Machine Learning Engineer Nanodegree
Capstone Proposal

Akshaya Rane

June 2, 2018

1 Domain Background

Community bike-sharing systems are public transportation programs which enable tourists as well as local residents to rent bicycles from one location and return it to a different location as needed. Their central motivation is to provide affordable access to bicycles for short-distance trips in an urban area as a way to enhance mobility, alleviate automotive congestion, thereby reducing air pollution, noise, and boost health. The first bike-share project began in Amsterdam in 1965 and today more than 500 cities in 49 countries host advanced bike-sharing programs with over 500,000 bicycles (1).

These systems are fully automated via a network of kiosk locations throughout a city and hence they generate digital footprints that reveal mobility in a city. Therefore machine learning can be used to study this data that can help in designing and planning policy in urban transportation, to optimize the service by forecasting rental demand thereby regulating the availability.

1.1 Motivation

My personal motivation for considering this program is because:

- There is great interest in these systems today due to their important role in traffic, environment as well as health issues in urban areas.
- I have used bike-sharing systems in cities like Boston and Vancouver and have preferred it over driving around the city, so I am very curious to investigate this system in detail.

Many major cities are aiming to become bike-friendly day-by-day. Bike sharing programs are emerging worldwide as can be seen in Figure 1[1]. This increasing trend indicates that we would soon see this service in most major cities alongside other public transport services. So, my personal goal is to use this project as a first step to create a machine learning framework using supervised learning techniques to forecast bike rental demand in a city. I describe the problem statement briefly in Section 2 and the datasets are described in Section 3. A solution statement, benchmark model, and corresponding evaluation metrics are discussed in Section 4, Section 5, and Section 6 respectively. Finally the project design is laid out in Section 7.

2 Problem Statement

The goal of this project is to predict bike rental demand in the Capital Bikeshare program in Washington, D.C. by combining historical usage patterns with weather data and annual holidays. In other words, we need to predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.

[1] http://www.earth-policy.org/data$_{c}enter/C23$

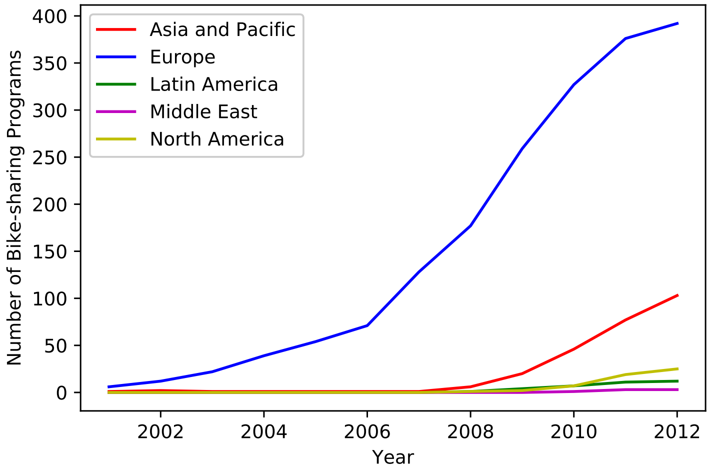

Figure 1: Number of bike-sharing programs across the world from 2000 to 2012.

3 Datasets and Inputs

This project will use the bike sharing dataset obtained from the UCI machine learning repository[2] which combines the hourly rental data between years 2011 and 2012 in Capital bikeshare system[3] with weather information[4], and holiday schedule[5].

This dataset contains a total of 17,389 instances that are hourly time stamps. Each instance is described by 16 features:

- dteday: date
- season: season (1:spring, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : if holiday:0,, else 1
- weekday : day of the week starting Sunday (0 to 6)
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0
- weathersit : 1 if Clear, Few clouds, Partly cloudy, Partly cloudy;
  2 if Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist;
  3 if Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds;
  4 if Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are derived via $(t - t_{\min})/(t_{\max} - t_{\min})$, $t_{\min} = -8, t_{\max} = +39$

[2] http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
[3] http://capitalbikeshare.com/system-data
[4] http://www.freemeteo.com
[5] http://dchr.dc.gov/page/holiday-schedule

- atemp: Normalized feeling temperature in Celsius. The values are derived via $(t - t_{\min})/(t_{\max} - t_{\min})$, $t_{\min} = -16, t_{\max} = +50$
- hum: Normalized humidity. The values are divided by 100 (max)
- windspeed: Normalized wind speed. The values are divided by 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered (target variable)

4 Solution Statement

This is a well-defined supervised learning problem, more specifically a regression problem, with hourly records of weather, holiday information as features and total number of bike rentals as target variable. I will consider all other features as input parameters initially. Then I would like to perform feature selection to determine the most important features and carry out the regression analysis. I will investigate the parameter space further to get the best test metric.

5 Benchmark Model

A few scikit-learn regressors (linear regressor, random forest regressor) were implemented initially. The performances are listed below:

Table 1: Benchmark performances

Algorithm	RMSLE
Linear regressor	0.6247
Random forest regressor	0.1245

6 Evaluation Metrics

Since the model will predict the number of bike rentals, I can compare those with the actual number of bike rentals. I think the root mean square logarithmic error (RMSLE) would be a good choice to evaluate the model which measures the ratio between the actual and predicted as:

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(log(p_i+1) - log(a_i+1)\right)^2} \tag{1}$$

where n is the number of hours in the test set, p_{i} is the predicted count, and a_{i} is the actual count.

7 Project Design

7.1 Programming Language and Libraries

- **Python 2 (Pandas, Numpy, Scipy, Matplotlib).**
- **Scikit-learn**.

7.2 Data preprocessing

The first step is collecting and extracting the data from UCI machine learning repository. The data files are in csv format with numeric values. I will load the csv file and display first few rows to get a glimpse of the data. The next step would be to check for missing values in the data.

7.3 Splitting the data

Since this is a time-series data set, I will split data as:

- The training set will comprise of the first 19 days of each month.
- The test set will be the 20th to the end of the month.

7.4 Data Visualizations and correlation analysis

The next step would be to visualize the input features to see how they affect the target variable (total number of bike rentals). This will also help in determining correlations between features and also in detecting outliers. An example plot is shown in Figure 2, which compares the 'feels-like' temperature with the total number of bike rentals. I have converted the normalized values to actual temperatures for better understanding. It can be noticed that the number of bike rentals peak at a temperature of ~ 25 C which is reasonable.

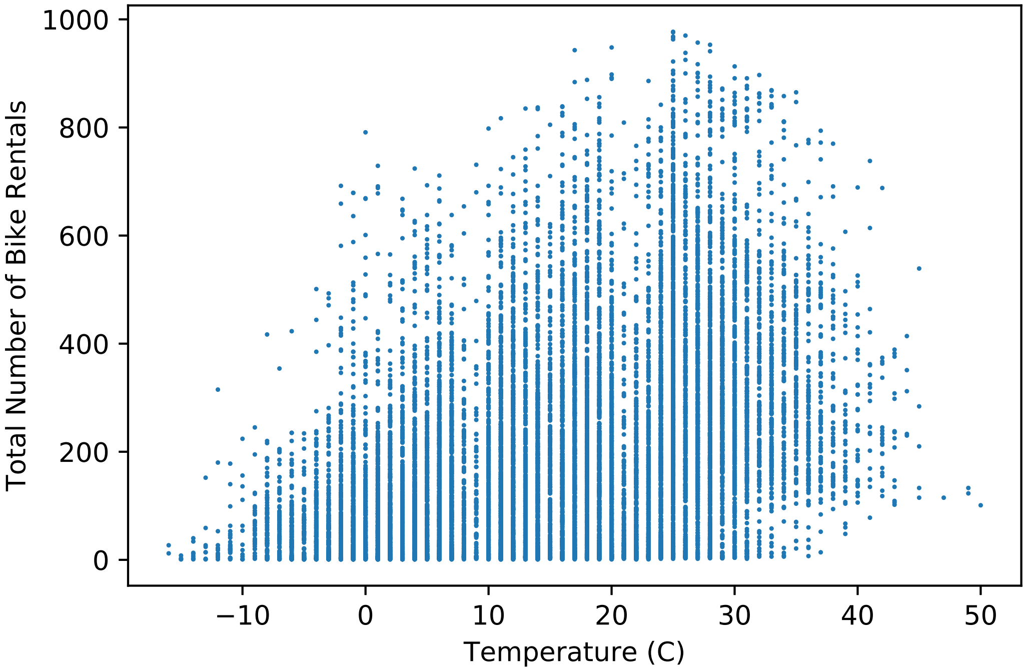

Figure 2: Total number of bike rentals as a function of 'feels-like' temperature.

7.5 Exploring models

I plan to explore various models (Linear regressor, decision tree regressor, ensemble models etc) to start with and then choose a best-fit model.

7.6 Optimization

Once the model is decided, I will tune the hyperparameters in order to optimize the model.

References

[1] Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide Earth Policy Institute; Larsen, Janet; 25 April 2013.

[2] Public bikesharing in North America: Early operator and user understanding http://transweb.sjsu.edu/sites/default/files/1029-public-bikesharing-understanding-early-operators-users.pdf

[3] Bike sharing demand-Forecast use of a city bikeshare system https://www.kaggle.com/c/bike-sharing-demand