House Price Prediction

This project builds and compares three machine learning models — Linear Regression, Random Forest, and XGBoost — to predict housing prices using data from the Kaggle competition House Prices: Advanced Regression Techniques. It demonstrates a complete pipeline: data retrieval, preprocessing, model training, evaluation, and result visualization.

Project Files

File/Folder	Description
`data_utils.py`	Downloads and unzips data from Kaggle API
`model.py`	Trains Linear, Random Forest, and XGBoost models
`images/`	Stores output plots (prediction vs. actual, importance)
`report.md`	Final report for academic submission
`README.md`	This file

How to Run

1. **Place your Kaggle API key in:** `~/.kaggle/kaggle.json`
   You can get your API key from your Kaggle account settings.
2. **Download the dataset:**

```
python data_utils.py
```

3. **Train models and generate plots:**

```
python model.py
```

4. **Check output plots in the images/ folder:**

- `lr_pred_vs_actual.png`
- `rf_pred_vs_actual.png`
- `xgb_pred_vs_actual.png`
- `rf_feature_importance.png`
- `xgb_feature_importance.png`

5. **Install required packages:**

```
pip install pandas numpy matplotlib scikit-learn xgboost kaggle
```

Model Comparison

Model	RMSE	R^2 Score
Linear Regression	42,509.32	0.7644
Random Forest	32,311.91	0.8639
XGBoost	32,172.62	0.8651

All three models performed well, with XGBoost achieving the best accuracy. Random Forest was close behind and provides meaningful feature importance. Linear Regression serves as a good baseline model.

Full Report

Read the full analysis and discussion in `report.md`

Author

[Caroline Kuo] (https://github.com/Tiff0416/house-price-prediction)