ST 516 Midterm Project Report

Group members:
Rezvan Mahdavi Hezaveh, Sajjad Taghiyeh, Mojtaba Sardar Mehni,
Hossein Tohidi and Kimia Vahdat

Executive Summary

In this project, we are attempting to find a model, which can best describe the relationship between number of riders (registered or casual) and other predictors such as weather characteristics and other features. After trying 6 different models for both registered and casual riders, we established that for both registered and casual riders, both PCA and GBM perform well (in terms of predictability) with mean squared error of 1293 and 2.08 for PCA models and 499 and 1.31 for GBM models, respectively.

Using variable importance graph of GBM, we found out that, for registered riders the most important predictors are year, working day and temperature; this means that from 2011 to 2012 the number of registered rides has significantly increased, which shows the growth of this business, in addition, registered riders use their bike more often during the week as a transportation mean so it makes sense that there is a significant relationship between working day and registered rides. Also, in warmer days there are more rides than colder days, which again makes sense because riding a bike in cold days is quite difficult. For casual riders, we have temperature and working days as the most important predictors; this means on sunny days or warmer days we have more casual riders also, during the weekdays the number of casual riders is lower than the weekends, which can be caused by the different kind of demand for riding a bike for fun as opposed to riding it as transportation means, which probably registered riders use it for.

If the priority of the management was a model with higher interpretability rather than predictability, we would choose the linear regression with Box-Cox transformed response variables. This model has a reasonable R squared (0.86 for Casual and 0.88 for Registered) and MSE (2.12 for Causal and 1312 for Registered) for both models (See figures S3 and S4 in the appendices for more detail).

Introduction

Bike sharing systems are a new generation of bike rentals in which the whole process from membership, rental and returning is automatic. In this project, we aim to better understand the behaviour of customers in bike sharing system in Washington DC using collected data for daily number of rides for both registered and non-registered users over a two years period. For each data point, timing of the rides and weather conditions are also collected since they may impact the customers' decisions to ride or not. We consider two types of models corresponding to registered and casual (non-registered) customers. First, the data was visualized and correlations were calculated and diagnostic plots were used to evaluate the normality and constant variance assumption. Box-Cox transformation was used for both models to satisfy the normality and constant variance assumption. Moreover, Cook distance is used to find the outliers and omitting them. As for finding the best prediction model, performance of multi-linear regression, Ridge, Lasso, PCA, Random Forest and GBM (6 models in total) are compared and 5-fold cross validation MSE was used as basis of our comparison. The results show that GBM and PCA show the best performance between all 6 models.

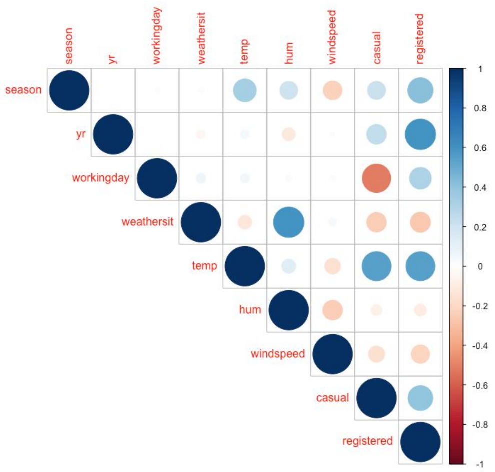
Data

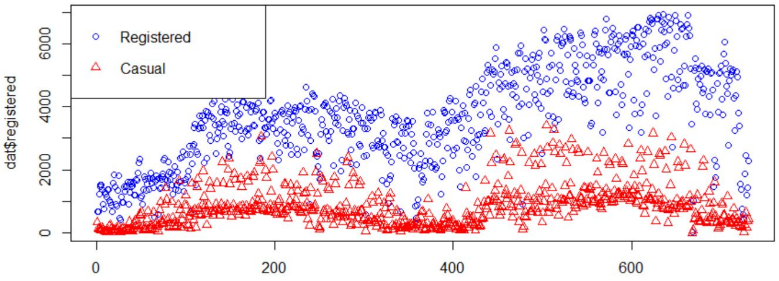
The given dataset contains the daily count of rental bikes between the year 2011 and 2012 in capital bikeshare system in the Washington DC. There are 7 independent variables, two dependent variables and 731 observations. Among the independent variables, season, yr, working-day and weathersit are considered as categorical variables; while temp, hum and windspeed are numerical values between 0 and 1. The dependent variables, registered and casual, are both numerical integer values.

In the first step, data has been checked for null values and fortunately no missing values found which made life way easier. Then, the 3 numerical variables got centered and their second order as well as their interactions were added to the model.

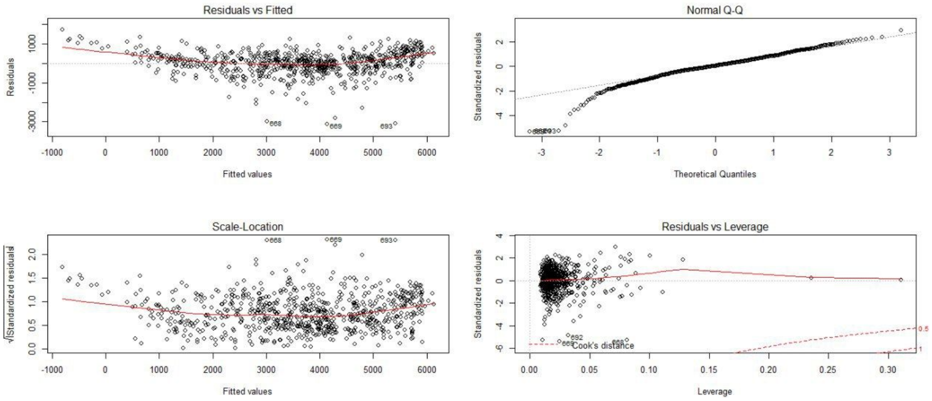
Next, the data was studied for finding the correlation between different variables. Note that the categorical variables were considered as numerical while generating the following graph.

As shown, casual (first dependent variable) has high correlation with temp and working-day and the registered has high correlation with year and temp. It seems that people (registered and casual users) are enjoying biking in hot weather. Also the correlation between the year and registered might explain the increase in the total registered-users, between 2011 and 2012. This assumption can be analyzed by looking into the following scatter plot which clearly shows a positive trend in the registered data. Note that in the rest of this study, we didn't consider the dataset as a time series and the season and year are considered as categorical variables.

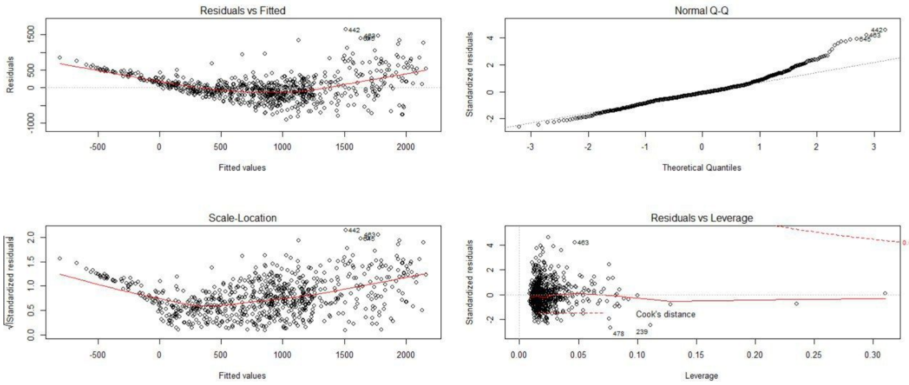




The next step was to investigate whether or not the given dataset has any outliers. As shown below, points 668-669 and 693 are kind of outliers in the casual data set and can be removed from the data set for the first model.



For the second model, points 442-463- 645 are kind of outliers and they are removed from the datasets for the second model.



Now, lets see whether or not these outliers can be explained by the historical events in the Washington DC area. For casual, point 442 is associated with March 17, 2012 which seems reasonable to be an outlier as it is the Saturday in cherry blossom season and the number of casual users are expected to be high. The second point is also a Saturday April 7th 2012, which is the World Health Day and many biking events occurred in DC area. So these dates are all associated with a significantly higher demand for casual bikers.

On the other hand, the outliers for the registered data are all associated with significantly lower demand in some specific dates. The first two points are related to one of the biggest hurricanes, Hurricane Sandy which impact the US east coast on Oct 29-30 2012. So, it is conceivable that the demands were pretty low on those days. The other outliers might be investigated further by looking into other data. For instance, the number of bike stations for that particular company grows significantly in 2011 which can also justify the aforementioned trend in the scatter plot.

Methods

We decide to use MSE and Rsquared for comparison of methods. To calculate MSE, we use 5-fold cross validation for all models.

Multiple Linear Regression

We started with simple linear regression for each dataset. For both datasets, the plots shows that the data is normally distributed but the errors are not zero-meaned and the variances are not constant. So, we decided to use Box-Cox transformation on both datasets. Before applying Box-Cox, we removed 3 data points from each dataset which seems to be outliers. These data points are far from other data points in Q-Q plot, residuals vs. fitted and scale-location plots. For casual dataset these data points are 442, 463 and 645. For registered data points are 668, 669 and 693.

After removing these outliers, we fitted linear models again. For casual dataset, the adjusted R square is changed from 0.7239 to 0.7325. For registered dataset, the adjusted R square is changed from 0.857 to 0.872. It shows that we got better models after removing outliers.

Linear model with transformation (Box-Cox)

We applied box-cox transformation on two simple linear models we had. The selected lambda for casual dataset is 0.222 and the best lambda for registered dataset is 0.667. Using this lambdas, we added a

transformed casual.box in casual dataset and registerd.box in registered dataset. Then, we fitted linear models with transformed values for both datasets.
We calculated MSE values of two fitted models with 5-fold cross validation. The MSE for casual dataset is 2.12 and the MSE for registered dataset is 1322.

Ridge Regression

We fitted Ridge regression model for both datasets and used 5-fold cross validation to optimize lambda within range of 10^{10} to 10^{-10}. The optimal value of lambda for casual and registered are 0.0292 and 0.583, respectively. The MSE for casual dataset is 2.12 and the MSE for registered dataset is 1322. The MSE values are the same as the MSE values of transformed linear models.

Lasso Regression

We fitted Lasso regression model for both datasets and used 5-fold cross validation to optimize lambda within range of 10^{10} to 10^{-10}. The optimal value of lambda for casual and registered are 9.48e-7 and 0.132, respectively. The MSE for casual dataset is 2.12 and the MSE for registered dataset is 1314. The MSE values are the same as the MSE values of transformed linear models.
Since the performances of the models did not change a lot by applying Ridge or LASSO, we can say that the least square model was performing well. In fact, if the interpretability of the model was our priority, the best model would have been the least square model with transformed response variables.

Principal Component Analysis (PCA)

We used 5-fold cross validation to find the best number of PCAs. The result shows that 13 PCAs explains more than 95% variances on both datasets. So, we used 13 PCAs for both datasets and calculated the MSE values. The MSE for casual dataset is 2.08 and the MSE for registered dataset is 1293. For both datasets, the MSE values are the best values among the used methods so far.

Random Forest

First, we started by tuning hyperparameters using 5-fold cross validation. For casual dataset, the best number of variables (mtry) is 5. For registered dataset, the best number of variables (mtry) is 9. We selected these numbers because they have the lowest RMSE in tune grid (based on printed values and plots, attached to the appendix). We also checked the convergence of the random forest and selected 1000 as the number of trees for both datasets. The MSE for casual dataset is 2.39 and the MSE for registered dataset is 1327. These MSE values are greater than MSE values of PCA so the PCA is the best model so far.
The variable importance plot for casual dataset shows that **working day** and **temp** are most influential variables. For registered dataset, **year** and **working day** are most influential variables.

Gradient Boosting Machine (GBM)

At last, we decided to use Boosting as a final technique. We tuned hyper-parameters for both datasets and the result is as follows:

- For casual dataset, n.tree=1000, int.depth=3, shrink=0.01, minobs=5.
- For registered dataset, n.tree=2000, int.depth=5, shrink=0.005, minobs=3.

Using these selected hyper-parameters, we applied Boosting and calculated MSE values. For casual dataset, the MSE is 1.31 and for registered dataset the MSE value is 499. For both datasets, these MSE values are the lowest values which shows that **Boosting** is the best method.
The summary of the model for casual dataset shows that **temp and working day** are most influential variables. For registered dataset, **year** and **temp** are most influential variables.

Results

According to the results presented in the table below, we have selected the GBM model as it has the smallest MSE error for both casual and registered data set and highest ability to explain the variation in the data (Rsquared is bigger than 0.9). In the model for registered data, year , temperature and working day are the most important predictors while in the other model for casual data, temperature and working day are the most important predictors. Shortcoming of our best model is that it is hard to interpret it compare to other models we developed in this project (e.g., Ridge, Lasso and Transformation).

	Model	MSE	Adj. RSquared	Num. Predictors
Casual	Transformation	2.12	0.868	16
	Ridge	2.12	0.868	16
	LASSO	2.12	0.868	15
	PCA	2.08	0.862	13 component
	Random Forest	2.39	0.841	16
	GBM	1.31	0.913	16
Registered	Transformation	1322	0.886	16
	Ridge	1322	0.886	16
	LASSO	1314	0.886	15
	PCA	1293	0.882	13 component
	Random Forest	1327	0.879	16
	GBM	499	0.954	16

Conclusion

In this project, we tried to understand the customers’ behaviour of bike sharing rental system in Washington DC. We started this analysis by performing a preprocessing procedure to validate the normality and constant variance assumption, by which we ended up using Box-Cox transformation as input to our models. We used 6 different modeling approaches, namely, multi-linear regression, Ridge, Lasso, PCA, Random Forest and GBM for both registered and casual (non-registered) customers. 5-fold cross validation MSE was the basis of our comparison between models, and our conclusion was that PCA and GBM will give us the best prediction models having the lowest validation MSE. Looking at GBM’s variable importance, for registered riders, year, temperature and working day were the variables with the most predictive power. On the other hand, in the casual riders’ group temperature and working days were the most important variables.

Appendix:

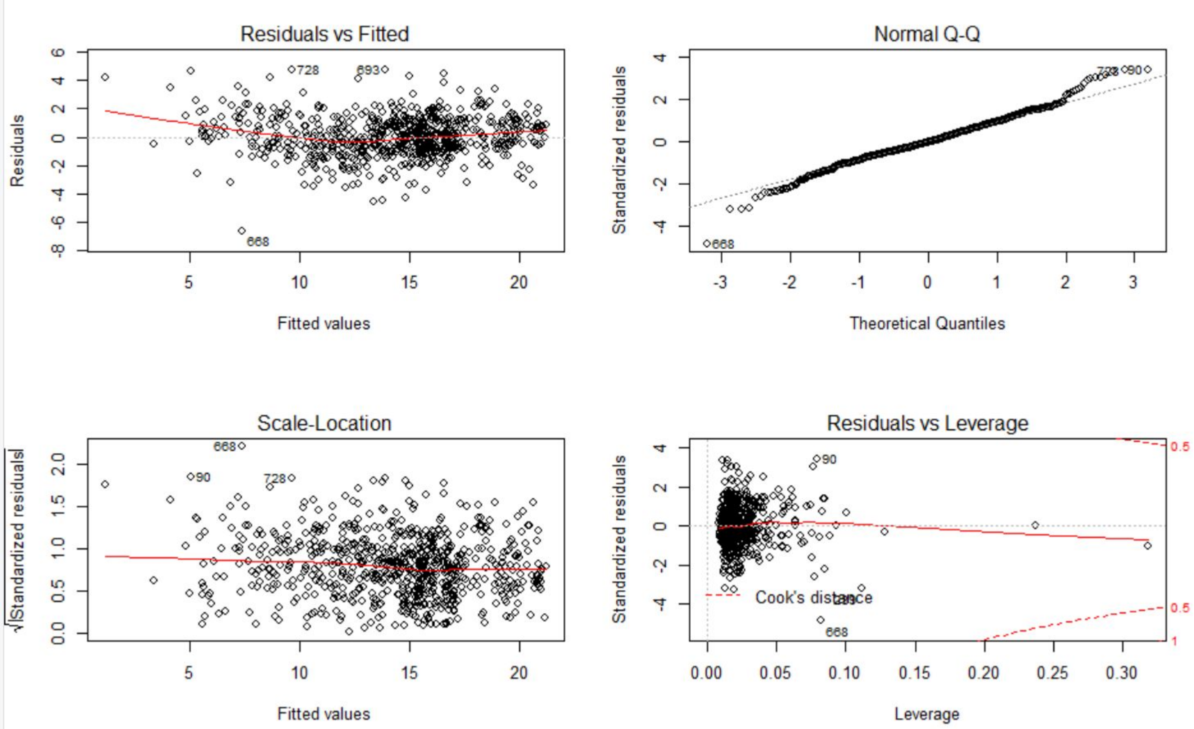


Figure S1. Diagnostic plot for linear model fit to the casual dataset after transformation

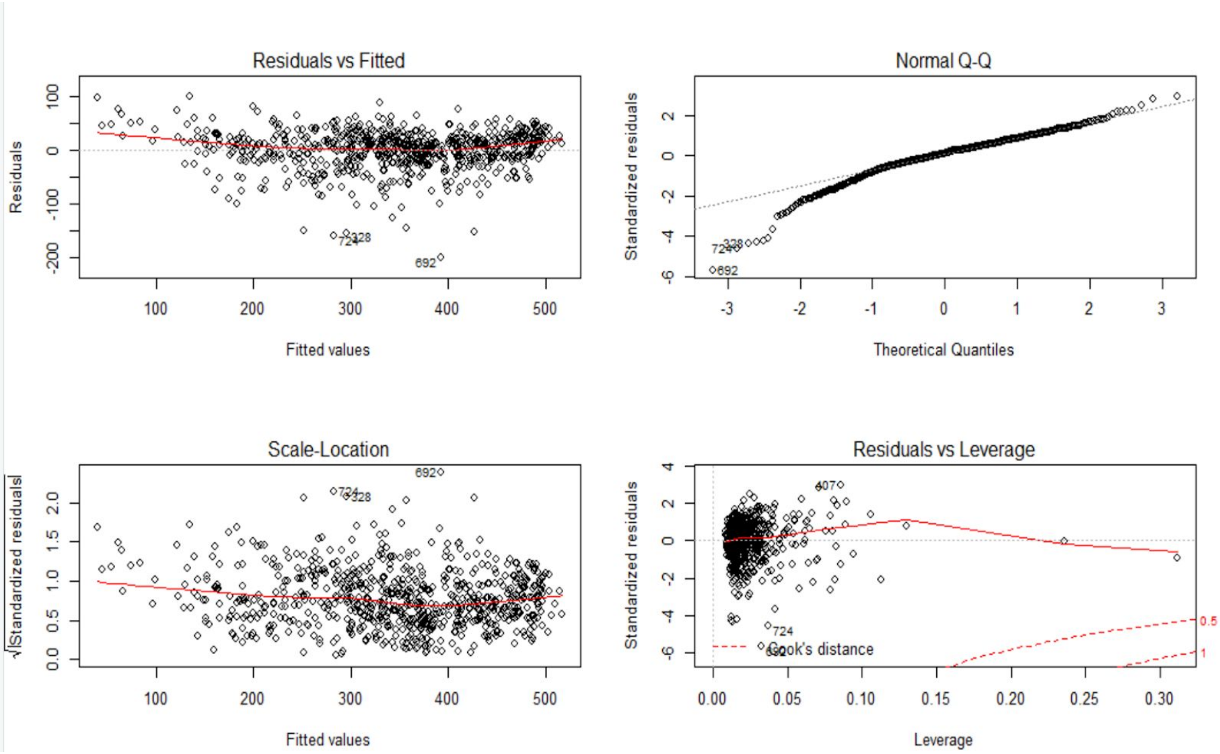
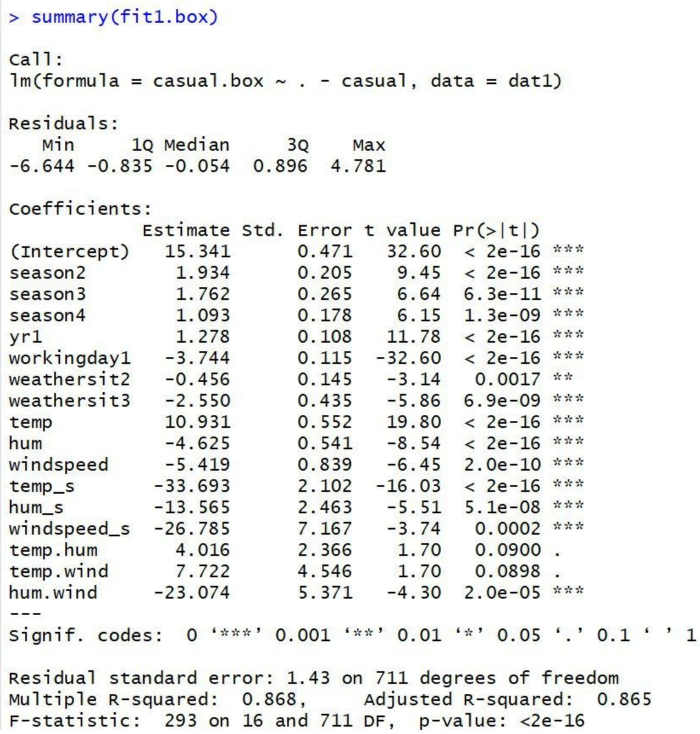
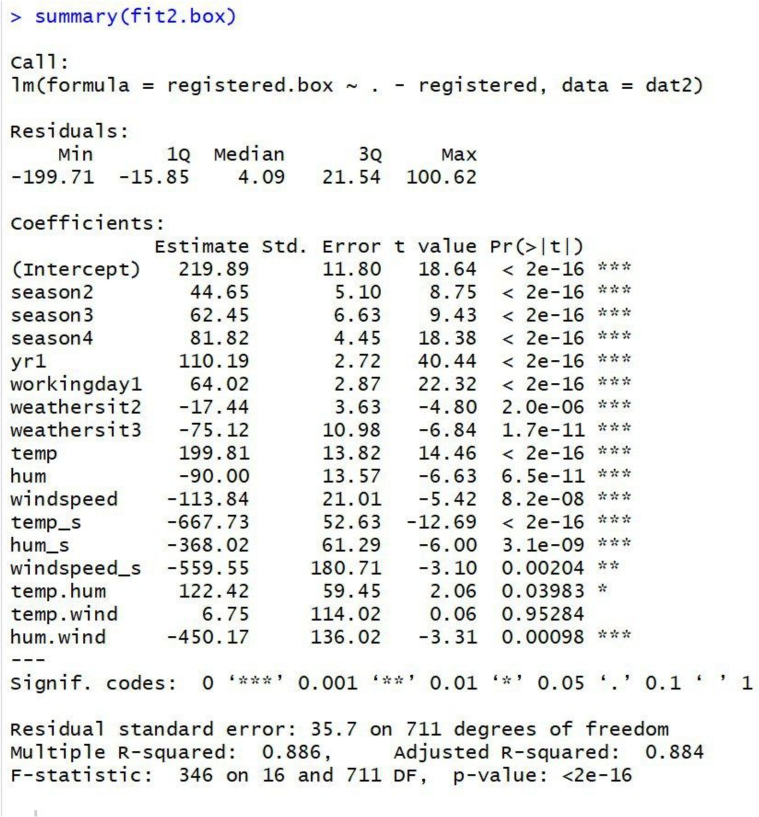


Figure S2. Diagnostic plot for linear model fit to the registered dataset after transformation



```
> summary(fit1.box)

Call:
lm(formula = casual.box ~ . - casual, data = dat1)

Residuals:
   Min     1Q Median     3Q    Max 
-6.644 -0.835 -0.054  0.896  4.781 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   15.341      0.471   32.60  < 2e-16 ***
season2        1.934      0.205    9.45  < 2e-16 ***
season3        1.762      0.265    6.64  6.3e-11 ***
season4        1.093      0.178    6.15  1.3e-09 ***
yr1            1.278      0.108   11.78  < 2e-16 ***
workingday1   -3.744      0.115  -32.60  < 2e-16 ***
weathersit2   -0.456      0.145   -3.14   0.0017 ** 
weathersit3   -2.550      0.435   -5.86  6.9e-09 ***
temp          10.931      0.552   19.80  < 2e-16 ***
hum           -4.625      0.541   -8.54  < 2e-16 ***
windspeed     -5.419      0.839   -6.45  2.0e-10 ***
temp_s       -33.693      2.102  -16.03  < 2e-16 ***
hum_s        -13.565      2.463   -5.51  5.1e-08 ***
windspeed_s  -26.785      7.167   -3.74   0.0002 ***
temp.hum       4.016      2.366    1.70   0.0900 .  
temp.wind      7.722      4.546    1.70   0.0898 .  
hum.wind     -23.074      5.371   -4.30  2.0e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.43 on 711 degrees of freedom
Multiple R-squared:  0.868,	Adjusted R-squared:  0.865 
F-statistic:  293 on 16 and 711 DF,  p-value: <2e-16
```

Figure S3. Summary of Box-Cox transformed model for Casual Riders



```
> summary(fit2.box)

Call:
lm(formula = registered.box ~ . - registered, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-199.71  -15.85    4.09   21.54  100.62 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   219.89      11.80   18.64  < 2e-16 ***
season2        44.65       5.10    8.75  < 2e-16 ***
season3        62.45       6.63    9.43  < 2e-16 ***
season4        81.82       4.45   18.38  < 2e-16 ***
yr1           110.19       2.72   40.44  < 2e-16 ***
workingday1    64.02       2.87   22.32  < 2e-16 ***
weathersit2   -17.44       3.63   -4.80  2.0e-06 ***
weathersit3   -75.12      10.98   -6.84  1.7e-11 ***
temp          199.81      13.82   14.46  < 2e-16 ***
hum           -90.00      13.57   -6.63  6.5e-11 ***
windspeed    -113.84      21.01   -5.42  8.2e-08 ***
temp_s       -667.73      52.63  -12.69  < 2e-16 ***
hum_s        -368.02      61.29   -6.00  3.1e-09 ***
windspeed_s  -559.55     180.71   -3.10  0.00204 ** 
temp.hum      122.42      59.45    2.06  0.03983 *  
temp.wind       6.75     114.02    0.06  0.95284    
hum.wind     -450.17     136.02   -3.31  0.00098 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.7 on 711 degrees of freedom
Multiple R-squared:  0.886,	Adjusted R-squared:  0.884 
F-statistic:  346 on 16 and 711 DF,  p-value: <2e-16
```

Figure S4. Summary of Box-Cox transformed model for Registered Riders

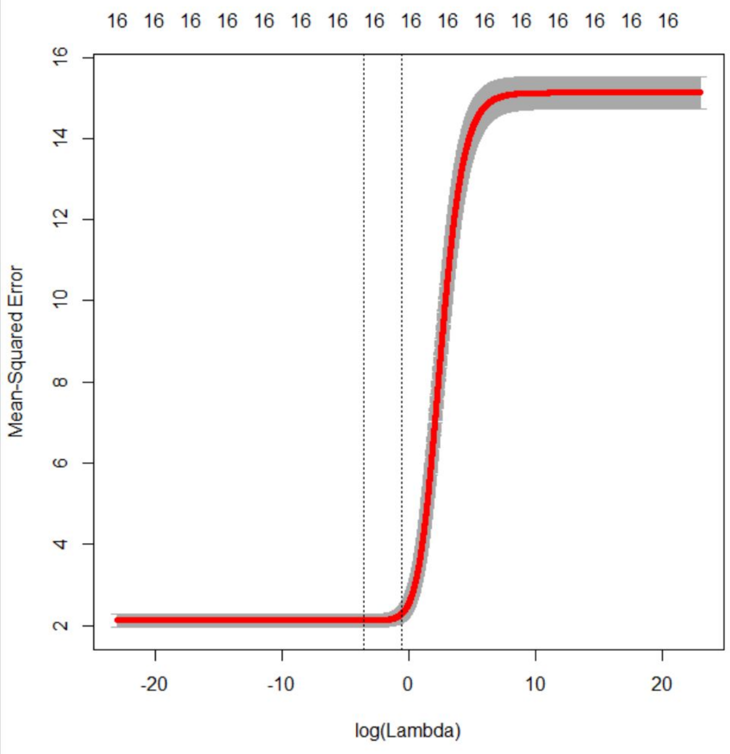


Figure S5. Selecting lambda for Ridge regression using the Casual dataset (best lambda = 0.0292)

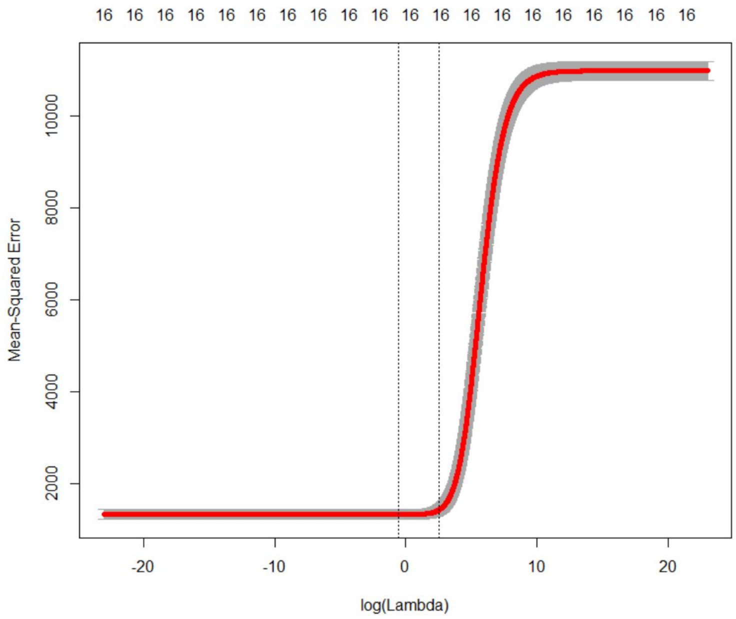


Figure S6. Selecting lambda for Ridge regression using the Registered dataset (best lambda = 0.583)

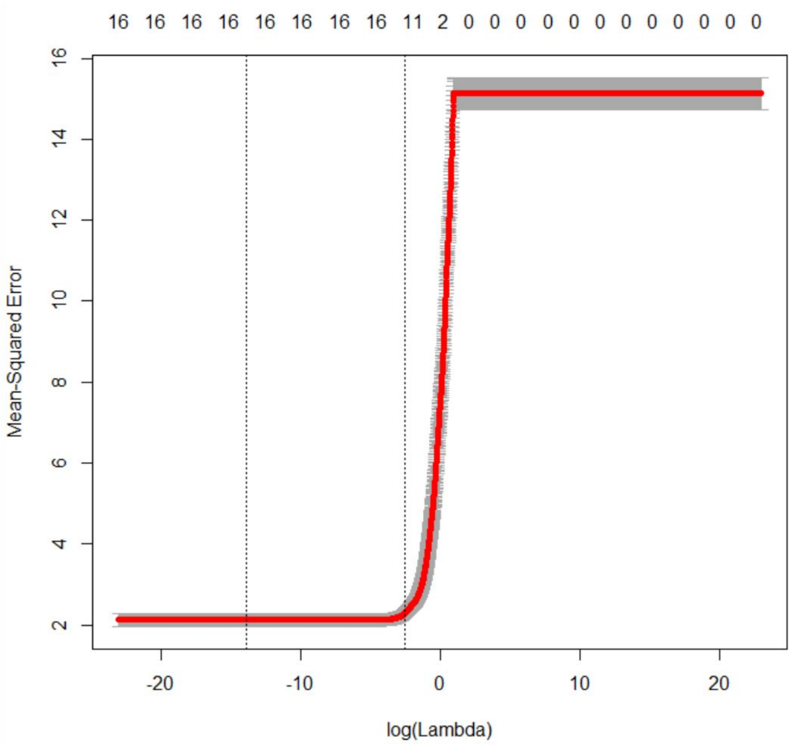


Figure S7. Selecting lambda for Lasso regression using the Casual dataset (best lambda = 9.48e-07)

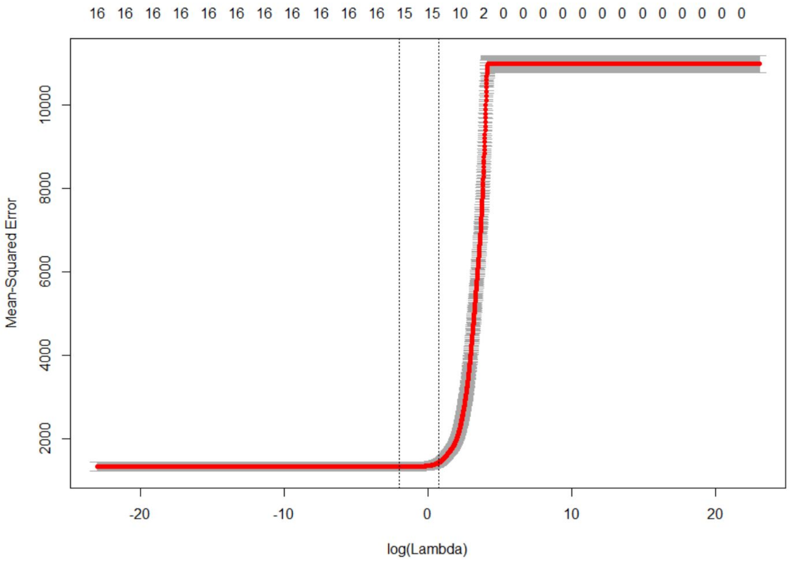


Figure S8. Selecting lambda for Lasso regression using the Registered dataset (best lambda = 0.132)

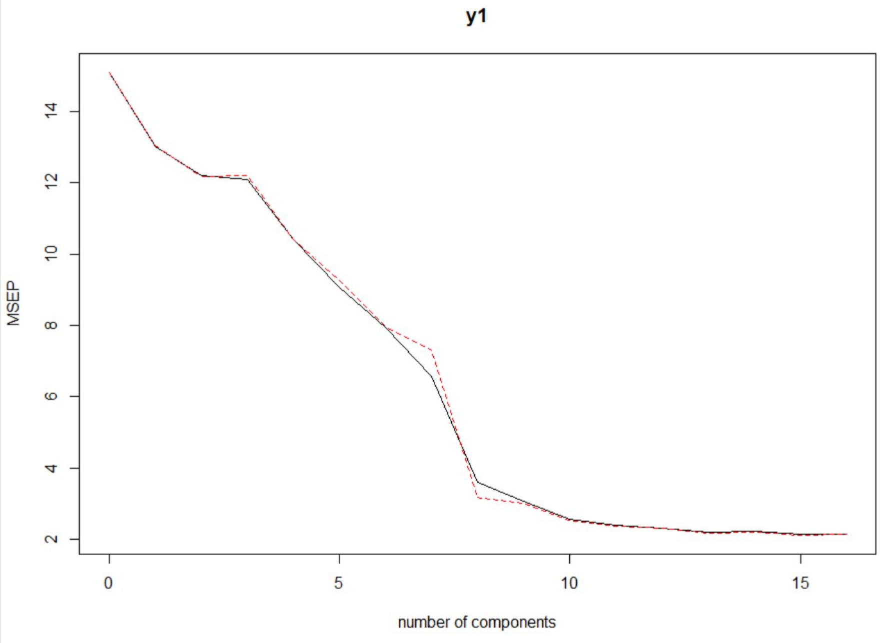


Figure S9. Number of components vs MSEP in PCA method using the Casual data (13 component is selected)

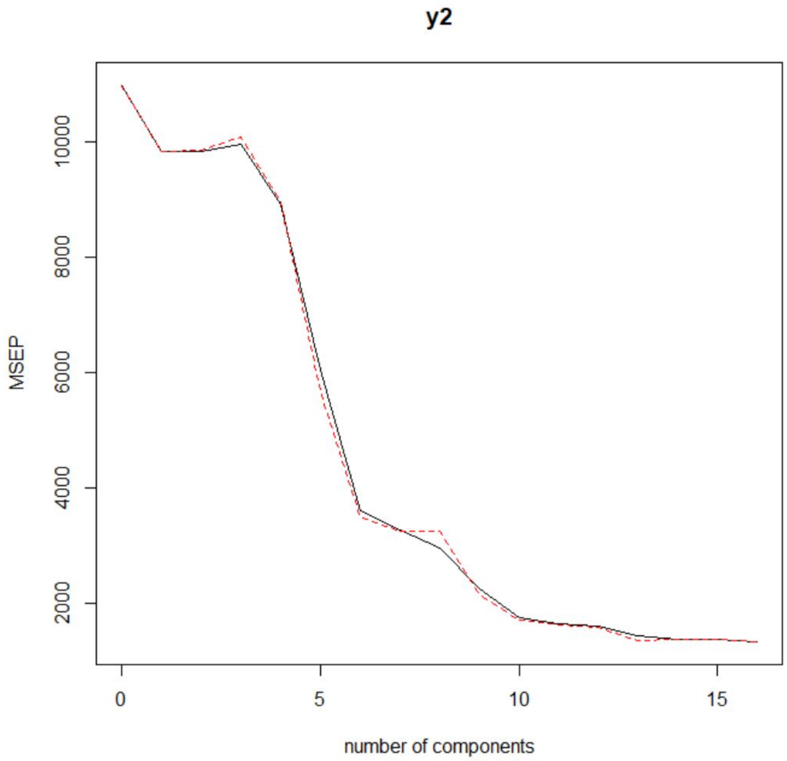


Figure S10. Number of components vs MSEP in PCA method using the Registered dataset (13 component is selected)

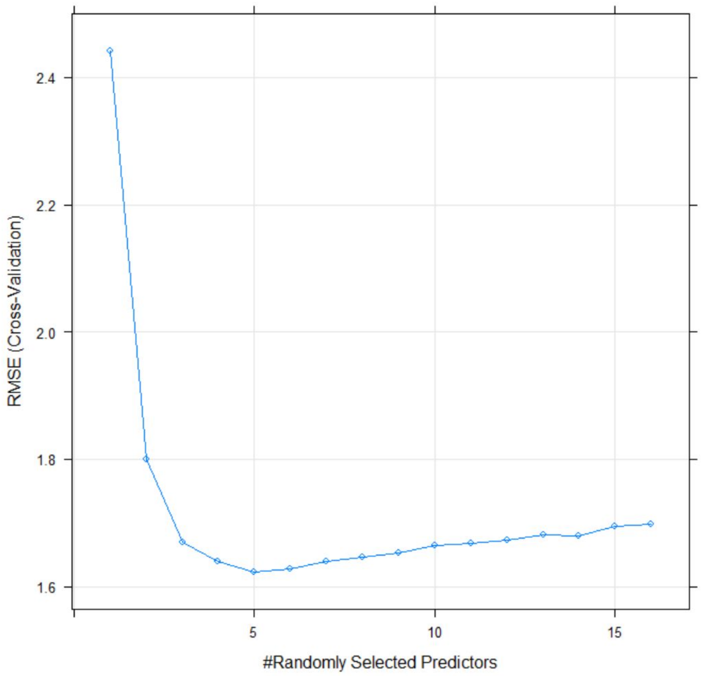


Figure S11. RMSE vs Randomly selected predictors in Random Forest method using the Casual data set (mtry =5 is selected)

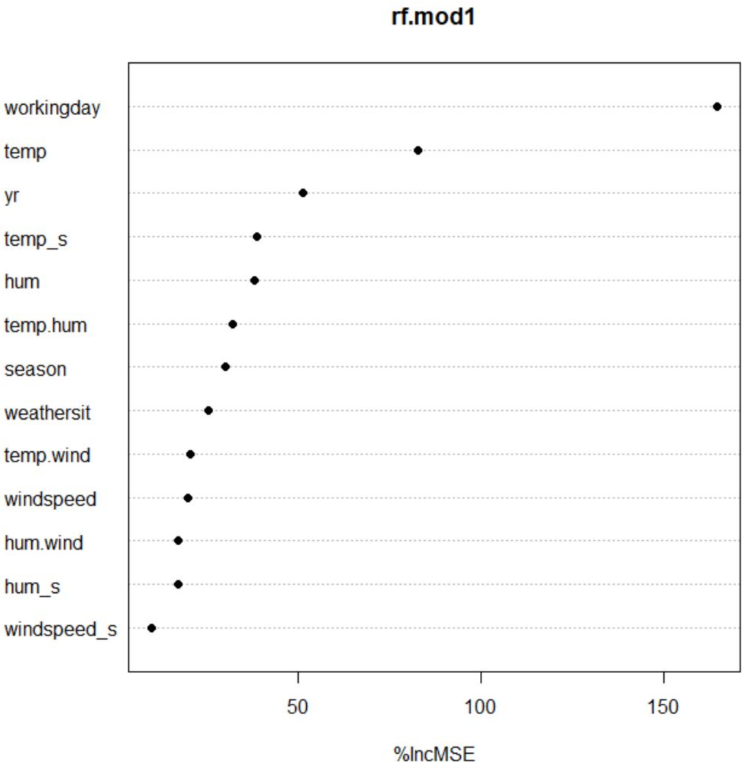


Figure S12. Effect of predictors on MSE in Casual dataset

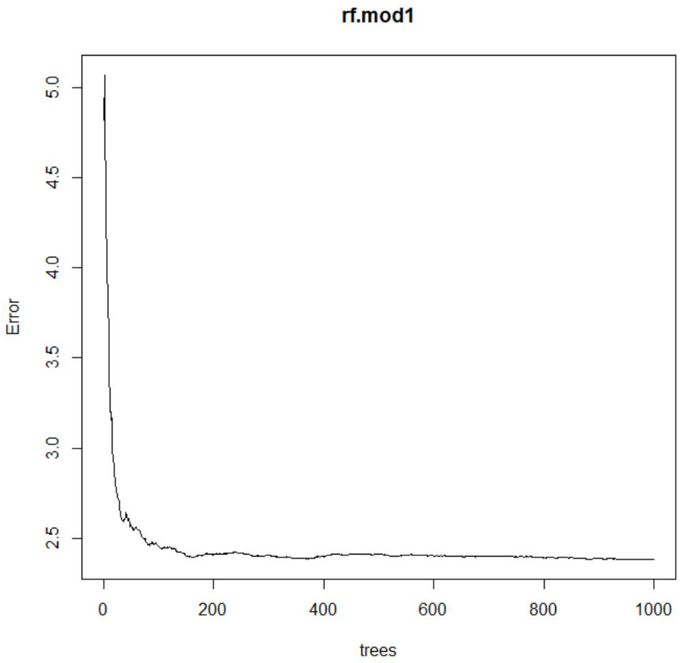


Figure S13. Selecting the number of trees in RF method using Casual data set (ntree=1000 is selected)

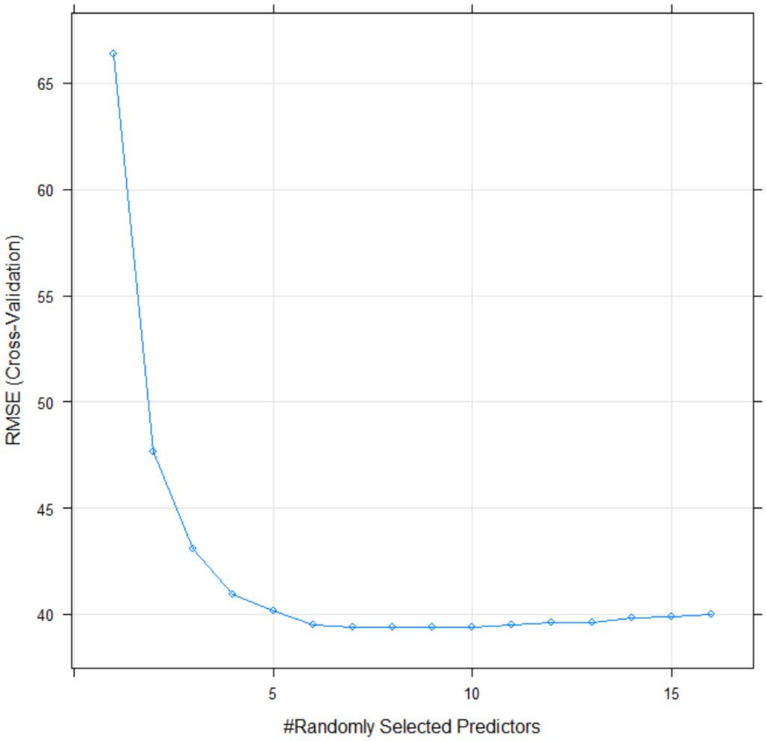


Figure S14. RMSE vs Randomly selected predictors in Random Forest method using the Registered data set (mtry =9 is selected)

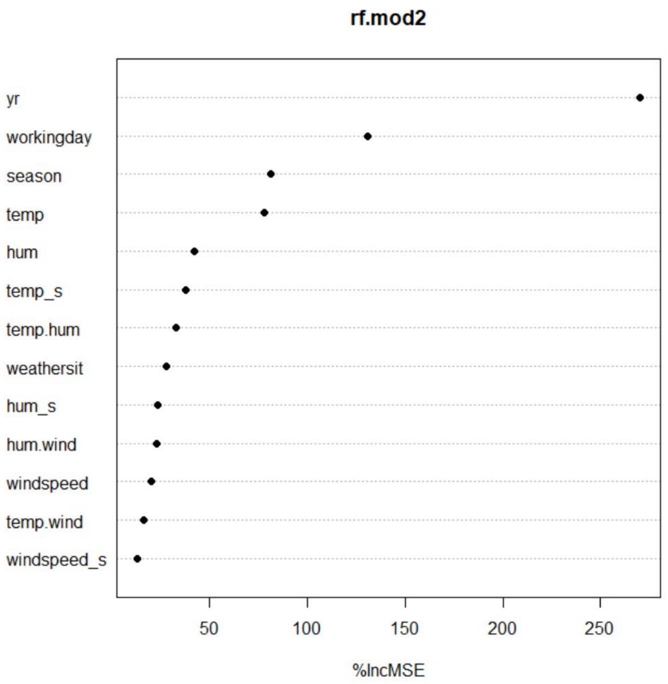


Figure S15. Effect of predictors on MSE in Registered dataset

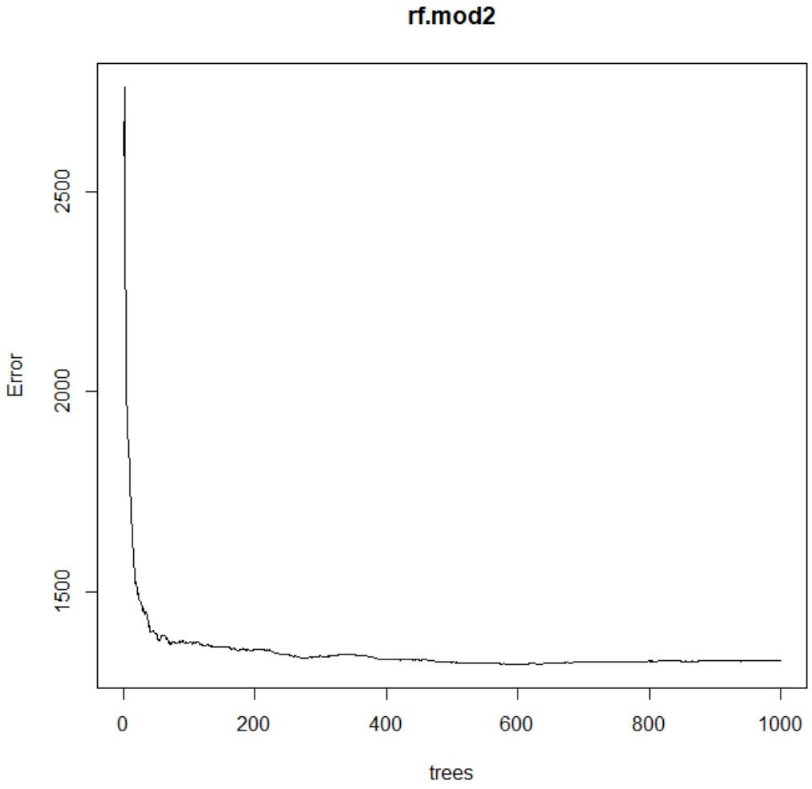


Figure S16. Selecting the number of trees in RF method using Registered dataset (ntree=1000 is selected)

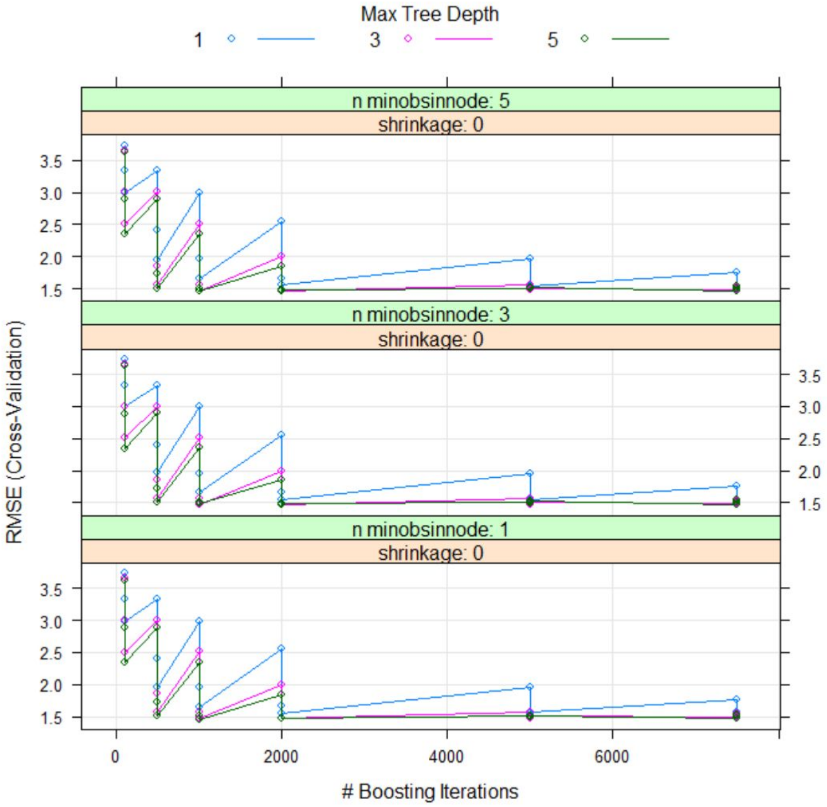


Figure S17. Selecting parameters for GBM: choose n.tree=1000, int.depth=3, shrink=0.01, minobs=5, casual data set

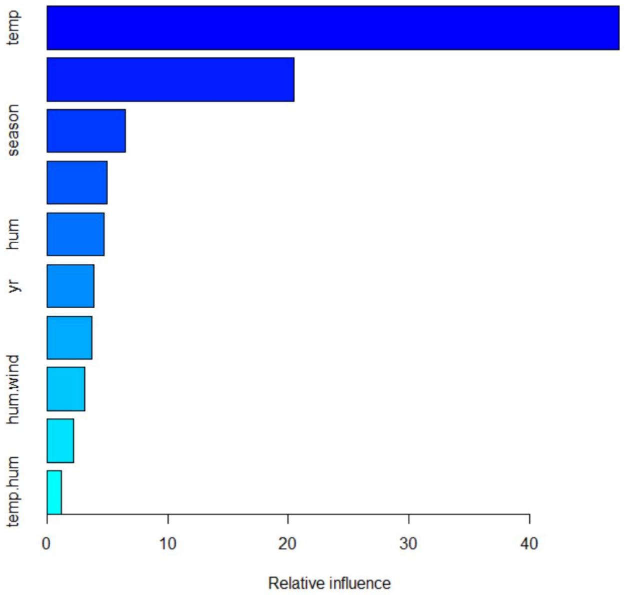


Figure S18. Relative influence of predictors using GBM for casual data set

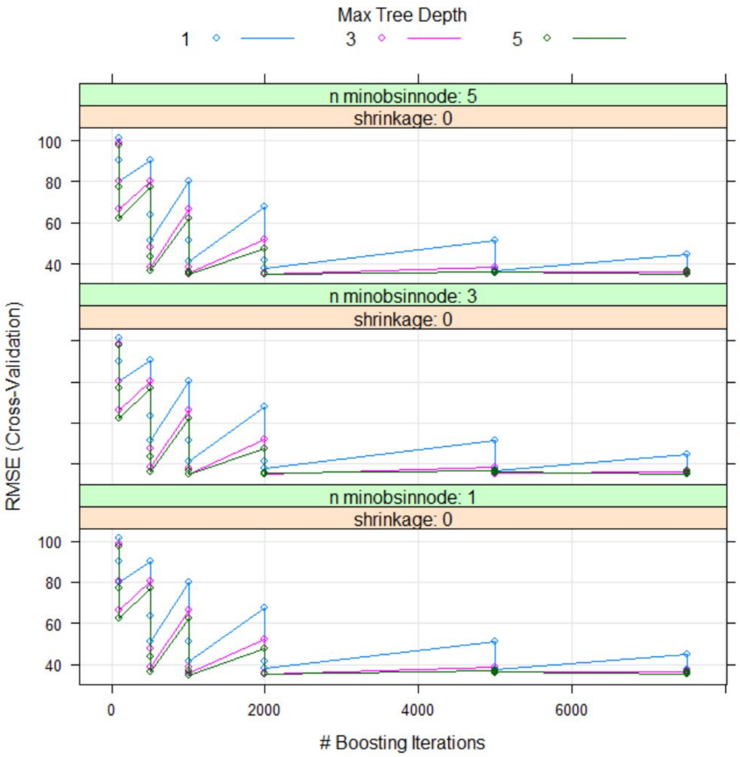


Figure S19. Selecting parameters for GBM: choose n.tree=2000, int.depth=5, shrink=0.005, minobs=3, Registered data set

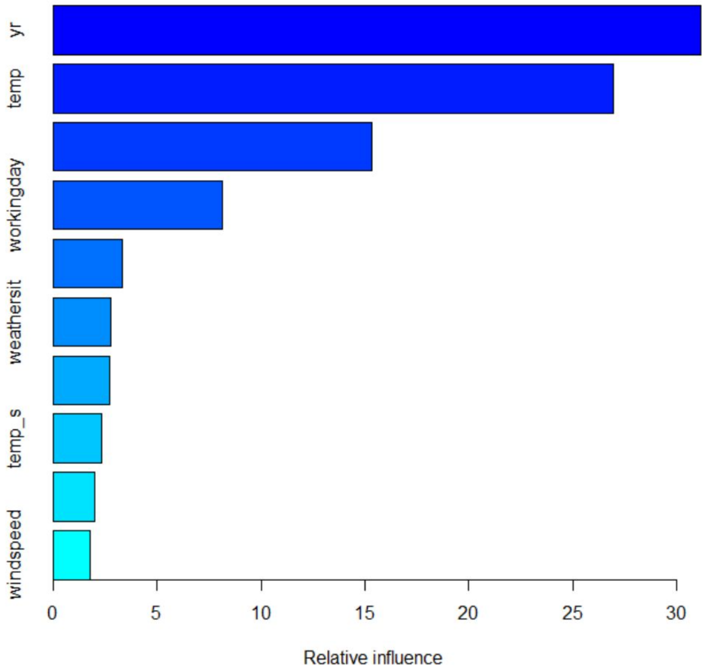


Figure S20. Relative influence of predictors using GBM for Registered data set