

INDIVIDUAL PROJECT REPORT

Moneka Bommasani

mbommas@g.clemson.edu

Introduction

I have used the Bike Sharing System in Washington, D.C from the beginning of 2011 to the end of 2012. I have performed exploratory analysis and built linear regression models to fit the data and analyze the relationship between the rental pattern with different factors like weather conditions.

Dataset

The dataset is a two-year log of bikes being rented in a bike sharing system in Washington, D.C., USA known as Capital Bike Sharing (CBS). The data is available in two comma separated files (.csv) – ‘day’ and ‘hour’ with the only difference being an extra “hr” attribute in the ‘hour.csv’. We will use the ‘hour.csv’ for our analysis.

There are 17 attributes with 17389 observations. Brief description of the attributes is as follows:

Instant	Index of observations (17,379)
Dteday	Date (Jan 1 st 2011 to 31 st Dec 2012)
Season	Season 1: Spring 2: Summer 3: Fall 4: Winter
Yr	The year 0: 2011 or 1: 2012
Mnth	Month of the year 1: Jan 2: Feb 3: Mar 4: Apr 5: May 6: Jun 7: July 8: Aug 9: Sep 10: Oct 11: Nov 12: Dec
Hr	Hour of the day (24-hour period – 0 to 23)
Holiday	1: Holiday 0: Not a holiday
Weekday	Day of the week 0: Sunday 1: Monday 2: Tuesday 3: Wednesday 4: Thursday 5: Friday 6: Saturday
Workingday	1: Working day 0: Holiday/Weekend
Weathersit	Weather id divided into 4 categories 1: Clear, Few clouds, Partly cloudy 2: Mist, Mist + Cloudy, Mist + Broken Clouds, Mist + Few Clouds 3: Light snow, Light rain + Thunderstorm + Scattered clouds, Light rain + Scattered clouds 4: Heavy rain + Ice pellets + Thunderstorm + Mist, Snow + Fog
Temp	Temperature in Celsius (Normalized: Divided by 41(max))

Atemp	Feels like temperature in Celsius (Normalized: Divided by 50(max))
Hum	Humidity (Normalized: Divided by 100(max))
Windspeed	Wind speed (Normalized: Divided by 67(max))
Casual	Count of Casual users
Registered	Count of Registered users
Cnt	Total count of bikes rented (Casual + Registered)

Online Resource

The UCI Repository: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

The Dataset zip file: <https://archive.ics.uci.edu/ml/machine-learning-databases/00275/>

Citation

Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

```
@article{year={2013},    issn={2192-6352},    journal={Progress in Artificial Intelligence},
doi={10.1007/s13748-013-0040-3}, title={Event labeling combining ensemble detectors and background knowledge}, url={http://dx.doi.org/10.1007/s13748-013-0040-3}, publisher={Springer Berlin Heidelberg},
keywords={Event labeling; Event detection; Ensemble learning; Background knowledge}, author={Fanaee-T, Hadi and Gama, Joao}, pages={1-15} }
```

Exploratory Data Analysis

I have done some descriptive statistics on the data and plotted histograms and QQ plots to check the normality of the variables.

Observations:


- No missing values in the dataset
- All the values in the data are not integers

🚦 Target Variable (cnt, casual, registered):

- Discrete data – cnt, casual, registered
- Maximum number of bikes rented in one day = 977
- Maximum number of bikes rented by registered users in one day = 886
- Maximum number of bikes rented by casual users in one day = 367
- There was at least one bike rented every day in 2011 and 2012
- The distributions were positively skewed as number of bikes cannot be less than zero and the frequency decreases as the count increases
- Normal Distribution - The distributions are skewed to the right so log transformation on the response variable (response variable+1 to take care of zero values) were taken. The log distributions seem closer to normal distribution.

Predictors:

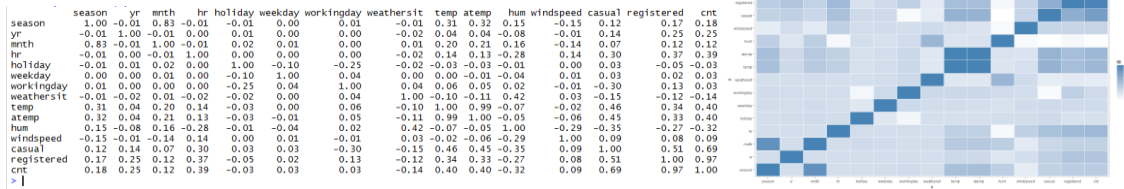
- Range of temp, atemp, hum and windspeed are below 1.00 as the data is normalized.
- Categorical Variables – season, holiday, weekday, workingday, weathersit, month, hr, yr (Computations cannot be done)
- Interval data – temp, atemp, hum, windspeed (Have a scale)
- 4 seasons have equal distribution
- 65.67% of the days were clear or partly cloudy whereas only 0.017% days were heavy rain with thunderstorms (Only 3 occurrences)
- 68.27% of the days are working days
- Number of holidays – 500 in 17379 days
- Temperature is highest in fall, followed by summer, followed by winter and then spring

 **Relationship between target and predictors from the plots** - I have used the boxplots to look at the relationship between the target and predictors. For the numerical variables (temp, atemp, hum and windspeed), I have fitted each a linear model to estimate the relationship.

Observations:

- Every day, there is a peak in the morning for the bike usage at about 8am and another peak in the afternoon at 5-6 pm
- Registered users have same trend as the total count for the hourly analysis
- Casual users have a huge difference in the hourly trend. The usage peaks only in the evenings (around 12pm – 7pm)
- Bike usage is slightly higher during weekdays rather than weekends for registered users mostly to commute
- Casual users have rented the bikes more on weekends when compared to weekdays for leisure purposes
- Bike usage is more from May to October and low from January to March probably because more users rent when it's hot rather than cold
- Bike usage has increased from 2011 to 2012
- People rent bikes more in summer and fall rather than spring and winter
- Bike usage decreases as the weather conditions become worse
- More registered users rent bikes when it's not a holiday which is opposite when compared to casual users who rent more on holidays
- More registered users rent bikes on a working day which is opposite when compared to casual users who rent more on non-working days
- There seems to be a positive relationship between the temp (and the atemp) and the bike rental i.e., with the increase in temp, the bike rental increased. We can say that people rent more bikes when climate is hot as compared to cold
- Influence of temperature for 2011 is more significant than for 2012
- There seems to be a negative relationship between the hum and the bike rental i.e., bikes have been rented more when the humidity is less
- There also seems to be a slightly positive relationship between the windspeed and the bike rentals.

Correlation Matrix – Removed the instant and date variables



Observations:

- High correlation (> 0.5) between:
(Month, Season)
(Temp, atemp)
(registered, casual)
(registered, cnt)
(casual, cnt)

Only one attribute from the pair is selected because of the multi-collinearity phenomenon.

- Considerable correlations between dependent and independent variables:

Dependent variable – count (registered and casual become redundant as they are highly correlated with the dependent variable)

Independent variables

Temp – High positive correlation

Mnt – Positive correlation

Hr – High positive correlation

Yr – Positive correlation

Hum – High negative correlation

Holiday – No proper correlation

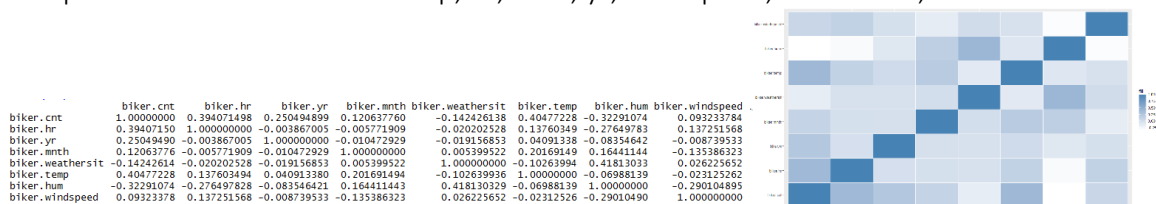
Windspeed – Positive correlation

Weekday – Negligible

Weathersit – Negative correlation

Workingday – No proper correlation

Independent variables now – temp, hr, hum, yr, windspeed, weathersit, mnt



- I did the same with **log(cnt)** being the **dependent variable** now and got temp, hr, hum, yr, windspeed, weathersit, season and workingday as the **independent variables** now.

Model Selection and Validation

Data science Model – Linear Regression

The data is quantitative. It is clear from the data that simple linear regression is not sufficient to predict the bike rental count. So a multiple linear regression model is fitted to analyze the count of the rentals.

- All the predictors are used to fit the model

- The actual bike count will deviate by 141.8 from the true regression line. The observed range of bikes is 1 to 977. A range of 976 being between max and min count. A deviation of 141.8 is a considerable error margin.
- About 38.89 % of variation in the bike rental count is explained by the predictors.

Now, I have taken the **log of dependent variable (cnt)** and fitted the model

- The actual log of bike count will deviate by 1.072 from the true regression line. The observed range is 0 to 6.88. A range of 6.88 being between max and min count. A deviation of 1.072 can be taken as not a considerable error margin.
- About 48.01 % of variation in the log of bike rental count is explained by the predictors.

So, log of the dependent variable definitely gives us a better model.

- Variables are removed using “Backward Selection” method till the p value for all the variables becomes less than or equal to 0.05. The model is fitted with the left variables.
 - The results are same as the above model.
- Independent variables which we have selected in the correlation matrix have been used to fit the model
 - The actual log of bike count will deviate by 1.074 from the true regression line.
 - About 47.77 % of variation in the log of bike rental count is explained by the predictors.

Comparing the models

I have done the anova test on the models to determine the best model. It turns out that the Model2 (Backward Selection model) is the best fit among all the models.

Residuals for the best model:

The residuals give differences between predicted values and the actual values of the count variable. Model is a good fit if they are closer to zero. Here, majority of residuals fall between +2 and -2 standard deviations. There are few which fall above +2 and below -2 standard deviations which suggests that this model is not predicting well in some cases. The QQ plot shows that the residuals follow a very close normal distribution and deviate a little bit at the ends.

Predictions

Now, the prediction formula is applied on the original dataset with the new regression model, and is being plotted. We can see that this model even though is not the best fit; performs well in most of the instances.

Conclusion

We have observed and analyzed the bike sharing system with exploratory data analysis. We then, looked at the relationships between the variables in the data set and eventually built a multiple linear regression model. The model which was developed performed fair enough at some instances, however there were some predictions that were quite different from the actual bike rental count recorded. The model could be improved further by taking interactions between the variables into consideration. Also, collection of more data over several years might help in building a more efficient model.