ANALYSIS OF CUSTOMER BEHAVIOR FOR WASHINGTON'S BIKE- SHARING SYSTEM

-Ashwin Balasubramaniam, Sakshi Sakshi, Shahid Akbar, Ved Vartak

Executive Summary

The objective of this project is to predict the number of rides for a given day. Two sets of models have to be fitted (Casual & Registered) on the Bikes Dataset collected by the Washington DC Transport Department to predict the same.

After analysis of various models, it was determined that for Casual Response the Second-order model with interaction terms that were significant (pertaining to p-values) and after exclusion of insignificant terms was found to be most optimal considering the interpretability and accuracy of the model. For Registered customers as well, the second-order model with interaction terms that were significant and after exclusion of insignificant terms was found to be most optimal.

Some of the key findings include:

- Temperature is the most influential factor whereas humidity and weather are found to have a negative correlation with respect to Casual and Registered Data.
- The correlation between working days and riders gives us an idea that it is greater during the weekends for registered customers whereas insignificant for casual riders.
- Riders are observed to be low during Season1(Spring) and Greatest during Season3(Fall).
- The ridership for casual customers grew by 50.35% and for the registered customers, it grew by 67.91% from the first to the second year.

1. Introduction

The main reason for the collection of bike-sharing data across the city is to better understand the behaviour of customers, which contributes towards the collective total number of rides recorded for a particular day by the city transportation department. It divides the Casual as well as the Registered customers differently pertaining to its characteristics. Factors contributing to customers opting to ride on a particular day have been taken into consideration and using various modelling techniques in the field of Data Analytics predictions for total rides on a single day has been made. Using various Linear regression, Multiple regression, Resampling methods like K-Fold cross-validation, Model Selection and Regularization methods like Best Subset Selection, Ridge and Lasso modelling approaches we are trying to figure out the model which provides us with the best accuracy and interpretability for predicting the number of rides on a single day. Combination of Factors such as R^2 which is the closeness of prediction to the fitted line in the fit model and the mean square error which gives us an idea of the closeness of our predicted values from actual. Using the most optimal combination of both the best possible model is selected for prediction. It was found that for fitting both the casual and registered datasets, the Reduced model with interaction terms proved to be the most optimal model. This was concluded on the basis of the Mean Squared

2. Data

The Transportation Department in Washington, DC conducted a survey to understand the behavior of customers for the bike-sharing system in the city, which resulted in this dataset. This dataset contains the daily count of rental bikes between the years 2011 and 2012 in the Capital rideshare system with the corresponding weather and seasonal information. The data consists of the following fields:

- instant: records index.
- dteday: date.
- season: denotes season through the following attributes:
    - 1: Spring
    - 2: Summer
    - 3: Fall
    - 4:Winter
- yr: year denoted by the following numbers:
    - 0: 2011
    - 1: 2012
- workingday: Denotes the working days of the riders through the following numbers:
    - 0: Holidays or Weekends.
    - 1: Other days.
- weathersit: Describes the weather situation through the following numbers:
    - 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: Normalized temperature in Celcius. The values are derived by:
  $(t - tmin)/(tmax - tmin)$, where, 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: Daily count of casual users.
- registered: Daily count of registered users.

We have to fit models, using the casual and register count as responses and rest of the variables as the predictors.

Preprocessing: Initial changes have been made on the dataset to ensure that the categorical variables didn't get processed as quantitative variables. The function as.factor() has been used to convert the variables: season, yr, workingday, and weathersit to categorical variables.
The columns instant and dteday have also been deleted for both casual and registered fits as the variables do not significantly contribute to the fits and create unwanted errors due to their formats (dteday).
Separate datasets have been created for both the fit models where casual doesn't have a registered column and vice versa.

3. Method

For fitting the models using casual and registered as the two responses, we employed various techniques, predominantly using the multiple regression models. The results from the relevant models have been tabulated in Table. 1 in the form of MSE values and R^2 for the number of casual customers. The given bar chart (Fig. 1) depicts the variation of MSEs across different models. The MSEs have also been calculated using the k- fold cross-validation method with 5 and 10 folds respectively.

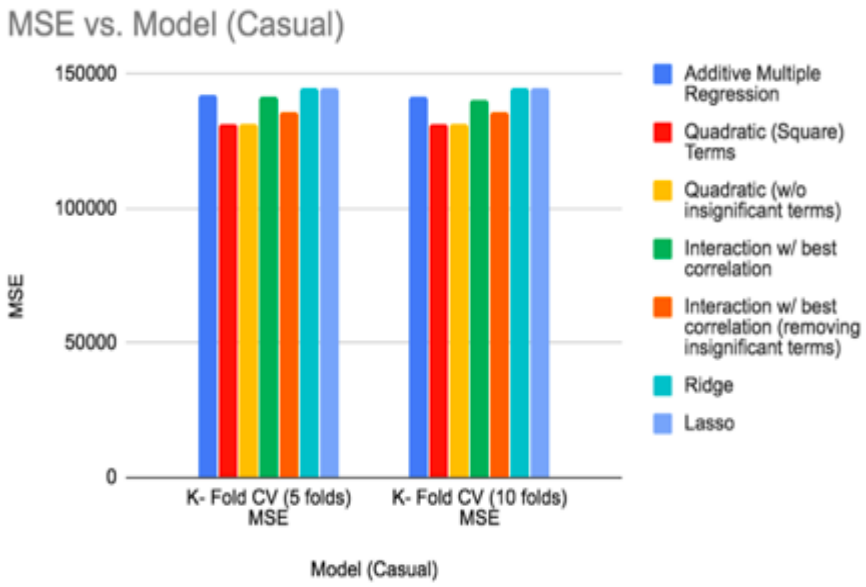

Fig. 1. Bar Chart depicting data in Table. 1

TABLE. 1. MSE and R^2 values for fitting models on Casual dataset				
Model (Casual)	**MSE**	**R^2**	**K-Fold MSE (5 Folds)**	**K-Fold MSE (10 Folds)**
Additive Multiple Regression	138413.131	0.702	142380.38	141870.83
Second Order Quadratic	127295.4169	0.7296	131627.24	131532.75
Second order Quadratic without insignificant terms	127571.1734	0.729	131729.05	131660.87
Interaction with best Correlation	121148.961	0.7426	141755.1	140474.66
Reduced Interaction with Best Correlation	121203.0531	0.7416	136146.55	135889.45
Ridge Regression	144561.1	-	144561.1	144561.1
Lasso Regression	144623.3	-	144623.3	144623.3
PCR	-	0.6494369	-	-

The first step was to fit an additive multiple regression model. The results of this fit have been given in Table. 1. The diagnostic plots for this model indicated that including quadratic terms could increase the fit for the given dataset.

The second step was to fit the data using square terms of the quantitative variables like temp, hum and windspeed. In this model, the quadratic variable I(windspeed^2) and windspeed were statistically insignificant as they had the p- values of 0.15392 and 0.67385 respectively. We deleted just the square term as the variable windspeed was statistically significant in the previous model.

The next step was including the interaction terms along with the reduced quadratic model. The variables that showed a somewhat high correlation were selected as interaction terms. The results of this fit have been depicted in Table. 1. The resulting fit model included several statistically insignificant terms. These terms were excluded to form a more interpretable model. This would be the Reduced Fit Model with Interaction terms. The results from this model had lower MSE (121203.0531) than most of the previous models and the fit explained 74.16% of the variance.

Some other methods like Lasso, Ridge Regression and Principal Component Regression were also performed on the dataset, the results of which have been included in Table. 1. These methods did not give satisfactory results (lesser MSE and higher R^2), hence they weren't considered for the model fit.

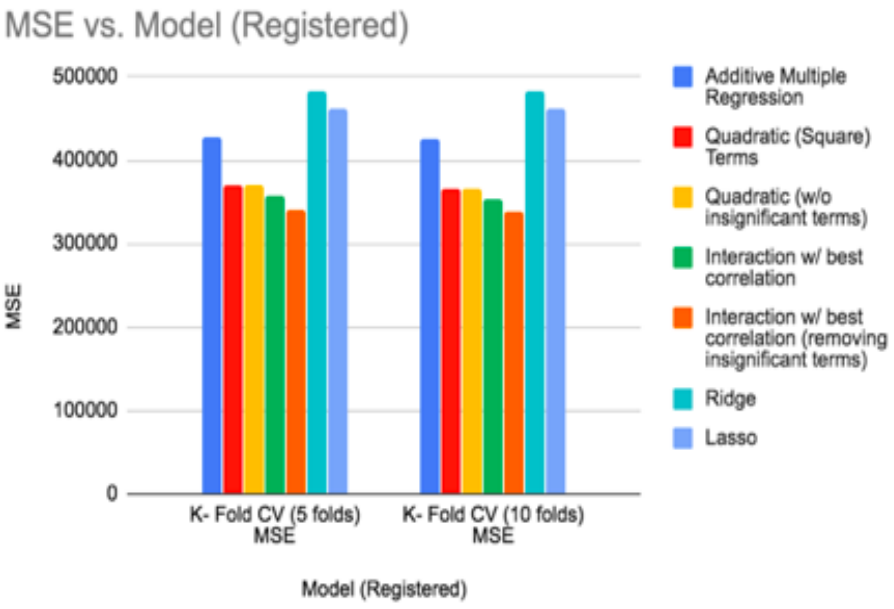

Fig. 2. Bar Chart depicting data in Table. 2

TABLE. 2. MSE and R^2 values for fitting models on Registered dataset				
Model (Registered)	MSE	R2 (Registered)	K- Fold CV (5 folds) MSE	K- Fold CV (10 folds) MSE
Additive Multiple Regression	410470.8467	0.8312	427331	425153.8
Quadratic (Square) Terms	347003.0088	0.8573	370492.7	365051.6
Quadratic (w/o insignificant terms)	347988.9841	0.8569	369940.9	365051.6
Interaction w/ best correlation	311022.2287	0.8721	357271.49	352308.03
Interaction w/ best correlation (removing insignificant terms)	312595.4724	0.8714	341352.58	337883.4
Ridge	483675.8	-	483740.3	483740.3
Lasso	452941.7	-	462523.5	462523.5
Principal Components Regression	-	0.522148	-	-

A similar procedure was followed for the registered count of customers as well. The results of these methods have been tabulated in Table. 2 and depicted by a bar chart in Fig. 2. Following this procedure, the selection of the most optimal model was carried out, by observing the MSE and R^2 values.

4. Result:

For the Casual response, the second-degree Quadratic equation with correlation terms was more appropriate for the given dataset. The conclusion was made based on the best R^2 value and lowest MSE value as compared to the other models which were fit. The k-fold cross-validation (both 5 and 10-fold) was performed to measure the training error and it was the least amongst the other models.
Given below is the summary for the selected fit model for Casual and Registered Customers respectively:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -129.53     285.30  -0.454  0.64996
season2         477.19     257.05   1.856  0.06381 .
season3        1081.43     528.34   2.047  0.04104 *
season4         -67.86     223.21  -0.304  0.76120
yr1             262.21      27.02   9.705  < 2e-16 ***
workingday1    -815.89      28.44 -28.686  < 2e-16 ***
weathersit2     270.75     227.04   1.192  0.23347
weathersit3    -171.72     410.05  -0.419  0.67550
temp           4859.17     699.41   6.948 8.42e-12 ***
hum             290.21     901.30   0.322  0.74756
windspeed     -1036.58     191.73  -5.407 8.79e-08 ***
I(temp^2)     -4034.07     988.34  -4.082 4.98e-05 ***
I(hum^2)       -206.04     802.95  -0.257  0.79756
season2:temp    872.62     518.37   1.683  0.09274 .
season3:temp    -99.85     848.90  -0.118  0.90640
season4:temp   1787.79     437.96   4.082 4.97e-05 ***
season2:hum    -799.29     253.89  -3.148  0.00171 **
season3:hum    -821.92     305.82  -2.688  0.00737 **
season4:hum    -818.55     304.17  -2.691  0.00729 **
weathersit2:hum  -542.97     343.54  -1.581  0.11443
weathersit3:hum  -178.97     456.57  -0.392  0.69518
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 353.3 on 710 degrees of freedom
Multiple R-squared:  0.7426,    Adjusted R-squared:  0.7353
F-statistic: 102.4 on 20 and 710 DF,  p-value: < 2.2e-16
```

Fig. 3. Summary for Best Fit (Casual)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1993.13     564.74  -3.529 0.000444 ***
season2         360.08     408.32   0.882 0.378154
season3        5259.76     845.36   6.222 8.38e-10 ***
season4        1189.00     360.74   3.296 0.001029 **
yr1            1683.44      43.29  38.886  < 2e-16 ***
workingday1     949.31      45.59  20.823  < 2e-16 ***
weathersit2    -287.35      57.74  -4.976 8.14e-07 ***
weathersit3   -1159.34     172.09  -6.737 3.34e-11 ***
temp           8130.24    1119.39   7.263 9.96e-13 ***
hum            5002.95    1415.75   3.534 0.000436 ***
windspeed       920.82    1252.98   0.735 0.462639
I(temp^2)     -4959.77    1584.88  -3.129 0.001823 **
I(hum^2)      -4266.07     991.96  -4.301 1.94e-05 ***
season2:temp    473.62     831.94   0.569 0.569334
season3:temp  -5732.27    1361.97  -4.209 2.90e-05 ***
season4:temp   1505.09     702.22   2.143 0.032425 *
season2:hum      93.22     400.16   0.233 0.815871
season3:hum    -718.97     486.05  -1.479 0.139523
season4:hum    -938.50     488.75  -1.920 0.055231 .
hum:windspeed -4805.68    1965.31  -2.445 0.014716 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 566.9 on 711 degrees of freedom
Multiple R-squared:  0.8714,    Adjusted R-squared:  0.868
F-statistic: 253.6 on 19 and 711 DF,  p-value: < 2.2e-16
```

Fig. 4. Summary for Best Fit (Registered)

Among the three quantitative predictors, temperature, humidity and weather, temperature is found to be the most influential predictor. Humidity and weather were found to be in a negative correlation with both casual and registered riders i.e., the number of riders as a whole increased as the humidity and weather conditions decreased (Weather is given in a scale of 4).

The correlation between the quantitative predictors and the response is given below:

TABLE. 4. Correlation between Quantitative predictors and Responses				
Predictors	Casual		Registered	
	Year1	Year2	Year1	Year2
Temperature	0.58	0.54	0.69	0.60
Humidity	-0.03	-0.07	0.01	-0.06
Windspeed	-0.19	-0.06	0.01	-0.07

1)Season

Once we dive into the correlation between the different seasons, it can be concluded that the number of riders is low during the spring season i.e., S1 and higher during fall season i.e., S3.

TABLE. 5. Number of Riders per season and %age distribution				
Season	Casual		Registered	
	Riders	% in a year	Correlation	% in a year
1 - Spring	60622	9.7	410726	15.6
2 - Summer	203522	32.82	715027	26.75
3 - Fall	226091	-0.06	835038	32.24
4 - Winter	129782	20.93	711831	26.63

2) Day of the week

If we looked into the correlation between working days and the riders it can be seen that the percentage of riders is higher for registered riders during weekends whereas it had an insignificant influence on the casual riders.

TABLE. 6. Distribution of Weekday and Weekend Riders				
Predictors	Casual		Registered	
	Riders	%	Riders	%
Weekend	316732	51.08	683537	25.57
Weekday	303285	48.92	1989125	74.43

TABLE. 7. Percentage increase from year 2011 to 2012		
Predictors	Casual	Registered
Year1	247252	995851
Year2	372326	1674521
% increase	50.35	67.99

The percentage increase from the first year to the second year in ridership has been indicated in Table. 7 for both the responses.

5. Conclusion:

In conclusion, based on the dataset given it can be summarised that there are various factors both categorical and continuous features which influence the number of riders in a day. Temperature and seasons played an important role in influencing the number of riders on a single day in the city of Washington, DC.

Weekdays were found to be preferred by almost 75% of the registered riders. Therefore, it would be preferable to increase the number of bikes during weekdays for the registered riders.

Based on the predictors, different models were trained and tested for the casual and registered responses. A model of 74% R^2 in response and a model of 87.14% R^2 in response was fitted for casual and registered response respectively. The model was then validated by a 5-fold cross validation method and the resulting mean square errors were noted. The statistical significance of each predictor and its influence on the response were tabulated which would help in prioritizing the features and therefore increasing the number of riders or planning the number of rides accordingly.