Loan Eligibility Prediction using Supervised Learning (Logistic Regression, Linear Regression, Naive Bayes, KNN, SVM, Gradient Descent)

Prepared by: Pavithra Shankar babu

Date: 05/01/2024

Executive Summary:

The primary prediction target of this project is to predict whether a loan application will be approved or denied based on the applicant's characteristics and financial history. This is a binary classification task, where the model will classify loan applications into two categories: approved or denied. Additionally, secondary prediction targets may include: Estimating the probability of default for approved loans: This can help the bank assess the risk associated with approving a loan and make informed decisions about loan terms and interest rates. Predicting the loan amount: Given a set of applicant characteristics, the model can predict the amount of loan that is likely to be approved, providing valuable information to both the bank and the applicant. By developing accurate predictive models, financial institutions can streamline their loan approval processes, minimize risk, and improve customer satisfaction.

This project aims to leverage machine learning algorithms such as Logistic regression, Linear regression, Naive bayes, Support vector machines (SVM), and Gradient boosting techniques to achieve these objectives. Later, the accuracy of all these algorithms are compared and the best one out of these is chosen.

Problem Statement:

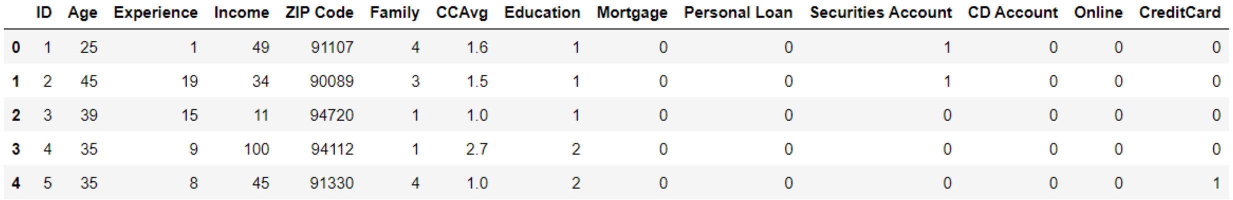
The primary challenge in loan eligibility prediction is to accurately assess the creditworthiness of loan applicants based on a variety of factors such as demographics, financial history, employment status, and loan preferences. Existing approaches often lack precision in predicting loan outcomes, leading to inefficient allocation of resources and increased risk of defaults. This project aims to overcome these challenges by developing advanced predictive models capable of accurately evaluating loan eligibility for individual applicants, thereby enabling financial institutions to make informed decisions and optimize their lending practices.

Objective:

The objectives of this project are to gather a comprehensive dataset containing demographic information, financial history, employment status, and other relevant factors related to loan applicants. Utilizing supervised learning algorithms, the project aims to construct predictive models capable of accurately determining loan eligibility for individual applicants. By incorporating feature engineering, data preprocessing techniques, and model optimization strategies, the goal is to improve the predictive accuracy and robustness of the models, thereby enabling financial institutions to make more informed and efficient lending decisions.

Dataset:

Dataset Link: https://www.kaggle.com/datasets/mahnazarjmand/bank-personal-loan



	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1

Before Modeling, the dataset will be undergone process like cleaning, EDA, pre-processing to improve accuracy.

Data Preprocessing:

Data preprocessing plays a crucial role in machine learning since the quality of data and the insights derived from it directly impact the learning capability of our model. Therefore, it is essential to preprocess our data before inputting it into our model to ensure its effectiveness. In data preprocessing, columns are analysed for wrong entries. Outliers and normality of the columns are also found for further execution. Additionally, I computed the correlation matrix to assess the relationships and dependencies between different numerical variables.

Model & Evaluation:

In this analysis, the performance of several machine learning models - including Linear Regression, Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes, and Gradient Descent - will be compared for the task of loan eligibility prediction. To assess the performance of each model, a series of steps will be followed. Each model will undergo k-fold cross-validation, with k set to 5, to estimate its accuracy on the training data. The mean accuracy and standard deviation of the cross-validation scores will be recorded.

Subsequently, the trained models will be evaluated on a separate testing dataset. The accuracy of each model on the test set will be computed and compared.

After completing the evaluation process, it may be observed that Logistic Regression achieved an accuracy of 0.97, Linear Regression had an accuracy of 0.84, SVM achieved an accuracy of 0.96, kNN had an accuracy of 0.98, Naive Bayes achieved an accuracy of 0.93, and Gradient Descent achieved an accuracy of 0.84 on the test set. These accuracy scores provide insights into the effectiveness of each model in predicting loan eligibility based on the provided features.

Result:

Upon evaluating the performance of various machine learning models for loan eligibility prediction, it was found that k-Nearest Neighbors (kNN) demonstrated the highest accuracy among the models considered. With an accuracy score of 0.97 on the test set, kNN outperformed other algorithms such as Linear Regression, Logistic Regression, SVM, Naive Bayes, and Gradient Descent. This result suggests that the kNN algorithm is particularly effective in capturing the underlying patterns within the data and making accurate predictions regarding loan eligibility based on the provided features. Consequently, kNN may be considered as a promising model for deployment in real-world loan approval systems, offering potential benefits in terms of improved decision-making and resource allocation for financial institutions.