Machine Learning Engineer Nanodegree

Capstone Proposal

Anton Mironenko
February 8th, 2019

Proposal

House prices prediction, from Kaggle competition

Domain Background

Quite often it is a challenge to predict house prices based on a big number of factors. Such a model with a decent accuracy would be a good tool for real estate agents. Machine Learning provides a lot of tools to solve the tasks like this.
There is a good corresponding competition on Kaggle with a clean dataset: https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

My personal motivation is to apply knowledge and techniques gained from Udacity Machine Learning Nano Degree to some practical project.

Problem Statement

Predicting house prices is a regression problem, where the output variable takes continuous values.

Datasets and Inputs

The following public dataset is going to be used:
https://www.kaggle.com/c/5407/download-all
It has 80 features, 1 dependent output variable (price), 1460 training data points and 1459 test data points.

Solution Statement

The solution will use regression methods with a help of Python and required machine learning libraries. It shall predict the price based on input variables.

Benchmark Model

The dataset is taken from Kaggle ongoing competition, so that its results can be taken as a benchmark model.
The goal is to provide a model with a metric fit in top 50% of Kaggle competition's results.

Evaluation Metrics

The Kaggle competition uses Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques#evaluation
The same metric will be used in the solution model.

Project Design

The following steps are going to be made to solve the Problem:

- Visualize and analyze the dataset, find correlations and other dependencies,

- Clean and normalize the dataset:
  - remove outliers,
  - fill missing data,
  - transform numeric features if needed,
  - perform one-hot-encoding for categorical features,
  - reduce dimensionality: filter out less important features.

- Try different regression methods, for example: LinearRegression, Lasso, Ensemble methods (AdaBoost, GradientBoost, RandomForest). Probably, some popular on Kaggle methods to be evaluated: XGBoost, LightGBM.

- Analyze obtained evaluation metrics, perform retrospective analysis.