MACHINE LEARNING INTERNSHIP QUALIFICATION TASK Report

-Yashaswe Amatya

Description:

You are given a dataset of the bank loans consisting of 15 columns and a corresponding target column. Your task is to build a machine-learning model that can accurately classify whether the personal loan was accepted or not based on the information provided.

Approach

I began the analysis by loading the dataset using pandas and conducting preliminary data exploration to understand its structure and content. Initial steps included displaying the first few rows, summarizing the dataset, and identifying missing values. Missing data was handled by dropping rows with any missing values to ensure data integrity.

In the data cleaning phase, specific characters ('#', '-') in the 'Gender' column were replaced with 'O', and categorical variables were converted into numerical values. Gender was encoded as M (0), F (1), and O (2), while Home Ownership categories were converted to Home Mortgage (0), Homeowner (1), and Rent (3). Personal Loan values with a space (' ') were converted to 0.

Data visualization involved creating count plots using seaborn to explore the relationships between education levels and personal loan status, as well as between gender and personal loan status. These visualizations provided initial insights into patterns and trends within the data.

Irrelevant columns such as 'ID' and 'ZIP Code' were dropped, and the dataset was split into features (X) and the target variable (Y). The data was then split into training and test sets with an 80-20 split.

Key Findings

Two machine learning models were trained and evaluated: a Support Vector Machine (SVM) and a Gradient Boosting Classifier. The Support Vector Machine (SVM) is a powerful supervised machine learning algorithm primarily used for classification tasks. It operates by finding the optimal hyperplane that best separates different classes in the feature space.

SVM aims to maximize the margin between the classes, thereby enhancing its robustness to unseen data points. Despite its effectiveness, SVM's performance can be impacted by the choice of kernel function and its hyperparameters, which may require fine-tuning for optimal results.

On the other hand, Gradient Boosting Machines (GBMs) are ensemble learning techniques that build multiple weak learners, typically decision trees, sequentially to correct the errors of the preceding models. Each subsequent model in the ensemble focuses on the instances that the previous models misclassified,

gradually improving the overall prediction accuracy. GBMs are renowned for their capability to handle complex datasets and are less prone to overfitting compared to other ensemble methods.

The SVM model achieved a test accuracy of 95.4%. In contrast, the Gradient Boosting model achieved a test accuracy of 98.0%, suggesting it to be a better model to predict the outcome using data. The Gradient Boosting model significantly outperformed the SVM model on the test data, suggesting it is more reliable for predicting personal loan acceptance.

Additional Task

Created a simple chatbot to take in user data and provide real time prediction of loan acceptance using Gradient Boosting Model.