Machine Learning Engineer Nanodegree

Capstone Proposal

Yibo Gong

July 18th, 2019

Proposal

Domain Background

I am very interested in the real estate market, which has been very hot in Toronto during the past few years, so for this capstone project I want to focus on something related to the realty market.

When a home buyer describes their dream houses, they usually will not begin with the height of the basement ceiling or the proximity to an east-west railroad. But, besides the number of bedrooms or a white-picket fence, there are a lot more factors that may impact a home price. So for this capstone project, I picked a project from Kaggle: **House Prices: Advanced Regression Techniques** (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview/description). I will apply the Machine Learning techniques I have learnt from the Udacity Nano Degree, to predict the home prices in Ames, Iowa, from 79 features described in the dataset.

The dataset I use (the Ames Housing dataset) was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often-cited Boston Housing dataset.

There are also some research has been done in this area. The below two papers
https://nycdatascience.com/blog/student-works/housing-price-prediction-using-advanced-regression-analysis/
https://arxiv.org/pdf/1809.04933.pdf
have shown some methodologies regarding how to process the data, and how to explore and analyze the data, and also the models can be applied on the data.

Problem Statement

The goal is to predict the sales price for each house. For each Id in the test set, the value of the SalePrice variable will be predicted. This will be a supervised regression learning. We can use Random Forest or Gradient Boosting to analyze the data. We can also apply some pre-processing on the data like normalization, or apply PCA to extract higher level features. The problem is also:

- Quantifiable – With the machine learning techniques, the problem can be expressed in math / logical terms
- Measurable – The metric we will use is RMSE
- Replicable – The model and parameters will be saved, so that the results can be replicated

Datasets and Inputs

The dataset used for this project is the Ames Housing dataset, which can be downloaded from Kaggle - *kaggle competitions download -c house-prices-advanced-regression-techniques*.

The data contains 79 features, 1 target (SalePrice) and 1460 data entries. Some of the important features include continues feature – LotArea, discrete feature – FullBath, nominal feature - Heating, ordinal feature – OverallQual.

I will use the train.csv (downloaded from kaggle) for training and testing. This data will be split into train and test set (8:2, random split); and then the train set will be used for train and validation for model optimization purpose.

Solution Statement

My approach to the project will be as below:

1. Step 1 – Data exploration and feature engineering
   a. Visualize the data and decide if there is any skewed data or not, and see if data normalization is needed or not
   b. Determine outliers if there is any
   c. Apply one-hot encoding if needed
   d. Use PCA to generate high level features which will include most of the data variance
2. Model selection with grid search and k-fold validation
   a. Decision Tree
   b. Gradient Boost
   c. Random Forest

Benchmark Model

I am thinking about using Decision Tree as the benchmark model and use the results from other models to compare to it, by using the same test data set and validation data set. The reason to use Decision Tree as benchmark is that, theoretically, Gradient Boost and Random Forest should have a better performance as they are ensemble models.

Evaluation Metrics

The evaluation metrics will be following:

1. root mean square error

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

   RMSE is the standard deviation of prediction errors. It tells you how concentrated the data is around the line of best fit. In this project, we will use RMSE to determine how good the model’s performance is.

2. model prcess timing (training time and running time)

Project Design

1. Exploring the data
   1. Find min, max, and histogram
   2. Find feature correlations
   3. Find missing data if any
2. Data preprocessing
   1. Find outliers and determine if we need to delete them or keep them
   2. For the missing data, analyze to delete the column or fill the missing data
   3. Transform skewed data
   4. Normalize data
3. Split data into training set and test set
4. Feature selection – PCA

5. Model comparison and selection

Use with grid search and k-fold validation, analyzing the model performance on the following models:

   d. Decision Tree
   e. Gradient Boost
   f. Random Forest

References

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