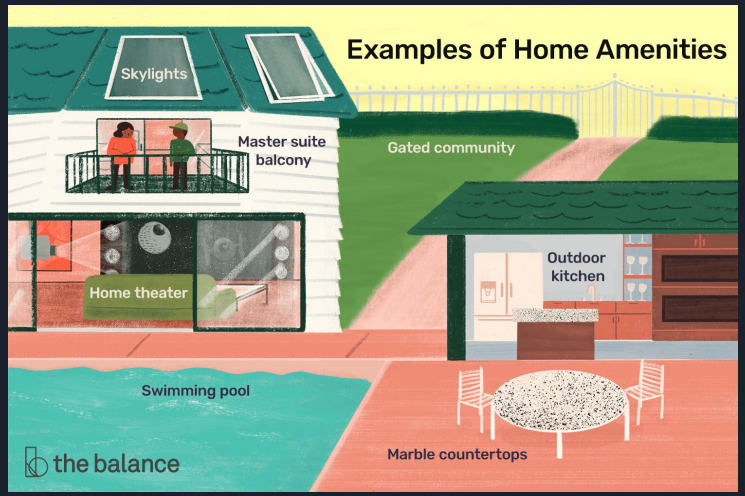
House Prices
Are you overpaying?

Jonathan Galvan and Krescens Kok

Introduction and Problem Statement



What do you look for in your dream house? Wood floors, skylights, enough rooms for your whole family? You can come up with a few criteria to determine this, and even more so, you can probably decide if a house is too expensive by taking just this few factors in mind.

Now, what if we had over 78 factors to consider (79 to be exact!) when deciding the price of a house?

In this project we aim to look at a large number number of house features and try to predict the house price.

Data Collection and Description

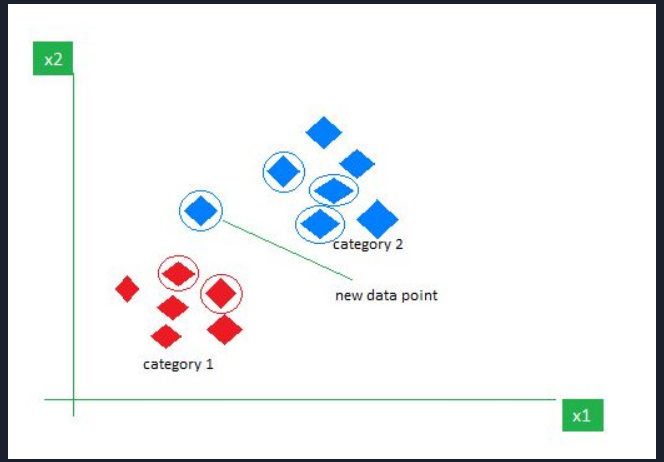
We used a database of houses in Ames, Iowa. The dataset consists of 79 characteristics for 1460 houses.

Some examples of the variables to be considered are:

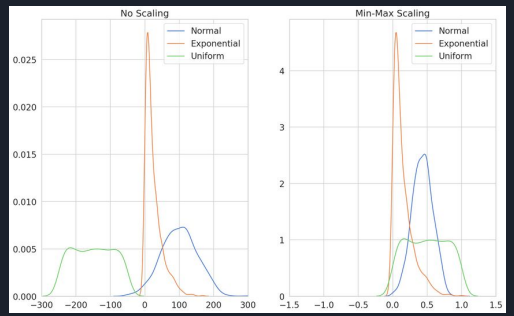
- **Geographical:** Zoning classification, Neighborhood, Lot shape, Land Slope, etc.
- **Size measures:** Lot Frontage Footage, Liveable square footage, Square footage of the basement, etc.
- **Quality measures:** Overall condition, Roof material, Exterior covering, Utilities present, etc.
- **Special features of the house:** Fireplaces, Pool, Fences, Elevator, Tennis Courts, etc.
- **Status of the Garage and Basement:** Number of cars fitting in the garage, Number of bathrooms in the basement, Square Footage of finished basement, etc.

Data cleaning and processing

- To fill out the missing data we analyzed houses sharing similar properties and used different imputing techniques to predict the correct values.
    - For example to extrapolate the Lot Frontage Footage we used a K-Nearest Neighbors imputer with respect to the Lot Area, Lot Configuration and the Square Footage of the 1st floor.



- To process categorical data we used two different approaches, the variables that had intrinsic orderings were processed using an ordinal encoder and the rest of them were processed using One-Hot encoding.



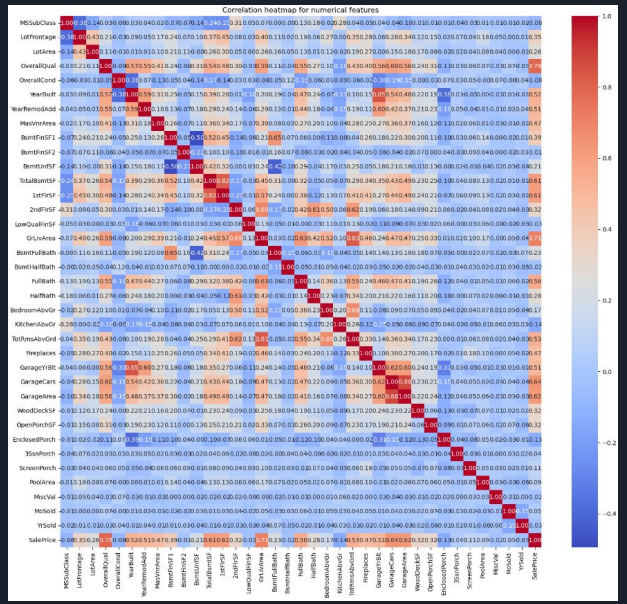
- The numerical data was processed through a MinMaxScaler.

Feature selection

After the data processing we ended up with more than 200 features so we had to reduce it somehow.

We did this by:

- Calculating correlations between features to detect multicollinearity.
- Merging features by creating new ones.
- Dropped features that had little correlation with the price or had so many categories that it was unlikely that they would help predict the price.

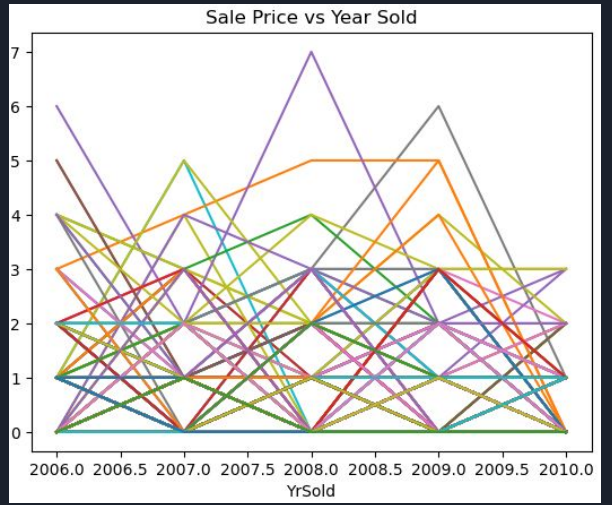


Modeling approach

Since we had information about the dates of sale of the houses, we briefly consider treating the data as a Time Series. We decided to change our approach and used ensemble regression methods and linear regression for comparison.

We used:

- Random Forest regressor
- An XGBoost regressor
- Least Squares regression



Results

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

The ensemble methods performed better, with both yielding errors smaller than 0.004, XGBoost came out ahead with an error of only around 0.0033 and $R^2 \approx 0.876$.

As a benchmark we used standard Linear Regression which yielded an error of approximately 0.0042 and $R^2 \approx 0.844$.

An interesting finding is that using around 30 parameters for the linear regression yields a smaller error than using all of them and explains almost the same amount of variance. The difference in mean squared error seems to indicate that regressing on all the features causes a strong overfitting.

Conclusion

So, can we predict house prices based on the features a house has?

- Yes! Using the ensemble methods, we can predict the sale price of a house.
- Based on the results of the XGBoost model, the most important features for this prediction includes:
  - GrLivArea, Overall_Rating, LotFrontage, LotArea, GarageYrBlt, etc...
- Logically, all these features makes sense as the bigger the house, the more expensive it is, especially with the Overall_Rating being included as well.