Predictive Modelling of Bike share dataset from UCI ML repository

MIDTERM PROJECT

ST-516

Team members:
Rehan Sheikh
Chaitanya Rajeev
Vishakha Patil
Rucha Girgaonkar

Contents

1. Executive Summary

- The objective of this study is to predict the number of rides on a given day for bike sharing system of transportation department of Washington, DC.
- Among the various models fitted for the data, we found that LASSO fitted on second order data is the best fit for prediction accuracy.
- From the analysis performed, we observed that the number of **casual riders** has **increased** by 50.76 % from 2011 to 2012 whereas the number of **registered riders** has **increased** 68.37% from 2011 to 2012.
- For both casual and registered users, there is a **positive relationship** between **temperature** and the **number of rides**.
- For both casual and registered riders, the number of rides is **highest** when the weather is **clear**, moderate when it's cloudy/misty and **lowest** during light **snow/rain**.
- The number of casual riders is more during weekend/holiday than it is on a working day. And, the number of registered riders is more during weekdays than on weekends /holidays. So, we can say that during weekends, casual riders and during weekdays, registered riders generally ride the rental bikes.

2. Introduction

Bike-sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return has become automatic. Opposed to other transport services such as bus or subway, the duration of travel, departure, and arrival position is explicitly recorded in these systems. We want to better understand the behaviour of customers for the bike-sharing system in the city.

Following is the objective of the study

1. Fit two (possibly sets of) models, one for the registered customers and another for the casual, non-registered riders to predict the number of rides on a given day
2. Prioritize the ability to accurately predict the number of rides for a given day, explicitly state what factors influence ridership and by how much.
3. Compare the number of ridership between 1st and 2nd year
4. Compare and differentiate the behaviour of registered and casual riders

3. Data

The dataset contains the daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. The following attribute information is given:

- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
  - 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 Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- 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

Exploratory Data Analysis for behaviour of casual and registered bikers with respect to different conditions:

- For both casual and registered riders, there is a **positive relationship** between **temperature** and the **number of riders**, i.e. as temperature increases, number of casual and registered bikers increase proportionally.
- For both casual and registered riders, the number of riders is **highest** when the weather is **clear**, **lower** when it is cloudy/misty and **lowest** during light **snow/rain**.
- The number of casual riders are **more during weekend/holiday** than it is on a working day. This tells us that casual riders may use the bikes for leisure purposes. On the other hand, the number of registered riders are more during weekdays than on weekends/holidays. This tells us that registered riders may use the bikes to commute to and from work more so than leisure.
- For both casual and registered riders, there seems to be **no relationship** between **humidity** and number of respective **riders**.
- The number of **casual riders increased** from 2011 to 2012. The number of **registered riders** also **increased** from 2011 to 2012 but at a higher rate.

4. Methods

4.1. Simple Additive Linear Model (fit1)

Casual	**Registered**
$R^2 = 0.715$	$R^2 = 0.846$
Diagnostics: • Both models showed Non-linear relationships from residuals plot. • Q-Q plot showed follows the normal line satisfactorily.	
104 fold cross validation used	104 fold cross validation used
Train MSE = 130071 | Test MSE = 134669	Train MSE =370742 | Test MSE = 384482
• We don't have errors from another model to make a judgement. • But since the residual plot shows non-linear relationship we move to fit a 2nd order model.	

4.2. Complete Second-Order Model (fit2)

Casual	**Registered**
$R^2 = 0.846$	$R^2 = 0.907$
Diagnostics: • Residual plots flattened significantly in both models as compared to fit1. • Q-Q plot showed follows the normal line satisfactorily with some long-tailed characteristics.	
Train MSE = 71433 | Test MSE = 86826	Train MSE = 222483 | Test MSE = 284702
• All errors benefit from a significant reduction from fit1 • Predictors are centered before modelling.	
• fit2 is much less biased than fit1 since test MSE for both responses has decreased. • Since we have 67 predictors, we will fit a LASSO and Ridge model over the 2nd order data.	

4.3. LASSO on 2nd Order Data (fit3)

Casual	**Registered**
$R^2 = 0.831$	$R^2 = 0.905$
Train MSE = 78162 | Test MSE = 82224	Train MSE = 230161 | Test MSE =273102
• Diagnostic plots inference remains the same as fit2	
• The LASSO model performs better than fit2 and there is further reduction in errors. • LASSO predictor count for casual:50 ; registered:54	

4.4. Ridge Regression on 2nd Order Data (fit4)

Casual	**Registered**
R^2 = 0.844	R^2 = 0.911
Train MSE = 72208 | Test MSE = 83856	Train MSE = 215352 | Test MSE =277131
• Diagnostic plots inference remains the same as fit2	
QQ Plot shows long-tail distribution.	QQ Plot shows long-tail distribution.
• The Ridge model gives similar errors to the LASSO model(fit3). • LASSO predictor count for casual:50 ; registered:54	

4.5. Best Subset selection on original data (fit5)

Casual	**Registered**
R^2 = 0.63	R^2 = 0.758
Train MSE = 139108 | Test MSE = 146016	Train MSE = 411938 | Test MSE = 428371
• Not a good model because of high errors. Mainly bad due to absence of 2nd order predictors.	

4.6. Log transform on 2nd Order data (fit6)

Casual	**Registered**
R^2 = 0.903	R^2 = 0.893
Train MSE = 70222 | Test MSE = 85858	Train MSE = 230718 | Test MSE = 301892
• This model was made because the diagnostics in fit2, fit3 and fit4 showed heteroscedasticity in the data. • The errors in the model are comparable to fit2, fit3 and fit4.	

4.7. Principal Components Regression on 2nd Order data (fit7)

Casual	**Registered**
R^2 = 0.846	R^2 = 0.907
Train MSE = 71433 | Test MSE = 86826	Train MSE = 222483 | Test MSE =284702
• This model was made to explore PCR and compare the errors. • The errors in the model are comparable to fit2, fit3 and fit4. But model isn't very interpretable.	

5. Results

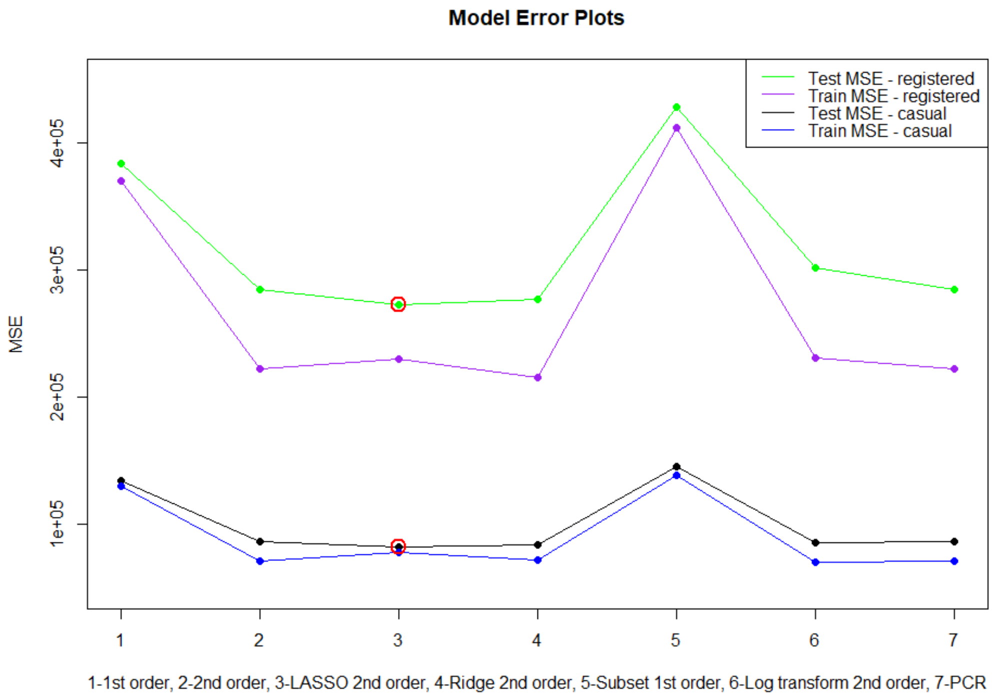
Model No.	**Description (seed=1000)**	**Errors - Casual Riders Prediction**		**Errors - Registered Riders Prediction**	
		Training MSE	**Test/CV MSE**	**Training MSE**	**Test/CV MSE**
fit1	Simple Additive Linear Model	130071	134669	370742	384482
fit2	Complete Second Order Model	71433	86826	222483	284702
fit3	LASSO on 2nd Order Model	78162	82224	230161	273102
fit4	Ridge Regression on 2nd Order Model	72208	83856	215352	277131
fit5	Best Subset selection on Linear Model	139108	146016	411938	428371
fit6	Log Transform on 2nd order model	70222	85858	230718	301892
fit7	Principal Components Regression on 2nd order data	71433	86826	222483	284702

- The LASSO on 2nd order data model is chosen for both casual and registered riders responses as it has the least CV error of all the models developed.
- The temp predictor has a high coefficient value and p-value < 0.05. This tells us that there is a strong correlation between temp and casual/registered riders
- The dteday and weathersit3 predictors are also significant predictors. This tells us that ridership increased steadily as days passed and rain/snowy weather has a significant negative impact on ridership in both casual and registered riders
- Also, there was increase in ridership during the summer season as compared to the other seasons.
- Test Error Casual(fit3) = 82224 ; Test Error Registered(fit3) = 273102
- The disadvantage with this model is that the interpretability is low because it has 50 and 54 predictors for casual and registered riders respectively.

6. Conclusion

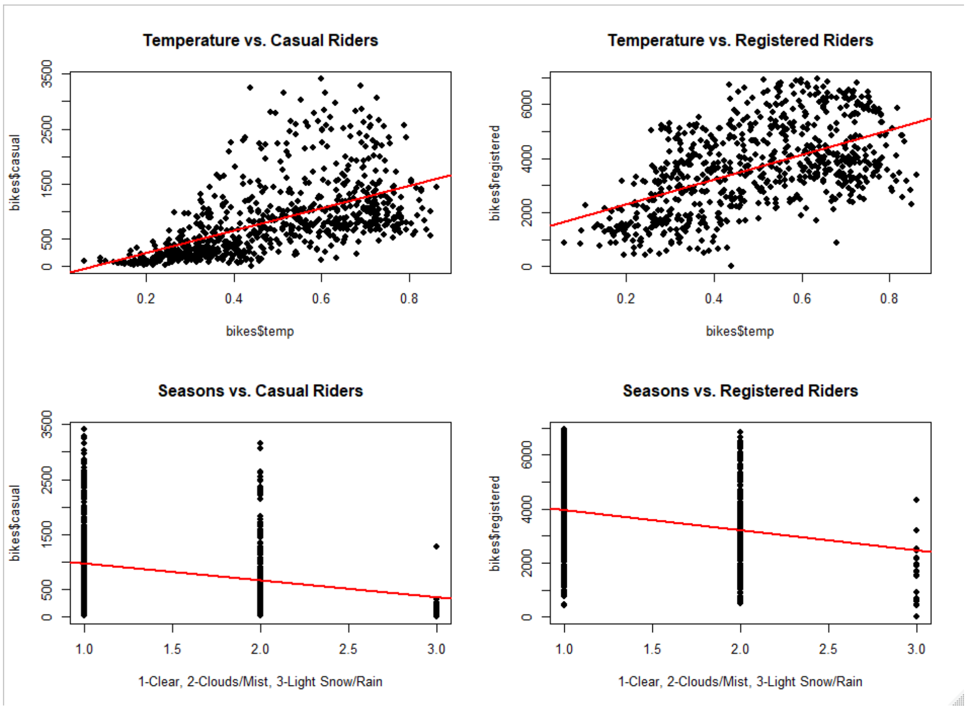
We initially started with a simple additive linear fit. But it had non-linear residual diagnostics and high test error. Hence, we decided to add non-linear predictors. Therefore, we developed 2nd order models along with LASSO and Ridge regression implementations. As expected, all the 2nd order models achieved a high reduction in test errors as compared to the first order model. We tried log transform on 2nd order model to remove heteroscedasticity that we observed with the other models. Analysts who are interested in having errors with constant variance can use this model. We also developed a Principal Components Regression model over the 2nd order data to compare errors. The errors turned out to be similar to the other 2nd order models but PCR models have low interpretability.

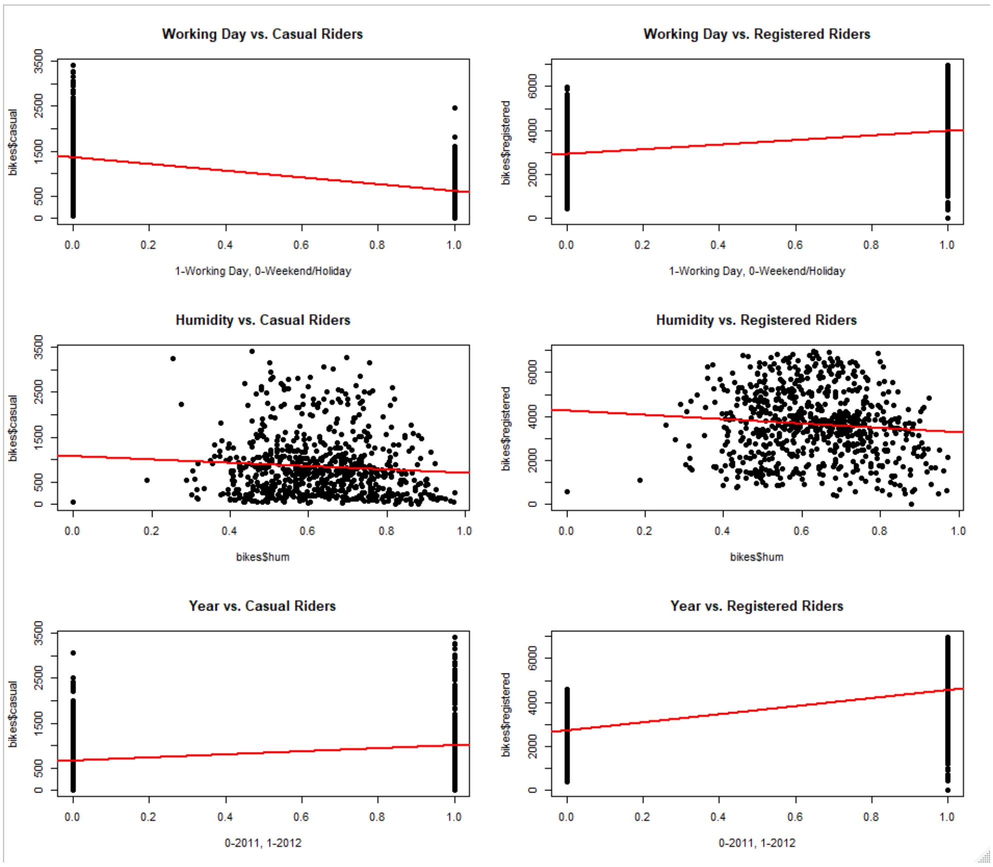
Finally, the LASSO model was selected because it had the lowest test errors for casual and registered riders. Also, it had the lowest no. of non-zero predictors among the other 2nd order models. This is due to the variable selection ability of the LASSO procedure. Furthermore, since the emphasis was on accuracy of prediction, this model is a good fit even though it is not as interpretable as the first order models. In future, we can use other algorithms such as regression trees, random forests, bagging and boosting and compare CV/test errors as well.



7. Appendix

- Residual plots, outlier plots etc. are implemented in the code file. The plots were excluded in the report.
- EDA graphs: All trends mentioned in the executive summary, Exploratory Data Analysis sections inferred from the following graphs and other calculations.


Temperature vs. Casual Riders
bikes$casual
bikes$temp
Temperature vs. Registered Riders
bikes$registered
bikes$temp
Seasons vs. Casual Riders
bikes$casual
1-Clear, 2-Clouds/Mist, 3-Light Snow/Rain
Seasons vs. Registered Riders
bikes$registered
1-Clear, 2-Clouds/Mist, 3-Light Snow/Rain


Working Day vs. Casual Riders
bikes$casual
1-Working Day, 0-Weekend/Holiday
Working Day vs. Registered Riders
bikes$registered
1-Working Day, 0-Weekend/Holiday
Humidity vs. Casual Riders
bikes$casual
bikes$hum
Humidity vs. Registered Riders
bikes$registered
bikes$hum
Year vs. Casual Riders
bikes$casual
Year vs. Registered Riders
bikes$registered