**Smart Lender - Applicant Credibility Prediction for Loan Approval**

Introduction

Loans are the core business of loan companies. The main profit comes directly from the loan’s interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don’t have assurance if the applicant is able to repay the loan with no difficulties.

In this project, we’ll build a predictive model to predict if an applicant is able to repay the lending company or not. We will prepare the data using Jupyter Notebook.

Literature Review

[1]Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval This work includes the construction of an ensemble model by combining different machine learning models. Banks struggle a lot to get an upper hand over each other to enhance overall business due to tight competition. Credit Risk assessment is a crucial issue faced by Banks nowadays which helps them to evaluate if a loan applicant can be a defaulter at a later stage so that they can go ahead and grant the loan or not. This helps the banks to minimize the possible losses and can increase the volume of credits

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.

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 [2] 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 [3] 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 [4] 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 [7] 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 [6] 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 [7] 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 [8] 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 [9] 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.[10] 2. Paper Name: Credit Risk Analysis and Prediction Modelling of Bank Loans Using R. Using the R package, this paper proposed a riskanalysis method for sanctioning a loan for customers. Data selection, pre-processing, feature extraction and selection,building the model, prediction, and evaluation were among the steps involved in developing the model.