Machine Learning Final Report Spring 2014

Capital BikeShare Data Set Analysis and Prediction

Pavan Reddy, Shreya Ghosh, Nianzu Ma, Paul Landes

Contents

1 Introduction

Bike sharing systems have made renting bikes efficient and quick with memberships, multiple bike locations and easy rental and return process. Through these systems, a user is able to easily rent a bike from a particular station and return back at another station. The biggest issue bike sharing systems face is due to adverse effects of harsh weather conditions on biking. This paper shows how different weather conditions would affect bike usage. The data used here has log of bikes rented in different seasons, under different weather conditions and on weekdays or holidays. We use different regression and classification algorithms to make relevant predictions for bike sharing systems, which would help them deal with extreme conditions like very few bikes were rented or all of the bikes were rented.

1.1 Data

In recent years bikes have turned out to be an efficient alternate mode of transportation in some cities, which lead to growth of bike-sharing systems. These systems are quite convenient for users, since the user do not have to worry about parking, they have options where you can become a member and often these systems let first 30-60 minutes of ride free. Currently there are over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles[1].

The data set being used in this paper was created from trip logs kept by Capital Bikeshare (CaBi) in Washington, DC, where Capital Bikeshare is currently the largest in the nation with over 1,200 bicycles at 140 stations. The data set put together by The University of California at Irvine has features like date, weather, weekdays, holidays and count of bikes rented. These features of the data make it apt for research such as finding patterns between different features and number of bikes rented.

The data set was created from three different sources. Bike trip logs were collected from Capital Bikeshare, merged with weather data from Freemeteo and a holiday schedule provided by The District of Columbia. The data set being used in this paper has hourly as well as daily usage data over the course of two years (2011 - 2012).

1.2 Data Definition

The compiled set of parameters features used follow in table table 1.

2 Problem Definition

The data has hourly and daily usage. In this paper we develop an understanding about how humidity, temperature, rainfall, snow and other weather conditions affect the number of bikes being rented. We also determine bike rental patterns of different kind of users[2].

[2] Registered and non-registered users

Variable	Description
instant	record index
dteday	date
season	season (1:springer, 2:summer, 3:fall, 4:winter)
yr	year (0: 2011, 1:2012)
mnth	month (1 to 12)
hr	hour (0 to 23)
holiday	weather day is holiday or not[1]
weekday	day of the week
workingday	if day is neither weekend nor holiday is 1, otherwise 0.
weathersit	intensity of weather (1 - 4)
temp	Normalized temperature in Celsius. The values are divided to 41 (max).
atemp	Normalized feeling temperature in Celsius. The values are divided to (max)
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	count of casual users
registered	count of registered users
cnt	count of total rental bikes including both casual and registered

Table 1: Data Features

This analysis would help bike sharing companies to predict what number to bikes they need to stock on any given day.

3 Feature Selection

WILL CHANGE We used various methods for data processing and choosing features to build the classification and regression models.

3.1 Feature Scaling

Feature Scaling is a method used for data processing to standardize the range of the feature variables. The range of the features are normalized to a value having unit variance. The values can standardized to either ranging from (0 to 1), this can be obtained by normalizing by the maximum value or ranging from (-1 to 1) which can be obtained by subtracting the value by the mean and then normalizing by the maximum value. For example, the temperature ranges from 0 to 40 deg C but it is standardized to a variable having unit variance ranging from (0 to 1). We used standardized features for variables temperature, feels like temperature, humidity and wind speeds. The consequence of using feature scaling is to make sure all the features have similar variances and ranges which inturn helps in faster convergence of gradient descent of the cost function for the regression models.

3.2 Dependent Variable Correlation matrix

	workingday	**temp**	**atemp**	**hum**	**windspeed**	**cnt**
workingday	1.00	0.06	0.05	0.02	-0.01	0.03
temp	0.06	1.00	0.99	-0.07	-0.02	0.40
atemp	0.05	0.99	1.00	-0.05	-0.06	0.40
hum	0.02	-0.07	-0.05	1.00	-0.29	-0.32
windspeed	-0.01	-0.02	-0.06	-0.29	1.00	0.09
casual	-0.30	0.46	0.45	-0.35	0.09	0.69
registered	0.13	0.34	0.33	-0.27	0.08	0.97
cnt	0.03	0.40	0.40	-0.32	0.09	1.00

Table 2: Features Correlation Matrix

Interpretation of table 2 are given below:

- 0 indicates no linear relationship.
- +1 indicates a perfect positive linear relationship: as one variable increases in its values, the other variable also increases in its values via an exact linear rule.
- -1 indicates a perfect negative linear relationship: as one variable increases in its values, the other variable decreases in its values via an exact linear rule.
- Values between 0 and 0.3 (0 and -0.3) indicate a weak positive (negative) linear relationship via a shaky linear rule.
- Values between 0.3 and 0.7 (0.3 and -0.7) indicate a moderate positive (negative) linear relationship via a fuzzy-firm linear rule.
- Values between 0.7 and 1.0 (-0.7 and -1.0) indicate a strong positive (negative) linear relationship via a firm linear rule.

According to the values in table 2 only temp (normalized temperature, value = 0.4), atemp (normalized feeling temperature, value = 0.4), and humidity (value = -0.32) has moderate positive-negative linear relationship with count (the total bike count of each hour). This indicates that these features implies an important role in regression and classification algorithms. This information helps us to do feature selection when building models for prediction. Figure 1 shows the temperature to bike rental count, which resulted in 0.4.

Note that, the correlation coefficient of windspeed and count is only 0.09. This number indicates a very weak linear relationship correlation. This challenges the common sense of

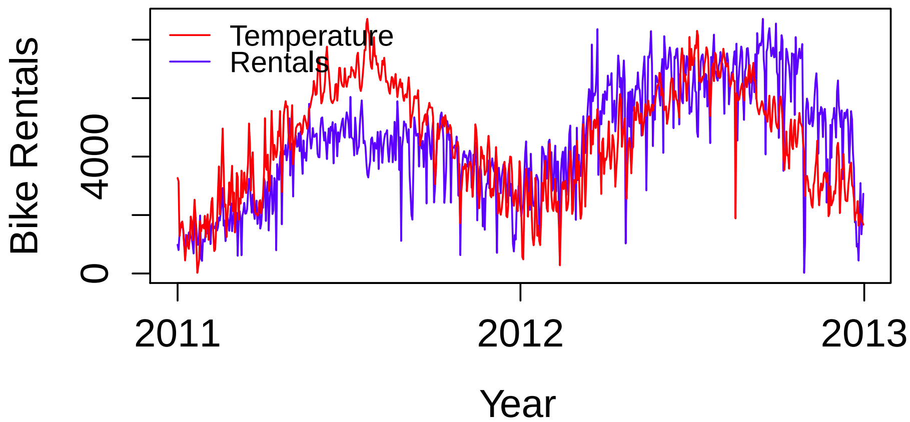

Figure 1: Temperature Bike Counts Correlation

those living in colder climates as windspeed, and therefore comfort of the ride, would effect the willingness commute by bike. However, the average in Washington D.C. the value is 8.2 MPH, this can be described as a "gentle breeze" defined by the Beaufort[3] wind force scale. Therefore, both the correlation coefficient value and real windspeed data show that windspeed will not affect the rental numbers of the D.C. area.

Figure 2 shows the correlation between wind speeds and rentals, which resulted in 0.09.

3.3 Stepwise Regression Feature Selection

Stepwise regression is a method used to automate the process of choosing the feature variables used in the regression models. This method allows selecting a subset of the predictor variables from a larger set of predictor variables by performing a stepwise regression. The stepwise regression can be performed by forward selection, backward elimination or both. We used backward elimination stepwise regression, this includes starting with a model that has all the predictor variables, and then deleting variables from the model that improves the model the most after its deletion, iteratively. Using stepwise regression, we were able to ascertain that features such a hour23 (23rd hour of the day), season_winter (winter season) are eliminated from the model. The predictor variables such as hour23 and season_winter have very low correlation coefficients, -0.1171 and 0.02942 respectively.

[3] http://en.wikipedia.org/wiki/Beaufort_scale

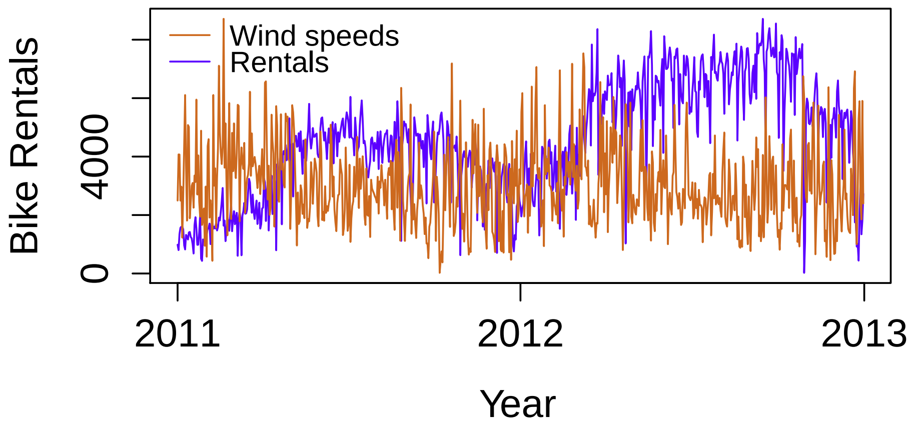

Figure 2: Correlation of Wind Speed and Rentals

4 Approach

4.1 Regression

4.1.1 Linear Regression

We built multiple linear regression models for both daily and hourly data. We built multiple regression models based on different feature selections. Based on the data procession, predictor variable analysis, feature selection methods, we built multiple linear regression models, which takes the general form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where pp is the number of variables. Examples of values of x are weather (i.e. rain), temperature, number of rentals, etc. The β values are the learned parameters.

4.1.2 Non-linear Regression

Based on the analysis of the predictor variables, we build non linear regression models using nonlinear combination of the model parameters. We transformed predictor variables such as the temperature, feels like temperature to their respective square values. Predictor variables such as wind speed which have lower impact on the model, based on the correlation analysis of predictor variables, were transformed to square roots of the respective values. Using these non linear combination of the predictor variables, we built non linear regression models for the hourly dataset.

4.1.3 Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning method that can be used to build regression models. We used SVM to build regression models for the hourly dataset. We tuned the SVM models with a multitude of parameters such as kernel functions, degree of the polynomial, constant for regularization term which intuitively is a parameter that helps mitigate the effects of overfitting, (gamma) the parameter that defines how far a training example's influence reaches etc. For kernel functions we experimented with SVM models that used different kernel functions such as linear, polynomial, radial and sigmoid functions.

4.2 Classification

4.3 Evaluation

4.3.1 Training and Testing

We split the entire dataset into a training and test set in $70:30$ percent ratio. The models were trained on the training set and we used these model to predict the total number of bike rentals in the testing set.

4.3.2 K Fold Cross Validation

We are using 10 fold cross validation to test accuracy of our models. It divides the data in 10 sets and treats 9 sets for training, 1 set for testing. This process is repeated 10 times and the output represents average of all the runs.

4.3.3 Regression

The regression model and predictions are evaluated using Coefficient of Determination or R^2 and Root Mean Squared Error (RMSE). R^2 is the measure of how well a model has fit the data. R^2 can be computed by

$$\begin{aligned} R^2 &\equiv 1 - \frac{SS_{res}}{SS_{tot}} \\ SS_{tot} &= \sum_i (y_i - \overline{y})^2 \end{aligned} \tag{2}$$

where $\overline{y}$ is the mean value; that is the total sum of squares and SS_res is the sum of squared residuals.

Root Mean Squared Error (RMSE) is a measure of the square root of the mean of the squared errors or squared deviations of the prediction values from the actual values. This value is an indicator to compare different models. A lower RMSE value indicates a better model. RMSE can also be used as a measure of how well a model fits the data and if the model is overfit or underfit. This is done by computing the RMSE of both the training set and the testing set. A model that is fit well will have comparable RMSE values. But, if the training RMSE is significantly lower than the RMSE of testing set then it indicates a model that is overfit, in which case we can use methods like regularization or increase and decrease the values of regularization parameters accordingly.

4.3.4 Classification

Evaluation for J48 and logistic regression models is done based on accuracy, precision, recall and f-measure. These measures can be computed from truth table.

5 Experimental Results

Cost	**Gamma**	**Kernel**	**Degree**	**Vector Count**	**RSME**
1000	0.0001	radial	N/A	8724	139.6982
10	0.01	polynomial	2	8807	133.0355
10	0.01	polynomial 3	8317	128.6375	
10	0.01	polynomial	4	8345	129.9472
10	0.01	sigmoid	N/A	12963	1427.655

Table 3: SVM Results

5.1 Implementation

Data processing, was implemented using Java, that is predictor variable transformations from nominal to numerical/boolean variables, generation of new variables, discretizing the variables.

We used Java and Weka API for building and evaluating classification models. We used Weka API for building J48 Decision trees, Logistic Regression and Support Vector Machines (SVM).

We implemented the regression models using R such as Multivariate Linear Regression and Support Vector Machines including methods such as regularization and stepwise regression for model selection. MASS and e1071 packages in R were used for the implementation.

6 Conclusion

After comparing results from regression and classification models, we observed few relations between features and nominal class. As discussed earlier windspeed does not affect prediction models, since Washington has average windspeed throughout the year. This assumption was later proved by correlation table. During step-wise regression the model removed winter since it did not have much effect on bike count class. This behavior was observed by classification models as well.

7 Future Work

As a progression to this project, we can add geo-location of each docking station, duration for each trip to our current data. Often bike sharing systems give offers where first 30-60 minutes of the ride is free, using this new data set we will be able to provide better free slots. Using geo-locations of docking stations we can alert bike sharing companies to restock bikes at a station.

References

[1] Uci machine learning repository: Bike sharing. http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#/.