Machine Learning Engineer Nanodegree

Capstone Proposal

Jiri Manak
July 27th, 2017

Proposal

Domain Background

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. 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

This problem is a regression predictive modeling problem.

First goal of the project is to create a model which is able to predict price of the house with defined features.

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 will use data publicly available on Kaggle: House Prices: Advanced Regression Techniques.

This collection consists of 1462 data points for training and 1461 data points for testing.

Testing set will be used for final evaluation of the model. Training set will be split to some number of folds to perform cross-validation.

Solution Statement

This project should extend my knowledge from the area of Regression Models of Supervised Learning which were not part of MLND. I would like to try and tune some of the models available in *sklearn* library. I also will create model based on Neural Network Regression using *Keras* and *TensorFlow*.

Regression Tutorial with the Keras Deep Learning Library in Python
http://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/

A Neural Network Model for Boston House Values
https://www.tensorflow.org/get_started/input_fn#a_neural_network_model_for_boston_house_values

Benchmark Model

- **Linear Regression** - I will use simple linear regression on square footage and price as a benchmark model.

- comparison between different models

Evaluation Metrics

- **Root-Mean-Squared-Error (RMSE)** - to stay compatible with Kaggle metrics I will evaluate the model on RMSE.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

where $\hat{y}_i$ is predicted value of observation i of regression dependable variable y_i computed for n different predictions

Project Design

- **Python 2.7**
- **scikit-learn** – open source machine learning library for Python
- **Keras**
- **Tensorflow**

Data exploration

The first step in my project will be data exploration. By reading the features description I will try to get more insight view on the features. I will explore which data are categorical and which are continues. Different visualization will help to see the dataset distribution and features correlation.

Missing data

Next step in data processing will be search for missing data and finding the right solution how to substitute them.

Models

As mentioned earlier, this project goal is also to get familiar with different regression models. My plan is to use following regressors

Random Forest
Gradient Boosting
Support Vector
Neural Network

Each model will be tuned to achieve best performance and results of all models will be compared. It will be also possible to compare results against leaders of the Kaggle competition.