Capstone Proposal

July 26, 2018

Proposal: Forecast Use of Capital Bikeshare Program

Domain Background

Washington DC is falling in love with bike share. Capital Bikeshare is metro DC's bikeshare service, with 4,300 bikes and 500+ stations across 6 jurisdictions: Washington DC; Arlington, VA; Alexandria, VA; Montgomery, MD; Prince George's County, MD; and Fairfax County, VA. Bike sharing is a means of renting bikes where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Bikeshare users are able to rent a bike from one location and return it to a different location on an as-needed basis.[1] It serves both work-related and personal travel needs. It is easy, fun, flexible, affordable and environment-friendly. Since its inception in 2010, Capital Bikeshare users have ridden more than 42 million miles, or 20 million trips, which reduced 28.64 million pounds of carbon dioxide, saved 1.72 million gallons of gasoline, and burned an astonishing 1.8 billion calories, according to the District Department of Transportation (April 26, 2018)[2].

Capital Bikeshare represents a small but growing component of the regional transportation system. Predicting bikeshare demand is important for effective operation and expansion of the existing bike sharing system. Knowledge of how demand fluctuates enables the bikeshare program management to keep the right amount of stock on hand. This project hopes to shed some light on the opportunities and barriers to strengthen the role of biking in the regional transportation system.

Many studies have been carried out to explore the usage of bike share systems in the world. For example, Kaltenbrunner et al. studied the spatial and temporal patterns of bike use over the hours of the day.[3] Zhou examined the spatial-temporal pattern of bike trips in Chicago.[4] Travel patterns between weekdays and weekends as well as between subscribers and casual users tend to be very different. Students from Stanford University also built different predictive models for bike sharing demand.[5] They tried methods including basic linear regression, Generalized Linear Models with Elastic Net Regularization, Generalized Boosted Model, Principal Component Regression, Support Vector Regression, Random Forest and Conditional Inference Trees. Random Forest and Conditional Inference Trees models reported the smallest RMSLE.

---

[1] https://www.kaggle.com/c/bike-sharing-demand
[2] https://ddot.dc.gov/release/capital-bikeshare-celebrates-20-million-trips-and-highest-daily-ridership-record
[3] http://www.dtic.upf.edu/~akalten/kaltenbrunner_etal2010PMC.pdf
[4] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137922
[5] http://cs229.stanford.edu/proj2014/Jimmy%20Du,%20Rolland%20He,%20Zhivko%20Zhechev,%20Forecasting%20Bike%20Rental%20Demand.pdf

Problem Statement

This project focuses on predicting the number of bike rentals for the Capital Bikeshare program, as part of a Kaggle competition.[6] It aims to understand key factors driving the hourly bikeshare demand using historical bikeshare usage data with weather data. The objective here is a typical demand forecasting problem. The predicted variable is the demand – number of bike rentals (demand), which is a real number rather than a class or category. Therefore this is a regression problem rather than a classification problem.

Target variable in this project will be hourly number of bike rentals (demand), which will be the variable to be predicted by the model. Input variables will be factors that drive the bike rental demand, such as the weather condition (rain, wind speed, humidity, temperature, etc.), time of the day, day of the week, season, holiday or weekday, etc. The selection of input variables could also be different for casual and registered users.

Datasets and Inputs

This project will use the data provided in Kaggle's competition, in which hourly rental data spanning two years (2011-2012) for the Capital Bikeshare program is collected. The training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. The goal is 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.
Date fields include:

- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday (binary indicator)
- workingday - whether the day is neither a weekend nor holiday (binary indicator)
- weather – four categories include:
    1) Clear, Few clouds, Partly cloudy, Partly cloudy
    2) Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    3) Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    4) Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp - temperature in Celsius
- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated (dependent variable)
- registered - number of registered user rentals initiated (dependent variable)
- count - number of total rentals

The goal is to predict the total number of bike rentals on an hourly basis. The last three variables are considered as dependent variables, while the remaining variables are independent

[6] https://www.kaggle.com/c/bike-sharing-demand

variables. These variables are appropriate given the context of the problem. For example, the sensitivity of bike use to weather conditions has been widely discussed in the literature.[7] In theory, bike usage can be affected by cold weather, precipitation, and excessive heat. It is also known that bike share demand shows seasonal patterns – peaking in summer/fall, and with a lull around holidays (especially for registered users). There is also a huge difference between casual and registered users. For regular registered users, there is a clear peak period travel. Commuting to and from work has dominated these two peak travel periods in the morning and in the afternoon. Casual users show a very different pattern. Number of trips by casual users gradually increases after 7 am in the morning, remain a relatively stable demand through the midday until PM rush period.

Solution Statement

A range of machine learning algorithms are considered to be used for demand prediction. The most common one is the method of regression. Other possible methods could be Ridge/LASSO Regression, Support Vector Regression, Gradient Boosting, and Random Forest. We will pick a few algorithms to evaluate. We will tweak the parameters to try to maximize each algorithm's performance for both the training and test sets. Models can be improved further by fine-tuning the modeling. Casual users and registered users should be predicted separately.

Benchmark Model

There is not an existing benchmark model. Instead, we will start with a multiple linear regression model with default parameters as a benchmark model. The performance of this model will serve as the benchmark for the more advanced approach later.

Linear regression is a linear approach to modelling the relationship between a dependent variable and one or more explanatory variables.[8] In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. We expect this linear regression model's performance would be relatively poor, as it may overfit the noise in the dataset.

A population model for a multiple linear regression model that relates an independent variable Y to independent variables X (X1,X2,…Xp-1) is written as:

$Y_i=b_0+b_1X_{i,1}+ b_2X_{i,2}+ \ldots + b_{p-1}X_{i,p-1} + e_i$

Evaluation Metrics

The results are evaluated on the Root Mean Squared Logarithmic Error (RMSLE). The RMSLE is calculated as:

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

[7] https://phillymotu.files.wordpress.com/2013/04/the-impact-of-weather-conditions-on-capital-bikeshare-trips.pdf

[8] https://en.wikipedia.org/wiki/Linear_regression

In the RMSLE equation, n is the number of hours in the test set, and p_i and a_i are the predicted count and actual count for a given hour respectively.

Project Design

(1) Download data from Kaggle.

(2) Data cleaning, exploration, feature observation (with some data visualization)
- Some new features need to be created based on existing ones. For example, datetime variable will be split into separate variables – year, month, day of the week and hour.
- Some other variables such as season and holiday need to be converted into new dummy variables using one hot encoding.
- Numbers of casual, registered and total users need to convert into logarithm form.
- Some variables clearly have strong correlations with each other. For example, working day vs. day of week, season vs. month. We need to use extra caution while selecting the features in each model.

(3) Developing models, model training and optimization of model parameters.
We will test a few different models, as discussed earlier. We will start with a benchmark model – a linear regression model with default parameters.
- Other than the models we discussed earlier, the XGBoost and LightGBM models could be good supervised learning approaches to try here.
- Since we will try multiple models, we will consider stacking, a model ensembling technique used to combine information from multiple predictive models to generate a new model, which often outperforms each of individual models due its smoothing nature and ability to highlight each base model where it performs best and discredit each based model where it performs poorly.[9]

(4) Analyzing and evaluating model performance (grid search and k-fold cross validation)

(5) Final conclusion.

[9] http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/