Machine Learning Engineer Nanodegree

Capstone Proposal

Niclas Geiger
June 10th, 2017

Proposal

Predict House Prices in King Country

Data Set Source: https://www.kaggle.com/harlfoxem/housesalesprediction

Domain Background

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.

Datasets 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.

Solution Statement

My solution will be a model which can predict the housing prize for a new entry (described by the aforementioned 19 features). I will try to use clustering algorithms like KNN, Support Vector machines to create my Model. The Problem Solution will use a supervised learning approach. I might also try feature extraction and space reduction algorithms on this problem.

Benchmark Model

The model will be benchmarked against a simple regression model using the living space as the only feature to make predictions.

Evaluation Metrics

My Evaluation metric will be the R2-Score using cross validation of the data set.

Project Design

First, I will start to load all the data and try to get an overview over it. This means finding averages, means and other values for features. Then I will create the Benchmark Model with the simple regression model. After this I will start to use different Clustering algorithms on several or all features and see how they work out. With the R2-Score I can evaluate how well this clustering is performing. Visualization in the process might also prove helpful to me. Feature Extraction or Feature Space Reduction Algorithms will also be tried if this seems helpful to me in the end. I am thinking about making big use of the latitude and longitude data which might help me create some kind of “heat map” for the more expensive or cheaper areas. This approach combined with a more kind of regression approach for the number of rooms and other values should give me good results as this seems like the approach currently used manually by real estate agents.