REPORT

Analysis of houses price determinants in Seattle area (USA 2015)

Author: Agata Traczyńska

Table of Contents

Executive Summary

In this report, we have built a pricing model for owners and real estate agents that helps in overcoming the problems of overpricing and underpricing by observing the factors that determine the price of the house. Through this methodology, we have identified key factors related to Metro Seattle house pricing such as: view from the windows, number of bathrooms and bedrooms, attractiveness of area, distance to the city center, living space, building condition and age of the building.

Problem Description

The purpose of this report is to build a recommendation price model for house owners and real estate agencies in Metro Seattle. We chose this topic in particular due to the concern amongst real estate agents that some of their properties get less interests among clients or no interests at all. We wanted to analyse what factors are driving the sold houses price rate. We hypothesized the following determinants: attractive view from the windows, property grade, living space area, distance from city center and age of the building will play a significant role in selling a house for a higher price along with other factors. Through our analysis we found that these factors did impact housing price rate. Keeping that in mind we further analysed and concluded that certain factors play a significant role in determining price of a property.

Data Used for Analysis

dataset	houses	Variable description	Type
Y	Price	Price at which the house was sold	continuous
X1	Bedrooms	Number of bedrooms	continuous
X2	Bathrooms	Number of bathrooms	continuous
X3	Sqft_living	Living spave area in squared feets	continuous
X4	Sqft_lot	Lot area	continuous
X5	Floors	Number of floors	continuous
X6	View	1 if the view from the windows is attractive	binary
X7	Condition	Property conditio (1-5)	ordinal
X8	Grade	Property grade (1-13)	ordinal
X9	Sqft_above	Living area above ground level	continuous
X10	Sqft_basement	area of basement	continuous
X11	Yr_built	Age of the builging	continuous
X12	Distance (km)	Distance from Seattle city center in km calculated from latitude and longitude.	continuous

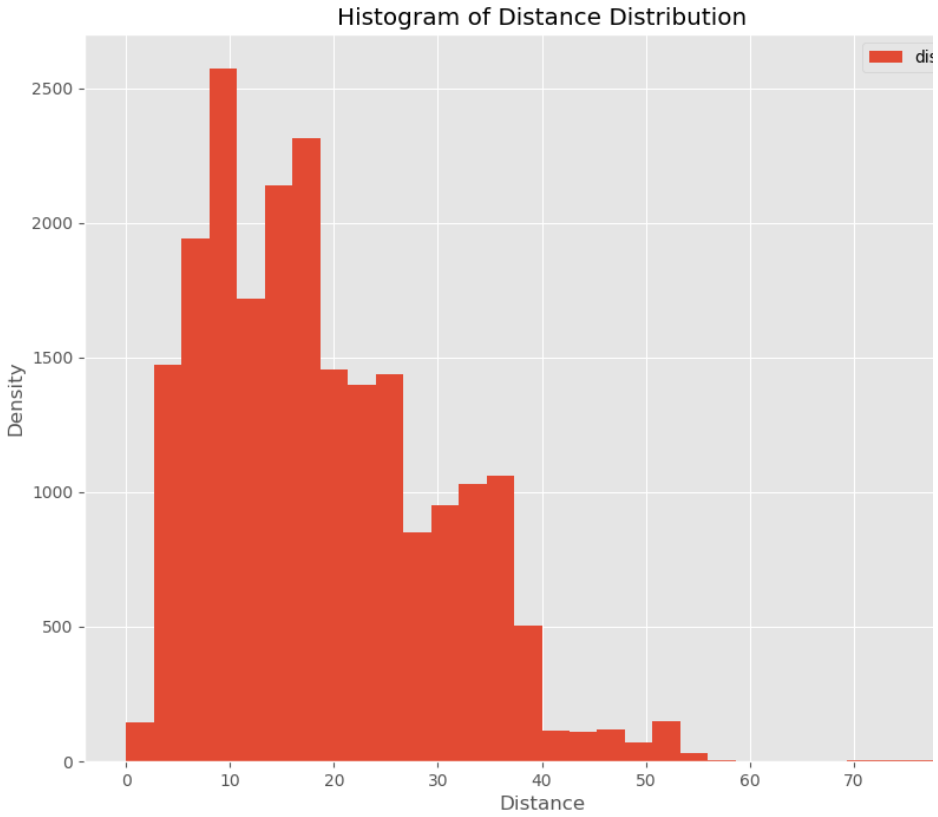
Analysis method in steps

1. Checking on histograms if every variable has a normal distribution. If not, checking if the logarithm of this variable will help.
2. Identifying outliers by plotting scatter plots of price.
3. Creating a heatmap of correlation to analyse the price correlation with its determinants.
4. Performing Lasso Regression to select important features of a dataset.
5. Performing 10-fold cross-validation on the regressor with the different specified alpha in order to select the best value for alpha.
6. Comparing R2 of three different Ridge Regression models to select the best one.

Checking the normal distribution of continuous variables

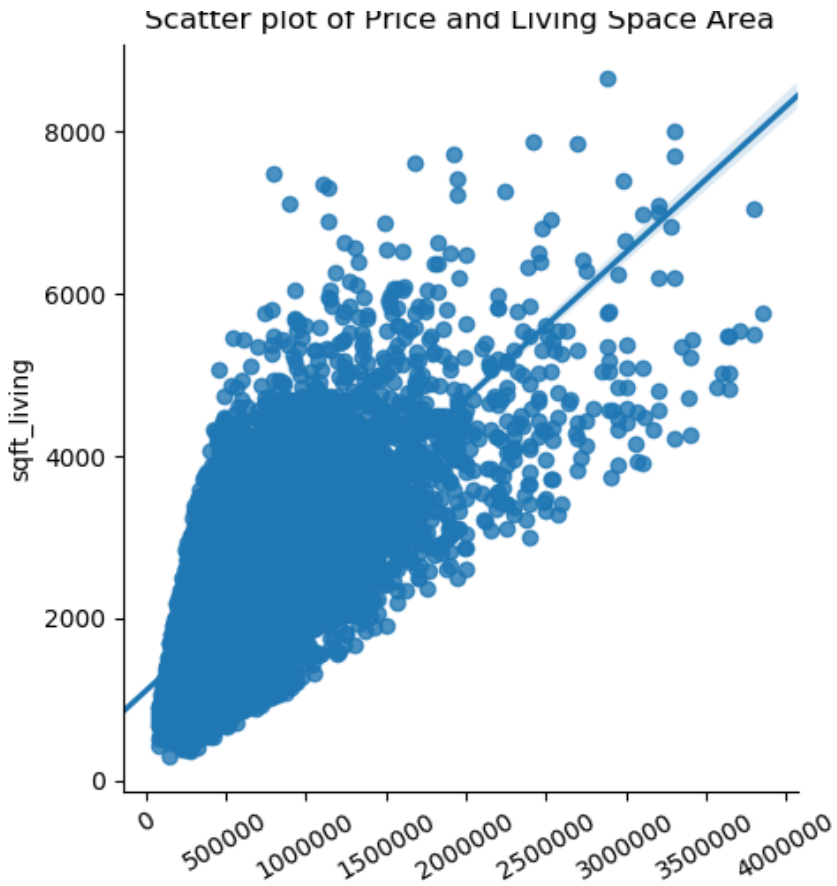
Histograms for all continuous and ordinal variables were created to check whether they have a normal distribution.

Next, if a variable has a positively or negatively skewed distribution, we check whether the logarithm of this variable helps to change the distribution for a normal one. In our case this approach didn't help, so all variables were left with their original distributions.



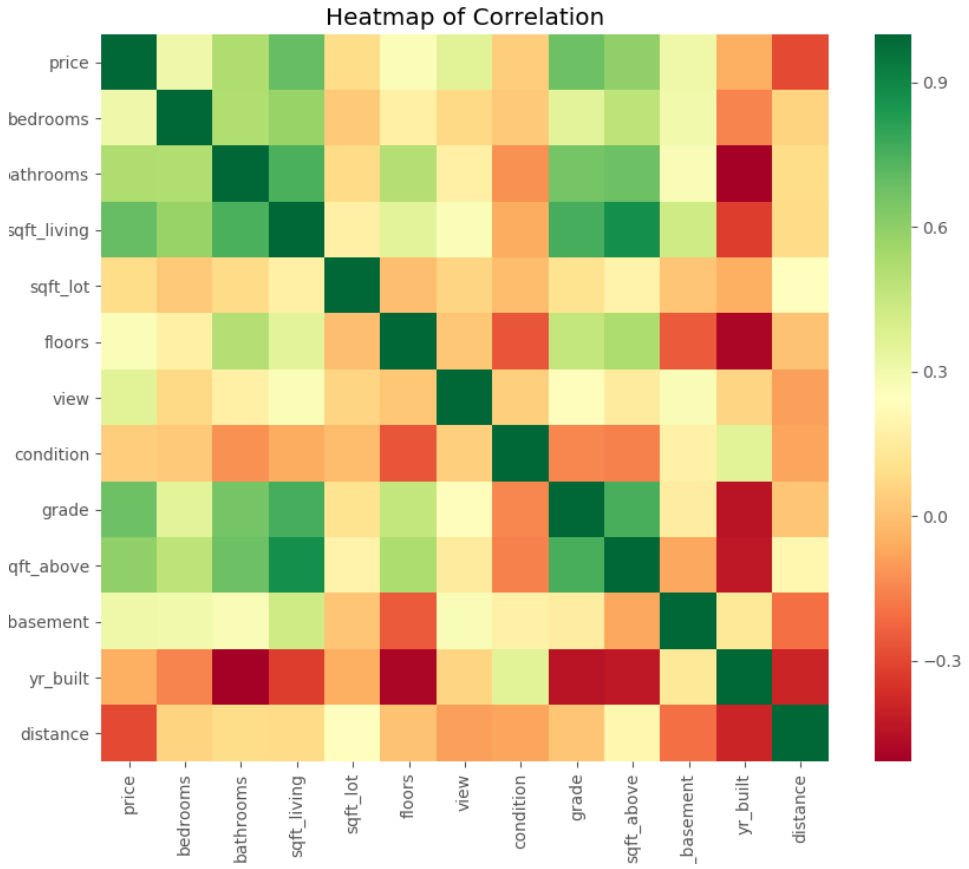
Identifying outliers

We deleted observations with prices higher than 4 million dollars and with living space area larger than 10 000 square feet. Together 13 observations were deleted as they could influence badly the results.



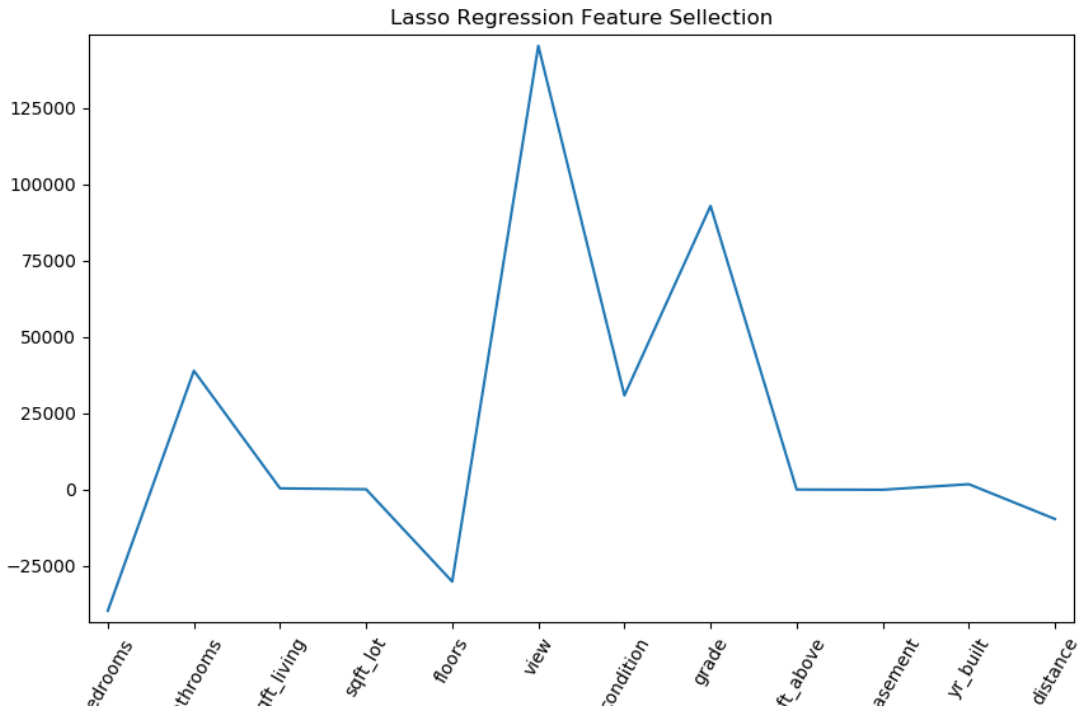
Analysis of Determinants Correlation

- Price is positively correlated with: number of bathrooms, living space, view from the windows, grade of attractiveness, area above the ground level.
- Price is negatively correlated with: condition, age of the building and distance from city center.
- Variables not correlated: sft_lot, floors, sqft_basement.



Lasso Regression for Feature Selection

- Lasso Regression can be used to select important features of a dataset. It shrinks the coefficients of less important features to exactly zero.
- Loss function = Ordinary loss function + $\alpha \sum_{i=1}^{n} |a_i|$
- From the graph we see that three variables have coefficients close to zero: sqft_lot, sqft_above, sqft_basement. Therefore, these three variables were removed for further modelling.

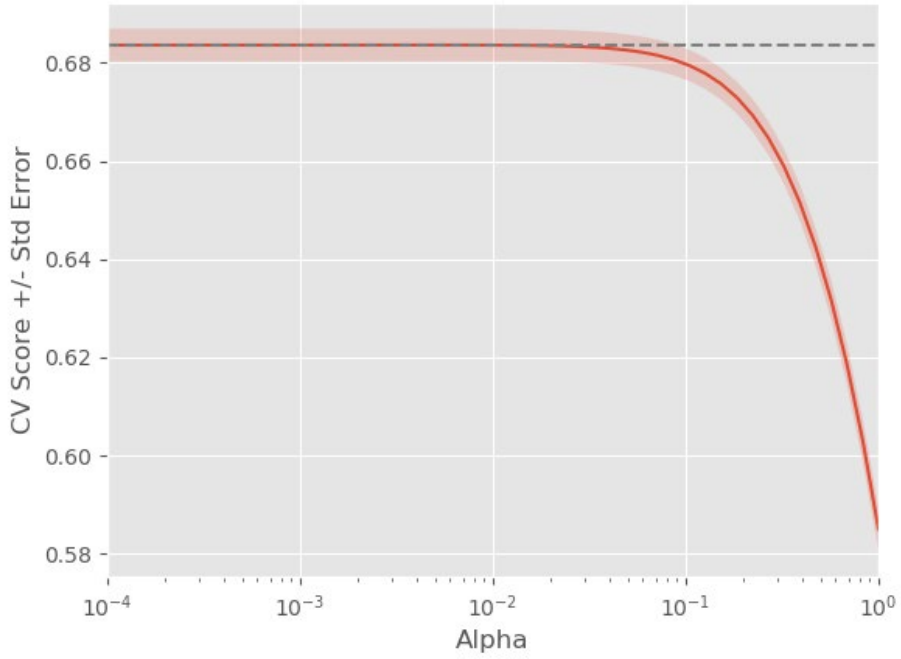


Regularization – choosing best alpha

Large coefficients can lead to overfitting. In order to penalize large coefficients we use regularization.

Here, we perform 10-fold cross-validation on the regressor with the specified alpha.

The graph shows how the R2 score varies with different alphas. It helps to select the right value for alpha. We have chosen alpha= 0.01 as it ensures high R2 score and penalizes enough the large coefficients.



Cross validation to compare the average R2 of different models

Ridge Regression -> Loss function = Ordinary loss function + $\alpha \sum_{i=1}^{n} a_i^2$

Normalization – all arguments on the same scale

We performed 10-fold cross validation to compare the average R2 of three different models. Model II with 8 variables was selected as this model is characterized by high average R2 and has all statistically significant variables.

Model	Model I. All variables	Model II. Without floors	Model III. Without floors, sqft_living
Average R2	0.69	0.67	0.63
Variables statistically not significant	Floors, sqft_living	-	-

Final Model Ridge Regression

Bedrooms
Bathrooms
Space living area in square feet
Attractiveness of view from the windows
Condition
Grade of the attractiveness
Age of the building
Distance from the city center

	Coefficients	Standard Errors	t values	Probabilites
0	-589580.1484	30428.452	-19.376	0.0
1	-20465.5575	3461.480	-5.912	0.0
2	41192.9813	5492.048	7.500	0.0
3	150.7025	5.417	27.821	0.0
4	141923.8182	8439.646	16.816	0.0
5	25796.8900	4009.051	6.435	0.0
6	101226.5658	3575.355	28.312	0.0
7	1511.2005	120.237	12.568	0.0
8	-7836.4295	248.229	-31.569	0.0

Results

- The largest influence on the price has a binary variable VIEW. A good view from the windows increases the price of a house on average of 141 000$ ceteris paribus.
- The second most important variable is GRADE of the attractiveness (1-13). A house with a grade one level higher costs on average by 100 000$ more ceteris paribus.
- One more bathroom increases the price of a house by 41 000$ ceteris paribus.
- The better the condition of the building, the higher (25 000$) the price is.
- The 100 square feet larger house costs on average 15000$ more ceteris paribus.
- Surprisingly, one more bedroom lowers the house price by 20 000$ ceteris paribus.
- Each kilometre further from the city center decreases the price on average of 8 000$.
- Unexpectedly, the older the house, the higher is the price (1500$) ceteris paribus.

Interpretation

Our findings shows that people are willing to pay more for houses:

- located in a green area with nice view from the windows,
- in attractive, clean and safe districts (grade of attractiveness),
- Equipped with more bathrooms,
- Located closer to the city center,
- In better condition,
- With larger living space,
- And the age of the building is not a disadvantage.

Recommendations

- Our findings proves the hypothesis that factors like an attractive view from the windows, property grade, living space area, distance from city center are driving the sold houses price rate. However, the age of the house does not influence the price negatively as we expected.
- Our findings can be used by owners and real estate agencies in Metro Seattle while choosing the price for the house to be sold on the market. Moreover, future investors can benefit from our model while deciding which property to invest in.
- The next approach that could be taken to evolve the model in the future is to build a quantile regression. Quantile regression will help to check whether factors influence the price differently depending on the level of the price.

Bibliography

- Datacamp courses: https://campus.datacamp.com/courses/supervised-learning-with-scikit-learn
- Sebastian Raschka: Python Machine Learning, Second Edition, Packt 2017