Predicting House Prices Using Random Forest Regression

Introduction

- **Objective:**

  - The goal of this project is to develop a Random Forest Regression model to predict house prices using the Ames Housing Dataset.
  - This dataset contains various features describing properties, including lot size, number of bedrooms, overall condition, and neighborhood.
  - The project involves data preprocessing, feature selection, model training, evaluation, and visualization of results to ensure accurate price predictions.

- **Dataset Used:** Ames Housing dataset

Dataset Link:

https://www.openintro.org/data/csv/ames.csv

- **Approach:**

  - Data preprocessing, including handling missing values and feature engineering.
  - Split the data into Training and Testing Sets.
  - Training a Random Forest regression model.
  - Evaluating model performance using appropriate metrics.

Data Preprocessing

- Handled missing values by dropping irrelevant data to avoid excessive data loss.
- Missing values in categorical and numerical features are replaced with Most Frequent and Median value with respect to columns.
- Machine learning models require numerical input, categorical variables are encoded using Label Encoding.
- Defining feature matrix and target variables.

Model Selection & Training

- Chose Random Forest Regression due to its robustness and ability to capture non-linear relationships.
- Split the data into training and testing sets (e.g., 80-20 split).
- Tuned hyperparameters using GridSearchCV (parameters such as the number of estimators, max depth, and min samples split).
- Trained the random forest model on the processed dataset.

Model Evaluation

- **Metrics Used:**
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - Mean Absolute Error (MAE)
    - R-squared (R^2)
- **Performance Analysis:**
    - Evaluated the model on the test set.
    - Compared predicted vs actual prices.
    - Assessed feature importance to understand key factors influencing price predictions.
- **Feature Importance:**
    - Overall Quality - A key determinant of house value.
    - Basement Area - Larger basements often contribute to higher prices.
    - Living Area (Above Ground) - The main living space size significantly impacts pricing.
    - Garage Area - A larger garage adds value to a house.
    - Number of Bathrooms - More bathrooms generally increase house value.

Results & Visualizations

- Feature importance plot highlighting the most influential variables.
- RMSE and R^2 scores to measure model effectiveness.

Dataset Columns:

- Dataset columns: ['order', 'pid', 'area', 'price', 'ms.subclass', 'ms.zoning', 'lot.frontage', 'lot.area', 'street', 'alley', 'lot.shape', 'land.contour', 'utilities', 'lot.config', 'land.slope', 'neighborhood', 'condition.1', 'condition.2', 'bldg.type', ' house2.style' , 'overall.qual', 'overall.cond', 'year.built', 'year.remod.add', 'roof.style', 'roof.matl', 'exterior.1st', 'exterior.2nd', 'mas.vnr.type', 'mas.vnr.area', 'exter.qual', 'exter.cond', 'foundation', 'bsmt.qual', 'bsmt.cond', 'bsmt.exposure', 'bsmtfin.type.1', 'bsmtfin.sf.1', 'bsmtfin.type.2', 'bsmtfin.sf.2', 'bsmt.unf.sf', 'total.bsmt.sf', 'heating', 'heating.qc', 'central.air', 'electrical', 'x1st.flr.sf', 'x2nd.flr.sf', 'low.qual.fin.sf', 'bsmt.full.bath', 'bsmt.half.bath', 'full.bath', 'half.bath', 'bedroom.abvgr', 'kitchen.abvgr', 'kitchen.qual', 'totrms.abvgrd', 'functional', 'fireplaces', 'fireplace.qu', 'garage.type', 'garage.yr.blt', 'garage.finish', 'garage.cars', 'garage.area', 'garage.qual', 'garage.cond', 'paved.drive', 'wood.deck.sf', 'open.porch.sf', 'enclosed.porch', 'x3ssn.porch', 'screen.porch', 'pool.area', 'pool.qc', 'fence', 'misc.feature', 'misc.val', 'mo.sold', 'yr.sold', 'sale.type', 'sale.condition']

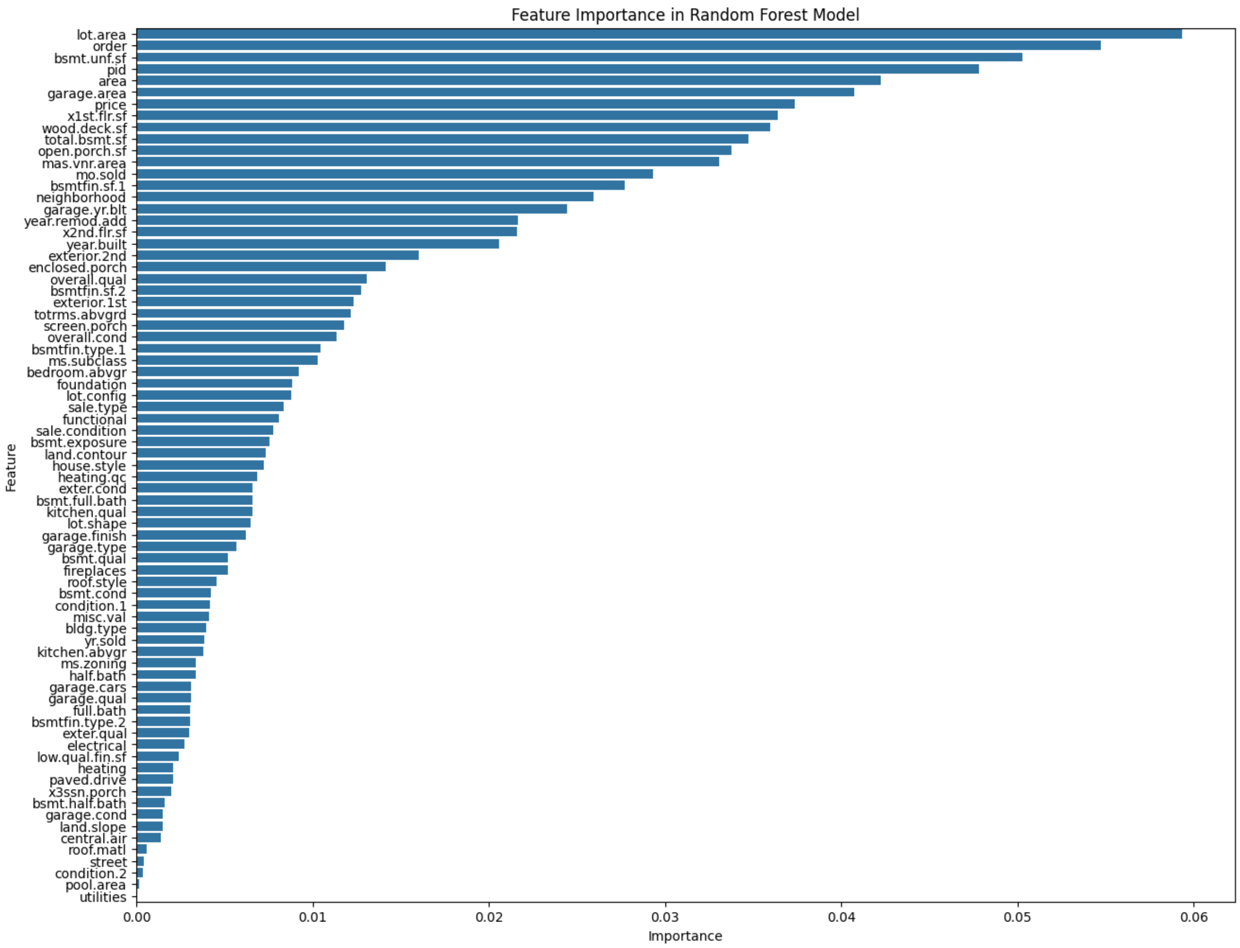
Model Evaluation:

```
Random Forest Performance:
MAE: 96677.68166127766
MSE: 12803570048.298616
RMSE: 113152.8614233799
R-squared: -0.014593773961467882

XGBoost Performance:
MAE: 98706.1640625
MSE: 13526952960.0
RMSE: 116305.42962390019
R-squared: -0.07191681861877441

LightGBM Performance:
MAE: 99139.07393181145
MSE: 13618635768.840046
RMSE: 116698.910744017
R-squared: -0.0791820569412327
```

Feature Importance:



Conclusion & Next Steps

- **Summary of Findings:**
  - Random Forest provided a reliable model for predicting house prices with reasonable accuracy.
  - Feature importance analysis indicated key drivers such as lot area, overall quality, and living area.

- **Future Improvements:**

  - Experimenting with ensemble models like XGBoost for improved accuracy.
  - Fine-tuning feature engineering techniques.
  - Using advanced techniques such as stacking or boosting for better generalization.

Links

Notebook Link:

https://colab.research.google.com/drive/1BB5kAfTf7JInPxbL-BvCnsjP0zPbpxzL?usp=drive_link

Drive Link which includes Dataset, Notebook:

https://drive.google.com/drive/folders/1j3i3hHZNRt2gkGmPeBSbQFfDgsvngH_0?usp=sharing

Preferred Platform (Used Platform to deploy this Assignment):

Google Colab: https://colab.research.google.com