Zillow

Team 45

Data Science Project

Business Understanding

For the use of our dataset, we are assuming the role of Zillow. A market leading housing platform designed to accurately price houses, called 'Zestimates'.

It is our goal to develop a prediction model that will allows us to accurately price houses looking at predictor variables such as:

- Bedroom Count
- Net Square Meters (net sqm)
- Center Distance
- Metro Distance
- Floor
- Age

It is the goal of our company, Zillow, to predict home prices that are reflective of their true market value. For example, if we overestimate / underestimate the price of a house by $10,000, the reputation and credibility of our company could suffer as a result. To mitigate this concern, we would guarantee to the customer with 95% accuracy of our home-price prediction model.

Creating Business Value

The goal of our prediction model would be to increase the reliability of our platform. Increasing the reliability would hopefully migrate customers from other platforms (Apartments.com, PropertyShark, Trulia, etc.) to our platform, thus increasing revenue.

Note on our model: Before implementation of this model in other markets, an understanding of that market's dynamics, trends, and geographical context would be necessary.

- For example: New York City vs. Wichita, Kansas

Data Understanding

Data Overview

- There are 7 columns of data
- 4,308 rows
- Combining for 30,156 observations
- Features / Attributes:
    - Bedroom Count
    - Net Square Meters (Net Sqm.)
    - Center Distance
    - Metro Distance
    - Floor
    - Age

- No missing values in our dataset.

- Lack of units (center_distance, metro_distance, price)

- Assumptions regarding data:
    - Center Distance and Metro Distance will both be measured in meters
    - Price is measured in Euros

Visualizations

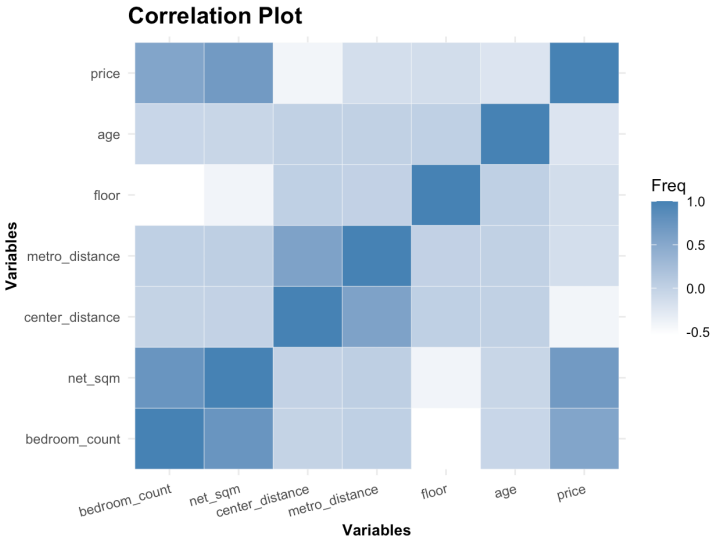


Figure 1. Correlation plot

- Strength of relationships:
  - darker blue closer to 1
  - white closer to –1

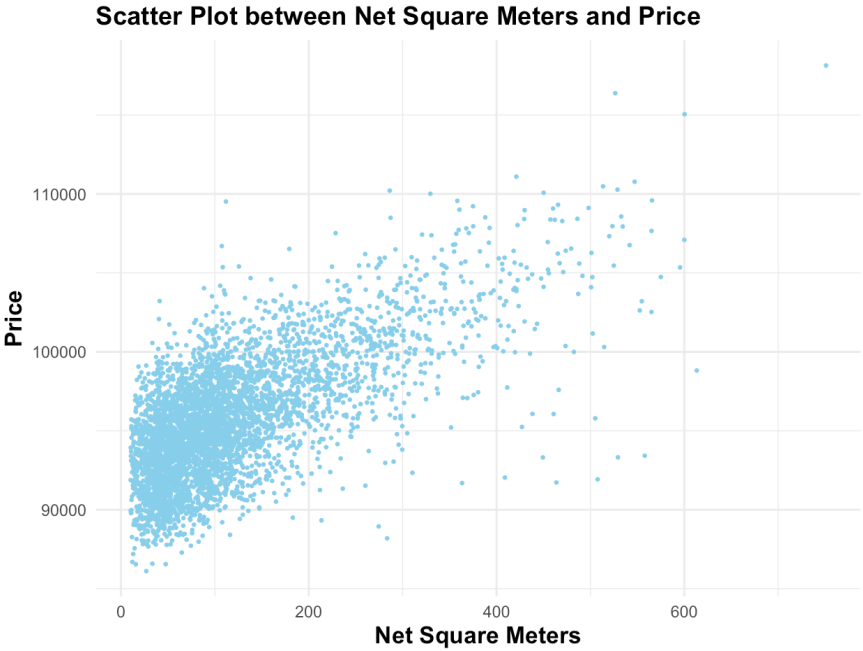
- Two highly correlated variables: counts of bedroom &net square meters

- Multicollinearity: the net_sqm & bedroom_count, metro_distance &center_distance

Visualizations

Figure 2.
Net Square Meters &Price

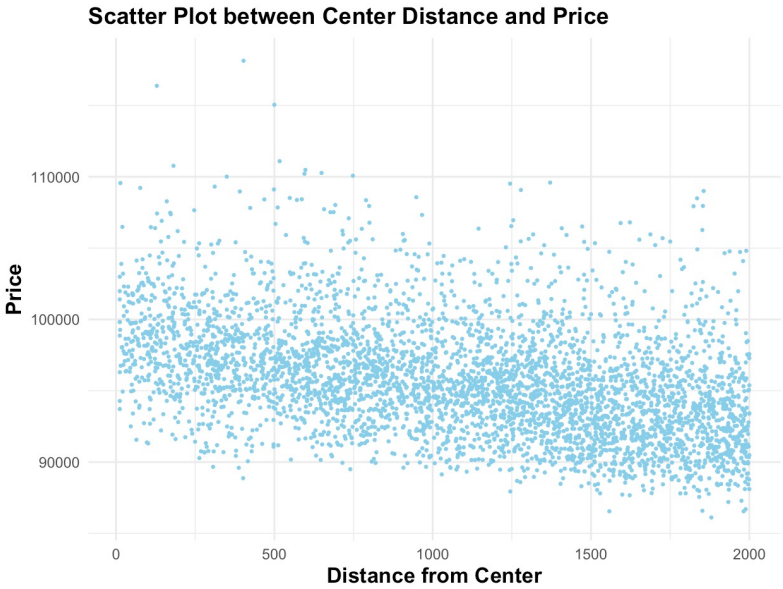
- Positive correlation



Visualizations

Figure 3.
Price V.S. Distance from Center.

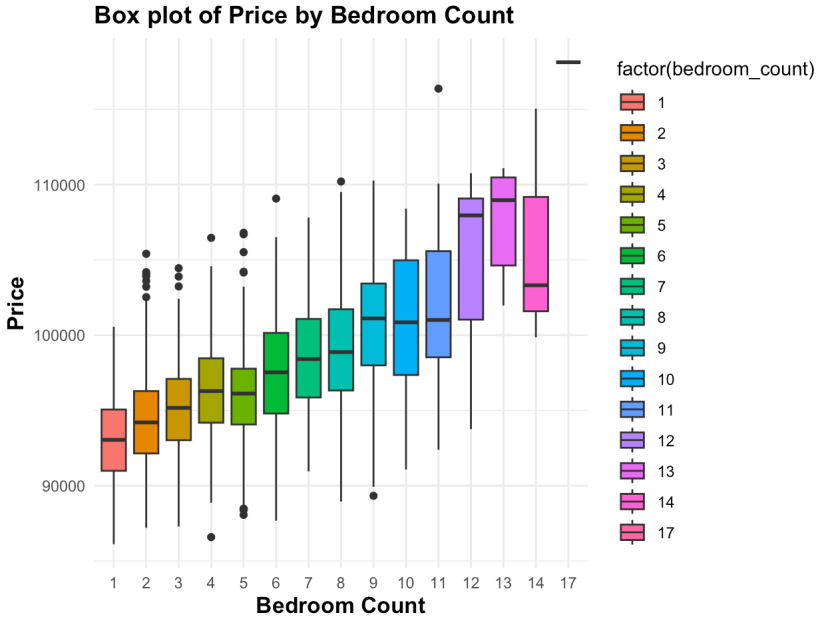
- Negative correlation



Visualizations

Figure 4.
Distribution of house prices based on the number of bedrooms.

General trend of increasing median prices, though there are also some outliers in the data.



Visualizations

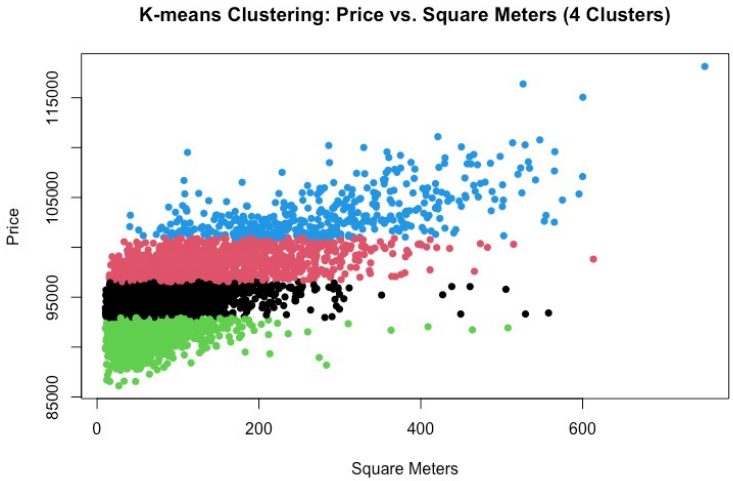


Figure 4. Price against Square Meters using K-Means Clustering

Data Preparation

Data Overview

Our dataset was chosen with consistency and simplicity in mind. The dataset has no missing values and is ready to run from download.

There were, however, adjustments made to our data in regard to training and testing data sets. We broke our main data frame 'housing_data' into two different data frames:

- 'housing_train_data'
- 'housing_test_data'

80% of the data from the original data frame was placed into 'housing_train_data' and the other 20% was placed into 'housing_test_data' .

This was done in order to test the OOS performance of the following models.

Modeling – Linear Model

Model Overview

A linear regression model was created using all predictor variables to quantify price. The coefficients are as follows:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.465e+04  1.455e+02 650.678   <2e-16 ***
bedroom_count    3.177e+02  2.342e+01  13.562   <2e-16 ***
net_sqm          2.482e+01  5.609e-01  44.247   <2e-16 ***
center_distance -3.369e+00  7.829e-02 -43.031   <2e-16 ***
metro_distance   6.789e+00  7.147e-01   9.499   <2e-16 ***
floor            1.208e+02  5.596e+00  21.581   <2e-16 ***
age             -2.661e+01  1.281e+00 -20.772   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

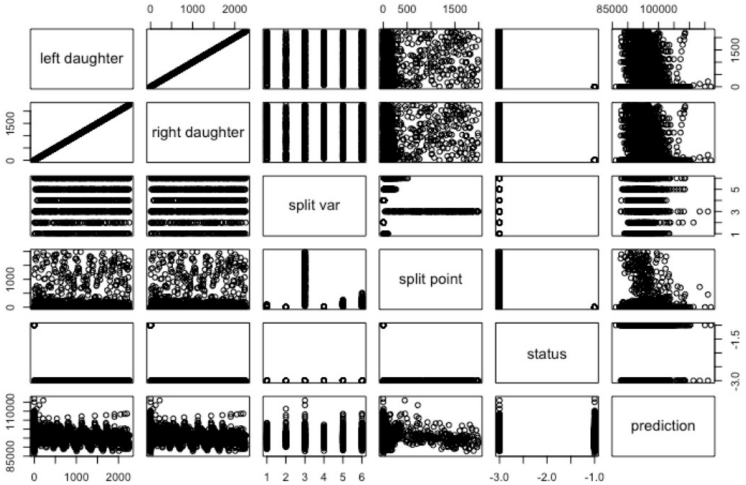
Linear Model with Interaction

We added all the interaction terms into the multiple linear model and the coefficients from the model are as below.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    9.452e+04  3.790e+02 249.433  < 2e-16 ***
bedroom_count                  4.862e+02  7.078e+01   6.870 7.61e-12 ***
net_sqm                        2.585e+01  1.977e+00  13.079  < 2e-16 ***
center_distance               -4.122e+00  2.956e-01 -13.944  < 2e-16 ***
metro_distance                 6.151e+00  2.946e+00   2.088 0.036869 *
floor                          1.589e+02  1.947e+01   8.158 4.74e-16 ***
age                           -2.759e+01  4.729e+00  -5.834 5.90e-09 ***
bedroom_count:net_sqm         -2.124e-01  1.464e-01  -1.451 0.146956
bedroom_count:center_distance -1.150e-01  4.966e-02  -2.316 0.020618 *
bedroom_count:metro_distance   1.119e+00  4.529e-01   2.471 0.013520 *
bedroom_count:floor           -1.563e+01  4.129e+00  -3.787 0.000155 ***
bedroom_count:age             -1.737e+00  8.617e-01  -2.016 0.043910 *
net_sqm:center_distance        3.920e-03  1.157e-03   3.388 0.000713 ***
net_sqm:metro_distance        -6.261e-02  1.029e-02  -6.083 1.31e-09 ***
net_sqm:floor                  1.024e-01  1.047e-01   0.978 0.328381
net_sqm:age                    5.509e-02  2.031e-02   2.713 0.006703 **
center_distance:metro_distance 4.137e-03  1.094e-03   3.781 0.000159 ***
center_distance:floor          8.421e-03  1.237e-02   0.681 0.495908
center_distance:age            4.463e-03  2.849e-03   1.566 0.117362
metro_distance:floor          -7.359e-03  1.129e-01  -0.065 0.948029
metro_distance:age            -1.152e-02  2.593e-02  -0.444 0.656896
floor:age                     -2.446e-01  2.018e-01  -1.212 0.225552
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Random Forest & Lasso

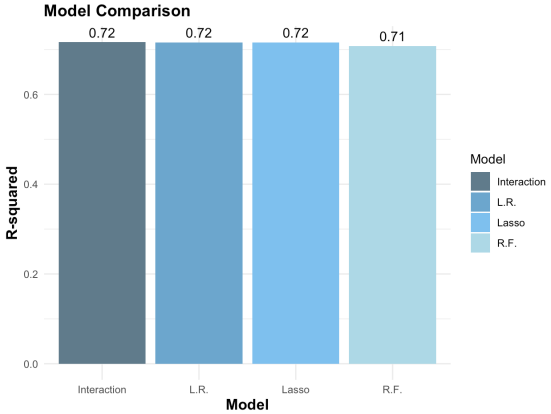
- Random Forest: created using all predictor variables to quantify price
- Lasso: excluded irrelevant features and handled the multicollinearity by finding the optimal lambda



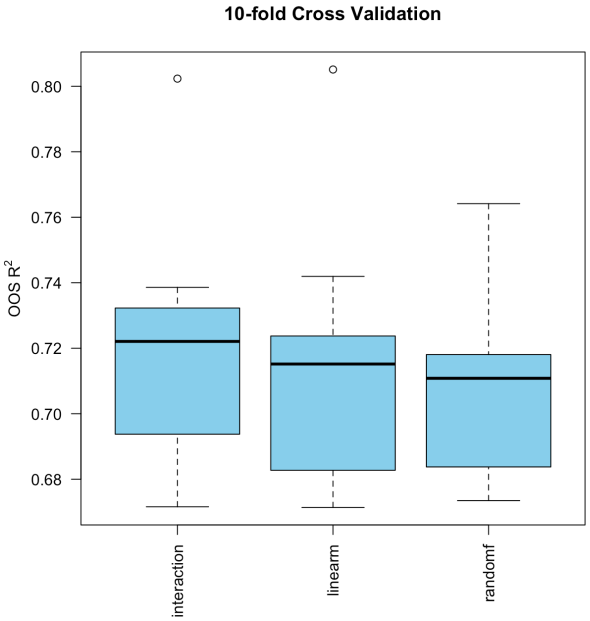
Model Evaluation

	Linear Model	Linear Model with Interactions	Random Forest	Lasso
R^2	0.7161	0.7164	0.7085	0.7160
MSE	4233551	4228730	4346620	4234962

**Table 1 Comparing the Out of Sample R^2 and MSE of each model based on the test data set.*



Model Evaluation




```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    9.452e+04  3.790e+02 249.433  < 2e-16 ***
bedroom_count                  4.862e+02  7.078e+01   6.870 7.61e-12 ***
net_sqm                        2.585e+01  1.977e+00  13.079  < 2e-16 ***
center_distance               -4.122e+00  2.956e-01 -13.944  < 2e-16 ***
metro_distance                 6.151e+00  2.946e+00   2.088 0.036869 *
floor                          1.589e+02  1.947e+01   8.158 4.74e-16 ***
age                           -2.759e+01  4.729e+00  -5.834 5.90e-09 ***
bedroom_count:net_sqm         -2.124e-01  1.464e-01  -1.451 0.146956
bedroom_count:center_distance -1.150e-01  4.966e-02  -2.316 0.020618 *
bedroom_count:metro_distance   1.119e+00  4.529e-01   2.471 0.013520 *
bedroom_count:floor           -1.563e+01  4.129e+00  -3.787 0.000155 ***
bedroom_count:age             -1.737e+00  8.617e-01  -2.016 0.043910 *
net_sqm:center_distance        3.920e-03  1.157e-03   3.388 0.000713 ***
net_sqm:metro_distance        -6.261e-02  1.029e-02  -6.083 1.31e-09 ***
net_sqm:floor                  1.024e-01  1.047e-01   0.978 0.328381
net_sqm:age                    5.509e-02  2.031e-02   2.713 0.006703 **
center_distance:metro_distance 4.137e-03  1.094e-03   3.781 0.000159 ***
center_distance:floor          8.421e-03  1.237e-02   0.681 0.495908
center_distance:age            4.463e-03  2.849e-03   1.566 0.117362
metro_distance:floor          -7.359e-03  1.129e-01  -0.065 0.948029
metro_distance:age            -1.152e-02  2.593e-02  -0.444 0.656896
floor:age                     -2.446e-01  2.018e-01  -1.212 0.225552
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Deployment

Would be used in the background of our user interface.

This model does not take into consideration factors like geography, tax codes, property, etc. that greatly affect home price.

- So, to deploy this model on a wider-scale, more information regarding the housing unit would be necessary.

Issues regarding the model:

- Monitoring and continuous learning: the model will not be static.
- Data storage and handling: ensuring safe and efficient storage becomes crucial as more data is added.

Ethical Considerations:

- Transparency: Users should be made aware that the prices displayed are *estimates* generated by a model and not definitive market values.
- Bias and Fairness: ensuring the model does not favor certain neighborhoods or property types over others.

Deployment continued

Associated Risks:

- Model inaccuracy: “All models are wrong, but some are useful”
  - Mitigation: Regularly update and test the model.

- Market dynamics: Housing market is influenced by hundreds, if not thousands, of factors. Many of which might not be included in future models.
  - Mitigation: periodically review the factors included in the model and adjust as needed.

Q&A

Team 45

Data Science Project