## Bike Renting Predictions

# Chandan Kapoor

**19 March 2020**

**Table of Contents**

Sr No. Title Pg No.

1. Introduction 3
   1. [Problem statement 3](#_TOC_250014)
   2. [Data 3](#_TOC_250013)
2. Methodology 4
   1. [Pre-Processing 4](#_TOC_250012)
      1. [Missing Value Analysis 5](#_TOC_250011)
      2. [Outlier Analysis 5](#_TOC_250010)
      3. Feature Selection 9
   2. Modeling 9
      1. Decision Tree Regression 12
      2. Random Forest Regression 13
      3. Linear Regression 13
3. Conclusion 15
   1. Model Evaluation…………………………………………………………………..15
      1. MAE……………………………………………………………………………..15
      2. MAPE……………………………………………………………………………15
      3. Root Mean Square Error (RMSE) 16
   2. Model Selection…………………………………………………………………………………………16

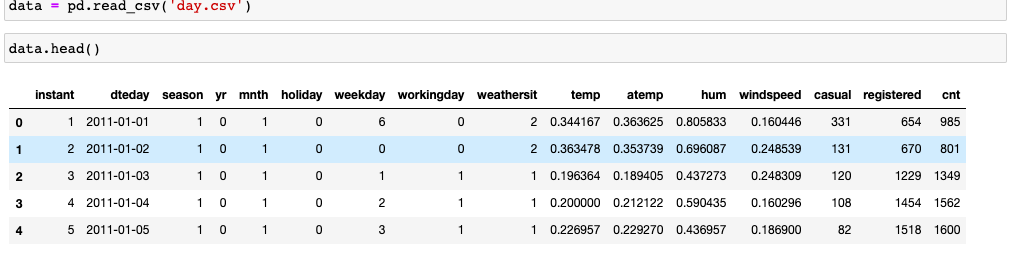
## Chapter 1 Introduction

### Problem Statement

In order to provide better accessibility to the customer a Bike Renting firm wants to know number of bikes rented depending on the environmental and seasonal settings so that approximate number of bikes can be made available. This will improve its operational efficiency which eventually lead in customer satisfaction and increase in sales over period.

### Data

Our task is to build a model which will predict the number of bikes rented on day to day basis based on environmental and seasonal changes. Given below is a sample data set which we will be using for the model development and prediction of number of bikes rented:



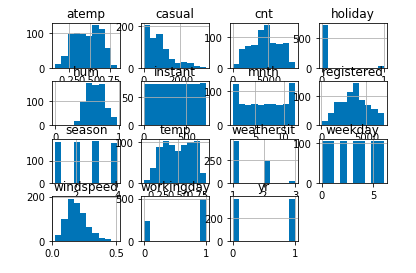
As you can see in the data, we have 7 categorical variables (season, year, month, holiday, weekday, working day, weathersit), 6 predictor variables (temp, atemp, hum, windspeed, casual, registered) and 1 Output or Dependent variable (cnt).

## Chapter 2 Methodology

### Pre Processing

Before starting predictive modelling, we need to check the data and understand various data definitions. Also, we need to check whether data is in proper format before developing model and feeding the data, if data is not in proper format, we need to clean the data in order to increase the efficiency of the model. To start the process, we have basic steps which needs to be completed such as data cleaning, visualization using graph or plots, standardizing if data is not uniformly distributed these steps are a part of Exploratory Data Analysis.

We have plotted various graphs in order to understand the uniformity and visualize data for better understanding. We also compare the independent variables with dependent variable in order to understand its contribution in the prediction of target or dependent variable. If we find two or more variables with similar data, we keep only one and remove other with similar data in order to lower the complexity of model.

**

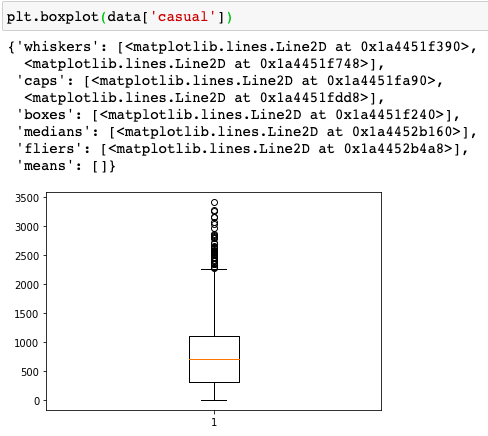
### Missing Value Analysis

We always check whether dataset contains any missing value. If it contains missing value, we need to impute missing observations in a variable if it has at least 70% of data. We impute the missing data with the help of mean, median in case of continuous variable and mode in case of categorical variable. We can also use K Nearest Neighbor (KNN) in order to impute missing values.

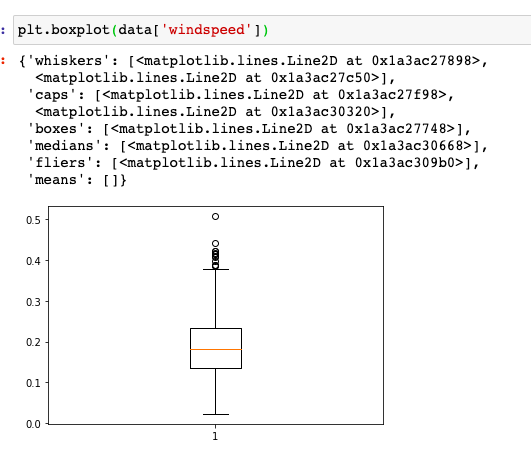
.

### Outlier Analysis

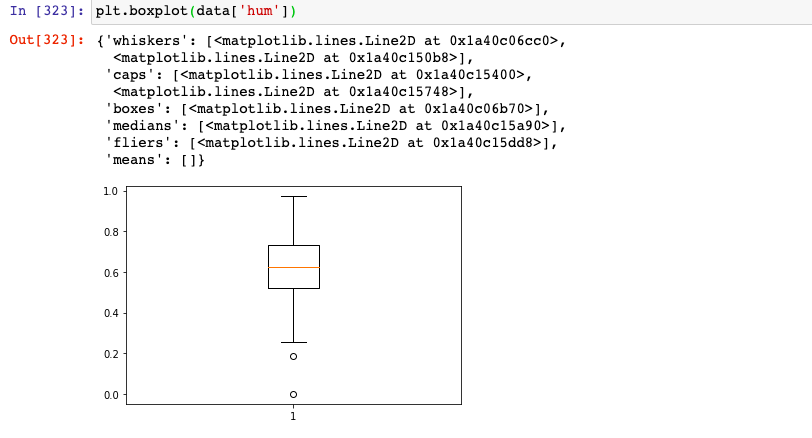
When I started detecting the variables the major outliers were present with casual, humidity and windspeed variables



Boxplot for casual



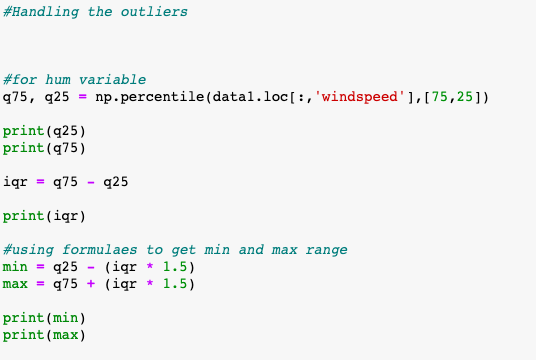
Boxplot of windspeed



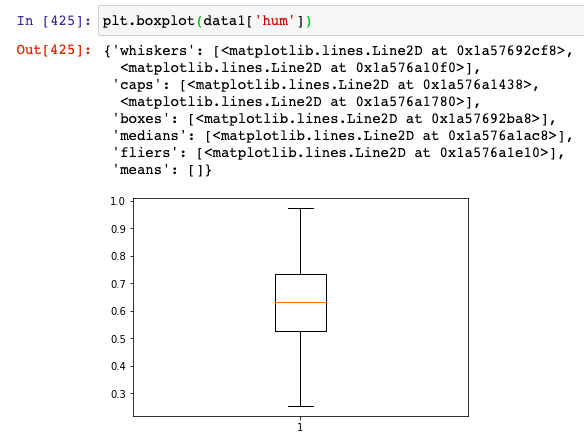
Boxplot for humidity

How did I implement my boxplots?

The formulae for IQR which is given helps us to divide the given data into different parts and makes us know what the values which are outliers in a dataset.



As we have successfully detected outliers in above variables in the dataset, we will now remove them from the data and re plot the graphs as below:



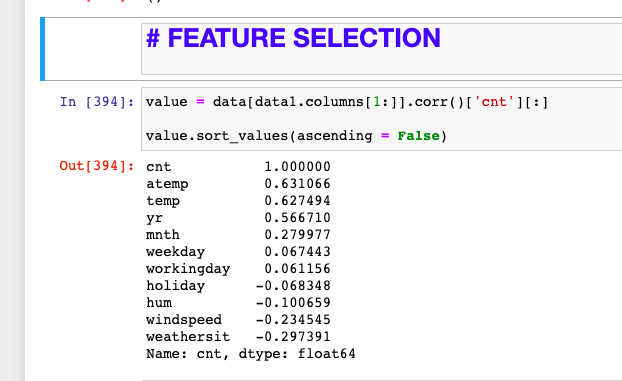
Boxplot for humidity

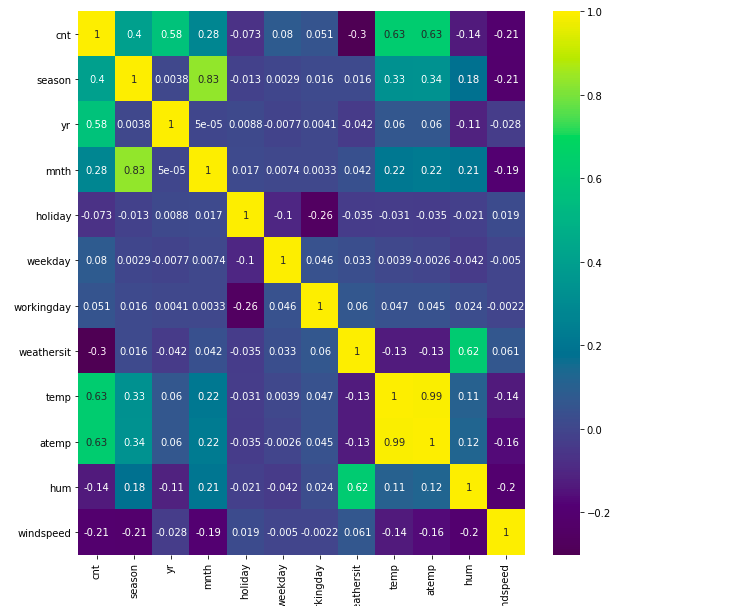
Our data is now free from outliers and we can proceed to next step that is feature selection.

### Feature Selection

Before running the any machine learning model, we need to check the variables which are used for the prediction. It happens that sometimes most of the variables in the dataset do not contribute much in the prediction of target or dependent data. In some cases, it can also happen two or more data may pass the similar information to the model which also may increase in the complexity of the model. In order to avoid these issues, we do feature selection and select variables which are contributing in the prediction of dependent variable.

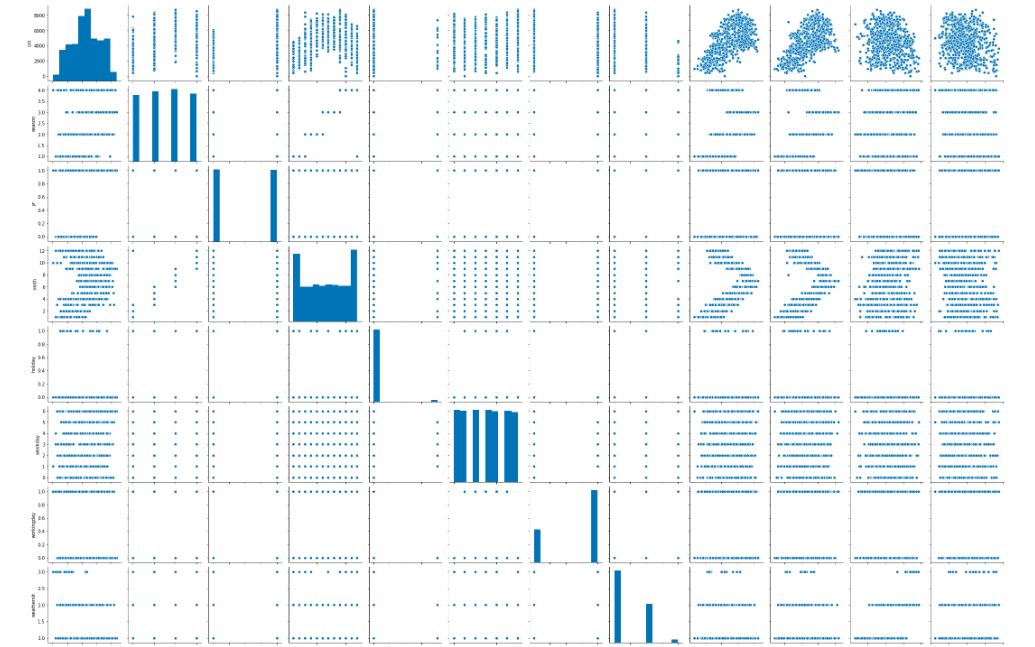
We will plot the correlation matrix between all the numeric variables to understand the relation between them as shown below:

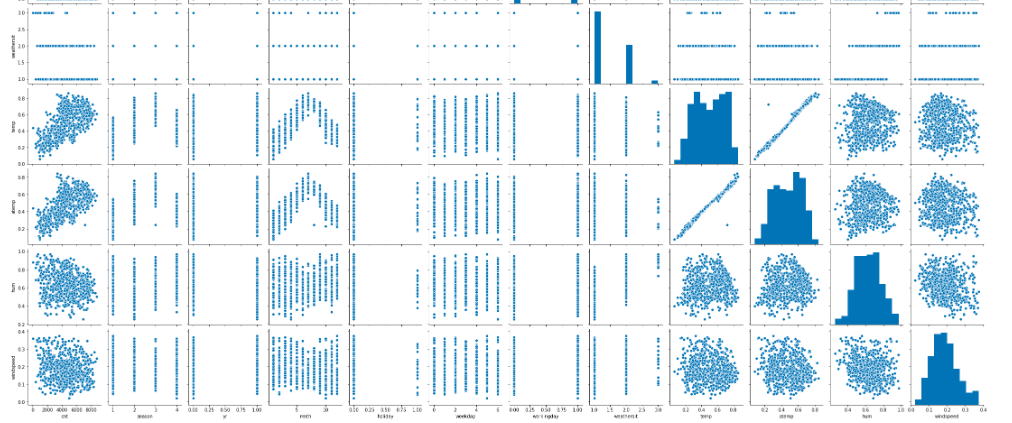
**



From the above matrix plotted amongst the numerical variables in the dataset we can see that independent variables temp and atemp are positively correlated and carry similar information so we will proceed with only one variable amongst temp and atemp.

Another way of checking feature is using Seaborn library which is well known visualization library. Seaborn has a defined function which is known as *pairplot()* using which we can plot multiple variables plot together and identify their trend. Below is the pairplot for bike rental dataset. As also seen in below figure we can look at the temp feature and atemp feature they are containing almost similar data as we checked earlier from the correlation matrix Therefore, we can exclude any one feature from temp and atemp. In our prediction we will be considering the temp variable.





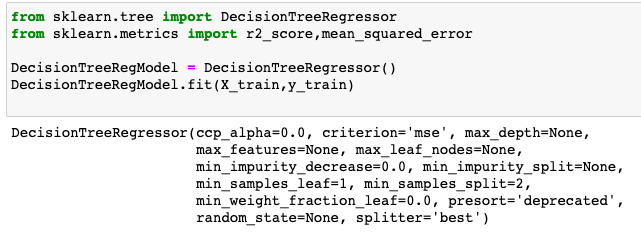
### Modelling

Once we clean the data in the data exploration phase, we can proceed with the model development. After data exploration initial phase, we came to know few points about our dataset such as categorical variables and continuous variables, dependent and independent variables. We also know our target or dependent variable is continuous since we need to predict the number of daily bikes rented. Therefore, our problem is a regression problem. We will be using Decision tree regression, Random forest and Linear Regression algorithm.

While in r I used decision tree svm and linear regression.

### Decision Tree Regression Algorithm

Decision tree is a predictive model based on branching series based on the Boolean tests. It can be used for both classification and regression problems. There are number of different types of decision trees that can be used in the machine learning algorithms. In machine learning algorithms in case of data mining we always must train the model using our data and post that we need to check accuracy of the model to predict the data. Therefore, we need to split our cleaned dataset into two parts train and test. From sk.learn.model\_selection we can use the defined function train\_test\_split to do the required task. It is always better to divide the data into 67:33 which is train and test data respectively.

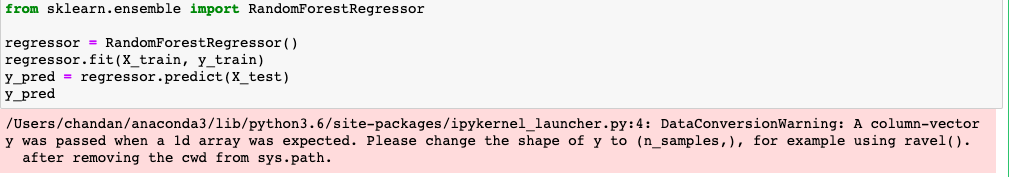


Below code in python is used to run the Decision tree algorithm on our bike renting dataset. Sklearn.tree is a package which has DecisionTreeRegressor function which we can use to implement the model. In next line of code we specify maximum depth as 15 wherein we inform the model that do not proceed beyond 15th level followed with the fit function which has independent variables and the dependent variable.

Once the model is trained, we run it on the test data to check the predictions and compare the predicted data with the actual dependent variable.

### Random Forest Algorithm

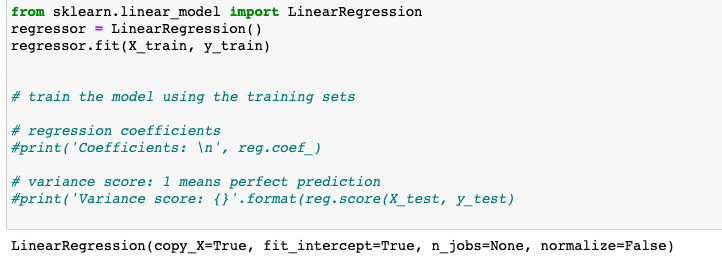
Algorithm uses the CART decision tree algorithm to generate the random forest. It is an ensemble which consists of many decision trees. Random forest uses the mean for regression problems, and it can be used for both classification and regression problems. Below is the Random forest model implementation python code. We can use the RandomForestClassifier function defined in the sklearn.ensemble library. In the function n\_estimators is the number of trees to be used in the forest. A sub optimal greedy algorithm is repeated several times using random selections of features and samples. The random\_state parameter allows controlling these random choices. Followed with these we need to provide the independent and dependent variable in the fit function to train the model.

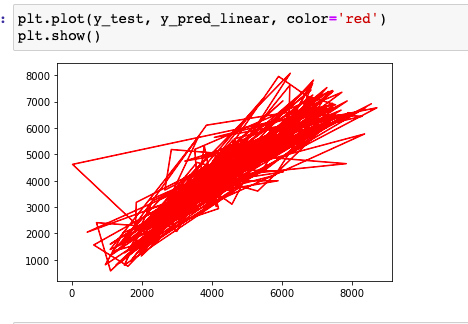


Similarly, as in the decision tree algorithm we predict the data using fit.predict() function by passing the dependent variables. Finally, these predicted values are compared with actual target variable to check the accuracy of the model.

### Linear Regression

Linear regression is one of the statistical models and it can only be used for regression problems. In linear regression we have two types of model’s simple linear regression model and multiple linear regression model. We can use the statsmodels.api library which has the linear regression function OLS defined. Below is the code used in python for implementation of Linear Regression.





*Fig 2.10 Summary of Linear Regression Model*

## Chapter 3 Conclusion

### Model Evaluation

After generating predictions from few models as above we need to finalize and select the best performing and efficient model. Mostly models are compared based on below criteria:

* + - Predictive Performance
    - Interpretability
    - Computational Efficiency

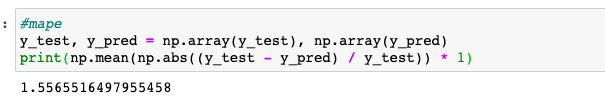
In our case of bike renting data since we have predicted the target variable, we will judge the models based on their predictive performance. Predictive performance can be measured using calculate average measures such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE).

### Mean Absolute Error (MAE)

MAE is the average of the absolute errors. We will not calculate this since people are more prone to understand data in terms of percentages instead of numbers.

### Mean Absolute Percentage Error (MAPE)

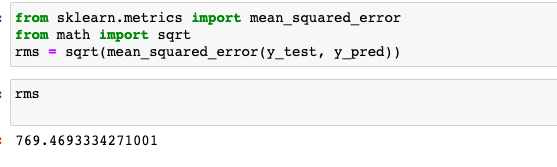
MAPE is the percentage version of MAE and it measures accuracy as a percentage of error. Below is the code snippet from implementation using python.

**

**Mape function**

### 3.1.2 Root Mean Square Error (RMSE)

RMSE is the squaring of errors, finding their averages and taking the square root. It is a time based measure. Below is the code snippet from implementation using python.

**

### Model Selection

We can see that all the models perform comparatively on average and therefore we can select either one without any loss of information.

References

Edwisor study material

Analytics Vidya

Data Camp