Predicting Housing Prices in Ames, Iowa - Team Poplar
Erdős Institute - Fall 2022

Michael Annunziata, Gaurish Korpal, Bartosz Langowski, Abraham Rabinowitz

Problem Description

The aim of this project is to accurately predict and explain the features that have the greatest impact on the sale price of houses in Ames, Iowa. We can help sellers understand what features of a house provide the greatest impact on value, and buyers attain a fair price and budget toward a home purchase. Our key performance indicators are the root-mean-squared error (RMSE) of our predictions, as well as the mean absolute error (MAE). Our main KPI is RMSE, as it is more sensitive to inaccurate predictions. We also aim for an interpretable model to understand how the house is priced.

Data Source and Exploratory Analysis

We analyzed the "House Prices - Advanced Regression Techniques" dataset from Kaggle which provides a description of the sale of 2930 individual residential properties in Ames, Iowa from 2006 to 2010. It contains 80 explanatory variables with numerical and categorical data such as square footage, neighborhood, and overall house condition. Our model is trained on the 1460 labeled instances provided in the training set.

Firstly, we removed features with many missing entries. We then looked at a correlation heat map to drop features that we considered redundant. Looking at some scatter plots of Sale Prices against highly correlated features made it clear that there was a linear relationship. To prepare for linear modeling, we applied a logarithm transform to the sale price as well as the skewed features. We transformed the categorical data into numerical data with the use of one-hot encoding.

Modeling Approach

Due to the strong linear relationship, we considered Ridge and ElasticNet regression model. Nested cross-validation showed that the ElasticNet performed quite well, with a log RMSE of about .0122. However, it removed many categorical variables and did not necessarily account for possible non-linear relationships. To capture these nonlinearities we also tried two gradient boosting models: XGBoost (XGB) and LightGBM (LGBM). Nested cross-validation showed they had a log RMSE of about .014 which was weaker by itself than a linear model but high enough to demonstrate that they were capturing some signal. Consequently, we decided to average XGB, LGBM, and ElasticNet. The combined model achieved a cross-validation score of .0117. As this was the best score, we chose it for prediction on unseen data.

Results and Conclusions

We tested our combined model on 1459 new instances in the test set, resulting in an RMSE of $27000 - a significant improvement over a linear baseline model with an RMSE of $45000. However, this stark difference is largely attributed to an outlier with a sale price far below the predicted value. By MAB which does not weigh this outlier as heavily, we see that on average our model is within $13000 of the correct price as opposed to $21500 for the linear model. Our model is better at pricing more expensive houses, though it struggles with extreme outliers. Further work can include combining outlier detection to flag houses that may be more appropriately priced by a domain expert. Finally, we use SHAP which quantifies the impact each feature values have on our model's predictions of sale prices. The features with the most value are mostly those dealing with the size and age of the house. However, features such as overall condition, time since remodeling, and kitchen quality can add as much as $25000 on whether renovations prior to sale would be worth undertaking.