House Price Prediction and Loan Eligibility Analysis

Contents

1 Abstract / Project Summary

This project aims to develop machine learning models for predicting **house prices** based on real estate features and assessing **loan eligibility** based on applicant financial data. Using **regression models** for house prices and **classification models** for loan eligibility, the project ensures accurate and explainable predictions. Data preprocessing, feature engineering, and model evaluation are key aspects of this study. The final models are deployed using **Flask/Streamlit** for user interaction.

2 Introduction

Real estate pricing and loan approvals are critical aspects of the financial sector. Accurately predicting house prices helps buyers make informed decisions, while an efficient loan eligibility model helps financial institutions assess applicants effectively. This project utilizes **machine learning algorithms** to analyze real-world housing and financial datasets, providing a data-driven approach to price prediction and loan approval processes.

3 Dataset Description

3.1 Loan Eligibility Dataset

- **Total Entries:** 614
- **Key Features:**
  - **Loan_ID:** Unique identifier for each loan application.
  - **ApplicantIncome, CoapplicantIncome, LoanAmount, Credit_History, Property_Area**
  - **Loan_Status:** Target variable (Y: Approved, N: Not approved)
- **Missing Values:** LoanAmount (22), Credit_History (50), other categorical fields.

3.2 House Price Prediction Dataset

- **Total Entries:** 21,613
- **Key Features:**
  - **Sale Price (Target Variable)**

  - **No of Bedrooms, No of Bathrooms, Flat Area (in Sqft), Lot Area (in Sqft)**
  - **Condition of the House, Overall Grade, Latitude, Longitude, Zipcode**

- **Missing Values:** Sale Price (4), structural attributes, and location data.

4 Methodology

4.1 House Price Prediction

Data Preprocessing:

- Handled missing values using mean/mode imputation.
- Scaled numerical features to normalize data.
- Encoded categorical features using one-hot encoding.

Model Selection:

- **Linear Regression**: Establishes a linear relationship between input features and house prices. It serves as a baseline model.
- **Decision Tree Regressor**: A tree-based model that splits data at decision nodes to capture non-linear relationships. It helps in handling complex interactions between features.
- **Random Forest Regressor**: An ensemble learning method using multiple decision trees. It reduces overfitting and improves predictive accuracy.
- **Gradient Boosting (XGBoost)**: A boosting algorithm that optimizes weak learners sequentially, improving performance by reducing errors iteratively.

Performance Metrics:

- **Mean Absolute Error (MAE)**: Measures the average absolute difference between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: Penalizes larger errors more than MAE, making it more sensitive to outliers.
- **R-squared ($\mathbf{R^2}$)**: Indicates how well the model explains variance in house prices.

4.2 Loan Eligibility Prediction

Data Preprocessing:

- Encoded categorical features such as employment type and loan purpose.
- Handled missing values using mode imputation.
- Normalized numerical features like applicant income and loan amount.

Model Selection:

- **Random Forest Classifier**: Uses multiple decision trees and majority voting to enhance classification performance. It is effective for high-dimensional datasets.
- **Gradient Boosting (XGBoost)**: A boosting model that builds trees sequentially to correct previous errors, increasing accuracy over time.
- **LightGBM (LGBM)**: A gradient boosting framework optimized for speed and efficiency, often outperforming traditional boosting techniques in large datasets.

Performance Metrics:

- **Accuracy**: Measures overall correctness of loan approval predictions.
- **Precision**: Evaluates how many of the predicted approvals were actually correct.
- **Recall**: Measures how well the model identifies approved applicants.
- **F1-score**: Harmonic mean of precision and recall, balancing false positives and false negatives.

5 Final Results

A comparison of the models used in both tasks is summarized below:

Model	R^2 Score
Linear Regression	0.78
Decision Tree	0.85
Random Forest	**0.91**
XGBoost	0.89

Table 1: Comparison of House Price Prediction Models

From the results, the Random Forest model performed the best for house price prediction, while LightGBM (LGBM) had the highest accuracy for loan eligibility prediction.

Model	Accuracy
Random Forest	87%
XGBoost	85%
LGBM	**88%**

Table 2: Comparison of Loan Eligibility Prediction Models

6 Conclusion

This project successfully developed machine learning models for house price prediction and loan eligibility assessment. The results indicate that advanced ensemble techniques such as XGBoost and Random Forest significantly improve accuracy. The model was deployed using Flask/Streamlit, allowing users to interactively check house price predictions and loan eligibility.

Future improvements include integrating real-time data sources and fine-tuning hyperparameters for better performance.

7 References

- Scikit-learn documentation: https://scikit-learn.org/
- Kaggle Housing Dataset: https://www.kaggle.com/datasets
- Flask Documentation: https://flask.palletsprojects.com/
- XGBoost Algorithm: https://xgboost.readthedocs.io/