Machine Learning I Project

Elgun Ismayilov

Introduction

1. Loan Default Prediction

Overview of the Problem:

- Banks face significant losses due to customers defaulting on loans.
- Impacts economic growth and financial stability.

  Objective:
- Build a model to predict loan defaults using client data.
- **Base Model**: Logistic Regression, with **SVC** and **Random Forest** for comparison.
- **Relevance/Importance**: Helps banks minimize financial losses.
- **Beneficiaries**: Banks, financial institutions, and the economy.

2.House Price Prediction

Overview of the Problem:

- Property prices are influenced by factors like location, size, and amenities.

  Objective:
- Predict property prices using real estate data.
- **Base Model**: Linear Regression, with **K-Nearest Neighbors** and **Decision Tree Regression** for comparison.

  Relevance/Importance: Helps buyers, sellers, and agents make informed decisions.
- **Beneficiaries**: Homebuyers, real estate agents, and market analysts.

Dataset and Data Preprocessing

1. Loan Default Prediction

Dataset:

- **Features**: Includes attributes such as funded amount, location, loan balance, income, credit score, etc.
- **Source**: Kaggle.

Data Preprocessing:

- **Label Encoding**: Applied to categorical features (e.g., loan status, education level) to convert them into numeric form.
- **Min-Max Scaling**: Used to scale numerical features (e.g., loan amount, income) to a [0,1] range for model compatibility.

2. House Price Prediction

Dataset:

- **Features**: Includes property size, number of rooms, location, neighborhood, amenities, etc.
- **Source**: Kaggle.

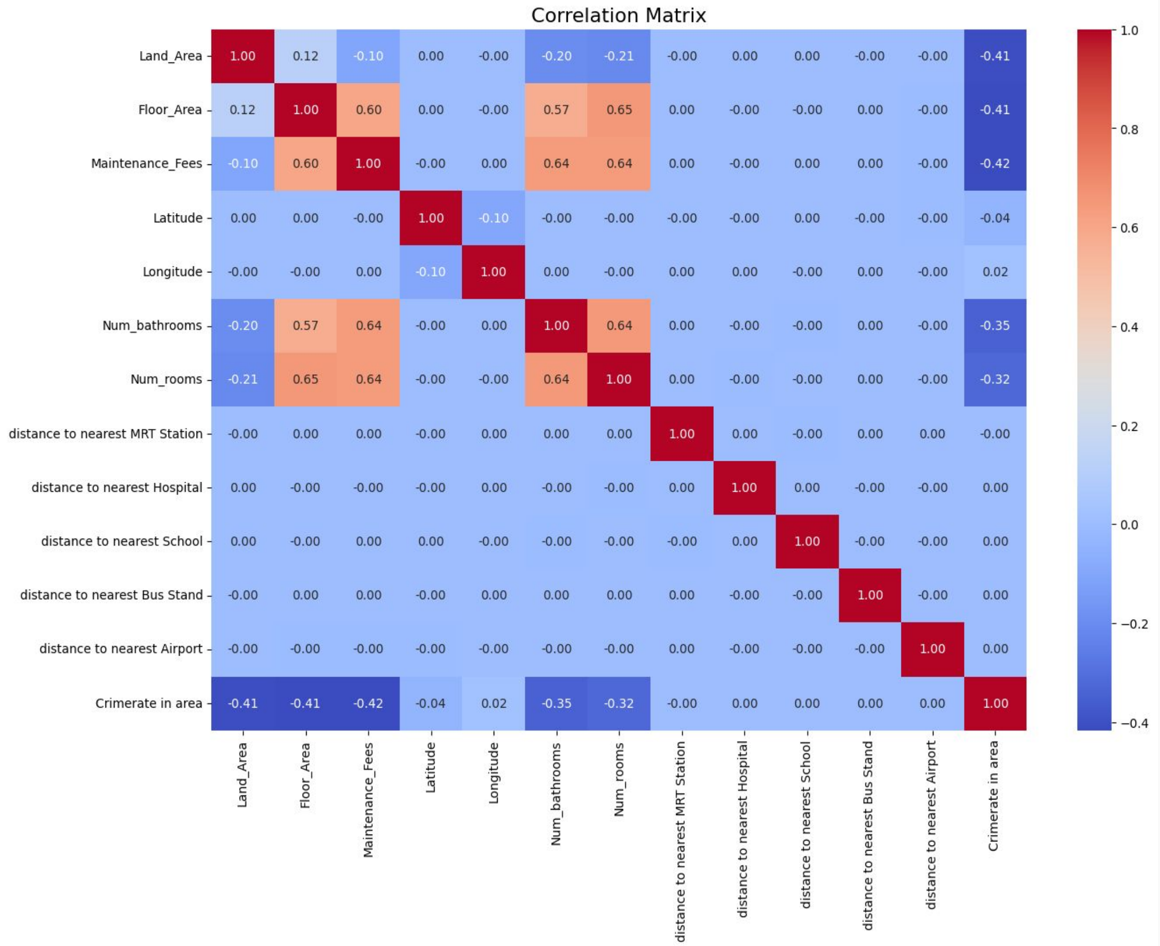
Data Preprocessing:

- **Label Encoding**: Categorical features (e.g., neighborhood type, house style) were encoded numerically.
- **Min-Max Scaling**: Scaled continuous variables (e.g., area, price) to standardize the input range for the models.

Exploratory Data Analysis

Key Observations to Mention:

- Correlation insights:
  - "Features like `Num_rooms` and `Floor_Area` are strongly correlated, which may influence models sensitive to multicollinearity."
  - "`Crime rate in area` has a negative correlation with important features like `Floor_Area`, indicating areas with larger properties tend to have lower crime rates."

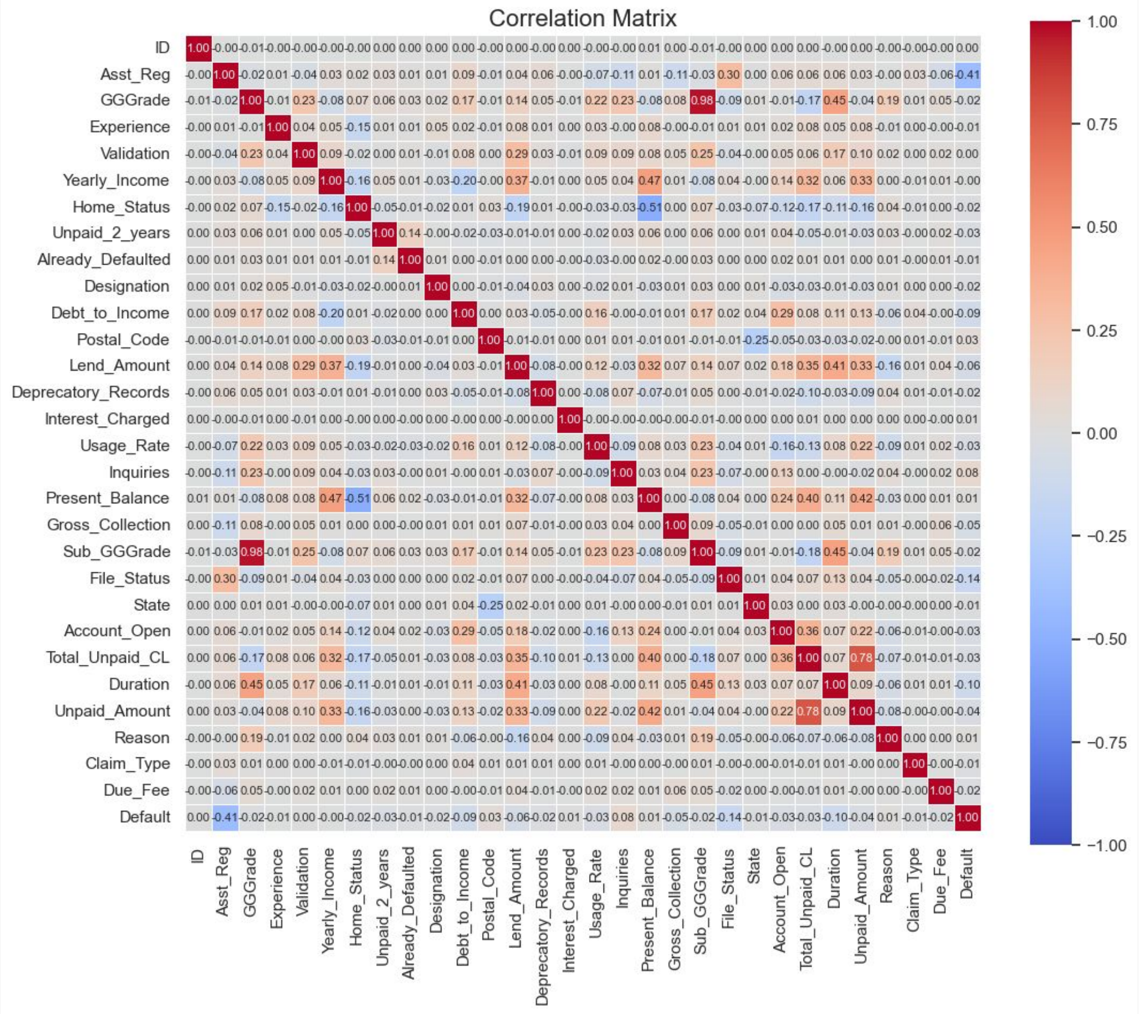


Exploratory Data Analysis

Correlation Matrix
This heatmap visualizes the correlation between various features in the dataset. Strong positive correlations are shown in red, while strong negative correlations are in blue. Key observations:

- Features like **GGGrade** and **Sub_GGGrade** are highly correlated (0.93), indicating redundancy.
- **Default** is negatively correlated with features like **ID** (-0.41).
- Most other features show weak or no correlation, suggesting low multicollinearity among them.

  This analysis helps identify significant relationships and redundant variables for predictive modeling.



Methodology

Loan Default Prediction

Algorithms Used:

- **Base Model**: Logistic Regression
- **Additional Models**: Support Vector Classification (SVC), Random Forest

Model Training:

- **Data Split**: The dataset was split into training (80%) and testing (20%) sets using **train_test_split**.
- **Training**: Models were trained using the training set with default hyperparameters initially.
- **Hyperparameter Tuning**: For Random Forest, parameters like `n_estimators` and `max_depth` were tuned using **GridSearchCV** for optimal performance.

Evaluation:

- **Metrics**:
  - Accuracy
  - F1-Score
  - ROC-AUC (for classification performance)

Methodology

2. House Price Prediction

Algorithms Used:

- **Base Model**: Linear Regression
- **Additional Models**: K-Nearest Neighbors (KNN), Decision Tree Regression

Model Training:

- **Data Split**: Split the dataset into training (80%) and testing (20%) sets.
- **Training**: Models were trained using default parameters.
- **Hyperparameter Tuning**:
  - For **Decision Tree**, tuned parameters such as `max_depth` and `min_samples_split`.
  - For **KNN**, tested different values of `k` (number of neighbors).

Evaluation:

- **Metrics**:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - R-squared (R^2)

Challenges Faced

Obstacles:

1. **Data Quality Issues**:
   - The dataset had **missing values** in some features, which could impact the quality of the model predictions.
   - **Imbalanced classes** in the loan default prediction dataset (non-defaulters outnumbering defaulters) caused models to bias predictions toward the majority class.
2. **Model Performance Issues**:
   - Some models, like **Random Forest**, exhibited overfitting, resulting in high performance on training data but lower performance on unseen test data.
   - **K-Nearest Neighbors (KNN)** had **long computation times**, especially during the grid search and prediction phase, making it inefficient for large datasets.
3. **Multicollinearity**:
   - The house price prediction dataset had some highly correlated features (e.g., property size and number of rooms), which created multicollinearity and instability in the regression models.

Results

1. Loan Default Prediction

Logistic Regression:

- **F1 Score**: 0.4016
- **Precision-Recall AUC**: 0.6274
- **Balanced Accuracy**: 0.6274

SVM (Support Vector Machine):

- **F1 Score**: 0.5249
- **Precision-Recall AUC**: 0.7538
- **Balanced Accuracy**: 0.7538

Random Forest:

- **F1 Score**: 0.3556
- **Precision-Recall AUC**: 0.6074
- **Balanced Accuracy**: 0.6074

2. House Price Prediction

Linear Regression:

- **Mean Squared Error (MSE)**: 0.001570
- **R^2 Score**: 0.9200

Decision Tree:

- **Mean Squared Error (MSE)**: 0.000661
- **R^2 Score**: 0.9663

K Neighbors:

- **Mean Squared Error (MSE)**: 0.005084
- **R^2 Score**: 0.7409

Conclusion

After evaluating the regression models based on **Mean Squared Error (MSE)** and **R^2 Score**, the following conclusions can be drawn:

- **Decision Tree** emerged as the **best performing model**, achieving the lowest **MSE (0.000661)** and the highest **R^2 Score (0.9663)**. This demonstrates its superior predictive accuracy and ability to capture the variance in the data effectively.
- **Linear Regression**, while simpler to implement, showed lower performance with an **MSE of 0.001570** and an **R^2 Score of 0.920011**, making it less suitable for this task.
- **K-Nearest Neighbors (KNN)** had the poorest performance with an **MSE of 0.005084** and an **R^2 Score of 0.7409**. Furthermore, its **high computational time** (approximately 7 minutes) during grid search and prediction makes it impractical for larger datasets.

Recommendation

The **Decision Tree** model is the recommended choice for its combination of accuracy, efficiency, and computational speed. **KNN**, while potentially useful for smaller datasets, requires careful consideration due to its computational overhead.

-

Conclusion

After evaluating **Logistic Regression**, **SVM**, and **Random Forest**:

1. **SVM** performed the best across all metrics with the highest **F1 Score (0.5249)**, **Precision-Recall AUC (0.7538)**, and **Balanced Accuracy (0.7538)**, making it the optimal choice for this classification task.
2. **Logistic Regression** provided faster computation but had lower performance in comparison to **SVM**.
3. **Random Forest** showed decent performance but required significantly more computation time.

Recommendation: **SVM** is the best model for accuracy, while **Logistic Regression** is ideal for speed, and **Random Forest** may be considered for specific use cases where robustness is key, despite longer training times.

Thank You!