Mini Project in

**Loan Prediction and Analysis**

#### By

**Kanishka Ghodke Roll No. 19IT2039**

**Aditya Sangade Roll No. 19IT2036**

**Manas Darekar Roll No. 19IT1064**

Project Guide

**Mrs. Nilima Dongre**



Department of Information Technology Dr. D. Y. Patil Group’s

Ramrao Adik Institute Of Technology

Nerul, Navi Mumbai 400706.

(Affiliated to University of Mumbai)

**2022**



**Ramrao Adik Institute of Technology**

#### (Affiliated to the University of Mumbai)

Dr. D. Y. Patil Vidyanagar,Sector 7, Nerul, Navi Mumbai 400706.

**CERTIFICATE**

This is to certify that, the Mini Project-2B based on ML titled

“**Loan Prediction and Analysis**”

is a *bonafide* work done by

Kanishka Ghodke Roll No. 19IT2039

Aditya Sangade Roll No. 19IT2039

Manas Darekar Roll No. 19IT2039

Project Guide

**Mrs. Nilima Dongre**

|  |  |  |
| --- | --- | --- |
| Mini-Project Co-ordinator | Head of Department | Principal |
| **(Mrs.Priyanka Shingane )** | **(Dr. Ashish Jadhav)** | **(Dr. Mukesh D. Patil)** |

**Mini Project Approval**

This is to certify that the Mini Project entitled ”Loan Prediction and Analysis” is a bonafide work done by *Kanishka Ghodke,Manas Darekar and Aditya Sangade* under the guidance of *Mrs Nilima Dongre*. This work has been approved as a Mini Project for Third year Information Technology.

Examiners :

1. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

2. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Supervisors :

1. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

2. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Date : Place :

**Declaration**

I declare that this written submission represents my ideas in my own words and where other’s ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Kanishka Ghodke Roll No. 19IT2039 \_\_\_\_\_\_\_\_\_

Aditya Sangade Roll No. 19IT2036 \_\_\_\_\_\_\_\_\_

Manas Darekar Roll No. 19IT1064. \_\_\_\_\_\_\_\_\_

Date : 22/04/2022

**Abstract**

Loans are the core business of banks. 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.

Loan Prediction Model :This is a classification problem in which we need to classify whether the loan will be approved or not. 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 (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation.

*Keywords*— Big data, Machine Learning, Regression

**Contents**

Abstract

1. Introduction 1
2. Literature Survey 2
3. Problem Statement 3
4. Objective 4
5. Proposed Framework 5
6. Implementation environment 6
7. Conclusion 7

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 survey

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 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 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 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 & Mezni 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 was to predict loan Status using Logistic Regression and a Binary Tree. 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 modelling 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 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 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 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 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. Aboobyda Jafar Hamid and Tarig Mohammed Ahmed focused on implementing data mining techniques using three models j48, bayesNet, and naïve Bayesdel 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 analysing 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.

**Problem statement**

We want to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, we have given a problem i.e. to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

It’s a classification problem , given information about the application we have to predict whether the they’ll be to pay the loan or not.

# Objective

We have built 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 and use various models to predict the target variable.

*A.Automation of loan prediction process*

User will be able to figure out whether he is eligible to take loan within seconds by entering basic details

*B.The previous datasets containing user information are taken*

We need basic user details like Gender, Marital Status, Education, number of Dependents, Income, Loan Amount, Credit History, and others. This data is raw and we process it further by cleaning

**Proposed Framework**

Data

For this problem, we have two CSV files: train and sample submission.

Train file will be used for training the model, i.e. our model will learn from this file. It contains all the independent variables and the target variable.

We clean and organize the data. On that sorted data we use various algorithms to find out whether the user eligible to get loan or not .

Loan Dataset : Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer’s loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.

Data cleaning and processing: In Data cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing , modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

Selected Algorithms

## Logistic Regression

We will start with Logistic Regression which is used for predicting binary outcomes.

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.  
Logistic regression is an estimation of the Logit function. The logit function is simply a log of odds in favour of the event.  
This function creates an S-shaped curve with the probability estimate, which is very similar to the required stepwise function

## Decision Tree Algorithm

Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**

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 Algorithm

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.

Random Forest is a tree-based bootstrapping algorithm wherein a certain no. of weak learners (decision trees) are 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. In the case of a regression problem, the final prediction can be the mean of all the predictions.

**Implementation environment**

We will be using Python along with the below-listed libraries.

**Specifications**

* Python 3.7
* 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.

##### **Conclusion**

* We did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed.
* We did bivariate and multivariate analysis to see impact of one another on their features using charts.
* We analysed each variable to check if data is cleaned and normally distributed.
* We cleaned the data and removed NA values
* We also generated hypothesis to prove an association among the Independent variables and the Target variable. And based on the results, we assumed whether or not there is an association.
* We calculated correlation between independent variables and found that applicant income and loan amount have significant relation.
* We created dummy variables for constructing the model
* We constructed models taking different variables into account and found through odds ratio that credit history is creating the most impact on loan giving decision
* Finally, we got a model with coapplicant income and credit history as independent variable with highest accuracy.
* We tested the data and got the accuracy of 83 %

##### **Bibliography**

* https://ieeexplore.ieee.org/document/9155614
* https://www.geeksforgeeks.com
* https://deepnote.com/@rhishab-mukherjee/Loan- Prediction-Project-TermPaper- VPSOpiywSu6FZeN2fK8fug
* https://www.javatpoint.com/machine-learning-decision- tree-classification-algorithm