Loan Approval Project:

- This repository contains an attempt to apply classification algorithms to the Loan Approval Prediction dataset from the Kaggle Tabular Playground Challenge (Loan Approval Prediction | Kaggle).

Overview:

- The task, as defined by the Kaggle Tabular Playground Challenge, is to develop a machine learning model to predict the likelihood of loan approval for individual applicants. The dataset contains information on applicant demographics, financial status, and loan details, such as age, income, employment length, credit history, loan purpose, and homeownership status. The goal is to classify each applicant as either "approved" or "not approved" for a loan, using the provided features to optimize prediction accuracy on an unseen test set.
- The approach in this repository formulates the problem as a binary classification task, using logistic regression as the primary machine learning model. The input features are preprocessed through encoding of categorical variables, normalization of numeric features, and removal of irrelevant columns like unique identifiers. The dataset is split into training, validation, and test sets to ensure robust evaluation. The logistic regression model is trained on the processed training data, and its performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Predictions are generated for the test set to create a submission file for the Kaggle leaderboard.
- Our best model achieved an accuracy of 90% on the validation set, with a precision of 67.04%, recall of 32.02%, F1-score of 43.34%, and an ROC-AUC score of 84.54%. The model demonstrates strong performance in predicting the majority class while offering moderate recall for the minority class. These metrics indicate a well-calibrated model that effectively distinguishes between approved and denied loans. At the time of writing, this performance aligns competitively with other submissions on the Kaggle leaderboard.

Summary of work done:

Data
Type: Tabular data

Input: The input dataset consists of a CSV file containing loan application data. Each row represents a unique loan application, and each column corresponds to a specific feature relevant to loan approval prediction.

Features:

- Numerical Features:
    - person_age: Age of the loan applicant.
    - person_income: Annual income of the loan applicant.
    - person_emp_length: Length of employment in years.
    - loan_amnt: Amount of the loan requested.

- loan_int_rate: Interest rate assigned to the loan.
- loan_percent_income: Percentage of income required to cover the loan payment.
- cb_person_cred_hist_length: Length of the applicant's credit history.

- Categorical Features (One-Hot Encoded):
    - person_home_ownership: Homeownership status (e.g., RENT, OWN, MORTGAGE, OTHER).
    - loan_intent: Purpose for the loan (e.g., EDUCATION, MEDICAL, VENTURE, etc.).
    - loan_grade: Loan grade assigned based on applicant's creditworthiness.
    - cb_person_default_on_file: Indicates whether the applicant has defaulted on previous credit obligations (Y/N).

Output:

- Target Variable:
    - loan_status: Binary variable (0 or 1) indicating whether the loan was approved (1) or denied (0).

This dataset was generated from a modified version of the Kaggle Loan Approval Prediction dataset, with train and test data created using a deep learning model to maintain a feature distribution similar to the original dataset. The goal is to predict the loan_status for unseen test data.

The dataset contains 50,827 instances split into 70% training (35,579), 15% validation (7,624), and 15% test (7,624). Each instance represents a loan application with 27 features, including numerical and one-hot encoded categorical variables.

Preprocessing / Clean up:

The data was preprocessed by handling missing values, encoding categorical variables, and normalizing numerical features. Categorical features like person_home_ownership, loan_intent, loan_grade, and cb_person_default_on_file were one-hot encoded. The target variable loan_status was converted to binary values (0 for "No" and 1 for "Yes"). Redundant columns, such as the unique identifier id, were dropped to avoid introducing unnecessary noise into the model.

Data Visualization:

Visualizations showed that most applicants were aged 20-40, with debt consolidation being the most common loan intent. Income had significant variability, with some high-income outliers. Correlations between features, like loan_amnt and loan_int_rate, indicated that larger loans generally have higher interest rates. These insights informed feature selection and model development.

Problem Formulation:

The task aims to predict loan approval status (loan_status) using features like person_age, person_income, and loan_amnt. Models tested include Logistic Regression, Random Forest, SVM, and XGBoost. Logistic Regression was chosen for its simplicity, while Random Forest and XGBoost were tested for better handling of complex relationships. Hyperparameters like regularization for Logistic Regression and n_estimators for Random Forest were tuned. Logistic Regression provided the best balance between accuracy and simplicity.

Training:

I trained the models using Python and libraries like scikit-learn and XGBoost on a local machine with an Intel i7 processor and 16 GB RAM. Training times varied, with simpler models like Logistic Regression being quicker. I used cross-validation and monitored loss curves for early stopping. Ensuring consistent feature engineering across datasets was key to successful training.

Performance Comparison:

The model's key performance metrics include accuracy (90.4%), precision (67.04%), recall (32.02%), F1-score (43.34%), and ROC AUC (84.54%). The ROC curve shows the model's ability to distinguish between classes, with a high AUC indicating good performance, though recall could be improved.

Conclusions:

The logistic regression model performed reasonably well, achieving good accuracy and AUC, but recall was relatively low, indicating it struggled with predicting the minority class (approved loans). Future improvements could focus on enhancing recall through techniques like resampling or using more complex models like random forests or XGBoost.

Future Work:

Next, I would experiment with more complex models like Random Forests or XGBoost to improve performance, especially in terms of recall for the minority class. Additionally, I would try hyperparameter tuning, resampling techniques (such as SMOTE), or feature engineering to enhance the model's ability to predict loan approvals. Future studies could explore deep learning models like neural networks, compare different ensemble methods, or investigate the impact of additional features such as applicant's credit score or external economic factors.

How to reproduce results:

To reproduce the results, install necessary libraries (pandas, numpy, scikit-learn, xgboost) and load the dataset. Preprocess the data by handling missing values, encoding categorical features, and scaling numerical ones. Train a model (e.g., Logistic Regression, Random Forest) on the training set, evaluate it on the validation set, and apply it to test data. Google Colab is recommended for cloud-based computation.

Overview of files in repository:

The repository contains the following files: Loan Approval Project.ipynb, the main notebook where all steps of data preprocessing, model training, evaluation, and submission generation are performed; train.csv, the training dataset used for model training; and test.csv, the test dataset used for model evaluation and generating Kaggle submissions. All tasks are handled within the single notebook.

Software Setup:

The required packages for this project include pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost, and imbalanced-learn. These can be installed using the following command in a terminal or Colab notebook: pip install pandas numpy scikit-learn matplotlib seaborn xgboost imbalanced-learn. Google Colab typically has most of these packages pre-installed, but they can be added if needed

Data:

The data for this project can be downloaded from the Kaggle Loan Approval Prediction Challenge page, where you'll find train.csv and test.csv files. After downloading, preprocessing steps such as handling missing values (e.g., filling or dropping them) and encoding categorical variables (using one-hot encoding) are necessary to prepare the data. You can also separate the features and target variable, ensuring that the test data undergoes similar preprocessing. These steps will make the data ready for training machine learning models and generating predictions.

Training:

To train the model, load and preprocess the training data from train.csv. Split it into training and validation sets using train_test_split. Initialize and fit your chosen model (e.g., Logistic Regression) on the training data. Evaluate the model using the validation set with metrics like accuracy, precision, and recall. Optionally, tune hyperparameters for better performance.

Performance Evaluation:

To evaluate the model's performance, use the trained model to make predictions on the validation set. Calculate key metrics such as accuracy, precision, recall, F1-score, and ROC AUC. You can also generate a confusion matrix to visualize true positive, true negative, false positive, and false negative counts. Additionally, use classification reports to assess model performance for each

Citations:

https://www.kaggle.com/competitions/playground-series-s4e10/data