Reddy, "Two Step Credit Risk Assessment Model for Retail Bank Loan Applications Using Decision Tree Data Mining Technique" an effective prediction model for predicting the credible customers who have applied for bank loan. Decision Tree is applied to predict the attributes relevant for credibility. This prototype model can be used to sanction the loan request of the customers or not.

The model proposed in "J.H. Aboobyda, and M.A. Tarig, "Developing Prediction Model of Loan Risk in Banks Using Data mining", Machine Learning and Applications" has been built using data from banking sector to predict the status of loans. This model uses three classification algorithms namely j48, bayes Net and naïve Bayes. The model is implemented and verified using Weka. The best algorithm j48 was selected based on accuracy,

The paper by "Atiya, Amir F. "Bankruptcy prediction for credit risk using neural networks: A survey and new results." explains the implementation of Artificial Neural Networks on the Bank dataset for predicting Bankruptcy.

Sarwesh Site, Dr. Sadhna K. Mishra (2013) proposed a method in which two or more classifiers are combined together to produce an ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique.

TABLE I
DATA SET DESCRIPTION

Variable	Description
Loan ID	Unique Loan ID
Gender	Male/Female
Married	Applicant Married(Y/N)
Dependents	Number of Dependents
Education	Applicant Education(Graduate/Under Graduate)
Self Employed	Self Employed(Yes/No)
ApplicantIncome	Applicant Income
CoapplicantIncome	coapplicant Income
LoanAmount	Loan Amount In (Thousands)
Loan Amount Term	Term of Loan in months
Credit History	Credit History meets guidelines
Property Area	Urban/Semi Urban / Rural
Loan Status	Loan Approved(Y/N)

III. DATA SET DESCRIPTION

The training data set is feed to machine learning model; on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new applicant is eligible for approval of the loan or not based upon the target function it learns from the training data sets. For this practice problem from Analytics Vidhya, we have been given three CSV files: train, test and sample submission.

TABLE I is the description for each variable.

- Train file:
  will be used for training the model, i.e. model will learn from this file. It contains all the independent variables and the target variable and it has around 700 data points.
- Test file:
  contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data and it has around 300 data points.
- Sample submission file:
  contains the format in which we have to submit our predictions.

IV. PYTHON PACKAGES

Following Python libraries I have used for working with data in Python.

Loan Prediction Methodology

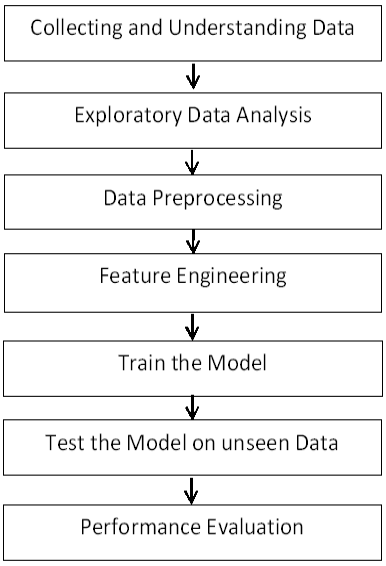

Fig. 1. Loan Prediction Steps

- Pandas : Data Manipulation and Analysis
- Numpy: Numerical Calculations
- Seaborn and Matplotlib: Data Visualizations.

V. LOAN PREDICTION METHODOLOGY

Figure 1 describes the prediction methodology. Firstly the information is collected and understanding the data is carried out then the exploratory data analysis has done on the data set , In third Step data preprocessing is taken place where missing value , outliers and stardization are happened. In the fourth step, the feature importance is carried out. It makes the model accurate and more efficient. In the fifth and sixth step, the machine learning approaches, were trained and tested on the unseen data. Finally, evaluation is done using Accuracy metric.

VI. EXPLORATORY DATA ANALYSIS(EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

I have done Univariate and Bivariate Analysis to understand each feature and relationship between target variable and independent variable. To perform EDA analysis I have removed missing data points from data set.

- Univariate analysis:
  It is the simplest form of analyzing data where I examine each variable individually. For categorical features I have used frequency table or bar plots which will calculate the number of each categorical value in a particular variable. For numerical features, I have used probability density plots to look at the distribution of the variable.

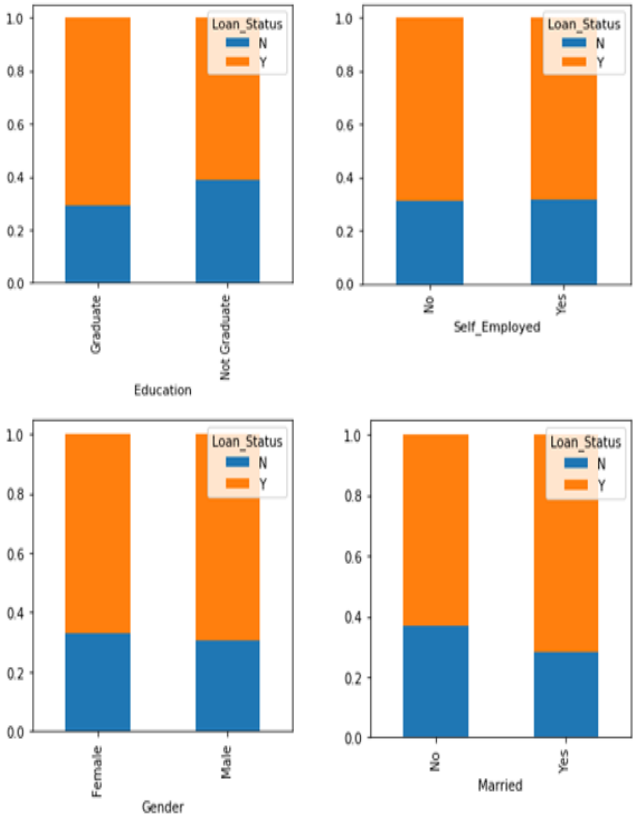

Fig. 2. Bivariate Analysis

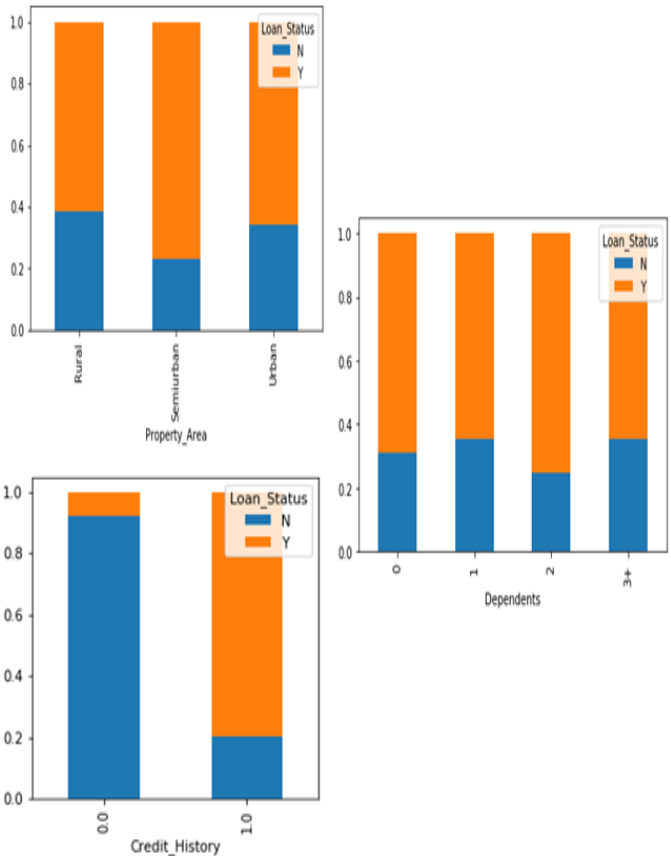

Fig. 3. Bivariate Analysis

- Bivariate Analysis:
  Fig. 2 and 3 shows bivariate analysis plots for categorical features. After looking at every variable individually in univariate analysis, I explored them again with respect to the target variable. I have used the stacked bar plot which will give the proportion of approved and unapproved loans for each categorical variable.Following are the observations.1). proportion of male and female applicants is more or less same for both approved and unapproved loans. 2).Proportion of married applicants is higher for the approved loans. Distribution of applicants with 1 or 3+ dependents is similar across both the categories of Loan Status. There is nothing significant we can infer from Self Employed vs Loan Status plot 3). we can infer that some one who is graduate has higher chances of loan approval than some who is not graduate. 4).It can be infered people with credit history as 1 are more likely to get their loans approved compared to 0. 5). Proportion of loans getting approved in semiurban area is higher as compared to that in rural or urban areas.
- Heat Map:
  Fig. 4 is the visualization of Heat Map. I have used the heat map to visualize the correlation between all the numerical variables. Heat maps visualize data through variations in coloring. The variables with darker color means their correlation is more. The most correlated variables are (ApplicantIncome - LoanAmount) and (Credit History - Loan Status). LoanAmount is also correlated with CoapplicantIncome.
- Pair Plot:
  Fig. 5 is the visualization of Pair Plot. A pairs 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. I have used the seaborn library to visualize pair plot.. The distribution curve on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship (or lack thereof) between two variables. The importance of credit history features can be inferred from the plot and it's effect on target variable.
- Note:
  Please see the jupyter notebook for all the analysis plots.

VII. Data Preprocessing

Before building any machine learning model data preprocessing is one of the critical step in data mining process which deals with cleaning of initial data set. Data preprocessing is the most time consuming phase of a data mining process. I have

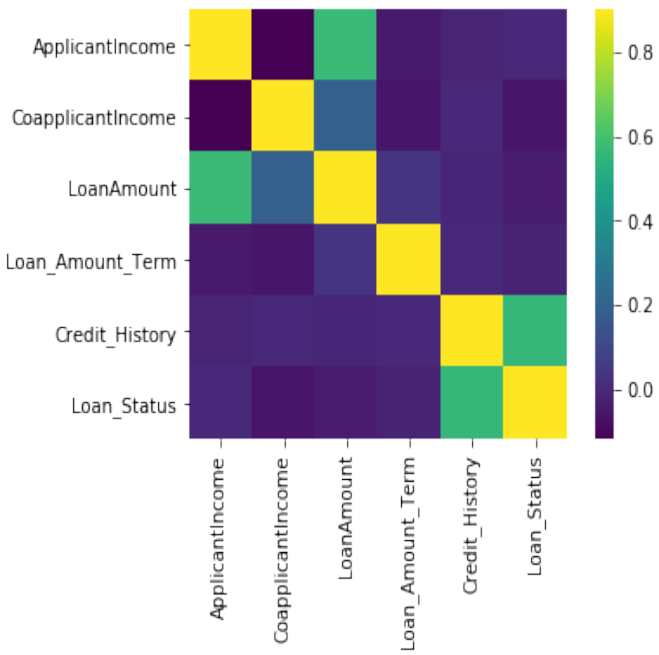

Fig. 4. Heat Map

Fig. 5. Pair Plot

preprocessed both train and test dataset using same process to make both data set consistent.

- As the dataset has many missing data points, I have replaced them by using mean, mode or median strategy depending on that particular feature characteristic.
- Based on the insights from Univariate data analysis, I discovered some of the numerical features have outliers and the distribution was right skewed, therefore applied log transformation to make it normally distributed.
- Feature Scaling has been implemented by Standardizing each features by removing the mean and scaling to unit variance. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected

VIII. Feature Generation and Removal

This section describes feature generation and removal process Based on the domain knowledge, I come up with new features that might affect the target variable as well as removed some of the features. following two new features are created:

- Income – Combine the Applicant Income and Coapplicant Income. If the total income is high, chances of loan approval might also be high.
- EMI – EMI is the monthly amount to be paid by the applicant to repay the loan. Idea behind making this variable is that people who have high EMI's might find it difficult to pay back the loan. I have calculated the EMI by taking the ratio of loan amount with respect to loan amount term.
- I drop the variables which used to create these new features. Reason for doing this is, the correlation between those old features and these new features will be very high and some of the machine learning assumes that the variables are not highly correlated. We also wants to remove the noise from the dataset, so removing correlated features will help in reducing the noise too. following features are droped ['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount', 'Loan Amount Term'].
- There are some unnecessary feature which are not important for training the model and will not affect the target variable so they are removed based on EDA.
- Following features are removed (Gender, Loan ID and Self Employed).

IX. Machine Learning Models

- Decision Tree:
  It is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.
- Logistic Regression:
  It is a type of supervised learning algorithm. It is used to predict a binary outcome given a set of independent variables. It is an estimation of Logit function. Logit function is simply a log of odds in favor of the event. This function creates s-shaped curve with the probability estimate.

- Random Forest:
  It is a tree based bootstrapping(random sampling with replacement) ensemble algorithm wherein a certain no. of weak learners (decision trees) is combined to make a powerful prediction model. For every individual learner, a random sample of rows and a few randomly chosen variables are used to build a decision tree model. Final prediction can be a function of all the predictions made by the individual learners.
- Artificial Neural Network (ANN):
  It is an efficient computing system whose central theme is borrowed from the analogy of biological neural networks. Every neuron is connected with other neuron through a connection link. Each connection link is associated with a weight that has information about the input signal. This is the most useful information for neurons to solve a particular problem because the weight usually excites or inhibits the signal that is being communicated. Each neuron has an internal state, which is called an activation signal. Output signals, which are produced after combining the input signals and activation rule, may be sent to other units.

X. Comparision of Models based on Accuracy, Advantages and Disadvantages

In this section, I am going to discussed about the results, which have been obtained by building models on Loan Prediction dataset, using various classification algorithms. I have used Scikit-learn which is an open source library for Python to build the predictive model.

The data is distributed among training and validation set with 80 and 20 percent ratio correspondingly for all the models. I have started to build the models using default parameters for all the models and trained the model using this parameters. I tested them first using validation set and then using test dataset. The accuracy achieved with default parameters is quite low.

Then I will try to improve the accuracy by tuning the hyperparameters for these models. I used grid search to get the optimized values of the hyper parameters. After Grid search I trained the model using these hyper parameters and tested on both validation and unseen data set. The model got the improved accuracy score.

Parameters settings for the different models and the Prediction accuracies obtained on the Validation and test dataset using various classification algorithms are shown in Fig.6.

As shown in the fig. 6 On the Validation set ANN is performing better compare to all the models with 85.36 percent accuracy followed by Logistic Regression with 84.55 percent, Random Forest with 82.11 percent and then Decision Tree with 79.67 percent. However, on the test dataset Logistic Regression is performing better compare to all the models with 79.16 percent accuracy followed by ANN with 78.47 percent, Random Forest with 77.77 percent and then Decision Tree with 76.38 percent.

Model	Parameters	Validation Set Accuracy	Test Set Accuracy
Decision Tree	Default	73.98	65.27
Decision Tree	min_samples_leaf=10, criterion='Gini' max_depth=2	79.67	76.38
Logistic Regression	liblinear Optimizer	84.55	79.16
Random Forest	Default	75.60	70.13
Random Forest	n_estimators: 40, criterion='Gini', and max_depth=4	82.11	77.77
Artificial Neural Network	hidden_layer_sizes=(20,20) , activation='logistic' ,solver='adam', max_iter=1000	85.36	78.47

Fig. 6. Parameter setting and Accuracy for machine learning models

Algorithms	Advantages	Disadvantages
Decision Tree	• Quite simple • Prediction is quite fast	•Over fits a lot (it generates high-variance models, it suffers less after the branches are pruned). • take a lot of memory (the more features have, the deeper and larger decision tree is likely to be)
Logistic Regression	•Probability/risk estimator. •Easy to implement and very efficient to train. • I got highest accuracy with this algorithm	•can't solve non-linear problems with logistic regression since its decision surface is linear.
Random Forest	•Robust to overfitting. •It has methods for balancing errors in data sets as I have imbalanced classes	•models generated with Random Forest take lot of memory due to large no. of estimators. •learning is slow (depending on the parameterization)
Artificial Neural Network	• Perform well for complex machine learning problems	•Very hard to simply explain. •parameterization is very complex (what kind of network structure should I choose? Hard to select What are the best activation functions for given problem

Fig. 7. Advantages and Disadvantages of each model

Fig.7 describes Advantages and Disadvantages for each algorithm.

XI. Future Research

- Make better visualizations between different variables to understand and generate new patterns or rules from the existing data.
- Create and add more features, try different models with different subset of features.
- Try with Boosting algorithms like XGBoost or AdaBoost.
- Making the given data more balanced (equal number of different outcomes) so the algorithm is able to distinguish better.

XII. Conclusion

After implementing given algorithms on the Loan Prediction Data, I observed that Logistic Regression has produced a better accuracy compared to the other implemented algorithms. I have realized machine learning plays a key role in finance institutions in reducing the time for processing a loan application and helping banks to target the right set of population. Further there are many domains in finance where machine learning can be used such as stock price prediction, Risk analysis, Fraud Detection etc.

References

References

[1] M. Sudhakar, and C.V.K. Reddy, "Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique", International Journal of Advanced Research in Computer Engineering Technology (IJARCET), vol. 5, no.3, pp. 705-718, 2016..

[2] J.H. Aboobyda, and M.A. Tarig, "Developing Prediction Model Of Loan Risk In Banks Using Data Mining", Machine Learning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016.

[3] International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016 Copyright to IJARCCE DOI 10.17148/IJARCCE.2016.53128 523 Loan Prediction Using Ensemble Technique Anchal Goyal1, Ranpreet Kaur2 Research Scholar, Computer Science, RIMT, Gobindgarh, India1 Assistant Professor, Computer Science, RIMT, Gobindgarh,

[4] The paper by "Atiya, Amir F. "Bankruptcy prediction forcredit risk using neural networks: A survey and new results."explains the implementation of Artificial Neural Networks on the Bank dataset for predicting Bankruptcy.