Machine Learning Engineer Nanodegree

Capstone Proposal

Jason Iijima

January 5th, 2018

Proposal

Domain Background

Bike share is a service where people can rent bicycles to ride around the city. They are typically charged by length of time they rent the bike. Early programs had stations where you can rent and return bikes from. Newer programs utilize GPS in smartphones to track the bike's location, as well as user routers. This allows bikes to be dropped off anywhere and not be tied down to a station. The benefit of a bike share program is that people do not need to allocate space in small city apartments for bike storage and you do not need to worry about bike maintenance.

Bike share programs have grown increasingly popular over the years. According to www.citylab.com, the US had their first bike share in Washington DC with 10 stations and 120 bikes. In 2013, New York launched a program with 6,000 bikes. Chicago and San Francisco launched their programs that year as well. By 2015, China has over 1 million registered bike share users.

Machine learning has been used to solve the problem of predicting bike share demand before. Stanford University has conducted a study using data from Washington DC bike share for the years 2011-2012. Results are presented here. They used a variety of approaches including, ridge regression, support vector regression, random forest and gradient boosted forest. They were able to attain a root mean squared logarithmic error of 0.31 for random forest and gradient boosted forest.

Problem Statement

Bike share programs have their share of issues. Logistically, bikes need to be transported to check out stations where the demand to rent a bike is high. Often times, bikes will be used for one way trips and without intentional placement of bikes by the service maintainer, the distribution of bikes will not be optimal in allowing availability in high demand areas. A simple example is the case of one station being at the top of a hill and another at the bottom. There will be a higher demand for bikes at the top, but many people will not want to ride the bikes up the same hill. If there are insufficient bikes at a station, the bike share program will lose business in the immediate loss of a rental, along with the user’s perceived future lack of availability of bikes.

The problem of predicting bike share demand by day and by station is a regression problem. We are looking to predict a numerical value of how many bikes will be rented at a given station based on each station’s usage metrics like location, whether the bike was returned to the original station, and weather information.

Datasets and Inputs

The data I will be using is the Austin Bike Share Trips as listed on https://www.kaggle.com/jboysen/austin-bike. I will also pull historical weather data from https://www.kaggle.com/grubenm/austin-weather/data. The bike share station data will contain name of the bike share station, station id, staton status, latitude, longitude, and location (combined lat/long). The bike share trip data will include bikeid, checkout time, duration in minutes, return station id, return station name, month of rental, start station id, start station name, rental start time (date and time), subscriber type (yearly, monthly, etc.), trip id and year. The weather data contains date, total daily precipitation, notes on weather, the max, average and min for temp, dew point, humidity percentage, sea level pressure, visibility in miles, wind speed (max and avg only).

There are 649,000 inputs of bike share usage over the course of 2013 - 2017. This is spread over 72 bike share rental stations. There are 1318 days worth of weather reports. The outcome will be an integer depecting the demand of bikes per day per station. The training and validation data will be split randomly from a subset of the data containing the years 2013-2016. The testing data will be all of the 2017 data. This will allow us to ensure there is no look ahead bias in the training.

Solution Statement

The goal of this project is to predict what the demand will be at the most used station for each day of the year in 2017 where we have a data. I will look at the total daily bike rentals for the most popular stations and check the deltas between rentals and returns at those stations. If the number of bikes at a station is less than 125% of the the reported demand for the next day, a staffer will be notified to bring x number of bikes to that station (there is a safety buffer of 25% to accommodate higher than expected usage of bikes due to variables that we do not have data on). The dataset includes information such as location of the bike station, time and date of rental, duration of use, etc. Outside information such as weather can also be used to predict demand. If the forecast is rain, it may be likely that the demand will be low. This will also be a variable in the prediction. To predict the demand, I will use a variety of approaches including linear regression, ridge regression, lasso regression, and support vector regressions. I will use all of the relevant variables and measure the output of each algorithm.

Benchmark Model

The benchmark model for this will a static models that takes the average number of daily rentals at each station over the course of a month. For example, if February had an average of 50 bikes rented at a station, that station would require 50 bikes docked at that station every morning in February. This is a very simple model, but the main drawback is that an average means it will overestimate some days and underestimate other days. The days where the number of bikes rentals were underestimated would lead to lost revenue.

Evaluation Metrics

One evaluation metric is the root mean squared logarithmic error between the predicted demand at a station and actual demand at a station. This will give us a level of accuracy in our prediction, but will have the benefit over a root mean squared error by negatively weighing predictions that are below the actual demand. From a business perspective, it is more detrimental to have less than you need because that means lost revenue. This evaluation can be used on the benchmark model as well. This will show us how much better a machine learning solution is over the benchmark model. The daily predicted demand should exceed the actual demand by 25% to prevent loss of revenue in unforseen events. The predicted value should also not exceed the actual demand by 50% to maximize bike usage and minimize stocking unused bikes. The evaluation metric will see how many days fit these two parameters.

Project Design

The first step is to explore the data and see what patterns stand out. See what the distribution is for each of the stations and how they differ amongst themselves. How much usage does a popular station get compared to one that has little usage? Do all stations grow in usage over time, or are some station locations poorly chosen and doomed to collect dust? What is the difference between bikes rented and bikes returned at any given station? This particular question will show us the magnitude of the issue I am trying to solve.

The next step is to process the data for modeling. Variables will be analyzed for correlations and variance. For example, if usage does not vary much between Monday-Friday but does vary between weekdays and weekends, we can reduce dimensionality by replacing a 7 option day of the week variable to a binary isWeekday variable. Principal component analysis can be done to reduce dimensionality as well.

Candidates for strategies to predict demand of bikes on a given station at a given day include linear regression, ridge regression, lasso regression, and support vector regressions. I will use all of the relevant variables and measure the output of each algorithm.