Bike Rental

Eeshan Gupta

20 *November* 2018

Contents

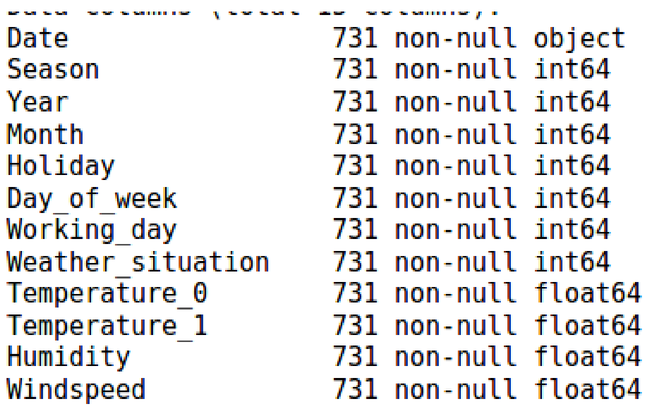
Chapter 1

Introduction

The number of rentals required by a rental company on a given day is an important feature to know. If the prediction of this value can be done, then the business can be developed according to this number. Gains can be maximized and losses can be minimized if the future requirement is already known to the management. This case study deals with a case of bike rental company, which procures bike to its customers.

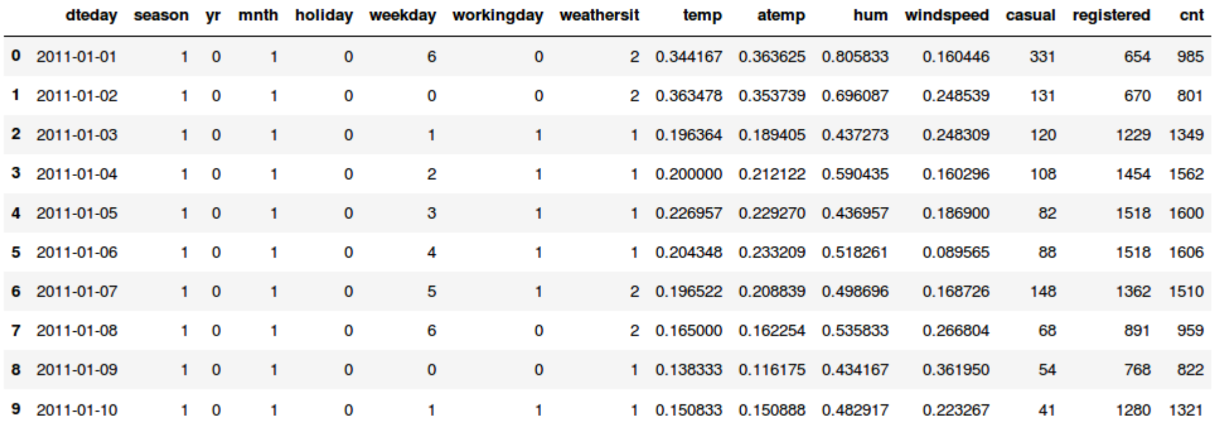
1.1 Problem Statement

The objective of this case study is to predict number of bikes rented out to customers daily. The number of bikes rented is predicted on the basis of environmental ans seasonal settings of the day. The prediction is based on machine learning algorithms which can predict a number based on the type of day. The type of day is decided by the following parameters:



```
Data columns (total 15 columns):
Date                  731 non-null object
Season                731 non-null int64
Year                  731 non-null int64
Month                 731 non-null int64
Holiday               731 non-null int64
Day_of_week           731 non-null int64
Working_day           731 non-null int64
Weather_situation     731 non-null int64
Temperature_0         731 non-null float64
Temperature_1         731 non-null float64
Humidity              731 non-null float64
Windspeed             731 non-null float64
```

Figure 1.1: Predicting Variables.



	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
5	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
6	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
7	2011-01-08	1	0	1	0	6	0	2	0.165000	0.162254	0.535833	0.266804	68	891	959
8	2011-01-09	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.361950	54	768	822
9	2011-01-10	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321

Figure 1.2: 10 Instances of the data.

The customers who rent the bikes are divided into two categories namely, Casual users and Registered users. Registered users represent the users who are regulars to the shop, while casual users are the one who randomly walked into the shop to rent a bike.

The following section explains the procedure that is followed during the development of the model.

Chapter 2

Methodology

2.1 Data

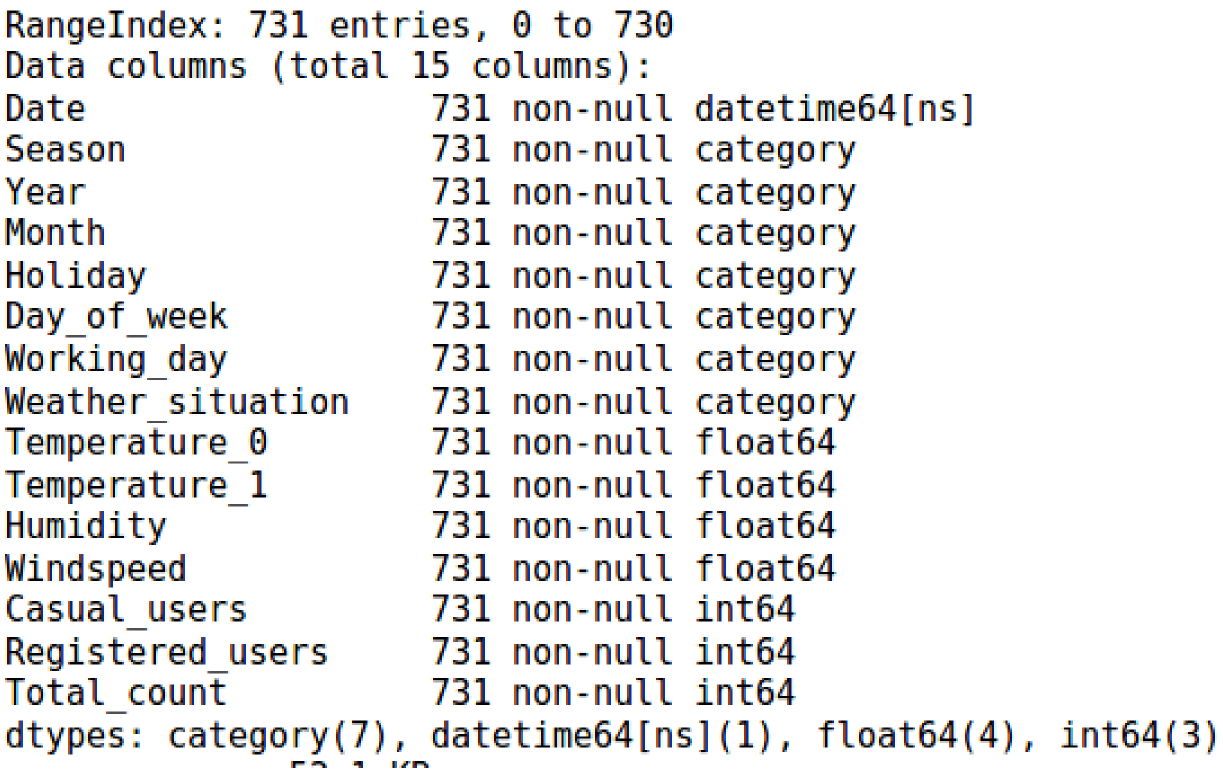
The data is provided as a single file which consisted of 731 instances. The seasonal and environmental setting of the day is given by features such as the season, year, month, whether the day is holiday or working day,the weather situations, the temperature of the day, the humidity and the wind-speed. Rest of the features describe the type of user and the number of bikes rented by each type of users. The summation of the bikes rented by each user type is total count of bikes rented, which is also the predictor.

2.2 Pre - processing

The provided data is required to pre-processed before we apply some machine learning algorithm to it. This is an integral part of the analysis.
In the analysis, the first step is to select the data-type of the feature correctly. Selection of a correct data-type will allow the analysis to be more precise and will help in understanding the underlying concepts more easily.

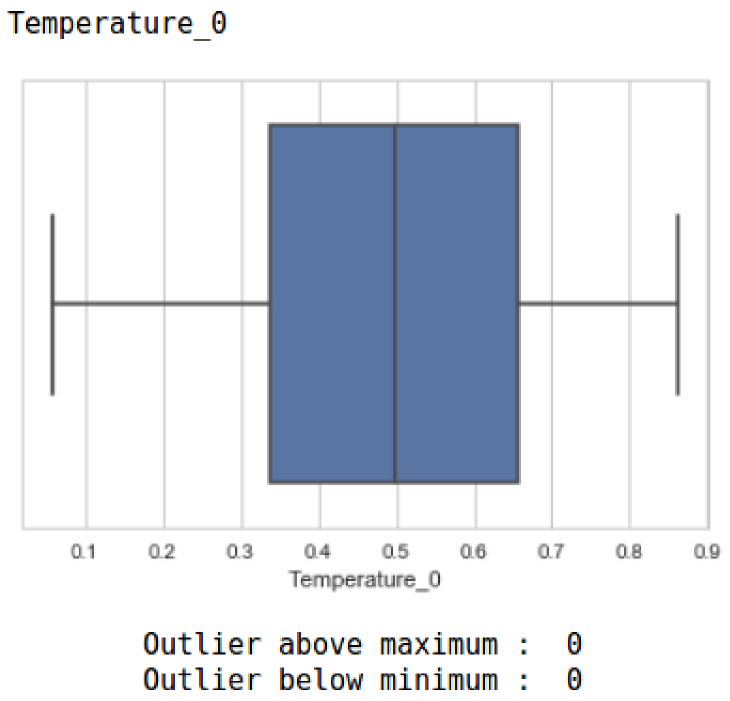
2.2.1 Missing Value Analysis

The missing values in the data are required to be found out and imputed before we proceed to any other step. These missing value play an important role, because if even a singe value is missing in any of the fields for any data point, all the learning will not be as good as it would be when there are no missing values.
For this data set we have no missing values present. So we proceed to the next step step.



```
RangeIndex: 731 entries, 0 to 730
Data columns (total 15 columns):
Date                     731 non-null datetime64[ns]
Season                   731 non-null category
Year                     731 non-null category
Month                    731 non-null category
Holiday                  731 non-null category
Day_of_week              731 non-null category
Working_day              731 non-null category
Weather_situation        731 non-null category
Temperature_0            731 non-null float64
Temperature_1            731 non-null float64
Humidity                 731 non-null float64
Windspeed                731 non-null float64
Casual_users             731 non-null int64
Registered_users         731 non-null int64
Total_count              731 non-null int64
dtypes: category(7), datetime64[ns](1), float64(4), int64(3)
```

Figure 2.1: Missing Values



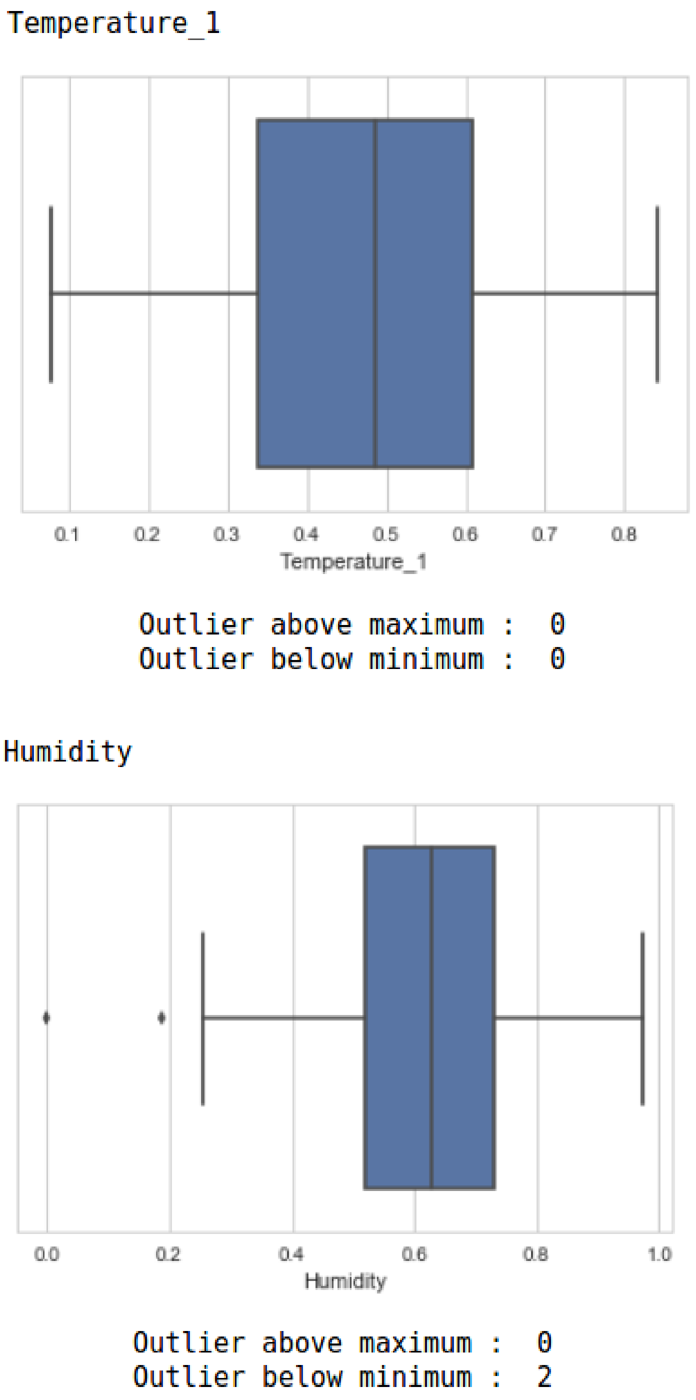
```
Temperature_0
```

```
Outlier above maximum :  0
Outlier below minimum :  0
```

Figure 2.2: Some visualization of outlier analysis

2.2.2 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.



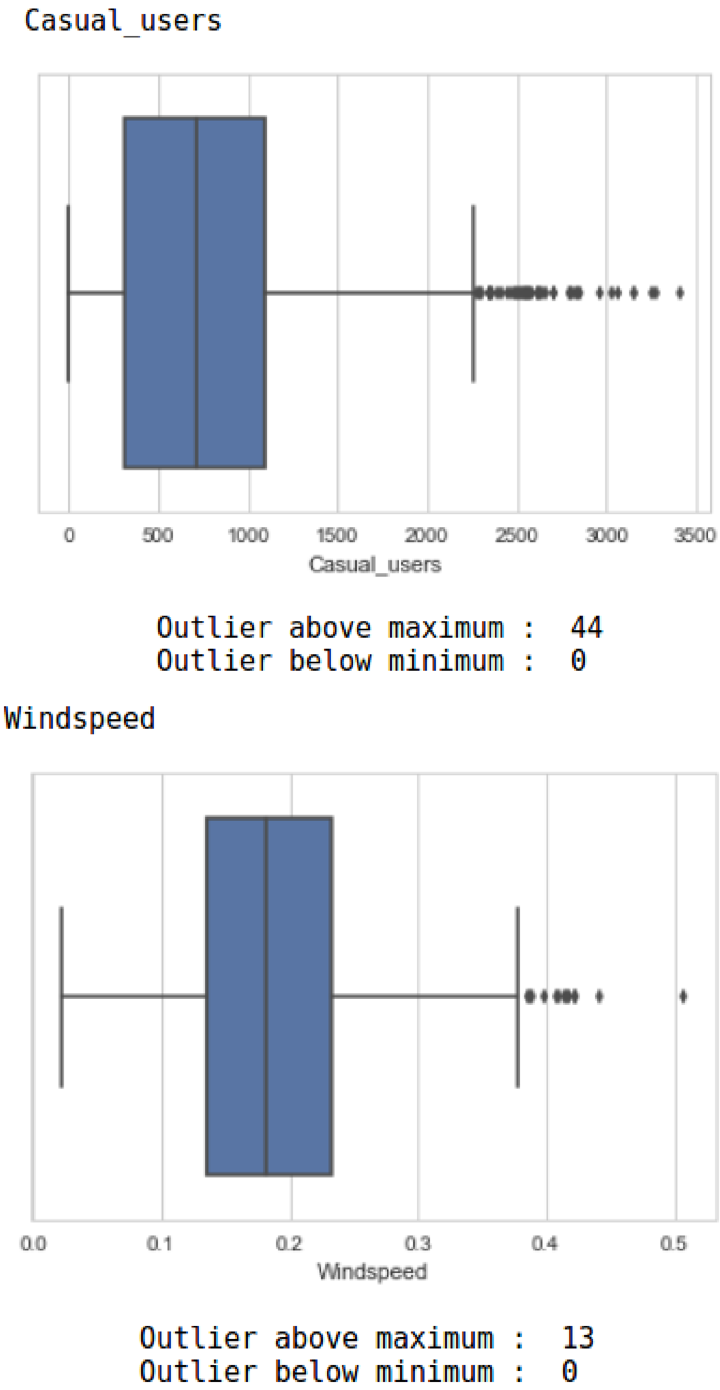
```
Temperature_1
```

```
Outlier above maximum :  0
Outlier below minimum :  0
```

```
Humidity
```

```
Outlier above maximum :  0
Outlier below minimum :  2
```

Figure 2.3: Some visualization of outlier analysis



```
Casual_users

        Outlier above maximum :  44
        Outlier below minimum :  0

Windspeed

        Outlier above maximum :  13
        Outlier below minimum :  0
```

Figure 2.4: Some visualization of outlier analysis

2.2.3 Feature Selection

In statistics, the variance inflation factor (VIF) is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.
Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multi-collinearity if there was no correlation with other predictors. A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.

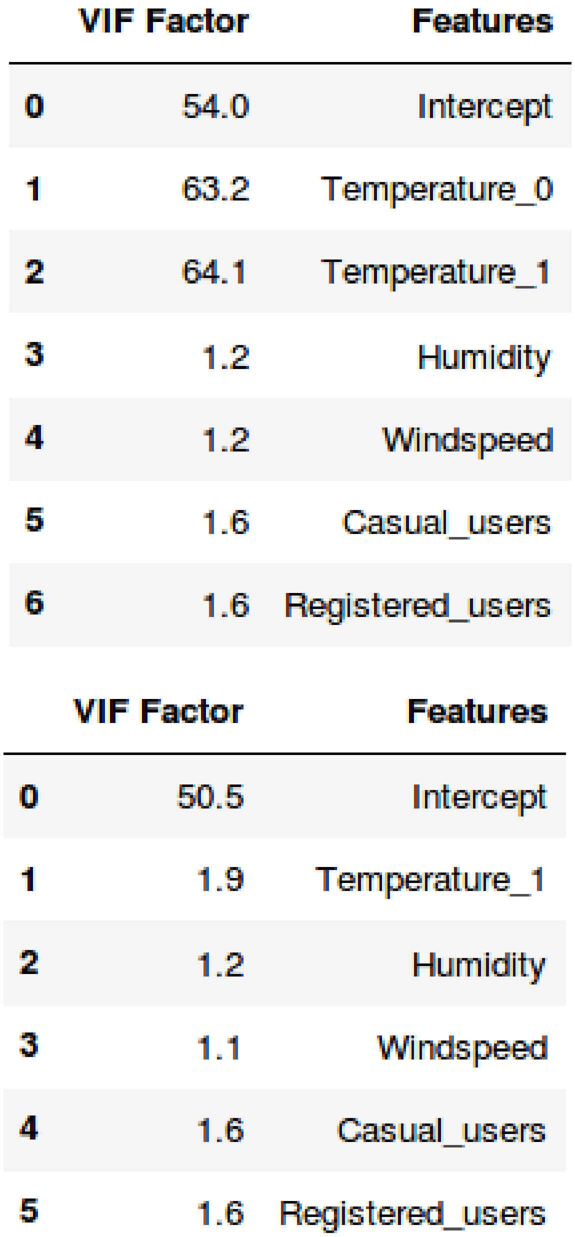
- Between 1 and 5 = moderately correlated.

- Greater than 5 = highly correlated.

The results of the VIF analysis are demonstrated in fig 2.5.

2.2.4 Feature Scaling

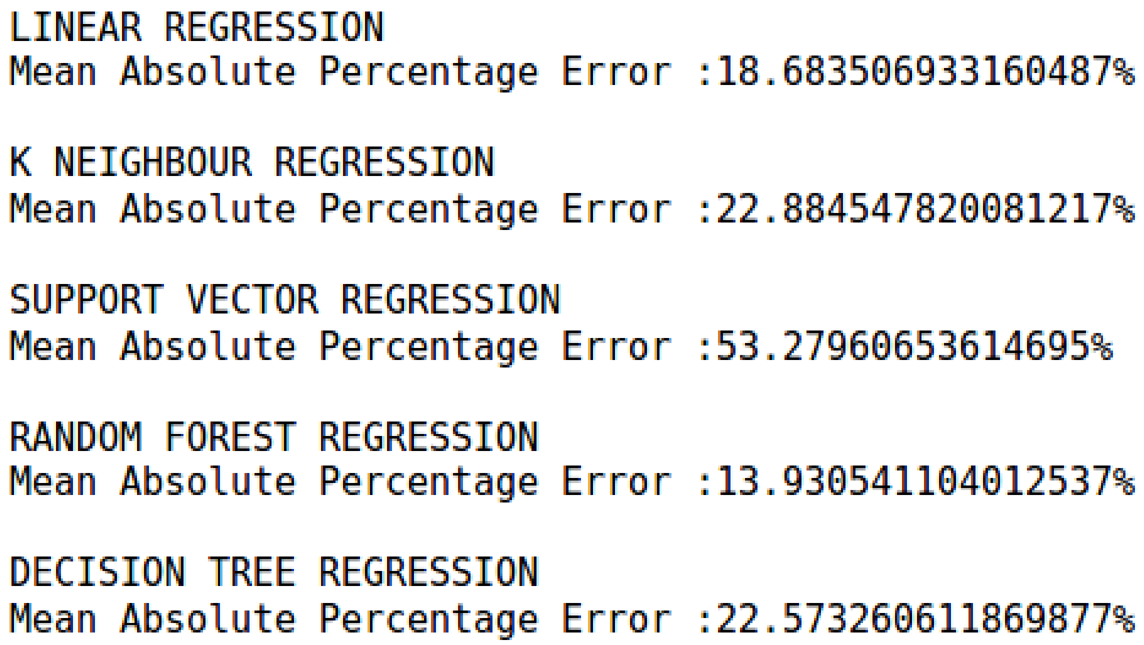
The numerical features are scaled using *Standardization* technique, where the mean is of the feature is brought down to 0 and standard deviation is the distribution is 1.



	VIF Factor	Features
0	54.0	Intercept
1	63.2	Temperature_0
2	64.1	Temperature_1
3	1.2	Humidity
4	1.2	Windspeed
5	1.6	Casual_users
6	1.6	Registered_users

	VIF Factor	Features
0	50.5	Intercept
1	1.9	Temperature_1
2	1.2	Humidity
3	1.1	Windspeed
4	1.6	Casual_users
5	1.6	Registered_users

Figure 2.5: Variance Inflation Factor.



```
LINEAR REGRESSION
Mean Absolute Percentage Error :18.683506933160487%

K NEIGHBOUR REGRESSION
Mean Absolute Percentage Error :22.884547820081217%

SUPPORT VECTOR REGRESSION
Mean Absolute Percentage Error :53.27960653614695%

RANDOM FOREST REGRESSION
Mean Absolute Percentage Error :13.930541104012537%

DECISION TREE REGRESSION
Mean Absolute Percentage Error :22.573260611869877%
```

Figure 2.6:

2.3 Model Selection

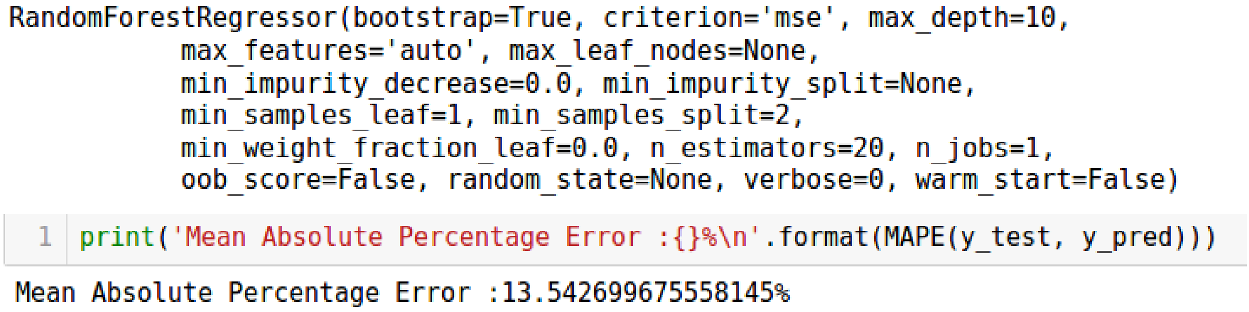
According to the problem statement, the environmental and seasonal settings of a day are given as the features to predict an estimate of number of bikes required in that particular day by the shop. This estimate of of number of bikes is a continuous variable and hence the problem becomes a regression problem. In a regression problem, the features are used to predict a numerical output, by means of learning the underlying relation of each feature.

In this analysis, five algorithms are chosen and compared for their learning ability. The comparison is done by calculating the *Mean Absolute Percentage Error* of each algorithm and then choosing the one with least error value.

- Linear Regression
- K Neighbour Regression
- Decision Tree Regression
- Support Vector Regression
- Random Forest Regression

The following figure 2.6 depicts the results of running the algorithm.

Random Forest Regression is the algorithm that is chosen to predict the estimate of bikes required at the shop for purposes of renting them out.



```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=10,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=20, n_jobs=1,
           oob_score=False, random_state=None, verbose=0, warm_start=False)
```

```
print('Mean Absolute Percentage Error :{}%\n'.format(MAPE(y_test, y_pred)))
```

```
Mean Absolute Percentage Error :13.542699675558145%
```

Figure 2.7: Random Forest Regression

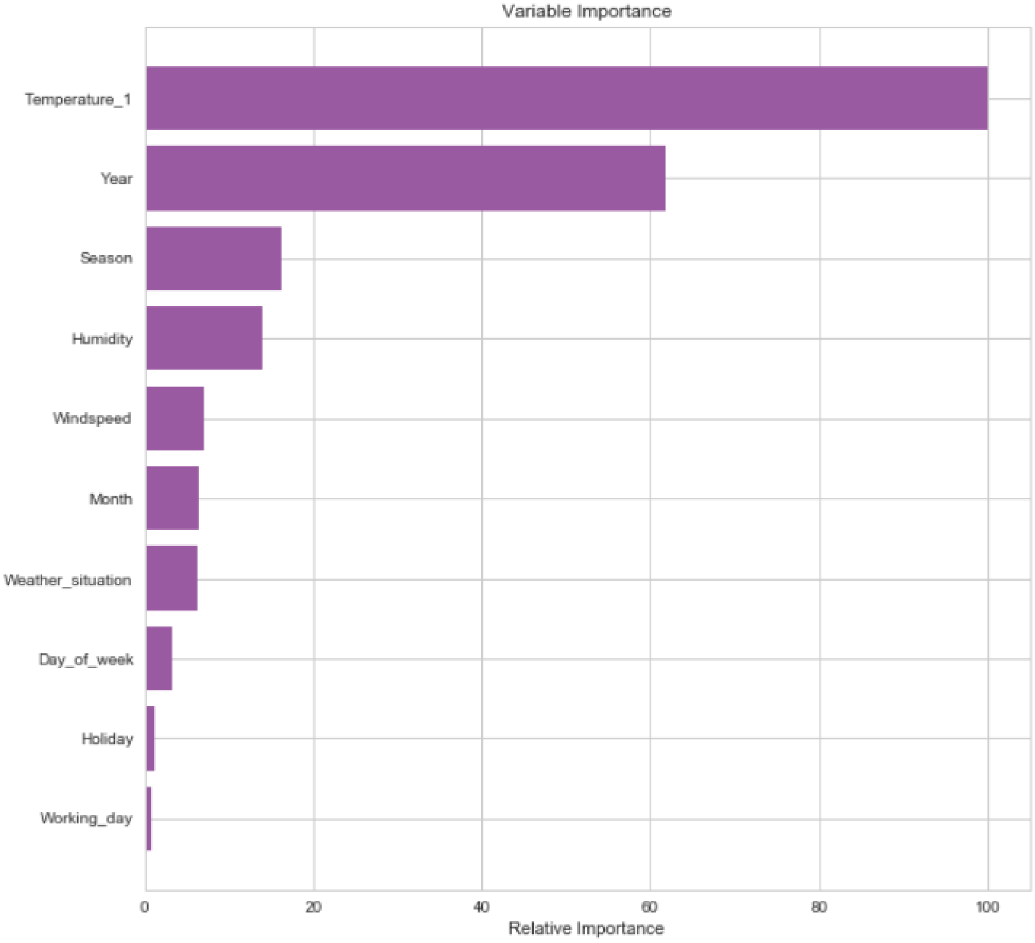


Figure 2.8: Variable Importance