# Machine Learning II 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 **MLP** and **LightGBM** 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 **XGBoost** and **CatBoots** 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**:

- **Size**: 87,501 rows and 30 columns
- **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.

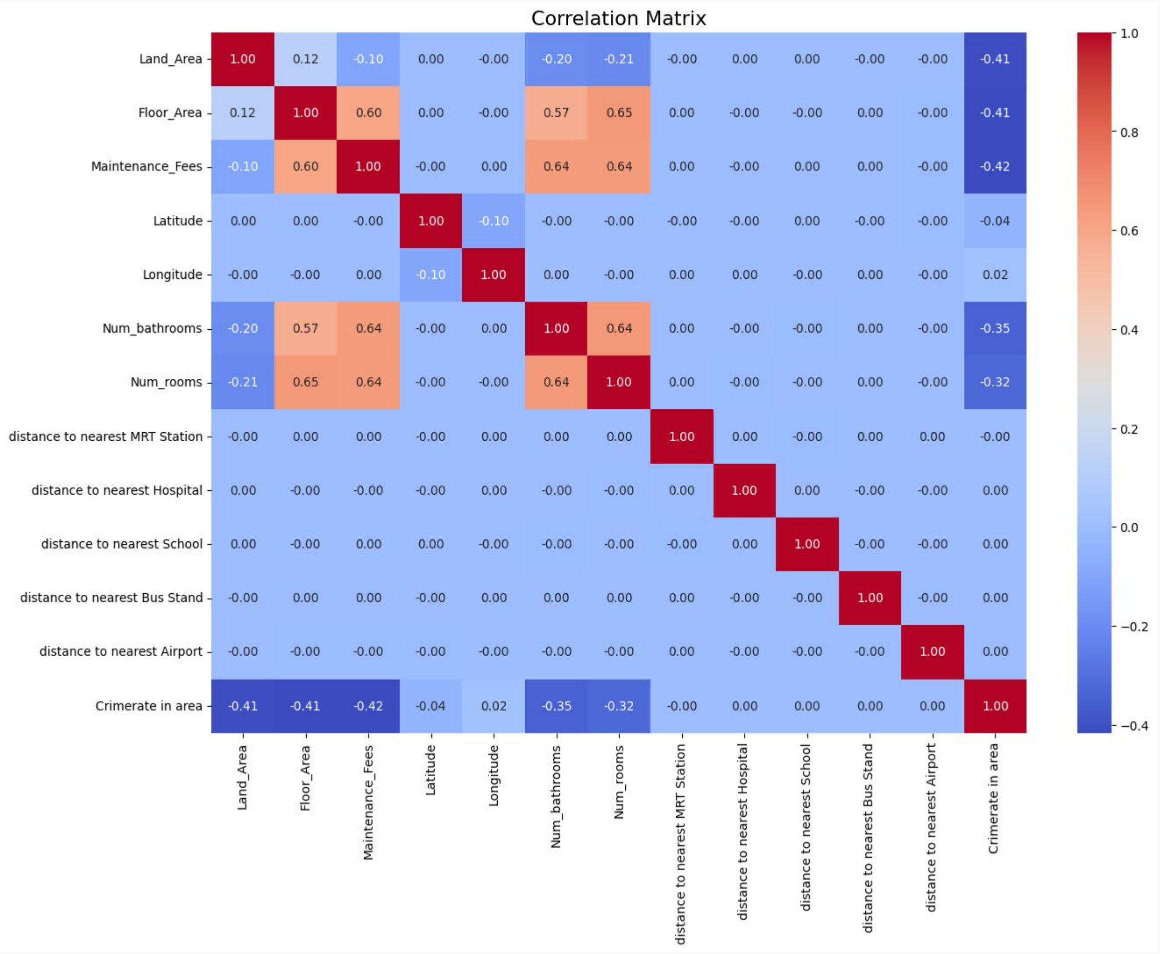
**Data Preprocessing**:

- **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.

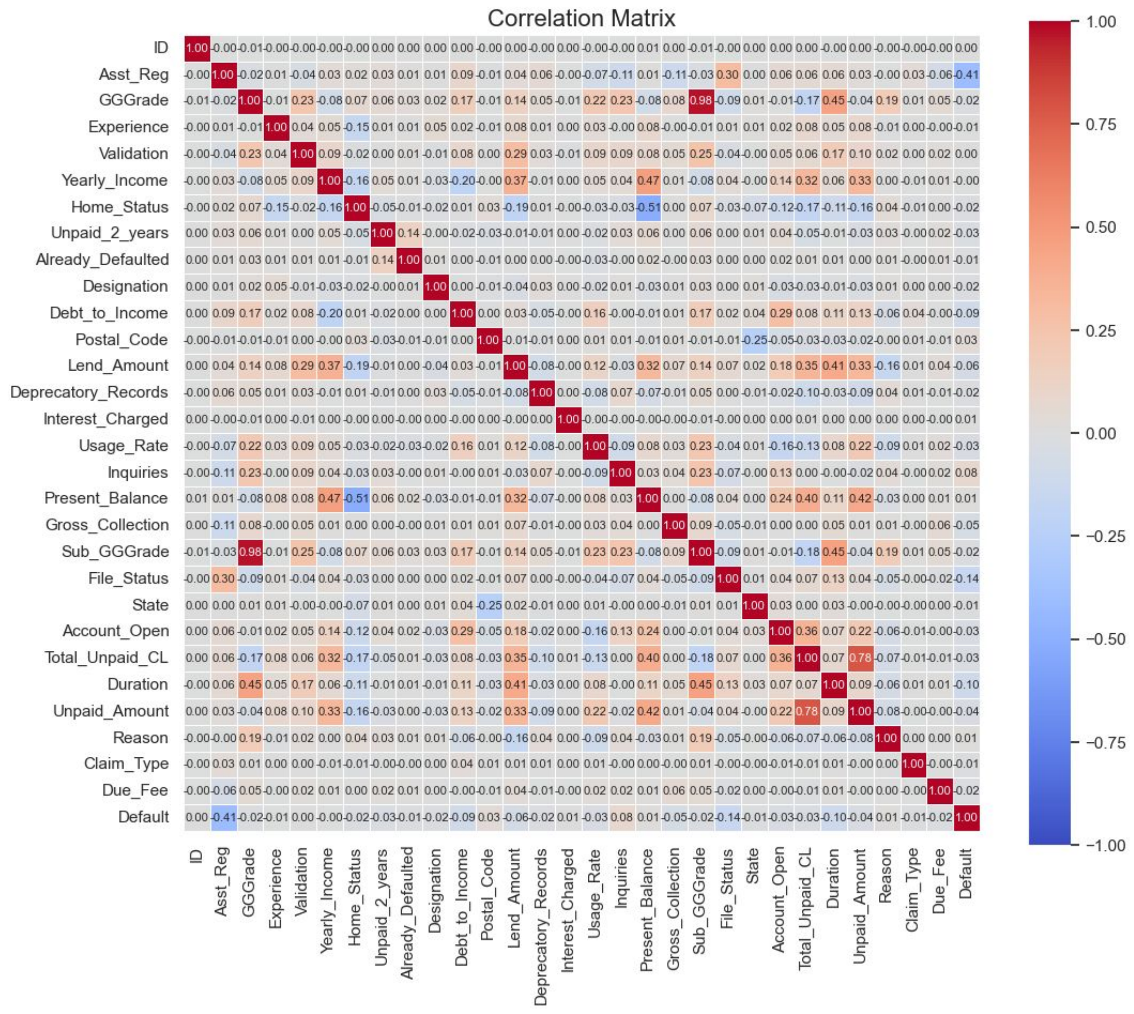
# Exploratory Data Analysis

**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."
- Impacts on model performance:
  - "Linear models might suffer from multicollinearity, whereas tree-based models like XGBoost can handle it well."



# 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**: XGBoost and CatBoost model

**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 XGBoost and Catboost model, 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**: LightGBM model MLP

**Model Training**:

- **Data Split**: Split the dataset into training (80%) and testing (20%) sets.
- **Training**: Models were trained using default parameters.
- **Hyperparameter Tuning**:
  - For LightGBM model and MLP, tuned parameters such as `max_depth` and `min_samples_split`.

**Evaluation**:

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

# Results

## 1. Loan Default Prediction

### Performance Metrics

- **Logistic Regression**
  - **Accuracy**: 83.28%
  - **F1 Score**: 0.6184
  - **Precision-Recall AUC**: 0.6274
- **XGBoost**
  - **Accuracy**: 83.95%
  - **F1 Score**: 0.6838
  - **Precision-Recall AUC**: 0.6242
- **CatBoost**
  - **Accuracy**: 84.03%
  - **F1 Score**: 0.6918
  - **Precision-Recall AUC**: 0.6243

## 2. House Price Prediction

### Performance Metrics

- **Linear Regression**
  - **MSE**: 0.001569
  - **$R^2$**: 0.920011
- **LightGBM**
  - **MSE**: 0.000657
  - **$R^2$**: 0.966484
- **MLP (Multi-Layer Perceptron)**
  - **MSE**: 0.000923
  - **$R^2$**: 0.952954

# Conclusion

## Conclusion

After evaluating multiple models for regression, the following insights were derived:

1. **Best Performing Model**:
   - **LightGBM** demonstrated the highest performance with an **$R^2$ score of 0.966484** and the lowest **Mean Squared Error (MSE) of 0.000657**, making it the most accurate model for this task.
   - Its efficiency and ability to capture complex patterns in the data make it an ideal choice.
2. **Competitor Models**:
   - The **MLP (Multi-Layer Perceptron)** model achieved a strong **$R^2$ score of 0.952954**, but its **long runtime (~10 minutes)** limits its practical usability for larger datasets or time-sensitive applications.
   - **Linear Regression**, while computationally efficient and simple to implement, achieved a significantly lower **$R^2$ score of 0.920011**, indicating it struggled to model the complexity of the dataset.

# Conclusion

Based on the results of the evaluation:

- **XGBoost** is the most effective model, achieving the **highest F1-Score** and demonstrating a strong balance between precision and recall.
- **CatBoost** offers comparable performance, making it a suitable alternative when categorical data optimization is critical.
- **Logistic Regression**, while less effective in this case, provides a simpler and computationally efficient option for basic use cases.

## Final Recommendation

For this task, **XGBoost** is recommended as the best model due to its superior performance across key metrics, making it highly reliable for loan default prediction.

# Thank You!