Evaluating Predictions on Bike Rental Frequency

Phanindra Kumar Kannaji, Venkata Rami Reddy Bujunuru
Graduate Student
UMBC - CSEE
{pkanna1, bo26494}@umbc.edu

***Abstract*—The purpose of this project is to develop a model which will predict the approximate number of bikes rented on a particular day or hour based on the Capital Bike share data. The predictions are made by combining the bike share data with daily weather data and holiday data of Washington DC. The models that are used for prediction is evaluated based on the accuracy of the predictions. Various models are evaluated such as Random Forest model, Bootstrap Aggregation, Conditional Decision Trees(C-trees), K-Nearest Neighbors (KNN), Support Vector Machines (SVM) for classification. From our observations, the best model for this bike share data is Random Decision Forest model on decision trees which is one of the ensemble learning methods for classification. The accuracy of this model is approximately 80%.**

I. Introduction

Bicycle sharing system started to grow in the year 1996 at Wisconsin, North America where red painted bikes are used for public service primarily students. In 2010, capital bike share started the bike share service in the area of Washington DC which is open to public. Heavy traffic in the capital areas and the demand for environment-friendly transport promoted the use of the bike share more frequently. The number of bikes that are rented are gradually increasing year by year. The usage of bikes depends primarily on weather and whether its a holiday or not.

The bike rental frequency is the number of bikes that are rented for a given hour. The elements/conditions of the weather that can be considered in deciding whether the bikes will be rented or not are :

- rain
- snow
- wind speed
- inclement weather conditions
- temperature
- humidity
- cloud cover etc..

The predictions related to any statistical models can be done by regression, classification, clustering etc,. which depends on the data. The bike share data is more suitable for classification type of models since the prediction is not an exact value but a class or a range of values. There are various machine learning models which can be applied to the bike share data to predict the busyness of the bike rentals during an hour. The accuracy of prediction is calculated for all these models and the best model with highest accuracy will be used to estimate the bike rental frequency in the future.

II. Motivation

There is a significant impact of the method of transportation an individual choose on the global climate. The heavy traffic and the fueled vehicles are one of the main reasons for the environmental pollution. A bike share program is the best approach for this problem as the bikes are rented at the starting point A and are left at a point B and within that paid hours, any other bike can be used to go to any other point C. The price that the user have to pay is based on the amount of time the bike is taken out of the slot and put back again. There is a demand and supply gap at some particular hours for example, when it is a holiday and sunny outside, the bikes rented will reach the saturation and leads to unavailability of the bikes. To resolve this problem, we created a model to predict the number of bikes rented on hourly-basis so that the demand and supply gap can be minimized and bikes will be available all the time which makes this a reliable mode of transportation.

III. Data Formation

The primary datasets involved in the data that have to be provided to the machine learning model are:

- Capital Bike Share data
- Weather data
- Holidays data

A. Bike Share Data

The bike share data[1] consists of the information regarding the bikes rented per bike rent instance from January, 2011 to June, 2016. Following are the features/variables of the dataset:

- Duration - Duration of trip
- **Start date** - Includes start date and time
- End date - Includes end date and time
- Start station - Includes starting station name and number
- End station - Includes ending station name and number
- Bike Number - Includes ID number of bike used for the trip
- Member Type - Lists whether user was a Registered or Casual member.

  Since the data contains all the bike rental instances, we performed a few operations on the data to calculate the hourly count and the time is split into year, month, day, day of the week, hour and the count for that hour is calculated by means of a GROUP-BY kind of functionality.

B. Weather Data

The weather data[2] is provided by Forecastio through an API offered by Darksky.net. The API will give the hourly data for a given latitude, longitude and day. Some important details provided by the Darksky.net are:

- **Apparent Temperature** - Feels like temperature during that hour
- **cloudCover** - The percentage of sky occluded by clouds, between 0 and 1, inclusive
- dewPoint - The dew point in degrees Fahrenheit
- **humidity** - The relative humidity, between 0 and 1, inclusive
- icon - A machine-readable text summary of this data point
- precipIntensity - The intensity (in inches of liquid water per hour) of precipitation occurring at the given time
- **summary** - A human-readable text summary of this data point
- **temperature** - The air temperature in degrees Fahrenheit
- **windSpeed** - The wind speed in miles per hour

The attributes provided by the weather data are stored from Jan, 2011 to June, 2016 to merge with the bike share data. The features considered in weather data are time, summary, wind speed, temperature, cloud cover and humidity. Summary provided in the dataset is a textual representation of the weather in the area and it is categorized into following:

- **1** - Clear, Breezy, Partly cloudy, hum,id, dry
- **2** - Foggy, overcast, Mostly cloudy, Drizzle, Windy
- **3** - Light snow, Snow, Light rain, Flurries, Rain
- **4** - Heavy rain, Heavy Snow and other inclement weather conditions

The summary is transformed into the above 4 classes and the possible values are a million as the text can be anything related to weather conditions. The combinations of the text is also possible. For example, a text can be Light snow and windy. In such cases the class with highest number will be picked as under fitting is much friendlier than over fitting in the weather because the combination of higher class and lower class will be dominated by the impact of the higher class on the number of bikes practically.

C. Holiday Data

The holiday data[3] is entered manually as the number of data points or samples that are effected by the holiday data is very less and is less than 12 per year. This entering of holiday data is done after merging both the weather and bike share data. The holiday data is added to increase the performance as the bike rental frequency increased almost to peaks during the holidays. Since the weekends are also considered as holidays, the weekends are also marked as holidays in the combined data set.

D. Combining the data

The common feature in both the weather and bike share data sets is the time. Since both the data sets are hourly based a join operation could be performed on both the data sets to merge them. First, the time in the weather data is also split into year, month, day, hour as done in bike share data. Then, the data sets are Left-outer-joined with left side of the join being the bike share data. The entries in the bike share data now have the weather data present. To this, the holiday data is added as another feature (0 or 1).

E. Classifying labels

The count of number of bikes rented per hour will be the final output of the other features in the dataset. Thus, for a given hour and weather conditions, the number of bikes rented have to be counted. With this approach, the error function have to be maintained as the exact prediction of the number of bikes is almost not possible. Thus we followed a classification kind of approach instead of regression by classifying the count of number of bikes into categories. We plotted the bike rental frequency and the number of times those frequencies occurred which is mentioned in Figure 1. Thus, the bike rental

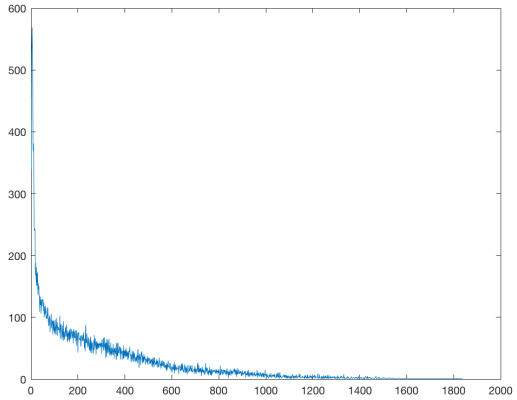

Figure 1: Bike Rental Frequency vs Count

frequency is classified into 7 classes for accurate prediction as mentioned in Table I.

Value	Frequency	Class
1	1 - 40	Can't see any bikes
2	40 - 200	Very few bikes
3	200 - 400	Less busy
4	400 - 600	Normal
5	600 - 800	Busy
6	800 - 1100	Very Busy
7	1100 - 2000	All bikes around

TABLE I: Classification of labels

REFERENCES

[1] Capital Bike Share Data-set http://www.capitalbikeshare.com/system-data
[2] Weather Data - ForecastIO - Washington DC https://darksky.net/dev/docs/response
[3] Holiday Schedule Data http://dchr.dc.gov/page/holiday-schedules
[4] Kaggle.com https://www.kaggle.com/c/bike-sharing-demand