PROJECT A11

repository https://git.io/JDCHZ

Hendrik Šuvalov
Raul Tölp

Predicting house prices

Kaggle competition

UNIVERSITY OF TARTU
Institute of Computer Science

01. Introduction

This project is based on Kaggle competition Ames Housing dataset, which was compiled by Dean De Cock. It consists of about 3000 observations which include 79 features along with the price, describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. Provided dataset proves that price negotiations influence the price much more than the number of bedrooms or a white-picket fence. With given features describing almost every aspect of residential homes, this competition challenges contestants to predict the final price of each home.

02. Objective

Our goal was to find out the main features that influence the house price and create a model that predicts it accurately.

Success Criteria

Success criteria can be directly measured by how accurate our results are once we submit these on Kaggle competition.

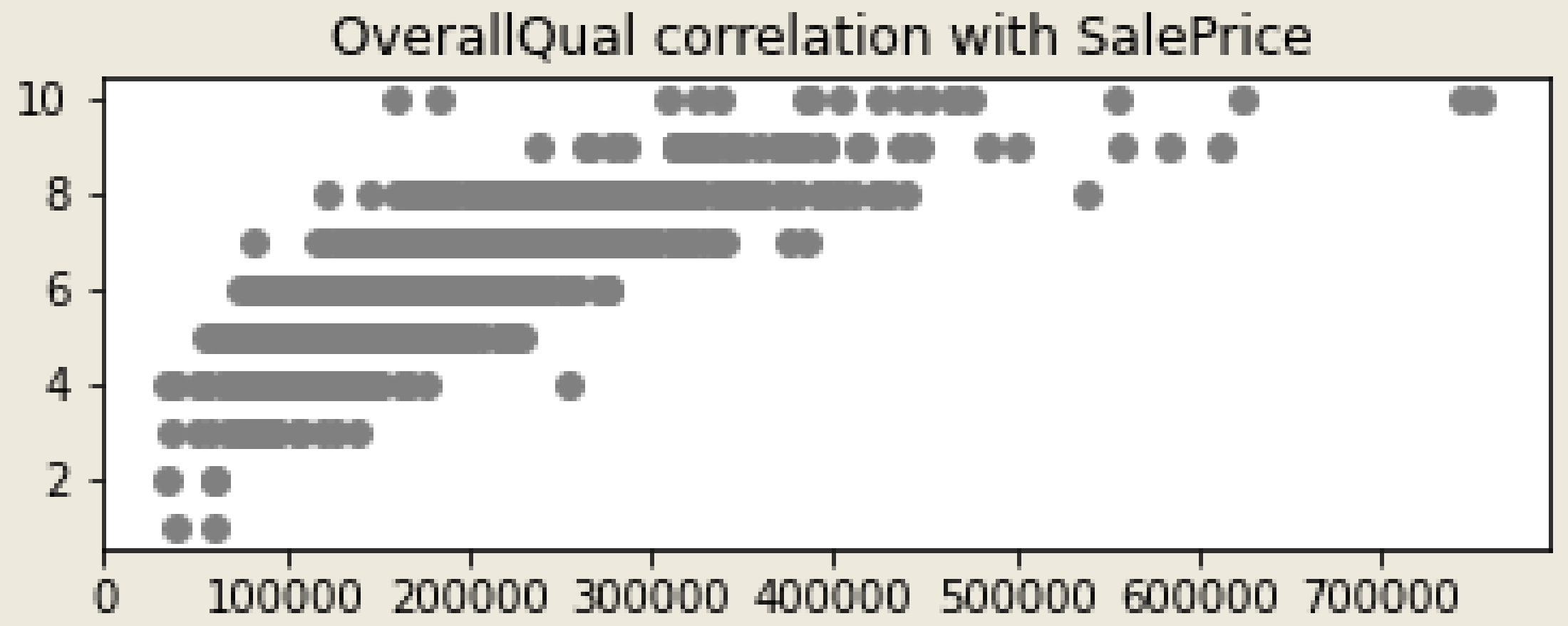
03. Methodology

Prediction methods used for cross validation score and RMSE:

- **Correlation**
- **Linear Regression (Lasso & Ridge)**
- **RandomForest Classifier / Regressor**
- **GradientBoosting Classifier / Regressor**

04. Analysis

Analysis was started with finding the best correlations between sales price and other features. Some features had to be normalized or encoded to one-hot vectors for modelling. We found that house sales price is mainly affected by overall quality, ground living area size, size of garage in car capacity, size of garage, size of basement area, size of the first floor.

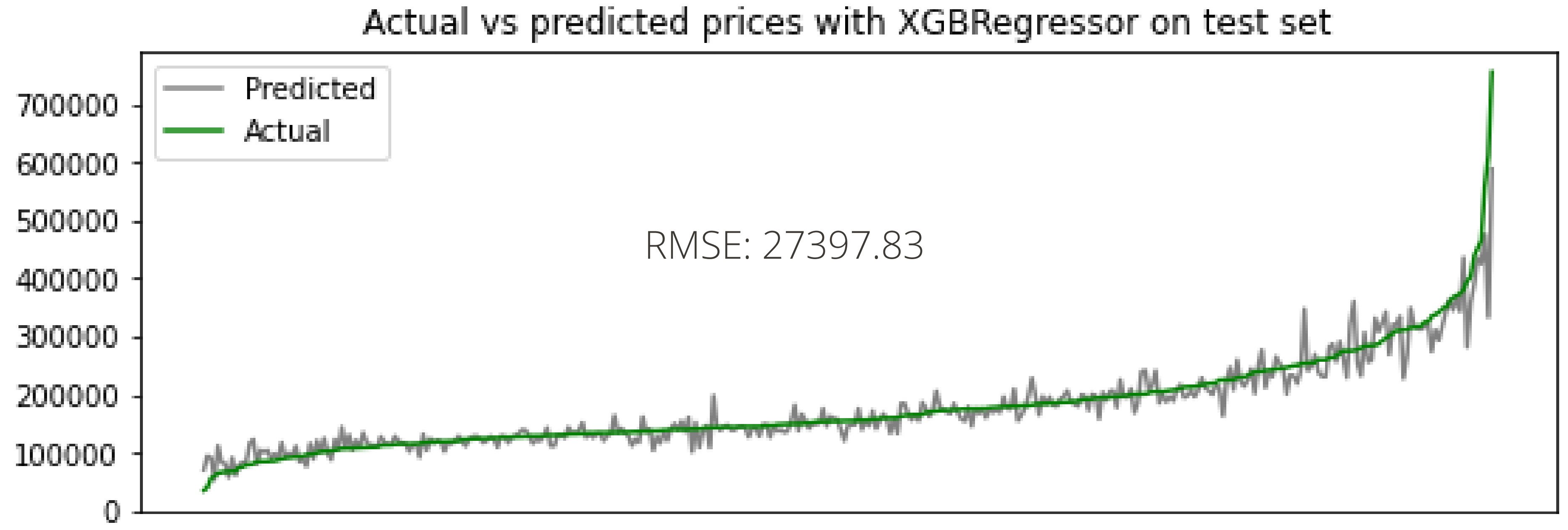


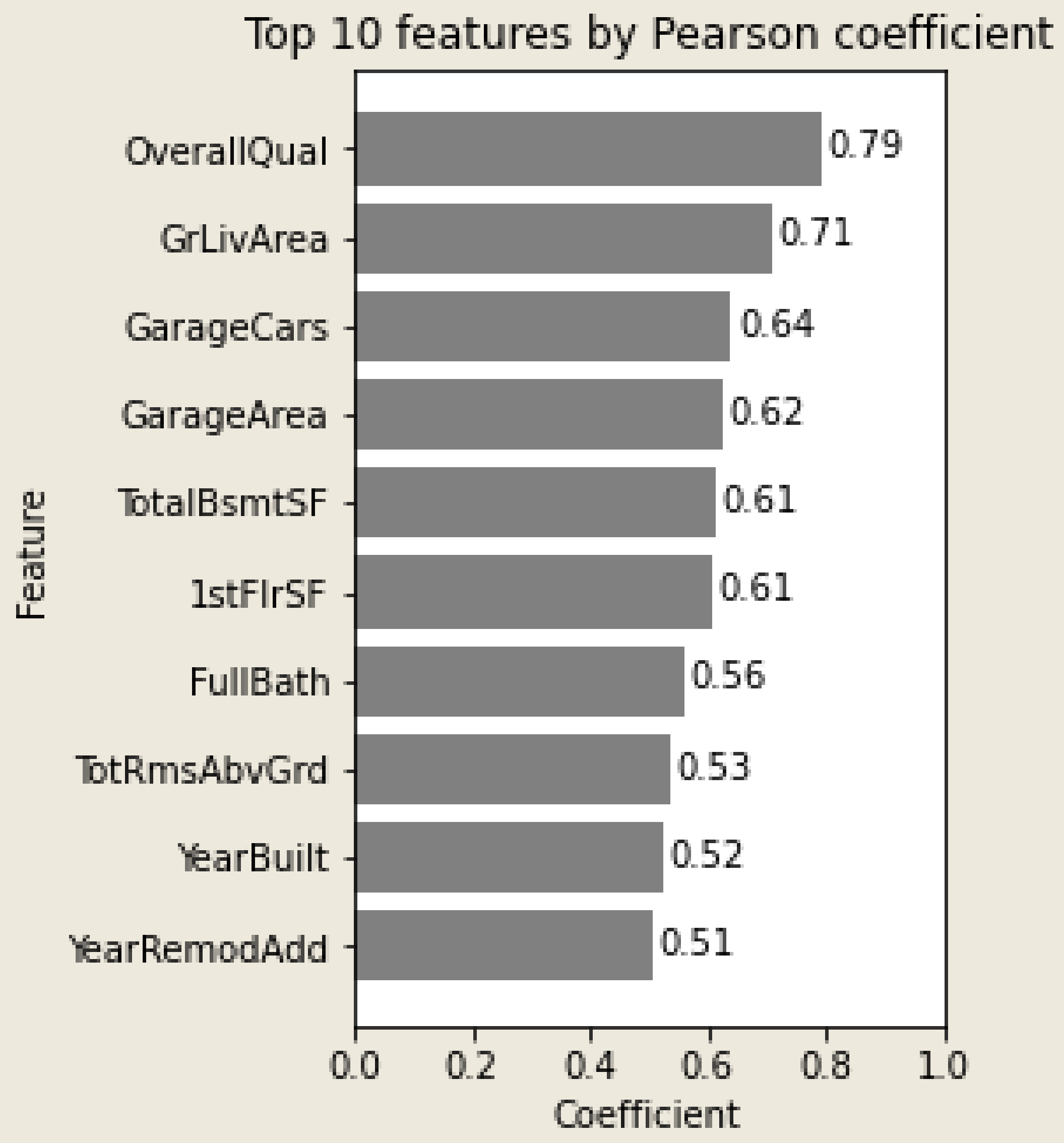
05. Results

Overall, regression models heavily outperformed classification models. We initially created models using parameters that intuitively made sense to us and submitted the predictions to Kaggle. For models that gave promising results, we then fine-tuned the parameters to get the best results possible and submitted multiple attempts with slightly differing parameters to get the best score from Kaggle.

The best results were achieved with Gradient Boost Regressor where the final Kaggle score was 0.13330. This result placed us in the top 32% of the Kaggle competition leaderboard. The second best model was a Random Forest Regressor that was slightly less accurate with a score of 0.15433 and the third one was Lasso Linear Regression model with 0.19229, although we didn't spend as much time fine-tuning them as we did for the Gradient Boost Regressor.

Below is a graph showing the differences in predicted vs actual prices on a 25% train-test split of the given training data:





06. Conclusion

We met our goal to finish in top of 1/3 of the Kaggle competition leaderboard. This kind of task helps to understand the data analysis and gave us a good overview of prediction methods and feature engineering options.