Summary Report: House Price Prediction

Introduction

This project aims to predict house prices using the **Ames Housing Dataset**. The approach includes **data preprocessing, feature engineering, exploratory data analysis, model training, and evaluation**. Two machine learning models—**Random Forest Regression** and **Neural Network Regression**—were compared based on their performance.

Data Preprocessing & Feature Engineering

Dataset Overview

- The dataset contains **2930 observations** and **82 features**.
- Only the most relevant features were selected.

Data Cleaning

- Missing values were identified and handled using **median imputation for numerical features** and **'Unknown' category for categorical features**.
- The dataset was transformed by **one-hot encoding categorical features** and **standardizing numerical features**.

Exploratory Data Analysis

Several visualizations were used to understand the dataset:

- **Correlation Heatmap**: Showed strong correlations between SalePrice and variables like Overall Qual, Gr Liv Area and Garage Cars.
- **Histograms and Scatter Plots**: Helped in identifying distributions and relationships between features.
- **Bar Plots**: Showed the effect of categorical variables like Neighborhood, Kitchen Qual and Bsmt Qual on sale price.

Model Training and Results

Random Forest Regression

- A **Random Forest Regressor** was trained on the dataset.
- Initial prformance:
    - **Mean Absolute Error (MAE)**: 17,131

    - **Root Mean Squared Error (RMSE)**: 29,788
    - **R-squared (R^2)**: 0.878

Hyperparameter Tuning

- **Hyperparameter tuning** done using Grid Search, leading to the best parameters:
    - **Best parameters**: max_depth=15, min_samples_leaf=1, min_samples_split=2, n_estimators=150
- After tuning, model performance:
    - **R^2 Score**: 0.882

Neural Network Regression

- A **Neural Network Model** was built using PyTorch:
    - **Input Layer**: Number of features
    - **Hidden Layers**: Two layers with **ReLU activation** and **dropout (0.2)**
    - **Output Layer**: Single node for regression
- Training was done for **150 epochs** with **Adam optimizer** and **MSE Loss**.
- Final result:
    - **R^2 Score**: 0.822

Model Comparison & Conclusion

- The **Random Forest model performed better** with an **R^2 score of 0.882**, compared to **0.822** from the Neural Network.
- A **bar chart comparison** confirmed that Random Forest was the more accurate model.
- **Final Conclusion**: Random Forest is the preferred model for predicting house prices due to its better accuracy and stability.

This report outlines the **key steps, insights, and model comparisons** for predicting house prices. The Random Forest model was found to be the best performer based on R2 score.