Loan Prediction

-Akshat Gaur

# **Introduction**

The two most pressing issues in the banking sector are: 1) How risky is the borrower? 2) Should we lend to the borrower given the risk? The response to the first question dictates the borrower's interest rate. Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e. the higher the interest rate, the riskier the borrower. We will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays the loan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment. Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default. The insured sum can cover the whole loan amount or just a portion of it. Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can be used to assess a customer's loan status and build strategies. This model extracts and introduces the essential features of a borrower that influence the customer's loan status. Finally, it produces the planned performance (loan status). These reports make a bank manager's job simpler and quicker.

# **Literature Review**

We start our literature review with more general systematic literature reviews that focus on the application of machine learning in the general field of Banking Risk Management. Since the global financial crisis, risk management in banks has to take a major role in shaping decision-making for banks. A major portion of risk management is the approval of loans to promising candidates. But the black-box nature of Machine learning algorithms makes many loan providers vary the result. Martin Leo, Suneel Sharma and k. Maddulety's [1] extensive report has explored where Machine Learning is being used in the fields of credit risk, market risk, operational risk, and liquidity risk only to conclude that the research falls short of extensive research is required in the field.

We could not find any literature review for loan prediction for specific Machine learning algorithms to use which would be a possible starting point for our paper. Instead, since loan prediction is a classification problem, we went with popular classification algorithms used for a similar problem. Ashlesha Vaidya [2] used logistic regression as a probabilistic and predictive approach to loan approval prediction. The author pointed out how Artificial neural networks and Logistic regression are most used for loan prediction as they are easier comparatively develop and provide the most accurate predictive analysis. One of the reasoning behind this that that other Algorithms are generally bad at predicting from non-normalized data. But the nonlinear effect and power terms are easily handled by Logistic regression as there is no need for the independent variables on which the prediction takes place to be normally distributed.

Logistic regression still has its limitations, and it requires a large sample of data for parameter estimation. Logistic regression also requires that the variables be independent of each other otherwise the model tends to overweigh the importance of the dependent variables.

A solution to this multicollinearity problem among the categorical explanatory variables is the use of a categorical principal component analysis which can be seen used by Guilder and Ozlem [3] on a case study for housing Loan approval data. The goal of Principal component analysis is to reduce the number of m variables where many of them would be highly correlated with each other, to a smaller set of n uncorrelated variables called principal components which account for the variances between the previous m variables. Methods such as PCA are known as dimension reduction of the data. It may be suitable for scaled continuous variables but it isn’t quite an appropriate method of dimension reduction for categorical variables. Thus, the authors here used a tweaked version of PCA for categorical data called CATPCA or categorical (nonlinear) principal components analysis which is specifically developed for where the dependent variables are a mix of nominal, ordinal, or numeric data which may not have linear relationships with each other. CATPCA works by using a scaling process optimized to convert the categorical variables into numeric variables.

Similar to PCA, Zaghdoudi, Djebali &amp; Mezni [4] compared the use of Linear Discriminant Analysis versus Logistic Regression for Credit Scoring and Default Risk Prediction for foreseeing default risk o small and medium enterprises. Linear Discriminant Analysis (LDA) is like PCA for dimensionality reduction but instead of looking for the most variation, LDA focuses on maximizing the separability among the know categories. This subspace that well separates the classes is usually in which a linear classifier can be learned. The classification of those enterprises correctly in their original groups through both these methods was inconsequential with Logistic regression having a 0.3% better accuracy score than LDA.

Another novel approach for T.Sunitha and colleagues [5] was to predict loan Status using Logistic Regression and a Binary Tree.&nbsp;Decision Tree is an algorithm for a predictive type machine learning model.

Classification and Regression Trees are referred to as CART (in short) introduced by Leo Breiman. It best suits both predictive and decision modeling problems. This Binary Tree methodology is the greedy method is used for the selection of the best splitting. Although Decision trees gave us a similar accuracy. The benefits of Decision Trees, in this case, were due to the latter giving equal importance to both accuracy and prediction. This model became successful in making a lower number of False Predictions to reduce the risk factor.

Rajiv Kumar and Vinod Jain [6] proposed a model using machine learning algorithms to predict the loan approval of customers. They applied three machine learning algorithms, Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) using Python on a test data set. From the results, they concluded that the Decision Tree machine learning algorithm performs better than Logistic Regression and Random Forest machine learning approaches. It also opens other areas on which the Decision Tree algorithm is applicable.

Some machine learning models give different weights to each factor but in practice sometimes loans can be sanctioned based on a single strong factor only. To eliminate this problem J. Tejaswini and T. Mohana Kavya [7] in their research paper have built a loan prediction system that automatically calculates the weight of each feature taking part in loan processing and on new test data the same features are processed concerning their associated weight. They have implemented six machine learning classification models using R for choosing the deserving loan applicants. The models include Decision Trees, Random Forest, Support Vector Machine, Linear Models, Neural Network and Adaboost. The authors concluded that the accuracy of the Decision Tree is highest among all models and performs better on the loan prediction system.

Predicting loan defaulters is an important process of the banking system as it directly affects profitability. However, loan default data sets available are highly imbalanced which results in poor performance of the algorithms. Lifeng Zhou and Hong Wang [8] in their call for paper made loan default prediction on imbalanced data sets using an improved random forests approach. In this approach, the authors have employed weights in decision tree aggregation. The weights are calculated and assigned to each tree in the forest during the forest construction process using Out-of-bag (OOB) errors. The experimental results conclude that the improved algorithm performs better and has better accuracy than the original random forest and other popular classification algorithms such as SVM, KNN, and C4.5. The research opens improvements in terms of efficiency of the algorithm if parallel random forests can be used for further work.

Anchal Goyal and Ranpreet Kaur [9] discuss various ensemble algorithms. Ensemble algorithm is a supervised machine learning algorithm that is a combination of two or more algorithms to get better predictive performance. They carried out a systematic literature review to compare ensemble models with various stand-alone models such as neural network, SVM, regression, etc. The authors after reviewing different literature reviews concluded that the Ensemble Model performs better than the stand-alone models. Finally, they concluded that the concept of combined algorithms also improves the accuracy of the model.

Data Mining is also becoming popular in the field banking sector as it extracts information from a tremendous amount of accumulated data sets.&nbsp;Aboobyda Jafar Hamid and Tarig Mohammed Ahmed [10] focused on implementing data mining techniques using three models j48, bayesNet, and naiveBayesdel for classifying loan risk in the banking sector. The author implemented and tested models using the Weka application. In their work, they made a comparison between these algorithms in terms of accuracy in classifying the data correctly. The operation of sprinting happened in a manner that 80% represented the training dataset and 20% represented the testing dataset. After analyzing the results the author came up with the results that the best algorithm among the three is the J48w algorithm in terms of high accuracy and low mean absolute error.

# **Methodology**

## **Machine Learing and Concepts**

Four machine learning models have been used for the prediction of loan approvals. Below are the description of the models used:

### **Logistic Regression**

This is a classification algorithm which uses a logistic function to predict binary outcome (True/False, 0/1, Yes/No) given an independent variable. The aim of this model is to find a relationship between features and probability of particular outcome. The logistic function used is a logit function which is a log of odds in the favor of the event. Logit function develops a s-shaped curve with the probability estimate similar to a step function.

### **Decision Trees**

This is a supervised machine learning algorithm mostly used for classification problems. All features should be discretized in this model, so that the population can be split into two or more homogeneous sets or subsets. This model uses a different algorithm to split a node into two or more sub-nodes. With the creation of more sub-nodes, homogeneity and purity of the nodes increases with respect to the dependent variable.

### **Random Forest**

This is a tree based ensemble model which helps in improving the accuracy of the model . It combines a large number of Decision trees to build a powerful predicting model. It takes a random sample of rows and features of each individual tree to prepare a decision tree model. Final prediction class is either the mode of all the predictors or the mean of all the predictors.

### **XGBoost**

This algorithm only works with the quantitative variable. It is a gradient boosting algorithm which forms strong rules for the model by boosting weak learners to a strong learner. It is a fast and efficient algorithm which recently dominated machine learning because of its high performance and speed.

### **Pandas**

### Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, json, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling.

### **Seaborn**

Seaborn is a python library for building graphs to visualise data. It provides integration with pandas. This open source tool helps in defining the data by mapping the data on the informative and interactive plots. Each element of the plots gives meaningful information about the data.

### **Sklearn**

This python library is helpful for building machine learning and statistical models such as clustering, classification, regression etc. Though it can be used for reading, manipulating and summarizing the data as well, better libraries are there to perform these functions.

# **Understanding the Dataset**

The machine learning model is trained using the training data set. Every new applicant details filled at the time of application form acts as a test data set. On the basis of the training data sets, the model will predict whether a loan would be approved or not. We have 13 features in total out of which we have 12 independent variables and 1 dependent variable i.e. Loan\_Status in train dataset and 12 independent variables in test dataset. The Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status are all categorical

# **Exploratory Data Analysis**

# **Univariate Visual Analysis**

## **Target Variable - Loan Status**

We will start first with an independent variable which is our target variable as well.

## **Predictor Variables**

There are 3 types of Independent Variables: Categorical, Ordinal & Numerical.

### **Categorical Features**

* Gender
* Marrital Status
* Employment Type
* Credit Hystory

****It can be inferred from the below bar plots that in our observed data:****

* 80% of loan applicants are male in the training dataset.
* Nearly 70% are married
* About 75% of loan applicants are graduates
* Nearly 85–90% loan applicants are self-employed
* The loan has been approved for more than 65% of applicants.

### **Ordinal Features**

* Number of Dependents
* Education Level
* Property or Area Background

****Our Visual Analysis below, indecates that:****

* Almost 58% of the applicants have no dependents.
* Highest number of applicants are from Semi Urban areas, followed by urban areas.
* Around 80 % of the applicants are Graduate.

### **Numerical Features**

* The Applicant's Income
* The Co-Applicant's Income

It can be inferred that most of the data in Applicant income is towards left which means it is not normally distributed. The boxplot confirms the presence of outliers. This can be attributed to income disparity in the society.

We can see that there are higher number of graduates with very high incomes, which are appearing to be the outliers.

CoapplicantIncome is lesser than applicantIncome and is within the 5000–15000, again with some outliers.

## **Bivariate Analysis**

Bivariate analysis is finding some kind of empirical relationship between two variables. Specifically the dependent vs independent Variables

### **Categorical Independent Vs Target**

#### **Gender Vs Loan\_Status**

There is not a substantial difference between male and female approval rates.

#### **Marriage Status Vs Loan\_Status**

Married applicants have a slightly higher chances of loan approval.

#### **Dependency Vs Loan\_Status**

Applicants with no dependents or 2 dependents have higher chances of approval. But this does not correlate well.

#### **Education Vs Loan\_Status**

Graduates have higher chance of loan approval compared to non-graduates.

#### **Employment Type Vs Loan Status**

Self\_Employed employees have slightly lower chances of loan approval but the situation is not that bad.

#### **Credit\_History Vs Loan\_Status**

It seems people with credit history as 1 are more likely to get their loans approved

Propotion of loans getting approved in semiurban area is higher as compared to that in rural or urban areas.

### **Numerical Independent vs Target**

We tried to find the mean income of people for which the loan has been approved vs the mean income of people for which the loan has not been approved but we don't see any changes in the mean income. So we make bins for the applicant income variable based on the values in ti and analyze the corresponding loan status for each bin

It can be infered that Applicant income does not affect the chances of loan approval which contradicts our hypothesis in which we assumed that if the applicant income is high, the chances of loan approval will also be high.

It shows that if the coapplicant's income is less the chances of loan approval are high. But this does not look right. The possible reason behind this may be that most of the applicants don't have any coapplicant so the coapplicant income for such applicants is 0 and hence the loan approval is not dependent on it. So we can make new variable in which we will continue the applicant's and coapplicant's income to visualize the combined effect of income on loan approval.

We can see that Proportion of loans getting approved for applicants having low Total\_Income is very less as compared to that of applicants with the Average, High and Very High Income.

It can be seen that the proportion of approved loans is higher for Low and Average Loan Amount as compared to that of High Loan Amount which supports our hypothesis in which we can considered that the chances of loan approval will be high when the loan amount is less.

# **Hypothesis Testing**

### **Gender vs Loan Status:**

As we saw in Bivariate Analysis that the conditional probability of getting a loan is equal across male as well as female. So now we need to conduct a statistical test to confirm this assumption. Null hypothesis : male and female has equal chances of getting a loan; [m1 == m2] Alternate hypothesis : not equal chances; [m1 != m2] p value &gt; 0.05 , we can not reject the null hypothesis and confirm that there is equal chance of getting a loan for male as well as female.

### **Property Area vs Loan Status:**

As we saw in Bivariate Analysis that the conditional probability of getting a loan is not equal across different property area. So now we need to conduct a statistical test to confirm this assumption. Null hypothesis: every property area has equal chances of getting a loan; [m1 == m2] Alternate hypothesis : not equal chances

### **Credit History vs Loan Status:**

As we saw in Bivariate Analysis that the conditional probability of getting a loan is not equal across different credit history. Null hypothesis : equal chances of getting a loan for different credit history Alternate hypothesis : not equal chances

### **Education vs Loan Status:**

We assumed that there is an association between getting a loan and Education, with a difference of just 9% in conditional probabilities. Null hypothesis : equal chances of getting a loan for different education background Alternate hypothesis : not equal chances

### **Dependents vs Loan Status:**

Earlier we found that except dependents = 0, others had an association with Loan Status. Null hypothesis : equal chances of getting a loan for different no. of Dependents Alternate hypothesis : not equal chances

### **Self Employed vs Loan Status:**

Earlier we assumed that there is no association between Self\_Employed and Loan\_Status. Null hypothesis : equal chances of getting a loan for Self employment Alternate hypothesis : not equal chances

**Conclusion.**

Loan Prediction model has showed excellent results for the model’s prediction.

It can be said that different algorithms have showed brilliant results.