Machine Learning Engineer Nanodegree

Capstone Proposal

Naoko Shimada
10/24/2018

Domain Background
My project is to build a model which predicts a demand of bicycle sharing. This project is proposed in Kaggle (https://www.kaggle.com/c/bike-sharing-demand/data) 3 years ago.

The city I live is considered as one of the most bike friendly city int the US. In fact, there has been significant development for building bike safe lanes within existing roads to support bicycle users and also to encourage more people to ride a bike in the city. This helped rental bike service flourish. When I started seeing more people are riding rental bikes, I noticed that not all renters were seemed tourists. Then, I began wondering if there was a different trend for bike demand due to a different interest, and motivation of renting a bike such as commute or sightseeing. If we can forecast when bikes are most needed, or when customers want to ride a bike for long distance, then the rental shops can plan for the demand by stocking bikes in needed location, or stocking different types of bikes (ex, an electric bike for long distance demand.) Thus, I wanted to practice how to build such a model using the data in Kaggle.

Problem Statement
Any economy has a demand and supply relationship. And rental bike business is not an exception. Being able to forecast how many bikes will be needed and when these bikes are needed, will help bike rental business. My model will find that what feature(s) will trigger more demand (how many bikes will be needed) so that we can plan to supply the demand (shops will stock more bikes). My assumption is that seasonality (time of the day and day of week) has a strong influence on rental bike demand. If so, I'll use the seasonality features to derive a model to predict the number of bike needed in a specific time/day period.

Datasets and Inputs
The dataset I will use is from Kaggle (https://www.kaggle.com/c/bike-sharing-demand/data). The data consist of the number of bike usage by every hour starting over 2 year period (from Jan/2011 until Dec/2012.) It also includes weather and temperature information of the time/day at which the bike usage was counted. One caveat of this data is that the data is already split into two parts: testing and training. The training set contains all features, but is comprised of the first 19 days of each month. The testing set contains all features, except the number of bikes (which we need to predict), and is comprised of the 20th to the end of each month.

Solution Statement
The goal is to forecast the number of bikes needed for each hour. Here, I plan to use an ensemble method (either (1) Adaboost or (2) Xgboost) with an underlining model of regression since my model needs to predict the number of bikes based on given features. (It is resemble to predict hosing price based on given features.) The date is brake down into hour of day and day of week(workday/weekends). So, I may try to use day of week in addition to hour of day to obtain a better prediction.

Benchmark Model
My benchmark model is chosen from the same bike-sharing demand project in Kaggle (https://www.kaggle.com/casalicchio/tuning-with-mlr). This model has the Root Mean Squared Logarithmic Error of 0.37977 using Xgboost in R.

Evaluation Metrics
I'll use the Root Mean Squared Logarithmic Error (RMSLE) to evaluate the performance of my model. The benchmark model has the highest score (0.37977) of RMSLE in the competition and I hope to get around RMSLE=0.5 with my model.

Machine Learning Engineer Nanodegree

Project Design

<tool of choice: Jupyter note book with Python>

1. **Data Exploration**

(a) Data summary: I'll run a summary stat to see data range of each feature. Also, I'll plot bike counts by hours to see if there is any trend.

(b) Feature evaluation: There are 2 basic category of given features: (1) seasonality related features (workday, holiday, and season) and (2) environment related features (weather, temperature,feel-temperature, humidity, and windspeed). First I'll check correlation between each feature vs bike counts with plotting pair wise scatter plot. Also I'll check features within the each category. If there are any highly correlated features within the category, redundant features will be eliminated from my model. I may try to create new features such as "month" and "day of week" and then compare them against "season" and "workday" features, respectively.

2 Building Model

(a) Data prep:
I'll use the training set to be my total set since the provided test set does not have the bike count. I'll use sklearn.model_selection package to split the training set into "my-training" set and "my-test" set. (This is not ideal since I'm building a model with a subset of the data, but this is the best I can come up with.)
As for "my-test" set, I'll delete the number of bike counts. Also, I may apply One-Hot-coding from sklearn to the features like "season" which has 1=spring , 2=summer, 3=fall and 4=winter.

(b) Model training:
Based on "my-training" set, I'll test two boosting algorithm with all features with Regression model. One is Adaboost and the other is Xgboost.

(c) Evaluation:
Apply the model to "my-test" set and calculate RMSLE and pick the lowest score of the model.

Reference

- Data: https://www.kaggle.com/c/bike-sharing-demand/data (train.csv)
- Benchmark Model: https://www.kaggle.com/casalicchio/tuning-with-mlr
- TOOL: Python /Jupyter notebook