Project 5

House Price Predictions (Kira)

Implementation details:

This folder contains 3 python notebooks, while evaluating please go through them in the following order:

1) **Exploratory analysis:**
   Contains detailed explanation of all the steps that I did in the Exploratory data analysis.
2) **Data Cleaning:**
   Contains detailed explanation of cleaning data and outlier detection and the feature engineering.
3) **ML_Models:**
   Contains all the models that I have tried.

Required libraries : xgboost,pandas,numpy,scikit-learn etc.

Comparison of different models:

Model	RMSE (Validation Data)	Kaggle Ranking (Test Data)
Simple Linear Regression	620.755	50234
Lasso Regression	0.1071	1534
Ridge Regression	0.1157	1564
XGBoost	0.0154	780(Top 2%)
Random Forest	0.0165	1045
Elastic Net	0.0111	1578

Best Model:

“XGBoost”

Reason I think it has a good performance is because. It does regularization which does not allow the overfitting. And it does the cross validation. And it finds out trends in the outliers and avoids them.It uses a boosting algorithm reducing the bias.

Lasso which also does regularization and feature selection also worked well on the test data but it could not beat XGBoost.

Random Forest which is an ensemble method also got a great score. I think it worked great because this uses bagging.