Machine Learning Engineer Nanodegree

# Capstone Project

Niclas Geiger

June 26st, 2017

# I. Definition

## Project Overview

Affordable housing becomes an increasing issue. The information of areas with cheaper prices and more expensive areas is often changing and hard to get. Therefore, machine learning can help improve this issue with predicting the prices for certain areas and houses. This will not only be beneficial for willing buyers but also for people looking to sell their house as they can see the best prize for this area and commodity in seconds.

## Problem Statement

To make a good prediction for housing prizes you need to look at several variables. The housing prize depends on the area you live in, the age of the house, the size and much more. All these attributes have to be looked at to make a precise prediction. Predicting the prize will most certainly look for similarities in other houses in the area to make an estimation. This is a regression problem using clustering to predict prizes for new entries.

## Dataset and Inputs

I will use the House Sales in Kings Country Dataset from kaggle. This dataset includes 21613 home sales in between May 2014 and May 2015 in Kings Country which also includes Seattle. Every entry has 19 features which include housing area, latitude & longitude and other important data.

The output data will be the predicted prize for the entered data. Data will be used for cross validating the algorithm. This means the data will be split into different training, validation and testing data for several optimization-runs.

## Metrics

For evaluating the result, I will use the r2-score which …

# II. Analysis

\_(approx. 2-4 pages)\_

## Data Exploration

|  |  |  |
| --- | --- | --- |
|  | **price** | **Sqft Living** |
| **Count** | 21,613 | 21,613 |
| **Mean** | 540,088.1 | 2079.9 |
| **Std.** | 367,127.2 | 918.44 |
| **Min** | 75000 | 290 |
| **25%** | 321,950 | 1427 |
| **50%** | 450,000 | 1910 |
| **75%** | 645,000 | 25550 |
| **Max** | 7,700,000 | 13540 |

### Price

The first values I will have a closer look at will be the price.

We have prices ranging from roughly $75,000 to $7,700,000. 50% of the prices is around $450,000.

The standard deviation is $360,000. The mean of the prices is $540,000.

If we look at the quarters we have:

- first 25% span over $235,000

- second 25% span over $130,000

- third 25% span over $195,000

- top 25% span $7,125,000

This clearly shows that the first 3/4 of all houses lie in a pretty similar range and are at least somewhat evenly distributed in their quarters. The top 25% are really far out regarding their prices which might be caused by a lot of expensive houses mixed with some extremely expensive houses.

### Living Sqft.

The second value I will examine is the living space (in square feet). This value spans in between 290 to 13,540 sqft.

The standard deviation is 918 and the mean around 2080 sqft. This points that most of the houses should be in between 1000 to 3000 sqft. When looking at the quarter distribution you get the following picture:

- first 25% is 1137 sqft. in distance

- second 25% spans 483 sqft.

- third 25% is 640 sqft.

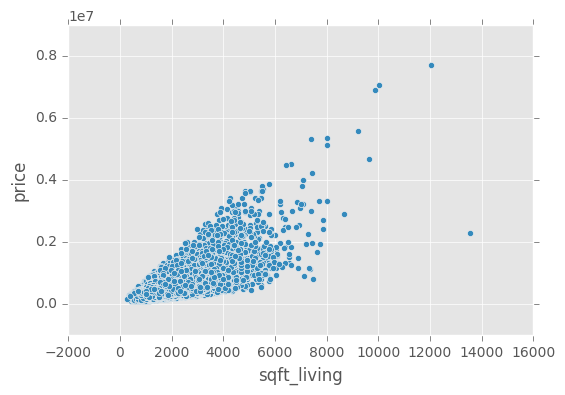
- the last 25% is 9990 sqft.

This huge difference proves the previous estimation as at least 75% of the houses are in between 290 to 2550 sqft. in living space. This also kind of resembles the discoveries of looking at the house prices.

## Exploratory Visualization

### Relationship between price and living space

I will print a plot to visualize the relationship between prices and living space. Usually the available space of a house is one of the biggest factors for the selling price.



As one can see here, this kind of hints to a rather linear relationship between both values.

### Relationship between prices and location

Another interesting factor for the prices I found was the location of the house. This knowledge is derived from my domain knowledge but can also be visualized by using a heat map over the prices on the geolocation of homes.

>TODO HEAT MAP OF HOUSE PRICE

## Algorithms and Techniques

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

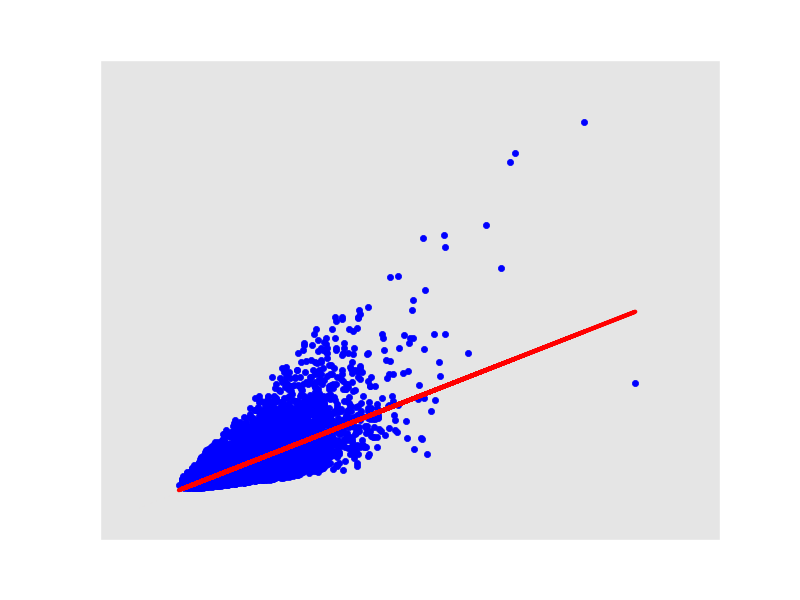
- \_Are the algorithms you will use, including any default variables/parameters in the project clearly defined?\_

- \_Are the techniques to be used thoroughly discussed and justified?\_

- \_Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?\_

## Benchmark

My Benchmark model will be a trivial linear regression over the living space of all houses and their corresponding prices. First, I will split the data into training and test sets using cross validation. Then I will train the linear regression on the training set. To get a comparable number I will then take the r2-score of the prediction the linear regression does on the test data compared to the actual data.



For the Simple Regression the R2 Score is: 0.4929.

# III. Methodology

## Data Preprocessing

### Feature Relevance

To try to find out which features might be redundant, we will compare the regression score for each feature. This will give the score we get for predicting a feature just by all the other features. If a feature has a regressor score > 0.95 then this is highly likely to be redundant. If a feature has a regressor score of < 0 then it is not really well described by the rest of the data and might therefore be an important one.

|  |  |
| --- | --- |
| **Feature** | **Regressor score** |
| Id | -0.13338979727 |
| Bedrooms | -0.160352462696 |
| Bathrooms | 0.50795150982 |
| Sqft living | 0.996110554063 |
| Sqft Lot | 0.269109294583 |
| Floors | 0.592701773203 |
| Waterfront | 0.293462831528 |
| View | -0.0166225785212 |
| Condition | -0.333728697699 |
| Grade | 0.588520208236 |
| Sqft above | 0.98419834569 |
| Sqft basement | 0.973475041586 |
| Year built | 0.639089125308 |
| Year renovated | -0.417516672101 |
| Latitude | 0.990196948347 |
| Longitude | 0.95619613868 |
| Sqft Living 15 | 0.568577429132 |
| Sqft Lot 15 | 0.182621375198 |

The result shows that sqft\_living, sqft\_above, sqft\_basement, zipcode, latitude and longitude can be very good described by the rest of the data when you leave them out. I think this might be explained, that all the square foot parameters describe the house and can therefore be explained by the number of rooms and other square foot values. zipcode can explain the latitude and longitude and vice versa. Bedrooms, view, condition and year renovated are all features which are hard to be described but just looking at all other features.

sqft above and basement are both already included in sqft\_living which is just the sum of it. To remove the above and basement feature I will check if the ratio of above to basement really has an impact on the price.

### Detecting Outliers

The second step to improve our learning behaviour is to find outliers and then remove them from the data set if needed. To detect outliers, I will compare us Tukey’s Method to compare the values of each feature for an entry to the difference between the 25th and 75th percentile for that feature. If the difference is more than 3 times higher, we consider an entry to be an extreme outlier. These extreme outliers will then be removed from the dataset.

The outlier detection resulted in 1613 data points removed for being extreme outliers.

### Calculating Feature Relevance

The next step involved finding feature relevance by calculating how much a feature could be described by using all other features. This was done by using a Decisiontree-Regressor and dropping the feature from the set. Then we split the set into 75% train and 25% test set using train\_test\_split. The score was the result of predicting the tested feature just using the Regressor. The results were as followed:

*Redundant features(>0) : ['sqft\_living', 'sqft\_above', 'sqft\_basement', 'zipcode', 'lat', 'long']*

*Significant features(<0) : ['bedrooms', 'sqft\_lot', 'waterfront', 'view', 'condition', 'yr\_renovated']*

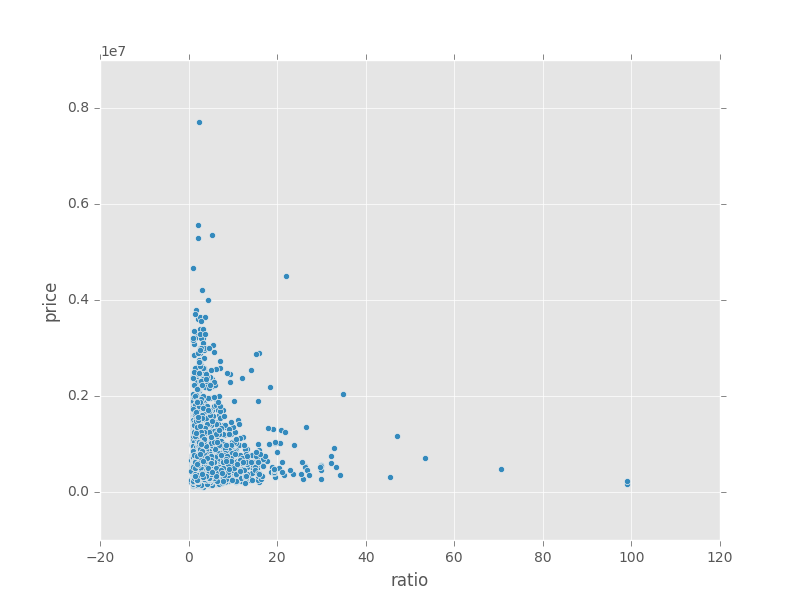


Figure 1 impact of above/basement-ratio to price

The result shows that sqft\_living, sqft\_above, sqft\_basement, zipcode, latitude and longitude can be very good described by the the rest of the data when you leave them out. I think this might be explained, that all the square foot params describe the house and can therefore be explained by the number of rooms and other square foot values. Zipcode can explan the latitude and logitude and vice versa. Bedrooms, view, condition and year renovated are all features which are hard to be described but just looking at all other features.

This result hinted that “sqft\_above” (score:0.98) and sqft\_basement” (score:0.99) could be really well described by using other features. Sqft above and basement are both already included in sqft\_living which is just the sum of it. To remove the above and basement feature I will check if the ratio of above to basement really has an impact on the price.

Therefore, I examined the impact of the ratio between those 2 to the price. No real correlation but some outliers could be found.

Because of that, both features were dropped from the data.

### Feature Performance

To further investigate on feature relevance I will calculate the entropy (information gain) which each future holds for the data set. This Code was inspired by one of the sklearn examples(http://scikit-learn.org/stable/auto\_examples/ensemble/plot\_forest\_importances.html#sphx-glr-auto-examples-ensemble-plot-forest-importances-py) and I use the ExtraTreesClassifier to describe the importance of features.

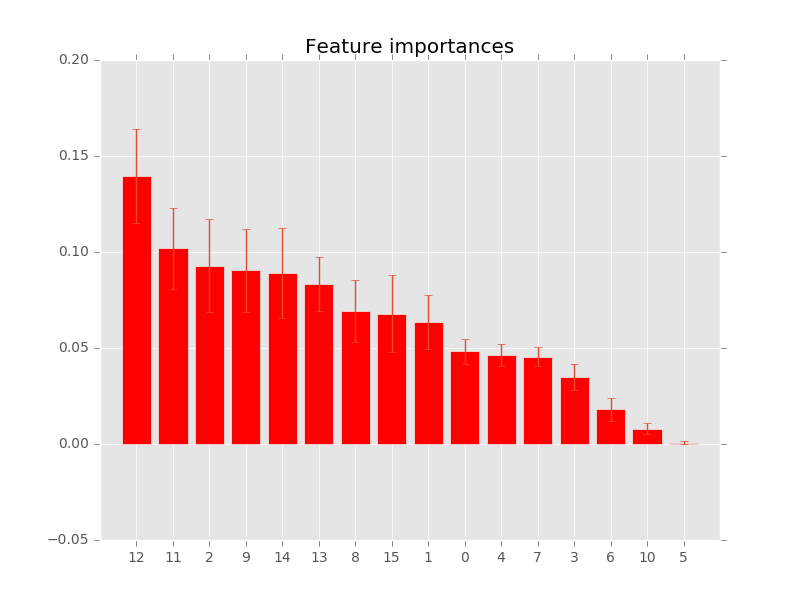


Figure Feature Performance

The Entropy chart is led by bedrooms with a bigger margin. After that bathrooms and square foot living comes in. With a medium value, we have square foot lot, floor, waterfront and view. These are followed by condition, grade, square foot above and square foot basement. Year built, zipcode and latitude have the same low entropy. Longitude and Square Foot Living 15 have even less. The least entropy with nearly zero has square foot lot 15.

#### what does that mean?

We can definitely see that even though the observation via heatmap of location dependant prices was interesting but longitude, latitude and zipcode all have really bad entropy. Therefore, they will be ignored in the following. The same goes for Square Foot Living 15 and Lot 15.

## Implementation

My Algorithm involved 2 Steps for training and for testing vice versa. The first step involved clustering the data and the second step was training several regressors, one for each cluster combination.

### Clustering

The idea was to take correlated features into a cluster thus decreasing the feature space.

Two approaches were chosen:

* Space Cluster: taking bedrooms, bathrooms and floors into one cluster
* Quality Cluster: taking waterfront, view, condition, grade, yr\_built, yr\_renovated

Space Cluster should sum up things regarding the space. Quality Cluster is about the quality of the house.

We find 5 different Clusters for the Space Cluster which are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster \ mean | condition | grade | yr\_built | yr\_renovated |
| Quality Cluster 0 | 3.5 | 7.7 | 1978 | 0 |
| Quality Cluster 1 | 3.2 | 7.7 | 1938 | 1996 |
| Quality Cluster 2 | 3.6 | 7.0 | 1952 | 0 |
| Quality Cluster 3 | 3.0 | 8.2 | 2004 | 0 |
| Quality Cluster 4 | 3.7 | 7.0 | 1918 | 0 |

When you look at the data, you see that the different 5 quality clusters mean:

* 0: An old aged house which is still in good condition and has average grades
* 1: A medium aged house, which is in good condition with better grades
* 2: An old house which was renovated recently and is thus in good condition with better grades
* 3: An medium ages house with average grades and rather good condition
* 4: A new house with good grades in a rather good condition

|  |  |  |  |
| --- | --- | --- | --- |
|  | bedrooms | bathrooms | floors |
| Space Cluster 0 | 1.9 | 1.22 | 1.14 |
| Space Cluster 1 | 2.9 | 2.5 | 2.1 |
| Space Cluster 2 | 3.2 | 1.6 | 1.1 |
| Space Cluster 3 | 4.3 | 2.7 | 1.7 |

When you look at the data, you see that the different 4 space clusters mean:

* 0: With an average of about 2 Bedrooms and 1 Bathroom spanning 1 floor, this is the typical small house
* 1: Averaging nearly 3 Bedrooms, 2 Bathrooms and spanning 2 Floors this is the normal "family house"
* 2: Having more than 3 Bedrooms, but only 1.5 Bathrooms and 1 Floor this is more of a bungalow
* 3: With on average over 4 Bedrooms and nearly 3 Bathrooms spanning nearly 2 floors on average these are the most spacious houses.

Clustering was done using KMeans.

### Regression

For every pair of quality and space clusters it’s own regressor was trained with the Linear Regression Algorithm. This was intended to have better linear Regression for better paired house. The Feature used for training was still sqft\_living as X and the price as y.

### Prediction

When predicting the prices for each entry the same approach was done. First clustering each value in the test set by their respective quality and space cluster. Then using the correspondent linear regressor on that value to predict it’s price.

## Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

- \_Has an initial solution been found and clearly reported?\_

- \_Is the process of improvement clearly documented, such as what techniques were used?\_

- \_Are intermediate and final solutions clearly reported as the process is improved?\_

# IV. Results

\_(approx. 2-3 pages)\_

## Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

- \_Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?\_

- \_Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?\_

- \_Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?\_

- \_Can results found from the model be trusted?\_

## Justification

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- \_Are the final results found stronger than the benchmark result reported earlier?\_

- \_Have you thoroughly analyzed and discussed the final solution?\_

- \_Is the final solution significant enough to have solved the problem?\_

# V. Conclusion

\_(approx. 1-2 pages)\_

## Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- \_Have you visualized a relevant or important quality about the problem, dataset, input data, or results?\_

- \_Is the visualization thoroughly analyzed and discussed?\_

- \_If a plot is provided, are the axes, title, and datum clearly defined?\_

## Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- \_Have you thoroughly summarized the entire process you used for this project?\_

- \_Were there any interesting aspects of the project?\_

- \_Were there any difficult aspects of the project?\_

- \_Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?\_

## Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- \_Are there further improvements that could be made on the algorithms or techniques you used in this project?\_

- \_Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?\_

- \_If you used your final solution as the new benchmark, do you think an even better solution exists?\_

-----------

\*\*Before submitting, ask yourself. . .\*\*

- Does the project report you’ve written follow a well-organized structure similar to that of the project template?

- Is each section (particularly \*\*Analysis\*\* and \*\*Methodology\*\*) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?

- Would the intended audience of your project be able to understand your analysis, methods, and results?

- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?

- Are all the resources used for this project correctly cited and referenced?

- Is the code that implements your solution easily readable and properly commented?

- Does the code execute without error and produce results similar to those reported?