Final Report Bike Rental Project

Abstract

The purpose of this study is to understand the influences of bike rental usage and also predict it in order to fulfill the demands. This study is based on the regression analysis methods with aim to create a prediction model for bike usage. The dataset used is from capital Bikeshare system Washington D.C.

Bike Rentals

Introduction/ Problem statement

In today's world Bike rentals have become so common that it has become the major mode of transport globally in every place. In this project we will be seeing how the Bike rentals go through during weekdays, weekends, Temperature, hourly cost of the rental. The main aim is to calculate the hourly bike rental for 3 months every hour.

Understanding the data /Hypothesis

Major Points to be considered which influences the bike rentals

- Hourly trend: There must be high demand during office timings. Early morning and late evening can have different trend (cyclist) and low demand during 10:00 pm to 4:00 am.

- Daily Trend: Registered users demand more bike on weekdays as compared to weekend or holiday.

- Rain: The demand of bikes will be lower on a rainy day as compared to a sunny day. Similarly, higher humidity will cause to lower the demand and vice versa.

- Temperature: In India, temperature has negative correlation with bike demand. But, after looking at Washington's temperature graph, I presume it may have positive correlation.

- Pollution: If the pollution level in a city starts soaring, people may start using Bike (it may be influenced by government / company policies or increased awareness).

- Time: Total demand should have higher contribution of registered user as compared to casual because registered user base would increase over time.

- Traffic: It can be positively correlated with Bike demand. Higher traffic may force people to use bike as compared to other road transport medium like car, taxi etc.

Methodology used for collection of data in terms of Time frequency

In this dataset we can see that the data was collected in the yearly format for the year 2011 and 2012. The time frequency of the data is every day, every hour from Jan 2011 to Sep 2012.

The test dataset contains data collected from October 2012 till 31st December 2012 of the hourly bike rental data.

Exploratory Data Analysis

Understanding the distribution of numerical variables and creating a frequency Table for numeric variables.

Few conclusions from the Analysis that can be drawn are season has four categories of almost equal distribution.

In aspect of weather, weather 1 has higher contribution mostly clear weather.

Also, we can infer that the working day and variable holiday is showing a Similar trend.

Variables temp, atemp, humidity and windspeed also seems to be naturally distributed.

Classification of bike rental based on every hour.

Hourly Trend:

The hourly trend of Bike every hour:

Next I have classified Bike Rentals on the basis of max and min based on the time.

Low: 0-6 Hrs.

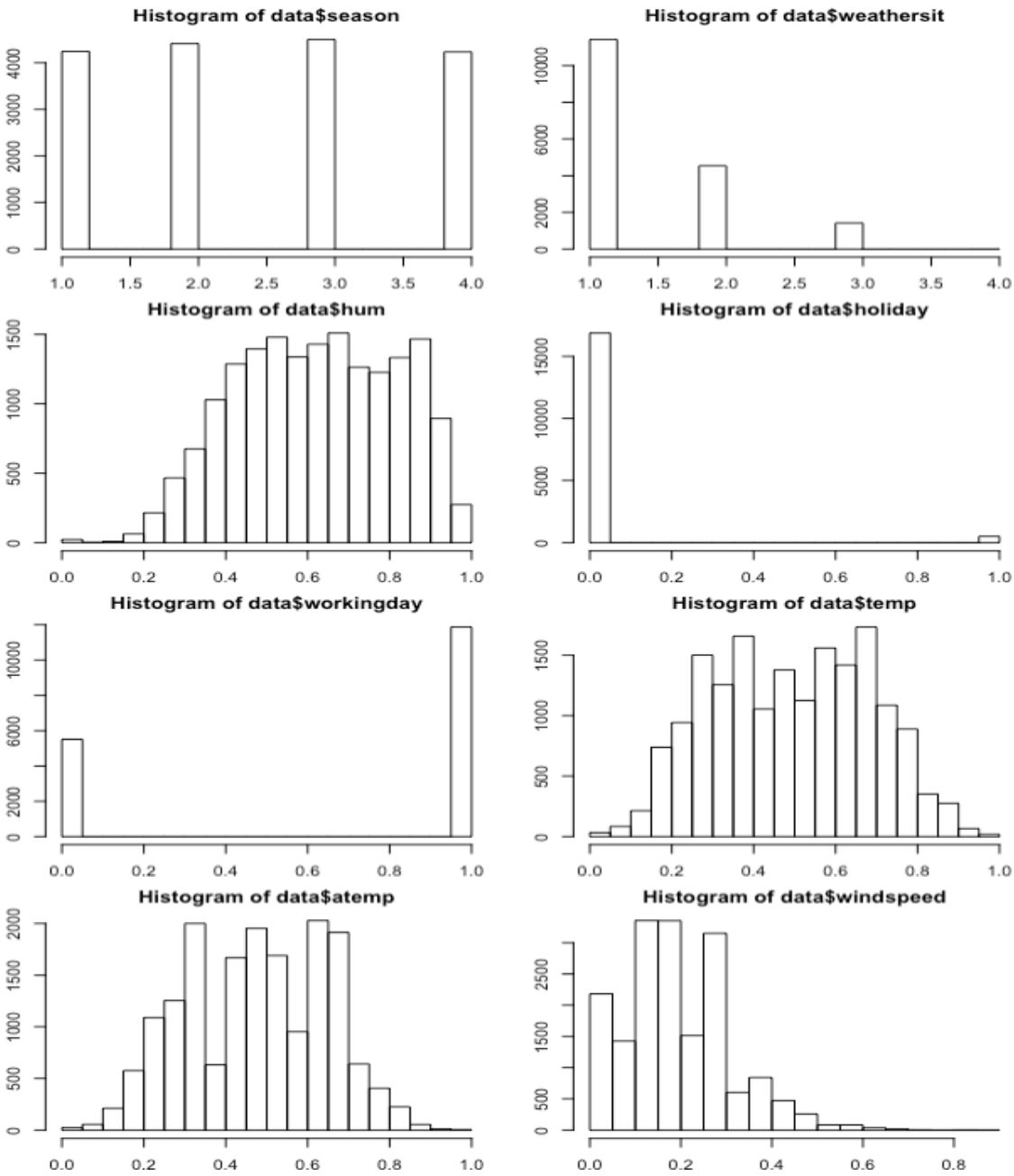
Average: 10-16 Hrs.

High: 7-9 and 17-19 Hrs.

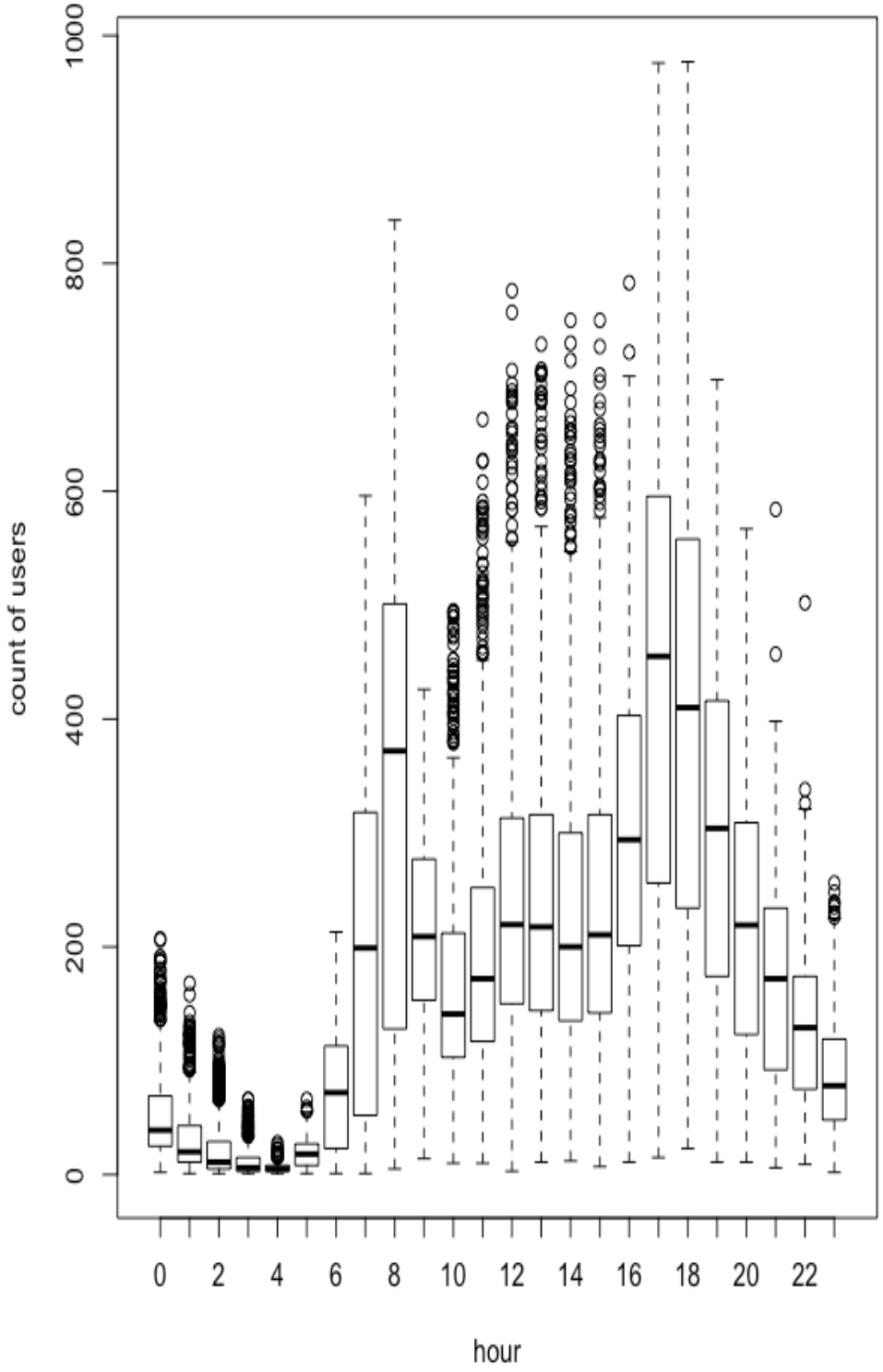
Exploratory Data Analysis & Data Preprocessing

Relationship among important variables.

Below Histogram plot depicts that season has four categories



Insightful Visualizations:



Data Preprocessing

Data Frame consist of variables that requires transformation into days, date and time.

In test dataset we have registered users, casual users and count which has been replaced from NAs.

I have also considered to change the discrete variables into factors i, e season, weather,

holiday and working day.

Hypothesis Testing using multivariate analysis has been carried out with respect to the given

Dataset.

Addition of new variables include Time, replacing dteday with date time has been carried

done.

Proper processing of outliers has been done on the given dataset.

Analytical Model Building (mention the alternate analytical approaches that they may see fit to be applied to the problem)

The analytical model that I have used in this problem is the linear regression model on the given dataset.

The alternate approaches that can be used in this problem is decision tree algorithm.

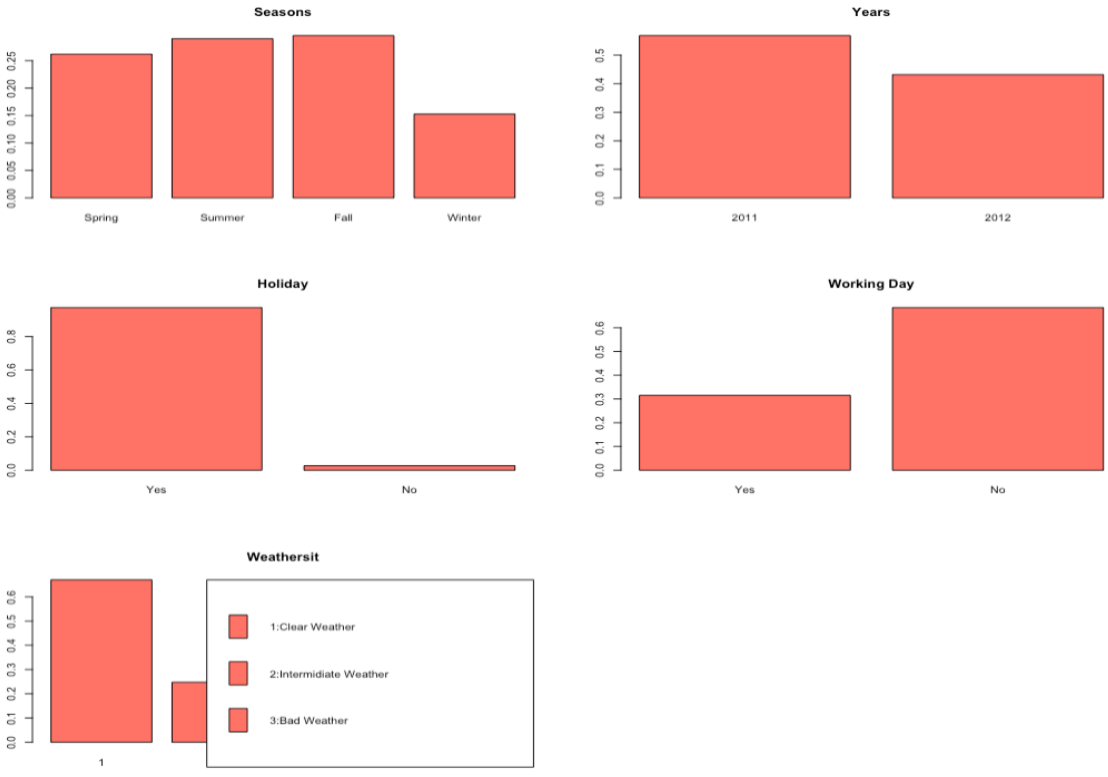
Also random forest can be used on the given dataset to solve the problem.

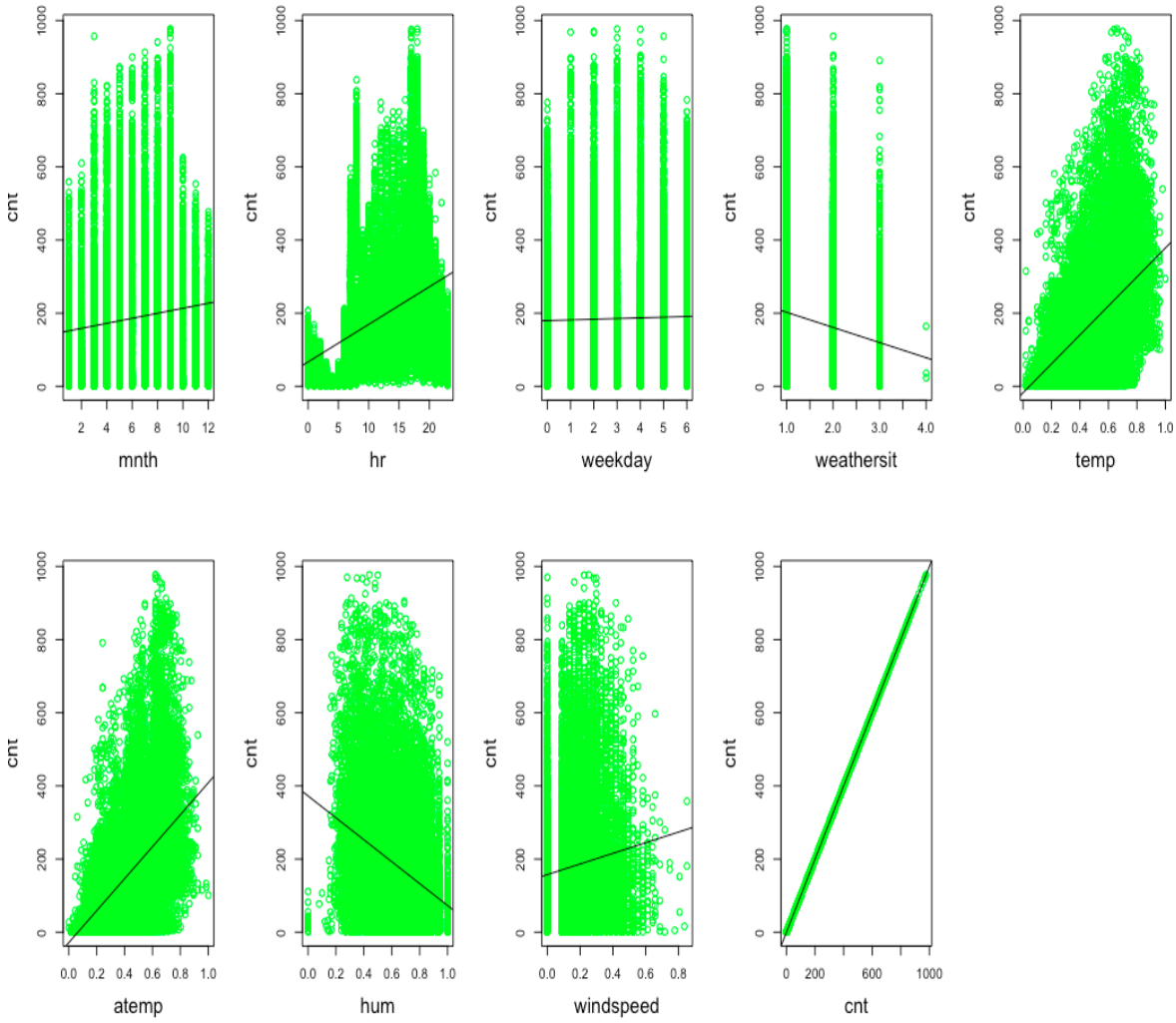
Also, I have carried out the linear regression model to validate the same with the test

Dataset.

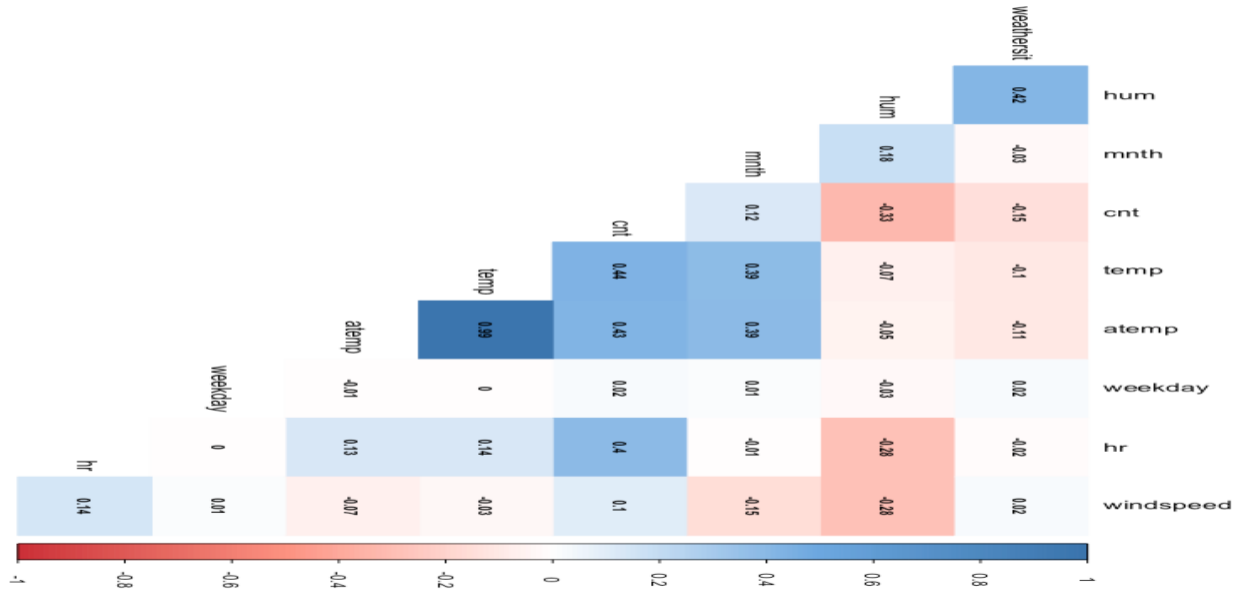
Modelling Process (validation & interpretation)

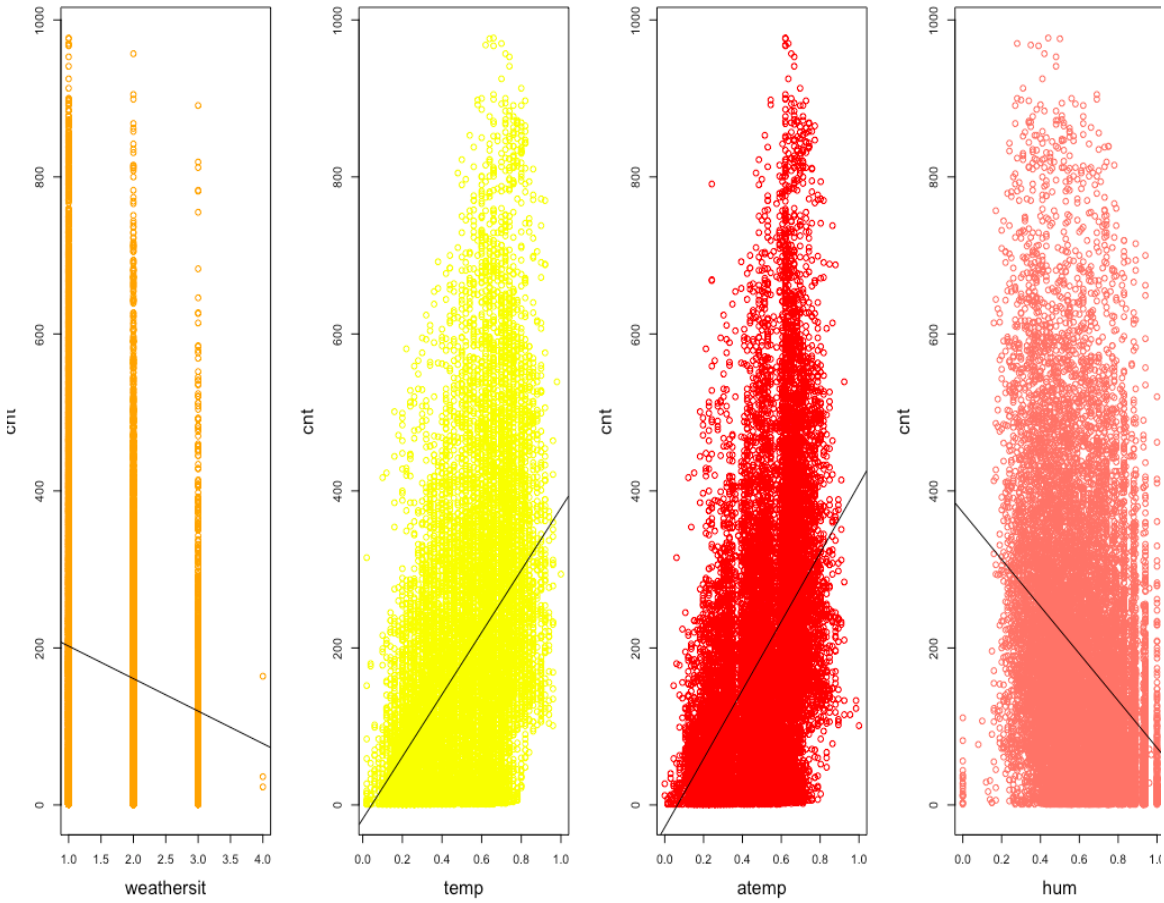
Model comparisons (confusion matrix, ROC, AUC)
Ensemble modelling, wherever applicable Interpretation from the best model
Business Insight

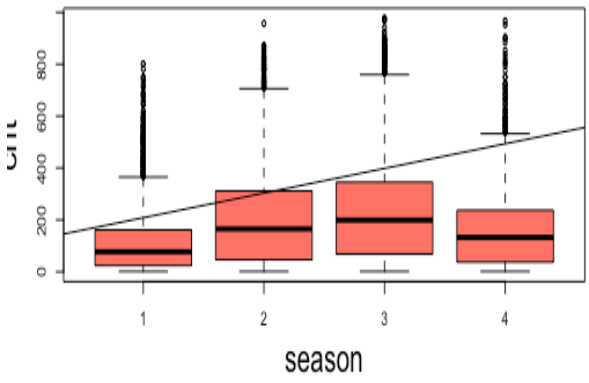


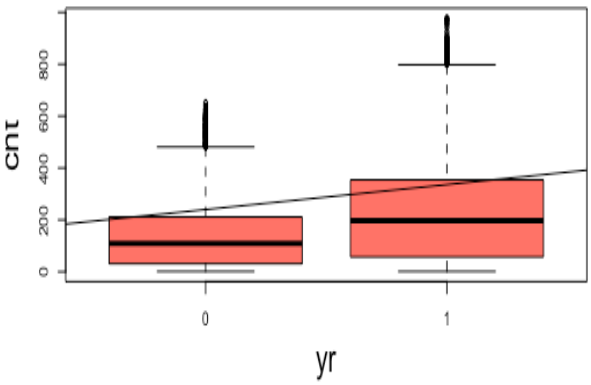

cnt
mnth
hr
weekday
weathersit
temp
atemp
hum
windspeed
cnt

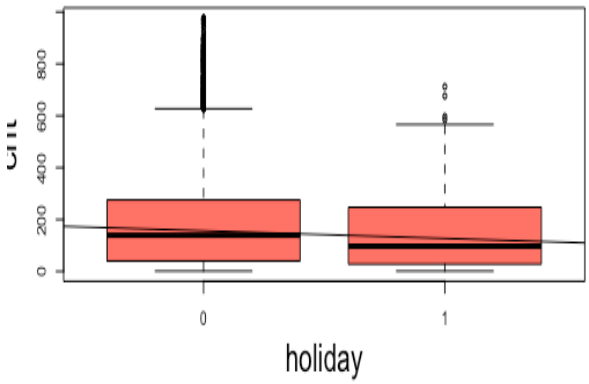
Correlation between the variables

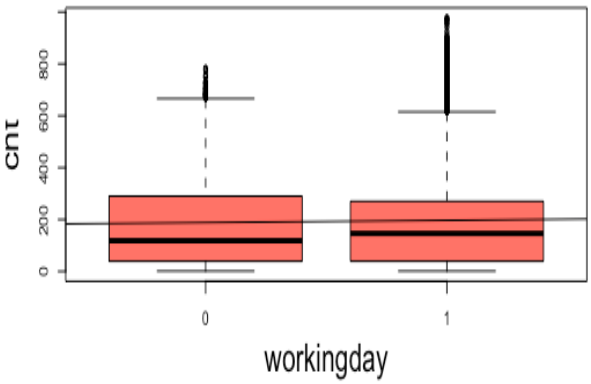


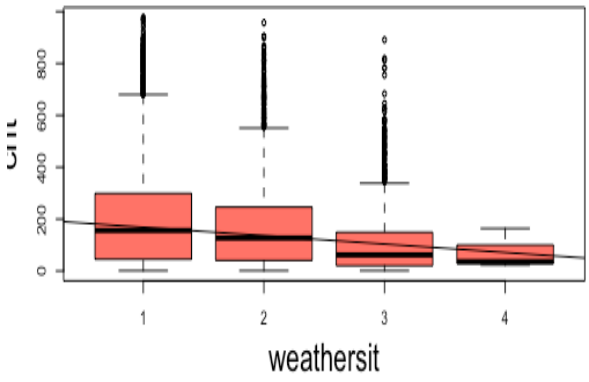

cnt
0
200
400
600
800
1000
1.0 1.5 2.0 2.5 3.0 3.5 4.0
weathersit
0.0 0.2 0.4 0.6 0.8 1.0
temp
0.0 0.2 0.4 0.6 0.8 1.0
atemp
0.0 0.2 0.4 0.6 0.8
hum

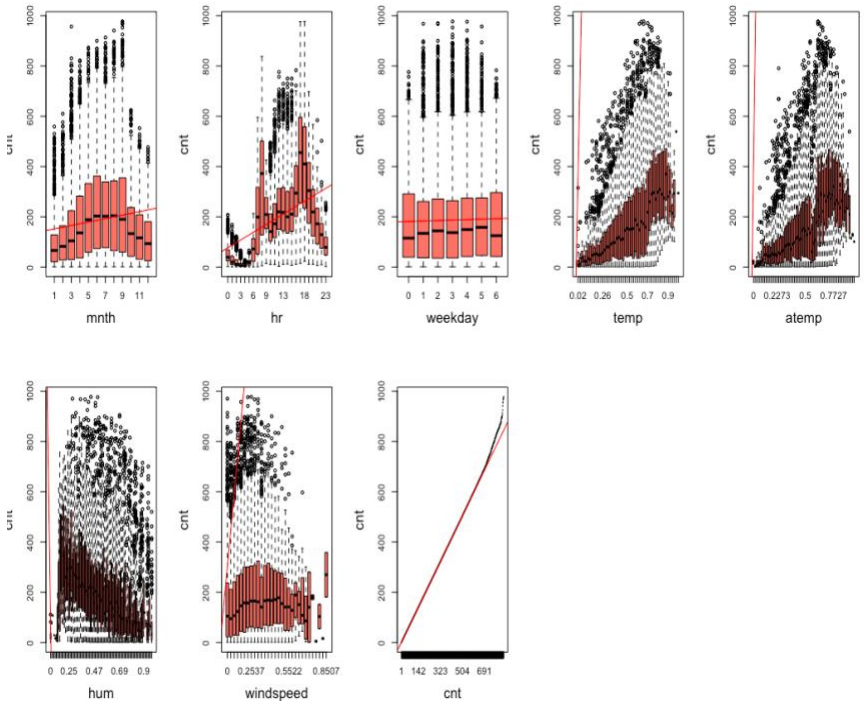

cnt
0
200
400
600
800
1
2
3
4
season

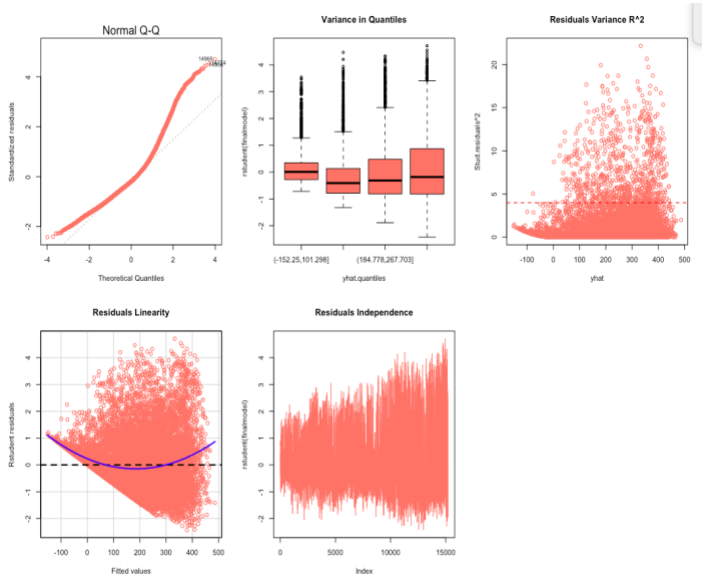

cnt
0
200
400
600
800
0
1
yr

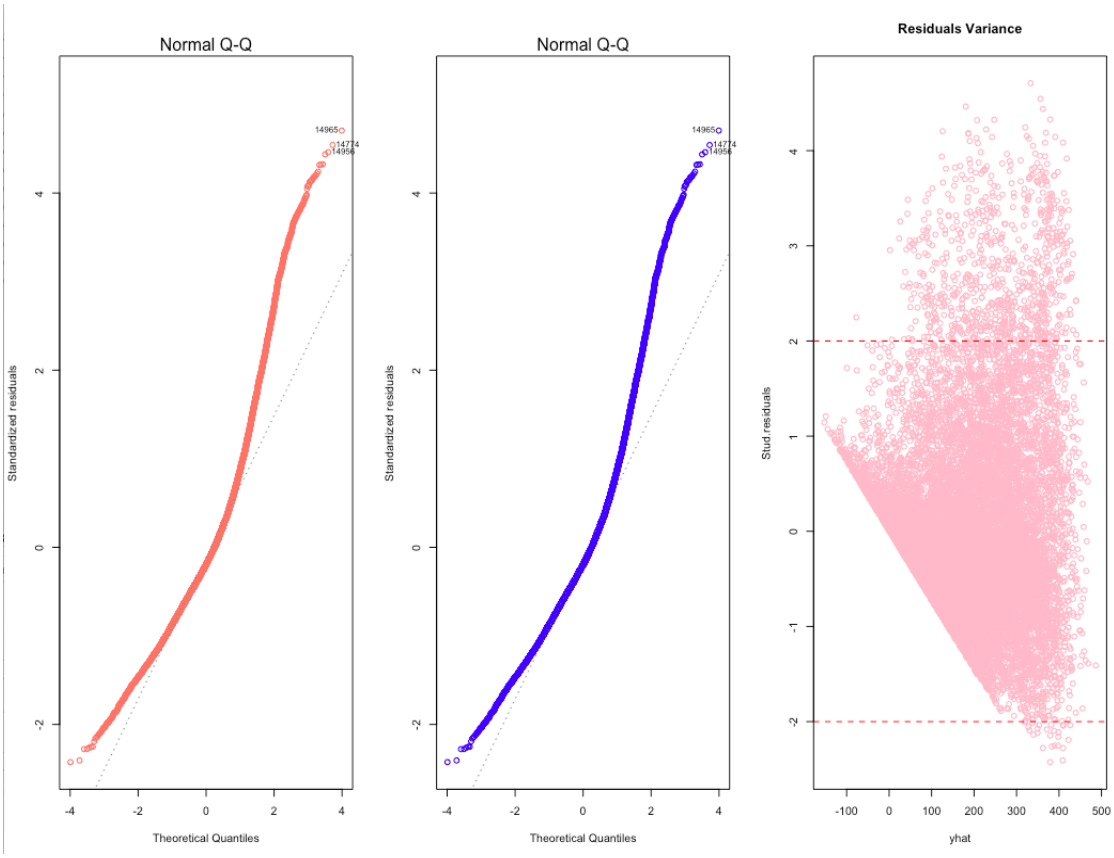

cnt
0
200
400
600
800
0
1
holiday


cnt
0
200
400
600
800
0
1
workingday


cnt
0
200
400
600
800
1
2
3
4
weathersit


cnt
mnth
hr
weekday
temp
atemp
hum
windspeed
cnt


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Variance in Quantiles
rstudent(finalmodel)
(-152.25,101.298]
(194.778,267.703]
yhat.quantiles
Residuals Variance R^2
Stud.residuals^2
yhat
Residuals Linearity
Rstudent residuals
Fitted values
Residuals Independence
rstudent(finalmodel)
Index



The above figures represents the various stages of the bike rental sharing visualization of Model building done via regression, Centralized model,

Also, I have used Anova test here have built three models out of which model three has one of the best accuracy so will be considering the same.

Heteroscedasticity also increases, and maybe it's a sign that we will have problem with model assumptions, if we will include temp or atemp as predictors. Humidity (hum) preserves also a lot of variance. Wind speed (windspeed) doesn't affect the bike sharing usage as we said in the previous paragraph; but also in the following graph we can observe this independence between cnt and windspeed because the fitting of the line doesn't follow the distribution of scatters. Also, it's important to mention that we have a lot of outliers and maybe influential points that will affect our models. Hence, we have to keep in mind these figures when we create our models. Afterwards, we have to check the distribution of bike share demand in every hour of a day. So, we created a boxplot graph. As we can see in the following figure (see Figure 5 – Total Usage in a Daily Bases) the distribution has high variance during a day. Although, we can divide the day in the following periods:

24:00 - 07:00 : Low bike usage ϖ 07:00 -16:00 : Middle bike usage ϖ 16:00 - 19:00 : High bike usage ϖ 19:00 - 24:00 : Middle-Low bike usage.

Model selection :

We have already observed the correlations and relations between variances and now we can construct our regression models and try to find the most appropriate of them for prediction of the total bike demand. As a response we have the cnt and all the other variables as predictors. We start our searching for a linear model ($Y = bo + b1x1 + \ldots bnxn$, with $\varepsilon \approx N[0,\sigma^2]$). Firstly, we constructed our first model with only numerical variables, initial model (lm() in R)(see Table 3- Initial Model). In this model we have temp, atemp and

windspeed as insignificant variables (because Pr>0.05 and we can't reject the null hypothesis, Ho = Coefficient equals to zero). Also, if we check Variance Inflection Factor – VIF based on Akaike criterion (vif() in R) we observe that temp and atemp have very high value (around 42.4), something that we expected (see Figure 3- Numeric Variables Correlation). In order to take a better decision we used Stepwise Regression method (step() function in R) and Backward Elimination in initial model and we created a new model, model 1, without atemp, windspeed, (see Table 4 - Model 1). This model has similar R^2 & R^2Adj ≈ 0.33; now we have low collinearity between variables, VIF (value ≈ 1). In this stage we added Factors to our model and we created the full model (see Table 6- Full Model). This model has higher R^2 & R^2Adj ≈ 0.39, but the R^2 increasing was expected, because we increased the number of predictors. In that model we have intercept, atemp, mth, windspeed, workingday and holiday as insignificant variables (Pr>0.05, we reject the null Hypothesis, Ho = Coefficient equals to zero) and also it's appeared again the high VIF between temp and atemp. So, we apply stepwise function (we prefer AIC criterion than BIC, because BIC has higher penalized and simplify a lot the model) in order to decrease the number of predictors. After Stepwise Regression method and Backward Elimination, we had the same model in both cases, model 3 (see Table 7- Model 3). This model has less predictors, method removed atemp, mth, windspeed, workingday and holiday; low VIF (because the atemp was excluded), R^2 & R^2Adj ≈ 0.39 and standard error ≈ 140 bikes. So, we can assume that model 3 has the best fitting for now. Also, the major business insights have tried building Random Forest, Featuring Engineering but was not successful as there was lot of negative AICs which was observed. The major drawback of the models was the Rsquare values.

Final Model and Interpretation

So, in the process of Model selection and interpretation we consider all the previously used data from descriptive analysis, pairwise methods and also the prediction abilities in the evaluation samples.

The most important is we have checked with Anova for the extra parameters from our models were not zero. Full model is very complicated with lot of insignificant variables.

The prediction abilities of this model are quite similar in the evaluation dataset.

No linear model with corrected assumptions, has the highest goodness of fit. Also, no linear model has extremely low standard error, which implies of overfitting in the errors of training sample. Though we had high goodness of fit in the evaluation sample, we are not sure about the ability to predict the samples.

In my view the appropriate model is the stepwise model for the analysis purpose. The model considered is the one with significant covariates, low VIF between variables and simple interpretation to describe the day for each season.

The coefficients of Weekday and Hour have different interpretation in comparison with the other numerical variables (Temp and Hum). These coefficients mean that if we change one day or one hour we will have considerable amount of bikes in a different day and more bikes in a different hour.

Temperature and humidity also play a major role. If we are in specific date considering year, season, day and hour and increase for one that means the increase in number of bikes. On the other hand, if we are in the same date and increase for 1 (100% humidity) then number of users will gradually decrease. So, the variability of predictions is 140 bikes which is not a low variance.

Conclusion

I would like to bring few major considerations into focus on the bike sharing where in weather conditions play a major role in bike usage. The most major factor is weather temperature and humidity. Also, common people don't like to use bikes for commuting in extreme weather temperatures, so this trend is also connected with seasons.

In an annual basis, mild winter looks the most suitable period for bike usage. Also, in extreme weather conditions people avoid using bikes. The major trend is seen where people tend to use bikes is during spring summer and gradually the trend decreases.

References

GitHub

Kaggle