House prices - Erdös Institute DS Bootcamp 2024

Krescens Kok and Jonathan Galvan

April 2024

Project Overview: The housing market consistently changes, especially from year to year and there are many other features of a house that can impact the cost of a home. We decided to see whether we could predict how much a house will sell based on features such as the quality, number of bedrooms, square footage, fireplaces, etc...

Stake Holders: This information is useful for: Realtors, Home buyers, Home sellers, Investors

Project Goal: Determine whether we can predict how much a house will sell based on numerous features.

Dataset: Our dataset consists of houses in Ames, Iowa that were sold throughout numerous years. The dataset includes 79 characteristics/columns for 1,460 houses, including both numerical and categorical variables. The characteristics could be grouped into separate categories such as : Geographical, Size Measures, Quality Measures, Special Features, Status of Garage and Basement.

Data Pre-Processing:

The missing values whose abscence indicate a feature that was not present were inputed with 'None' or with zero. Ohter numerical values were filled using a KNN-imputer. The categorical data was processed using Ordinal Encoding for the data with intrinsic ordering and One-Hot encoding for the one without intrinsic orderings. The numerical columns were also processed using a MinMax Scaler to account for differences in scale and outliers.
More details about our pre-processing can be found here: [Data-Cleaning].

Approach:

We tried to predict the logarithm of the house prices to uniformize the contribution of individual errors.

- We found that using a RandomForestRegressor resulted in a mean squared error of .00357 and the r^2 value was .868 (picture here).
- We also tried using the XGBoost model, and also resulted in pretty good metrics with a mean squared error of .00335 and an r^2 value of .876 (picture here).
- Lastly, we used a linear regression to use for comparison and found that the mean squared error was .0089 and an r^2 value of .844 (picture here).

Results

Looking at the metrics of all 3 models, we concluded that the XGBoost Regressor was the best, resulting in the lowest MSE and highest r^2 value.
We also were able to determine the most important features that would help with this model, by looking at the weight. The top 3 most important features include:

- GrLivArea
- Overall_Rating
- LotFrontage

It makes sense that these features would be important in predicting a house price since usually, the bigger the house/lot area, the higher the sale price.

Future Improvements:

Future iterations of this project could be to try to understand even better the feature selection to see exactly how much do they contribute to a price of a house and try to create a minimal model that predicts prices with good accuracy using a very small number of features.

Conclusion:

Back to our research question and project goal of can we determine whether we can predict how much a house will sell based on numerous features? We determined that we can indeed predict the ‘SalePrice‘ based on about 30 features. Using all the features with the Linear Regression resulted in a larger mse, indicating that there was overfitting involved.