**Project Name: Bike Renting**

**-Abhishek R**

**Index**

**1)Problem statement**

**2)EDA**

**3)Preprocessing**

**a)Missing Data Analysis**

**b)Outlier Analysis**

**c)Feature Scaling**

**d)Feature Selection**

**4)PCA**

**5)Modeling**

**6)Hyperparameter Tuning**

**7)Conclusion**

**Problem statement**

The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

The details of data attributes in the dataset are as follows –

1)instant: Record index

2)dteday: Date

3)season: Season (1:springer, 2:summer, 3:fall, 4:winter)

4)yr: Year (0: 2011, 1:2012)

5)mnth: Month (1 to 12)

6)hr: Hour (0 to 23)

7)holiday: weather day is holiday or not (extracted fromHoliday Schedule)

8)weekday: Day of the week

9)workingday: If day is neither weekend nor holiday is 1, otherwise is 0.

10)weathersit: (extracted fromFreemeteo)

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

11)temp: Normalized temperature in Celsius. The values are derived via

(t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale)

12)atemp: Normalized feeling temperature in Celsius. The values are derived via

(t-t\_min)/(t\_maxt\_min), t\_min=-16, t\_max=+50 (only in hourly scale)

13)hum: Normalized humidity. The values are divided to 100 (max)

14)windspeed: Normalized wind speed. The values are divided to 67 (max)

15)casual: count of casual users

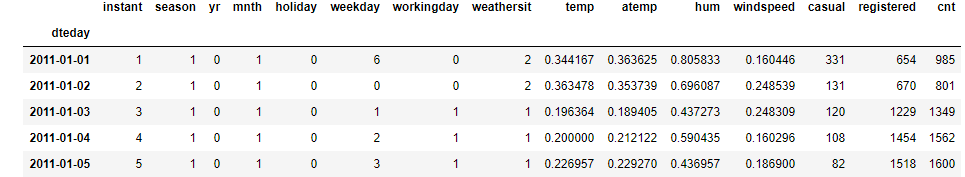
16)registered: count of registered users

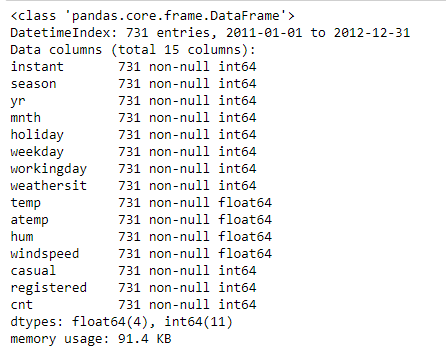
17)cnt: count of total rental bikes including both casual and registered

**EDA**

After importing the data, first lets look at the sample of data set that we have received and perform some basic Exploratory data analysis. EDA is a term for certain kinds of initial analysis and findings done with data sets, usually early on in an analytical process. Some experts describe it as “taking a peek” at the data to understand more about what it represents and how to apply it. Exploratory data analysis is often a precursor to other kinds of work with statistics and data.

Data was imported as “dateday” as index, lets take a look at first five rows of data to get started with EDA:

****

****

From the above table we can see that our data has 731 rows and 15 columns. Most of the columns are int64 or float64. We will have to classify them properly. Also we can see that there are no missing values.

By looking at the unique values the following variables should be classified as categorical:

season, yr, mnth, holiday, weekday, workingday, weathersit

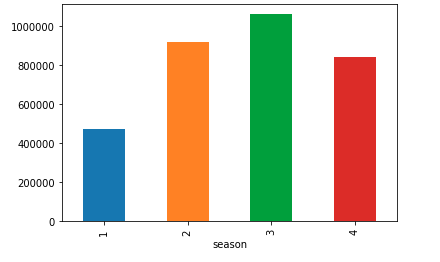
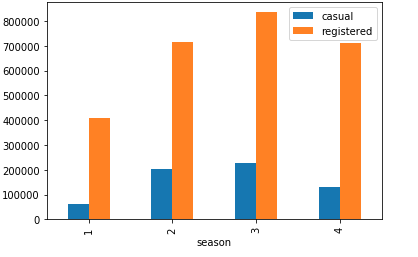
The following variables should remain numerical:

temp, atemp, hum, windspeed, casual, registered, cnt

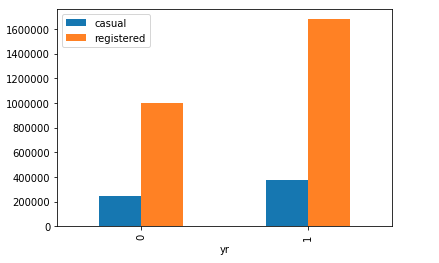
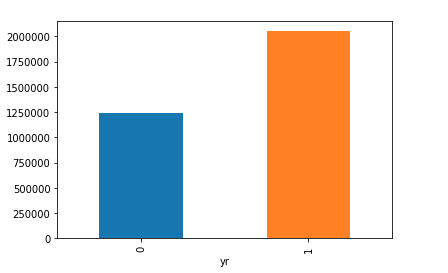
Also we can see that “casual”, “registered”, “cnt” are our target variables and the remaining columns are our features. Next we’ll see what we can make out of data by visualizing it.

**Visual EDA**

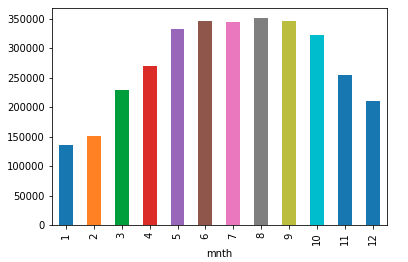
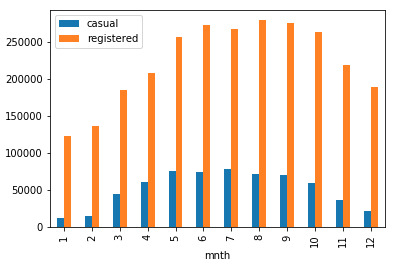
Perhaps one of the greatest disparities between those who live in the world of data science and those who don’t is the ability to translate the data so that everyone can understand it. One of the best ways for data scientists to present analysis of the data to those outside the industry is by generating visualizations. By visualizing data we can understand the data even better. Lets visualize and try to get a better sense of data:

****

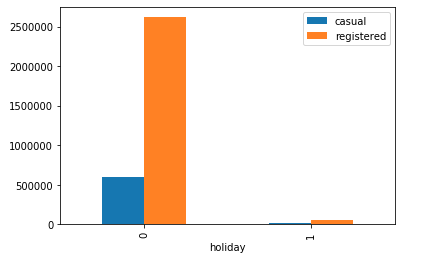
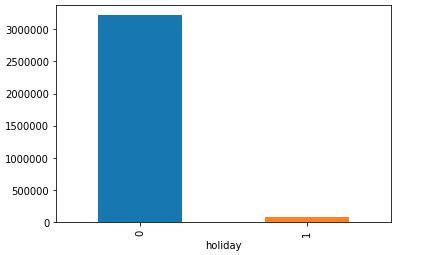
We can notice that winter count of rental bikes was highest in fall followed by summer and winter with springer having lowest count of rental bikes.

****

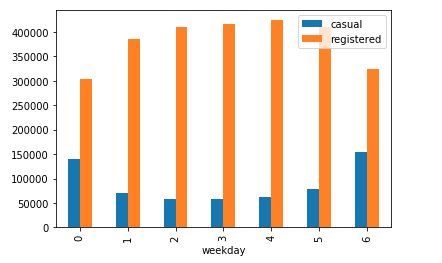
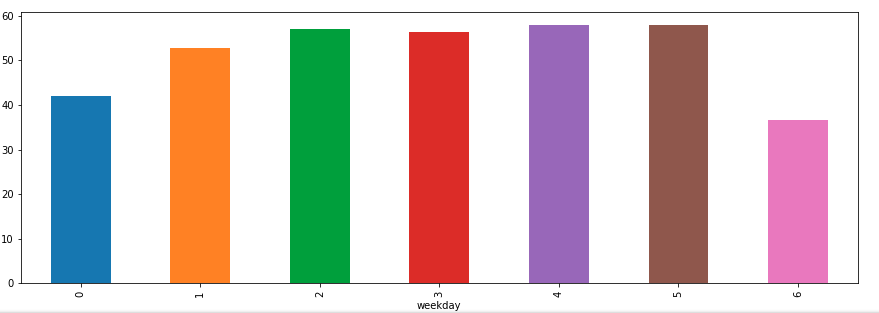
We can notice that 2012 has an increase in count of bike rentals as compared to year 2011.

****

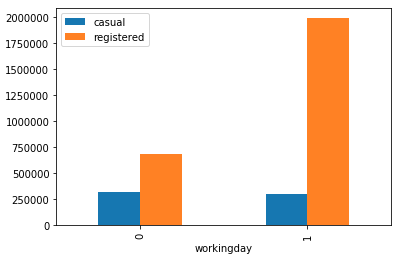
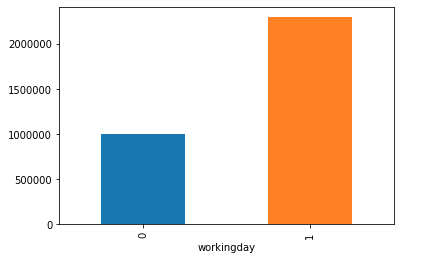
We can notice that August seems to be the month with highest count of bike rentals. More importantly count of bike rentals seems to be high from May to October.

****

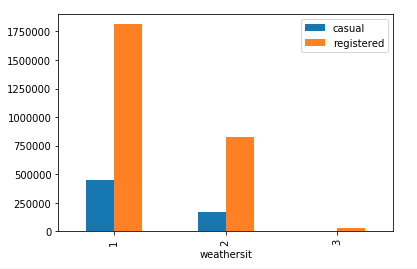
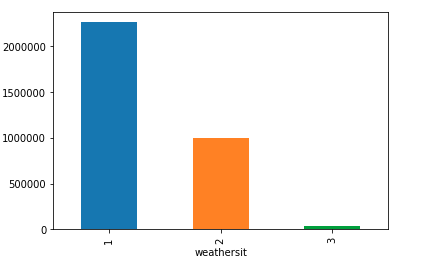
We can see that count of bike rentals on holidays are very less compared to bike rentals on days which were not holidays.

****

We can notice that count of bike rentals are comparatively less on weekends as opposed to other weekdays.

****

We can see that count of bike rentals are very high on weekdays as compared to count of bike rentals on weekends and holidays.

****

We can see that count of bike rentals are high when weather is “Clear, Few clouds, Partly cloudy, Partly cloudy” followed by “Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist” weather.

Count of bike rentals are very less when the weather is “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds”.

**Missing Value**

Having missing data in our dataset may affect the conclusions that can be drawn from the dataset.

Missing values are a common occurrence, and we need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Below are common options to deal with missing values in dataset:

1)Drop the values

2)Fill the missing value with mean or median

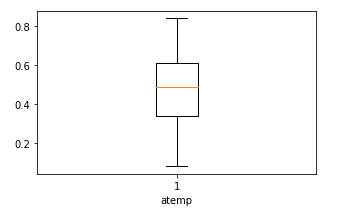
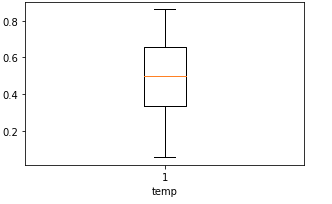
3)Fill the missing value with KNN imputation

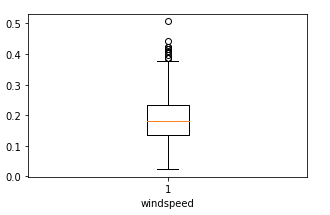
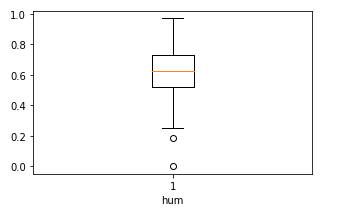
However in our case there are no missing values in data.

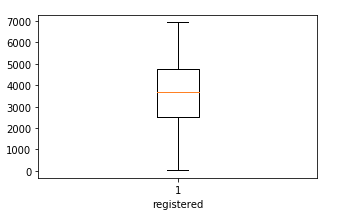
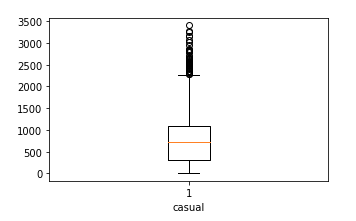
**Outlier Analysis:**

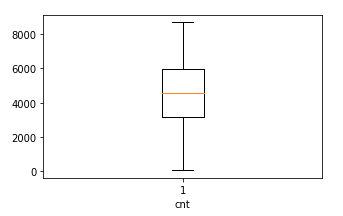
Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

We can make use of boxplots to detect outliers:









We can notice that there are outliers in casual and windspeed columns. In this case as an outlier treatment we have gone ahead and removed all the rows with outliers.

After treating outliers we are left with 676 rows.

**Feature Scaling**

Feature scaling is a method used to standardize the range of independent variables or features of data.

We can see that following columns have already been normalized:

temp: (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39

(only in hourly scale)

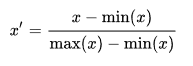
atemp: Normalized feeling temperature in Celsius. The values are derived via

(t-t\_min)/(t\_maxt\_min), t\_min=-16, t\_max=+50 (only in hourly scale)

hum: Normalized humidity. The values are divided to 100 (max)

windspeed: Normalized wind speed. The values are divided to 67 (max)

We can however notice that target variables have not been normalized. For the sake of further analysis, target variables have been normalized with below formula:



where x x {\displaystyle x} is an original value, x! x ′ {\displaystyle x'} is the normalized value.

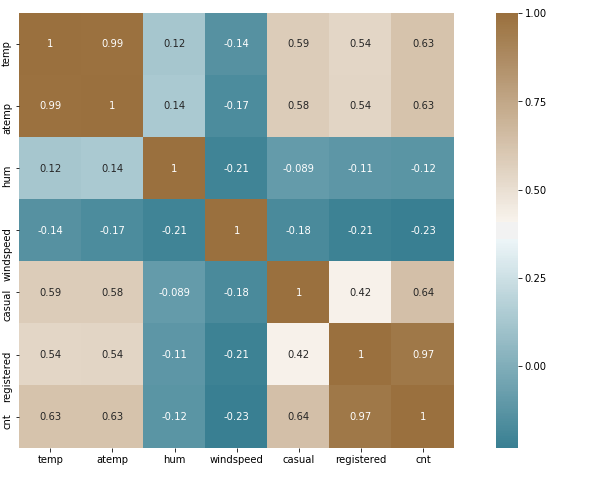
**Feature selection**

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

For this we will perform Correlation Analysis. Correlation tells you the association between two continuous variables. We can find if any two variables are highly correlated by viewing a heat map.

By looking at the heat map below it can be observed that “temp” and “atemp” are highly correlated and “cnt” and “registered” columns are highly correlated with each other.

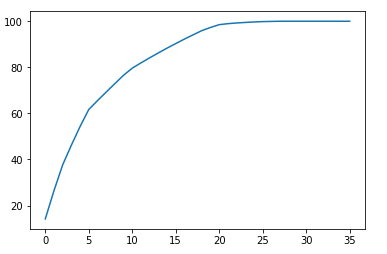
In this case as part of feature selection instead of keeping only one column in cases of highly correlated pair of columns. Hence “temp”, “casual” columns were removed.

****

**PCA**

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.



We can see that 20 variables are explaining more than 95% of data. Hence first 20 principal components are selected to be fed to the model that we are going to design.

**Model Selection**

Here three models have been taken for consideration, the output of which will be compared.

**1)Decision Tree**

A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node. Decision trees can handle both categorical and numerical data.

**2) Random Forest**

The **random forest** model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

*g*(*x*)=*f*0(*x*)+*f*1(*x*)+*f*2(*x*)+...

where the final model *g* is the sum of simple base models *fi*. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called [**model ensembling**](http://en.wikipedia.org/wiki/Ensemble_learning). In random forests, all the base models are constructed independently using a **different subsample** of the data.

**3)Linear Regression**

In statistics, linear regression is a linear approach to modelling the relationship between a dependent variable and one or more  independent variables. The case of one independent variable is called simple linear regression. For more than one independent variable, the process is called multiple linear regression. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula:

y = c + b\*x

where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

**Hyperparameter Tuning**

When creating a machine learning model, you'll be presented with design choices as to how to define your model architecture. Often times, we don't immediately know what the optimal model architecture should be for a given model, and thus we'd like to be able to explore a range of possibilities. In true machine learning fashion, we'll ideally ask the machine to perform this exploration and select the optimal model architecture automatically. Parameters which define the model architecture are referred to as **hyperparameters** and thus this process of searching for the ideal model architecture is referred to as hyperparameter tuning*.*

**Conclusion**

Since this is a regression problem, the metric we consider is Root mean square error.

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

After performing hyperparameter tuning the results of each model are as below:

**1)Decision Tree**

#Root Mean Squared Error = 0.11456320372193872 (python)

#Root Mean Squared Error = 0.12764151(R)

#R^2 Score = 0.7248744303357981 (python)

# R^2 Score: 0.69098383(R)

#MAPE: 18.812862571564235 (python)

#MAPE: 22.84871% (R)

**2) Random Forest**

#Root Mean Squared Error = 0.08564579859271436 (python)

#Root Mean Squared Error = 0.08795391 (R)

#R^2 Score = 0.8462366680827034 (python)

# R^2 Score: 0.85053461 (R)

#MAPE: 14.396712183975607 (python)

#MAPE: 15.04173% (R)

**3)Linear Regression**

#Root Mean Squared Error = 0.10414518826440598 (python)

#Root Mean Squared Error = 0.10402858 (R)

#R^2 Score = 0.7726373756387745 (python)

# R^2 Score: 0.79094500 (R)

#MAPE: 18.268886233633456 (python)

#MAPE: 20.06494% (R)

As the errors of Random forest are comparatively less we can freeze our Random Forest model.