**Machine Learning Engineer Nanodegree**

**Capstone Project**

Jiri Manak  
August 8th, 2017

**I. Definition**

**Project Overview**

A purchase of a house is usually one of the largest investments in life of most people. For a proper personal finance planning a knowledge of property price is a first step towards this goal.

The price of a house is determined by number of features. The number of square meters, number of rooms and house location are the common attributes. Anyhow, price of the house can be influenced by some specific features. For a buyer, it would be interesting to know not only an average price of the property with required attributes but also which features influence the price most. Comparing features, which have the strongest impact on the price with the personal preferences, allows buyer to make a tradeoff between the dream house and budget.

**Problem Statement**

We have a collection of data about houses which were already sold. Base on this data, can we predict how much will cost my dream house? Or a property owner may ask: „How much is my house worth? “

This problem can be solved by creating of a regression predictive model and the solution for it is to create such a model.

Second goal is to create list of the features sorted by its importance to influence the price of the house.

Third goal is to compare results from different regression models.

**Datasets and Inputs**

Collect data from usual sources as web portals, advertisements, real estate agencies offerings will be an enormous task. Therefore, for purpose of this project I used data publicly available on Kaggle: House Prices: Advanced Regression Techniques.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

This collection consists of 1462 data points for training and 1461 data points for testing. There are almost 80 features for each record which describe the house from more aspects.

**Metrics**

To stay compatible with Kaggle metrics I will evaluate the model on RMSE.

* **Root-Mean-Squared-Error (RMSE)** - the logarithm of the predicted value and the logarithm of the observed sales price.

where is predicted value of observation *i* of regression dependable variable computed for different predictions

For evaluation and comparison of different models I have also used

* **R^2 score – coefficient of determination –** which isa proportion of the variance in the dependent variable that is predictable from the independent variable. An R2 of 0 means that the dependent variable cannot be predicted from the independent variable. An R2 of 1 means the dependent variable can be predicted without error from the independent variable.

where is predicted value of observation *i* of regression dependable variable computed for different predictions

<https://en.wikipedia.org/wiki/Coefficient_of_determination>

http://stattrek.com/statistics/dictionary.aspx?definition=coefficient\_of\_determination

**II. Analysis**

**Data Exploration**

#### Features

Each house in the dataset is described by 80 features. These features characterize almost every aspect of the property which can differentiate particular house from the others.

#### SalePrice

Sale price of the house is the target variable. Base on these values I will try to find model which will predict these prices with highest possible accuracy.

The basic statistic the sale price shows following:

*Maximum price: $755,000.00*

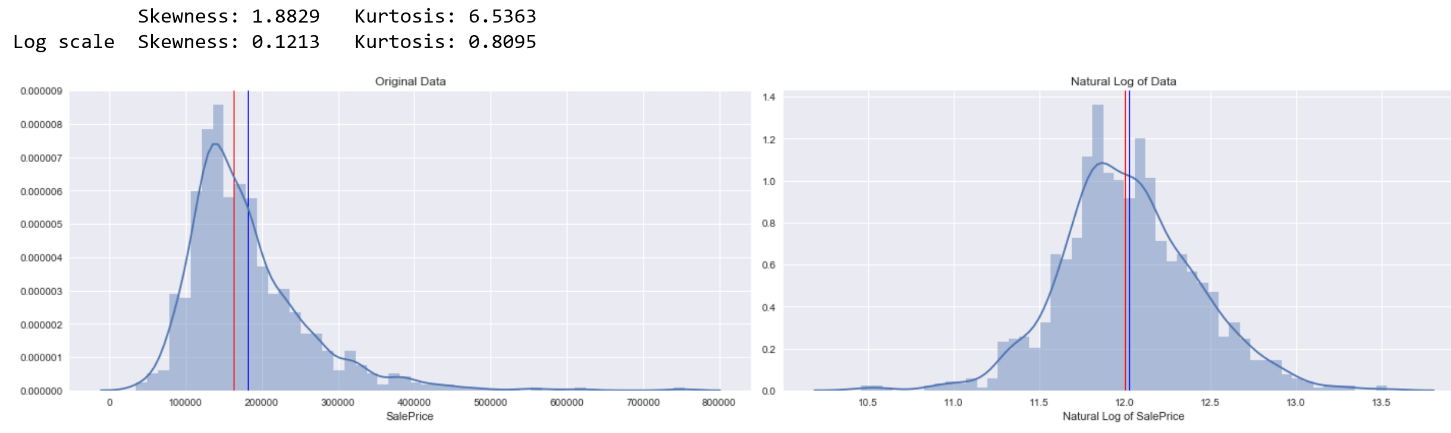
*Minimum price: $34,900.00*

*Mean price: $180,921.20*

*Median price $163,000.00*

*Standard deviation of prices: $79,415.29*

The distribution of SalePrice deviates from normal distribution, shows skewness to the left (towards lower prices) which is not a surprise, because maximal price of a house has no limits.



#### Categorical and Continuous values

Simple request on features data type stated, that 27 of features are numerical and 43 are categorical.

The features describing measurable attributes as for example square footage are described by ontinues values, or it can hold number specifying of an amount of objects.

Typical continues variable features are for example:

* GrLivArea – Above grade (ground) living area square feet
* LotArea: Lot size in square feet
* BsmtFinSF1: Type 1 finished square feet
* TotalBsmtSF: Total square feet of basement area
* PoolArea: Pool area in square feet
* Fireplaces: Number of fireplaces

There is also a number of categorical values which can hold any value from specified set of values.

Typical categorical varibales are Overal Quality

* KitchenQual: Kitchen quality
* FireplaceQu: Fireplace quality
* Functional: Home functionality rating

#### Missing Data

Exploring the data by looking for missing values shows, that there is a quite a large number of missing data. Anyhow, in feature description it is stated, that values are usually missing if the object, which feature describes doesn’t exist. For example, a pool. If property has a pool, then feature ‘PoolSF’(pool square footage) is filled out with a number of pools area in square feet. If property hasn’t any pool, then the feature value is missing. In this case we can simply replace missing value by zero. For categorical features we can use value *‘None’*.

There are really only few features in dataset with missing values. These values were replaced by the most frequent value for particular feature in the dataset.

#### Outliers

There are more possible ways how to identify outliers. Removing them might by risky because it can influence the model in negative way. Therefore conservative approach has been taken. Correlation graph of ‘GrLiveArea’ with respect to ‘SalePrice’ shows two data points in upper left corner indicating large living area with unusual low price. These data points are identified as outliers.

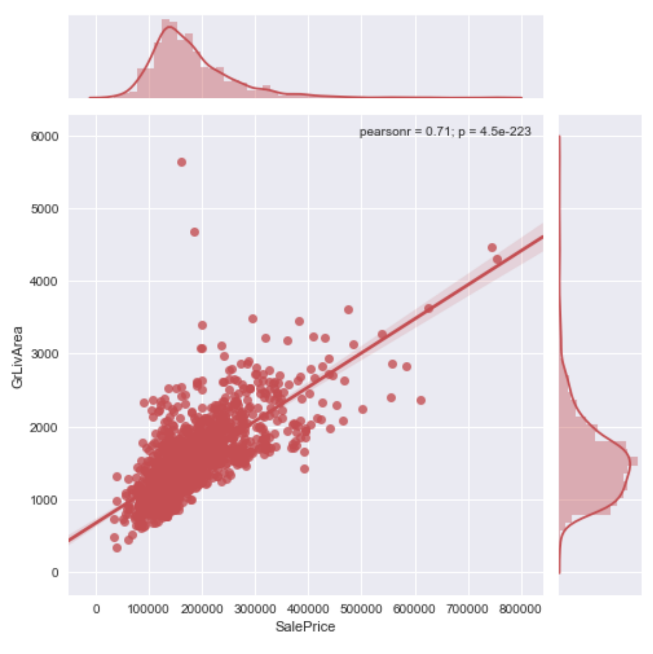
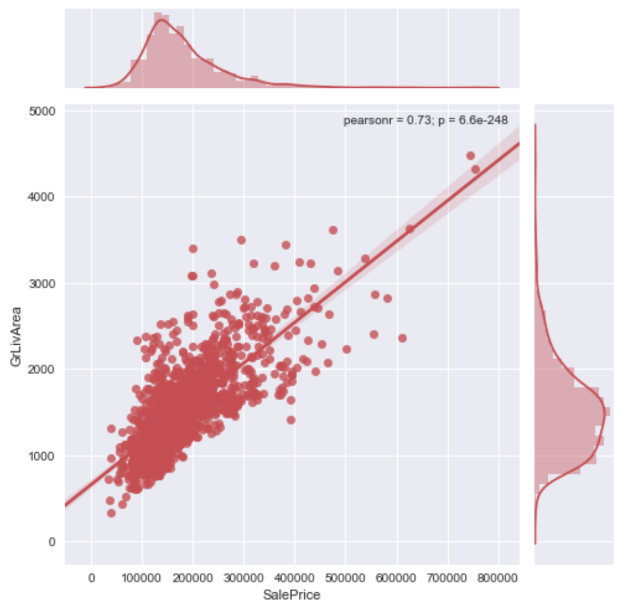
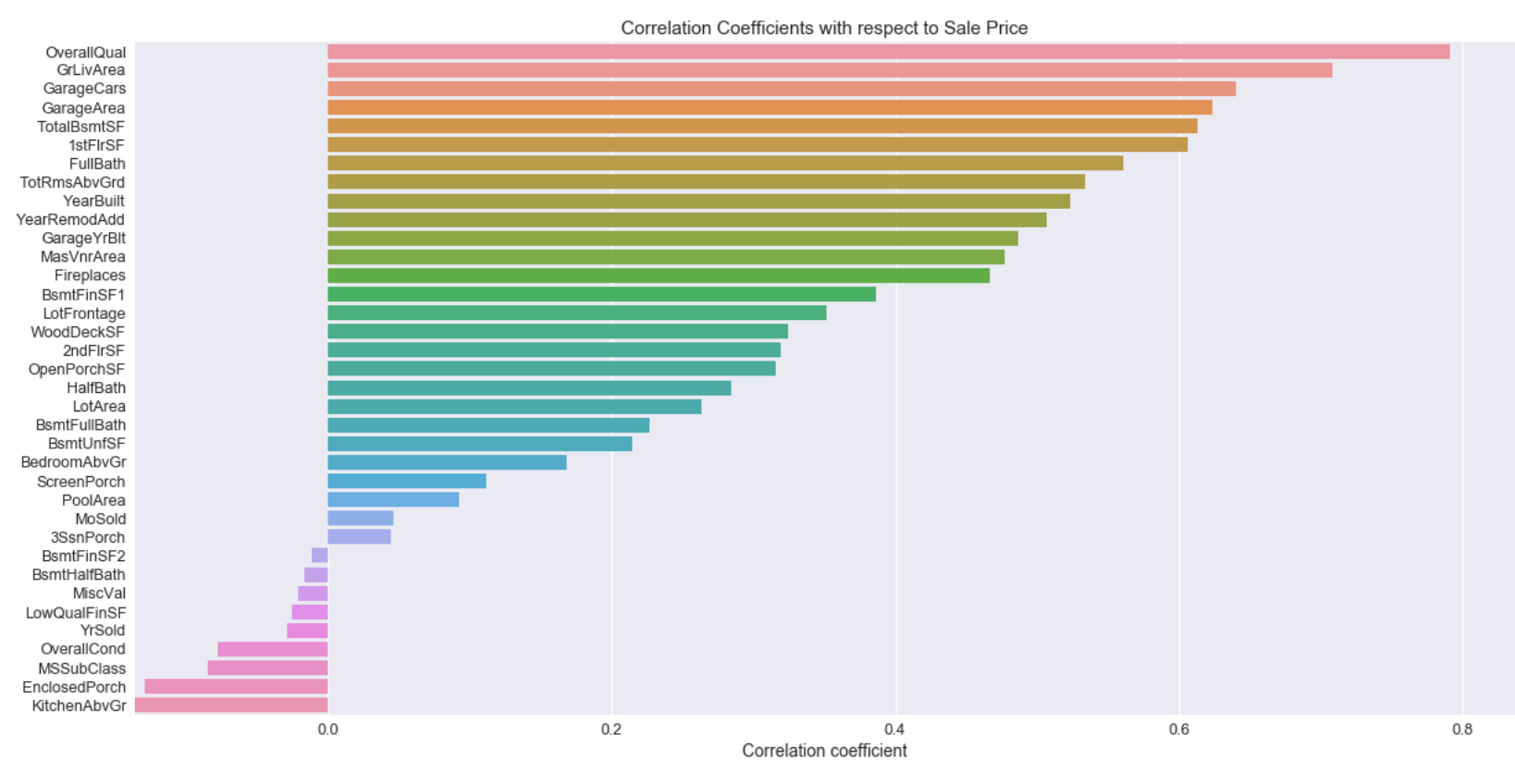


Figure . Without outliers

Figure . With outliers

**Exploratory Visualization**

Basic data investigation shows features correlation with respect to SalePrice. Graph shows that there is a number of features which correlates to SalePrice.



Top 5 features

**OverallQual** Overall material and finish quality

GrLivArea: Above grade (ground) living area square feet

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

TotalBsmtSF: Total square feet of basement area

Overal quality is categorical value. Next four features describing a size of a property part. This is only evidence that size of house have the highest influence on the 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**

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* *Has some result or value been provided that acts as a benchmark for measuring performance?*
* *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

**III. Methodology**

*(approx. 3-5 pages)*

**Data Preprocessing**

**remove outliers**

**separate target variable**

**Join train and test data**

**handle missing data**

**adding total SF**

Missing data

#### Removing Outliers

#### Missing Data

Exploring the data by looking for missing values shows, that there is a quite a large number of missing data. Anyhow, in feature description it is stated, that values are usually missing if the object, which feature describes doesn’t exist. For example, a pool. If property has a pool, then feature ‘PoolSF’(pool square footage) is filled out with a number of pools area in square feet. If property hasn’t any pool, then the feature value is missing. In this case we can simply replace missing value by zero. For categorical features we can use value *‘None’*.

There are really only few features in dataset with missing values. These values were replaced by the most frequent value for particular feature in the dataset.

**Implementation**

Created methods for data preprocessing

missing data

cathegorical

remove\_outliers

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

**Refinement**

GridSearch

**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?