

Prediction of bike rental count

Ajaydarshan | Data scientist course | 02-Jan-2019

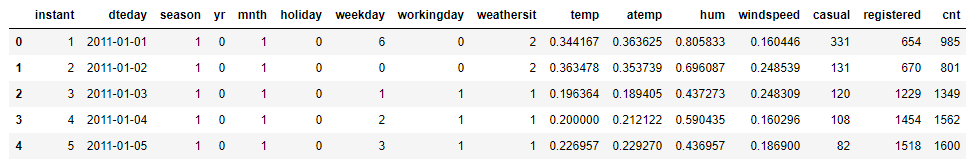
# Chapter 1 Introduction

### Problem statement

The objective of this project is to predict bike rental count on daily basis based on environmental and seasonal setting.

### Data

Sample of data is given in below figure.

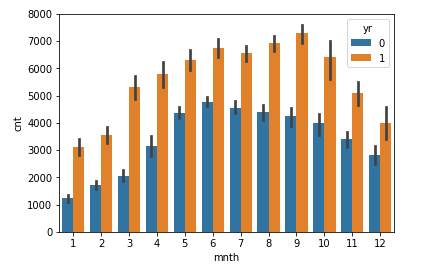
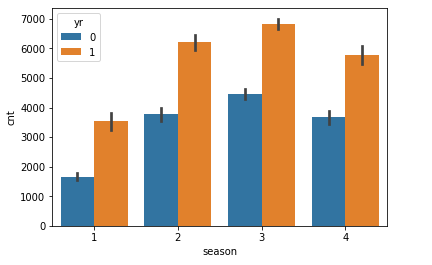


As seen from figure it contains total of 16 variables. Among them 13 are predictor variables and 3 are dependent variables. Here, our task is to predict “cnt” variable. So we have to build regression model.

# Chapter 2 Methodology

### Pre-Processing

First process in predictive analysis is exploratory data analysis. This involves exploring data, cleaning data and visualizing it through plots. Bike rental count depends on day, month, season, working day or weekday. Below figures shows the average rental count for different months, days, season etc., for two years.



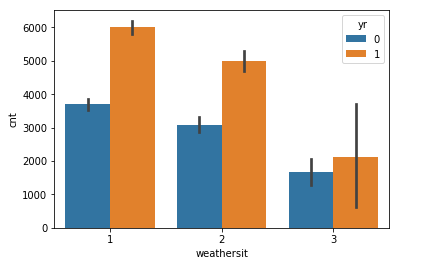
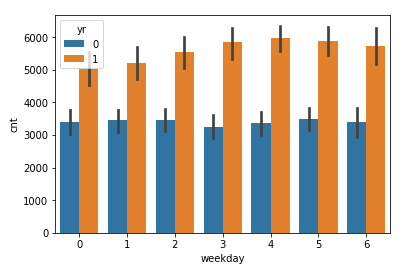


Figure 2.1 Average rental count for predictive variables for 2 years

In below figures we will see the variation of cnt for different temp, hum and windspeed for different season and year.

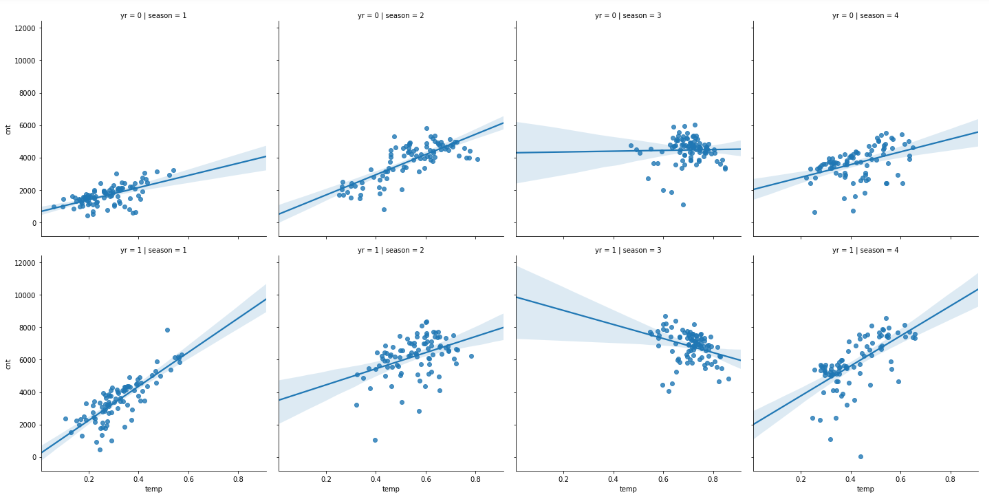


Figure 2.2 Variation of cnt w.r.t temperature

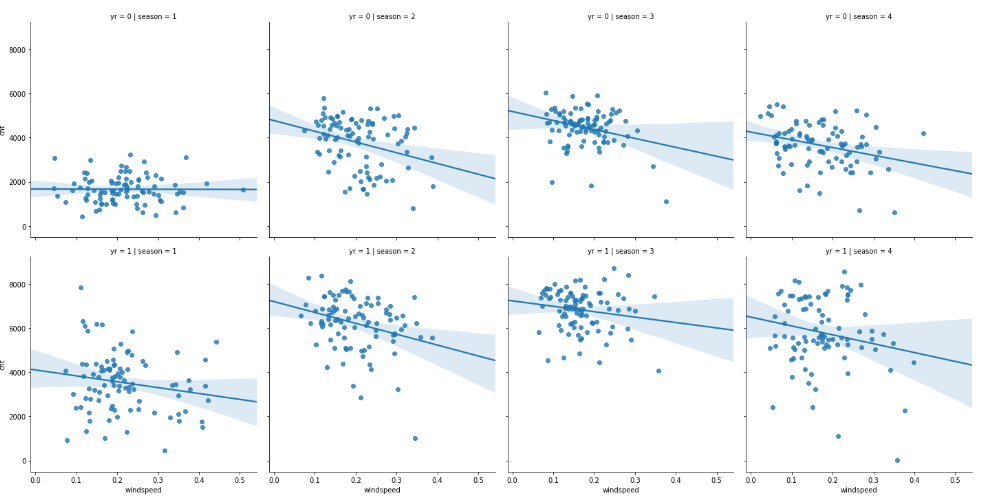


Figure 2.3 Variation of cnt w.r.t windspeed

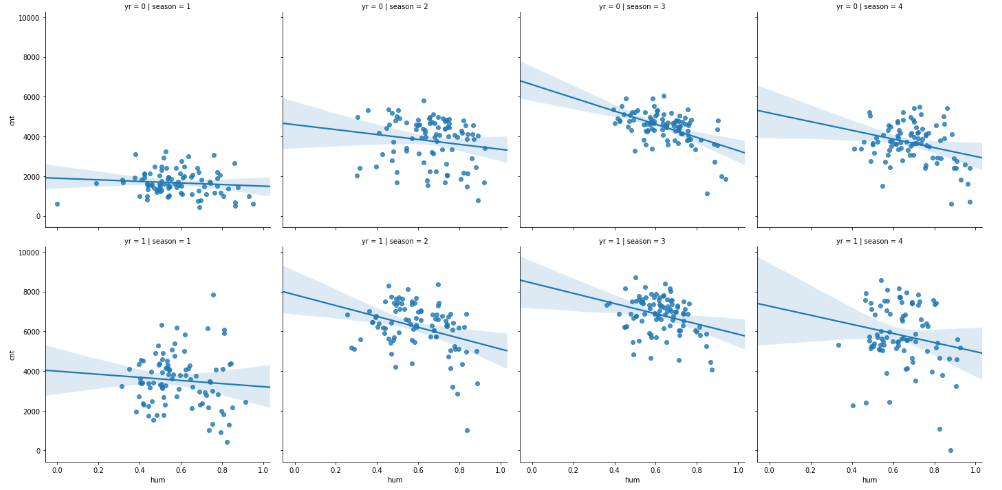


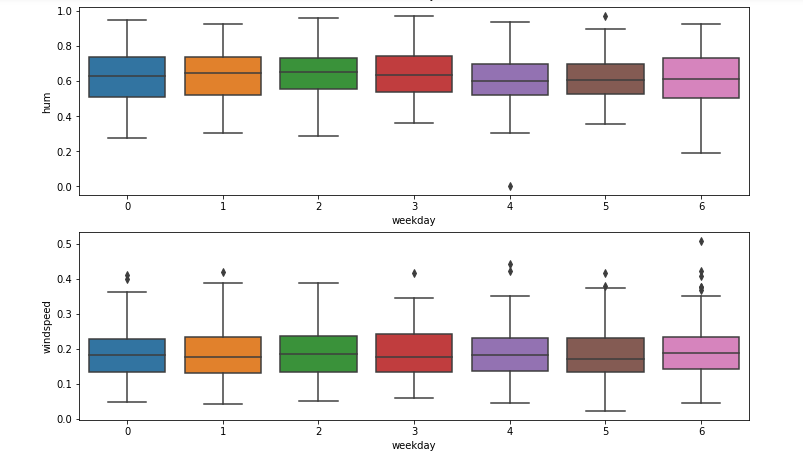
Figure 2.4 Variation of cnt w.r.t hum

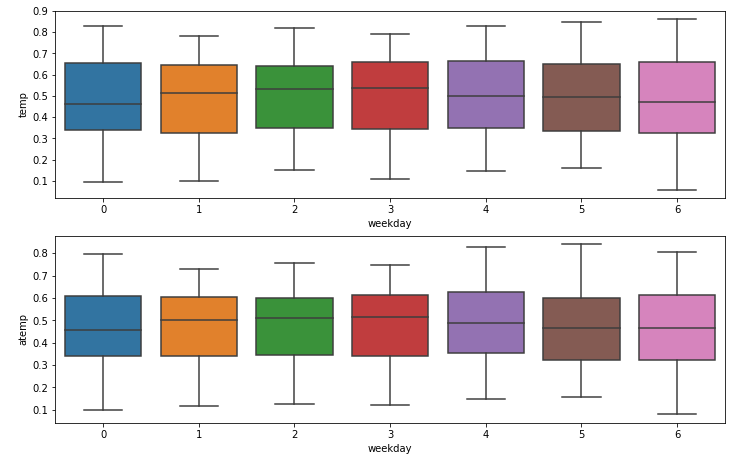
After visualizing data, next step involves missing value analysis. In the data provided as there are no missing values, we will proceed with the outlier analysis.

##### Outlier analysis

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. An outlier can be detected by plotting box plot on numeric variables.

Below figures shows outlier analysis done for numeric variables of data.





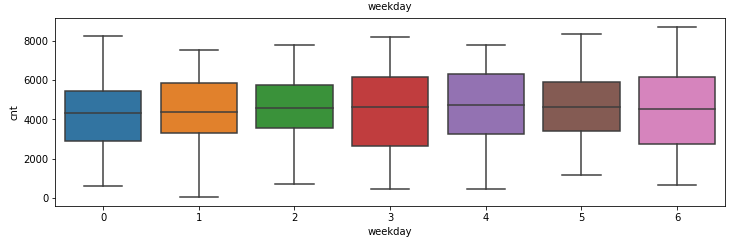


Figure 2.5 Box plots for numeric variables of data

For this data, as data provided is less, we will detect outliers and fill with NA and then impute using KNN imputation.

##### Feature selection

Feature selection is the process of selecting a subset of relevant features (predictors) for use in model construction. There is a possibility that many variables in our analysis are not important at all or may cause multicollinearity problem. Here, we use correlation analysis for numeric variables and chi square independence test on categorical variables.

In correlation analysis, we remove variables, if they are highly correlated.

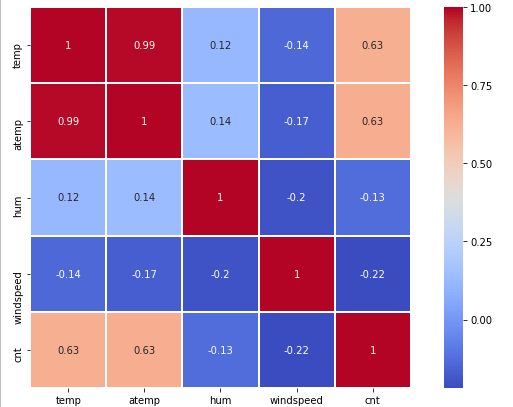


Figure 2.6 Correlation analysis on numeric variables.

As shown in figure 2.6, temp and atemp are highly correlated. We will remove this variable.

Chi square analysis assumes two hypotheses.

Null hypothesis: Two variables are independent.

Alternate hypothesis: Two variables are dependent.

If two variables have p-value less than 0.05(for 95% confidence) we reject Null hypotheses and say two variables are dependent and vice versa.

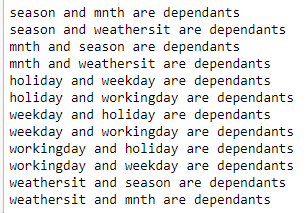


Figure 2.7 Chi square analysis results (in python)

As seen from Figure 2.7, all categorical variables depend on one another. This is expected as month, day, season and other variables are related in real world. So, no need of removing any variable from model.

But here, for building model, workingday variable has been removed as linear regression model has not developed coefficients for this variable.

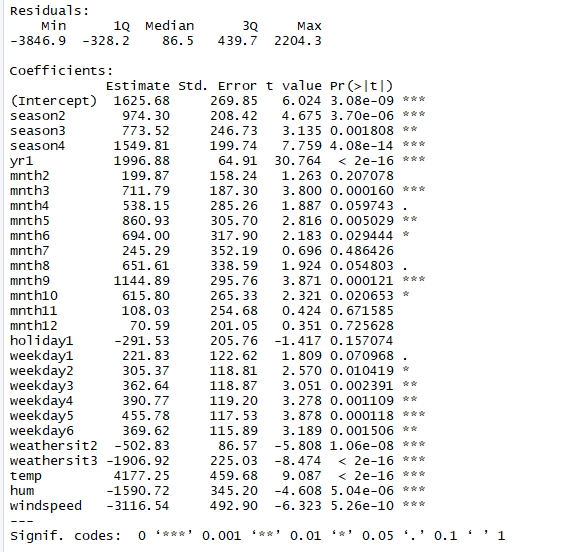
### Modeling

##### Model selection

As this is regression problem, we will first select linear regression model and then evaluate the performance.

##### Linear regression

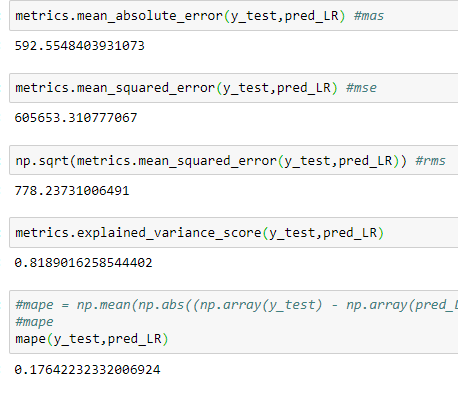
Below figure shows the summary of model applied on data





As we can see Adjusted R squared value is 0.8413, this model can explain 84% of data.

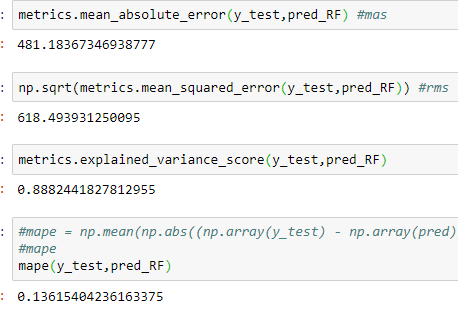
So, we can consider this model for evaluating with test data. After applying on test data, model performance is evaluated with mae, rmse, mape and mse. Evaluation results are shown below,



MAPE is 0.1764, accuracy of model is 83%. This model performed better on test data. But still we can consider other model and evaluate the results.

##### Random forest

This model is applied on data and evaluated results are shown below



As seen from results above, accuracy of model is 87%

# Chapter 3 Conclusion

### Model evaluation

As this is regression model, we will use RMS (root mean square), MAE (mean absolute error) and MAPE (mean absolute probability error) for evaluating the results.

RMS is calculated using below formula:



MAE is calculated as below



MAPE is calculated as below



### Model selection

As observed from analysis results, both the models performed good. But random forest performed better than linear regression model. So, we will consider random forest model for this data.

Below figures shows scatter plot for true values against predicted values for both models.

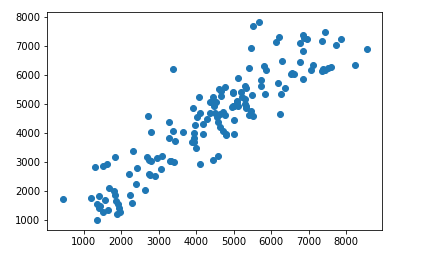


Figure 3.1 Predicted vs True values for linear regression

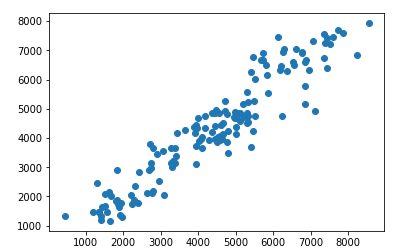


Figure 3.2 Predicted vs True values for Random forest

Below figures show distribution plots for difference between true and predicted values.

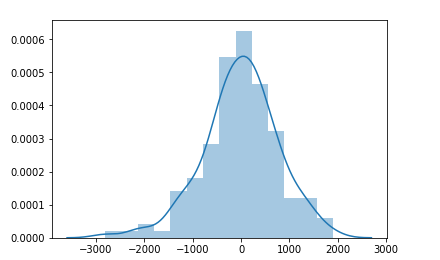


Figure 3.3 Distribution plot of (y\_test - pred) for Linear regression

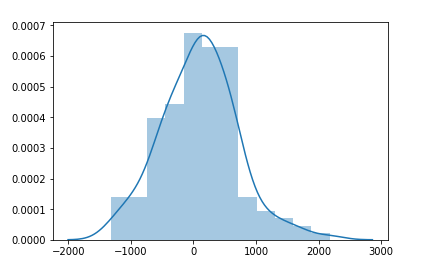


Figure 3.3 Distribution plot of (y\_test - pred) for Random forest