PREDICTION ANALYSIS OF BIKE RIDERSHIP

NC STATE UNIVERSITY

By

Jimit Shah (200318691)

Sai Shashank (200320973)

Shounak Deo (200321421)

Sushrut Ghotankar (200266022)

Yuvaraj Vivekanandan (200252347)

Supervised by

Dr. Dan Harris

Executive Summary

In this project, we predict the number of rides for registered and non-registered, casual riders on a given day for the city Washington, DC. The data has been collected for couple of years and the prediction is made out the of dataset. The prediction model used in this project is **k-folds cross validation(5-folds)** because of its high interpretability and moderately high stability in prediction.

Key Findings:

1. The ridership has found to have increased over the year by ~65 % .
2. The ridership was found to be higher on sunny and pleasant days(high temperature) as compared to other days irrespective of the season.
3. Registered riders are less affected by external factors like season, working day as compared with the casual riders.

Introduction

The goal of analyzing the bike sharing dataset is to understand the behavior customers and find the significant factors which drive the decision of the customers to ride the bike or not. Application of this analysis will help us to predict the number of rides on any given day based on the significant predictors.

The predictors in this dataset include following categorical predictors, Year, Season, Working/Non-Working Day, Weather Situation, and following numerical predictors Temperature, Humidity, and Windspeed.

We will fit various regression models such as Linear, K-fold, Ridge, Lasso, PCR, and Random Forest, on both Casual and Registered riders using the above-mentioned predictors to successfully predict the number of riders based on best MSE estimate and adjusted R^2. The final model or the set of models will be chosen, based on those estimates.

Data

The dataset "Bike-Sharing-Dataset" was obtained by the UCI Machine Learning Repository. This is a collection of databases, domain theories and data generators which are used by the machine learning community for empirical analyses. This dataset contains the daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. Capital bikeshare has over 350 stations in Washington, D. C. Bike sharing systems are a new way of traditional bike rentals. We have two sets of riders – Registered riders (one's who have an active membership and registration) and Casual riders (one's who rent it if they have a need and are not registered).

The dataset can be found in the link - https://www.capitalbikeshare.com/system-data.

Since the given dataset is cleaned in our case, we do not require any data massaging. In order to get the initial view on the data, we split the data into two tables – One with Registered drivers and one with casual drivers.

The dataset consists of mixed data types (binary,categorical and numeric). The categorical variables include 'weathersit', 'season', 'yr' and 'workingaday' as a binary. The predictors 'temp', 'hum' and

'windspeed' are a class of continuous numeric data types. The response variables 'casual' and 'registered' are also continuous numeric class variables which are the number of casual and registered riders for a given day.

From the correlation matrices obtained for casual and registered riders with continuous predictors (temperature, windspeed and humidity), the variables are not significantly correlated. Multicollinearity does not affect the variables.

The analysis was determined by a Pearson Correlation test on the continuous predictors Temperature, Humidity and Windspeed. When we plotted the correlation matrix for the above dataset, the following results were obtained.

	temp	hum	windspeed	registered
temp	1.0000000	0.1269629	-0.1579441	0.5400120
hum	0.1269629	1.0000000	-0.2484891	-0.0910886
windspeed	-0.1579441	-0.2484891	1.0000000	-0.2174490
registered	0.5400120	-0.0910886	-0.2174490	1.0000000

	temp	hum	windspeed	casual
temp	1.0000000	0.12696294	-0.1579441	0.54328466
hum	0.1269629	1.00000000	-0.2484891	-0.07700788
windspeed	-0.1579441	-0.24848910	1.0000000	-0.16761335
casual	0.5432847	-0.07700788	-0.1676133	1.00000000

Methods

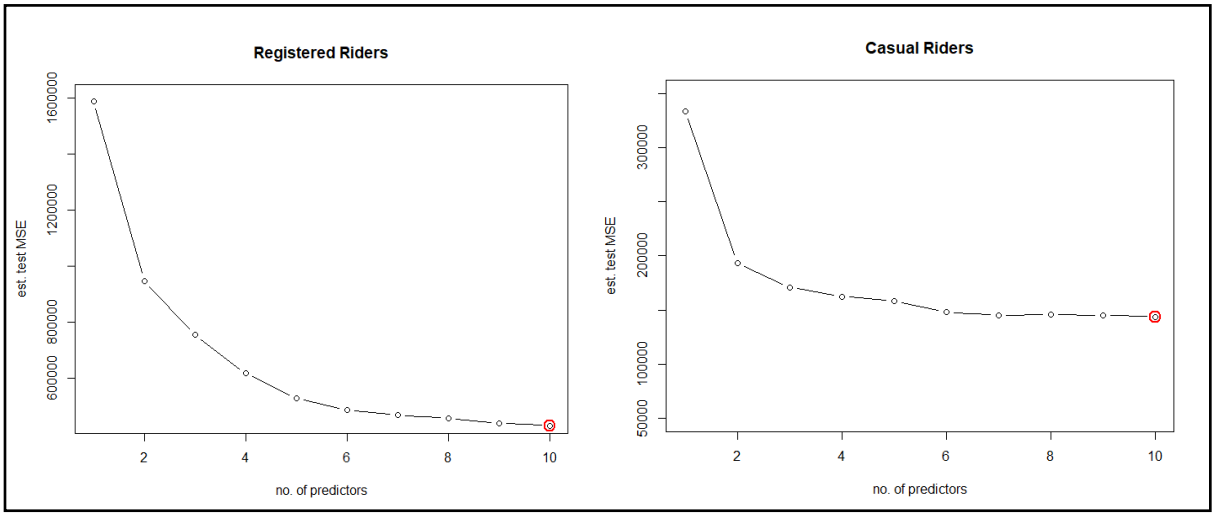
Linear-Regression Model

In this method of analysis, we have bifurcated the data in 2 sets - registered and casual. We fit a linear model with 10 predictors (including the categorical variables). We observed that for registered ridership, all the 10 predictor variables were significant but for casual ridership, except season 3 – Fall, all other predictor variables were statistically significant.

Since our dataset was normalized on its own, we didn't add any quadratic terms to the model. We generated interaction terms and fit a complete second order linear model with 13 predictors (including the categorical variables) for casual riders and registered riders and then the observation shows that there is negligible difference in the adjusted R^2 as compared with the linear model.

k-folds cross validation(5-folds)

In this model, we took 5-folds cross validation technique to perform the analysis. Extending the linear model, in k-folds, we split the dataset in the 5 set of folds and performed the linear regression technique on each model. Post that, we take the average value of MSE and R^2 values we obtained from the model.



From the above plot, for registered riders, the estimated mean square error plot starts to reduce as we increase the number of predictors. Starting from 6, the average decrease in MSE is less.

Similarly, in case of casual riders, the estimated mean square error plot starts to reduce as we increase the number of predictors. Starting from 6, the average decrease in MSE is meagre.

As we observed from correlation matrix, the temperature plays an important role in all the 10 models for both casual and registered ridership.

From the summary of cross-validation model, the below one's are not significant for 6-predictor model

Registered – Season 2, Season 3, humidity and windspeed

Casual – Season 3, Season 4, Weathersit 2 and Weathersit 3

Ridge and Lasso Regression

In Ridge regression, we add the constant term lambda with the root mean square value and try to optimize the resultant equation for all predictor variables.

In Lasso regression, we add the constant term lambda with the absolute values of predictor coefficients in order to get the resultant. The method optimizes the resultant non-linear equation.

In ridge and lasso technique, we performed the analysis for different set of combinations like having only continuous variables excluding categorical variable, including the categorical variable, having squared term and build models for each combination. We observed that, with the presence of all predictor variables, the model performed the best.

PCR and Random Forest Analysis

PCA reduces the dimensionality of data containing a large set of variables. This is achieved by transforming the initial variables into a new small set of variables without losing the most important information in the original data set. These new variables correspond to a linear combination of the originals and are called principal components. In our dataset, we have 10 different dimensions.

When we applied the principle component analysis to our dataset – Casual and Registered, we observed that the performance of the model improved slightly as compared with the linear model.

Random Forest (RF) is one of the many machine learning algorithms used for supervised learning, this means for learning from labelled data and making predictions based on the learned patterns. Random Forest can be used for both classification and regression. Random Forest makes predictions by combining the results from many individual decision trees.

In our model, we added only the continuous predictor variables for the random forest technique. This led to the drop-in variability value as compared with the rest of the models.

Results:

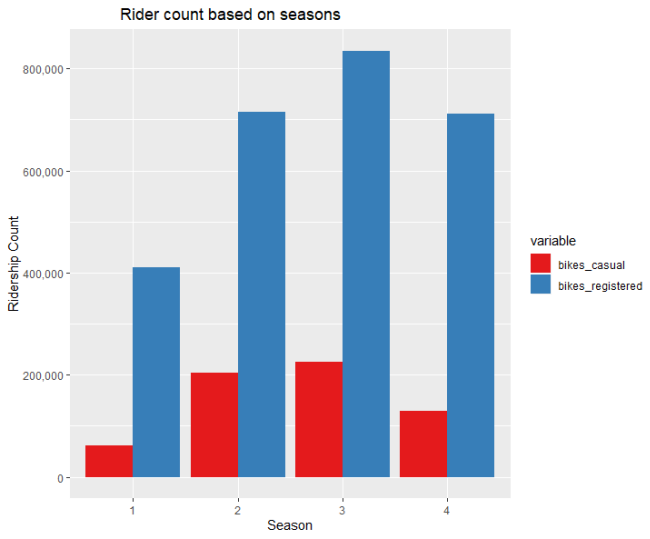
Model	Registered Riders		Casual Riders	
	Adjusted R^2	**MSE**	**Adjusted R^2**	**MSE**
Linear	0.8288	410470.8	0.7019	138413.1
K-Folds(5-folds)	0.8296	4306479	0.6858	143825
Ridge	0.8311	425984	0.7	142797
Lasso	0.8311	426012	0.706	142880
PCR	0.8311	410470.8	0.706	138413.1
Random Forest (only Continuous Variables)	0.42	142351	0.3393	311043

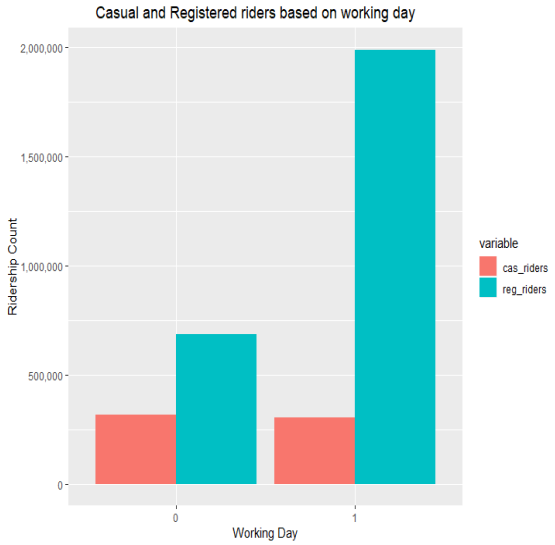
From the above table, we observe that the Ridge regression model gives the best prediction model followed by k-folds cross validation technique. If accuracy is our only concern, then we can opt with the ridge regression model for our dataset. In order to have better interpretability, we are going for the second most stable model – k-folds cross validation method.

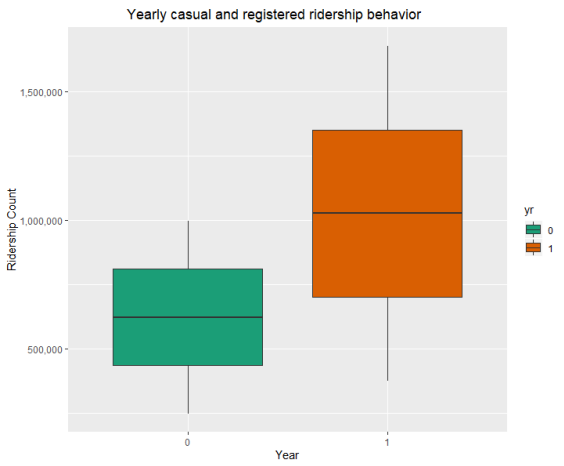
In k-folds technique, the prediction error is less, and the model is clear on performance. As we observed in the previous plot, we can consider 6 predictor variables instead of 10 in order to have easier understandability.

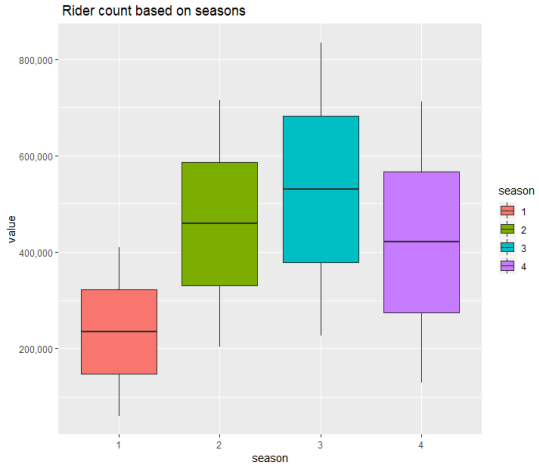
The following analysis is performed for ridership behavior based on different Seasons, Working day and Years.

1. In Seasons, fall season is seen as the best for riders. The maximum number of riders are present in fall season, both casual and registered, as compared with the rest. Hence there is a season variability in demand in our dataset.
2. In Working day, registered riders tend to drive more on working day as compared with holiday. But in case of casual riders, they are least affected with the presence of working day or not.
3. In Year, clearly there is an increase in ridership from year 0 to year 1. As the ridership program has become popular, many people have started using the bike rides in the later increase. This trend suggests that there can be an increase in ridership for the upcoming year.









Conclusion

The project provides a good insight on what factors affect ridership given the set of predictor variables. In our dataset, we observed that temperature played an important role in prediction of ridership. This makes sense as the riders are expected to drive on a pleasant day than a snowy day.

All variables provided in the dataset include external and environmental factors. Availability of specific data can vastly impact the scope of this project. For the company to expand its operations, data relating to population demographics (age, sex, occupation etc.) can be used to target specific audiences. Additionally, location-wise data can provide an insight into where each of the groups are most likely to be located. Using this to maximize ridership bikes can be strategically located at these spots. As ridership is influenced primarily by temperature and nice weather conditions, the company could plan to build temporary rental stations during the peak demand season in order to maximize the gain from this condition.