Bike Rental Prediction

-by Vishnu Derkar

Content

		Page No.
1.	**Introduction**	
	1.1 Problem Statement	3
	1.2 Data	3
2.	**Methodology**	4
	2.1 Pre-Processing	4
	2.1.1 Using Pandas Profiling	4
	2.2 Exploratory Data Analysis	5
	2.2.1 Plotting Target Variable against other variables	5
	2.2.2 Plotting Data against Time	6
	2.2.3 Plotting data against Temperature	8
3.	**Model Evaluation**	9
	3.1 Model Preparation	9
	3.2 Mean Absolute Error	9
	3.3 Model Selection	9

Introduction

1.1 Problem statement:

The Bike sharing system are new generation of bike rental. There are around 500 bike sharing programs around world consisting 500 thousand bikes. Being able to monitor the duration of time travelled, departure and arrival position turns it into a virtual sensor network can be used for sensing mobility in the city. It is expected to detected most important events in the city. The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

1.2 Data:

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

instant: Record index

dteday: Date

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

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

mnth: Month (1 to 12) hr: Hour (0 to 23)

holiday: weather day is holiday or not (extracted from Holiday Schedule)

weekday: Day of the week

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

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

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)

atemp: Normalized feeling(apparent) 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) casual: count of casual users

registered: count of registered users

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

First few rows of the data is shown below:

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0	3	13	16
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	8	32	40
2	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0	5	27	32
3	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0	3	10	13
4	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0	0	1	1

Methodology

2.1 Pre Processing:

Data pre-processing is an important step in analysing the data for building a good model. As seen in this data we know that the attributes like temp, atemp and hum are normalised. Also the variables such as casual and registered count would not help us in predicting the daily bike usage but be more useful in data analysis of the given dataset. There are a total of 731 records of the year 2011 and 2012 combined.

2.1.1 Using Pandas Profiling:

Pandas profiling is an open source Python module which can quickly do exploratory data analysis with just few lines of code. Using Pandas Profiling we have found the following overview of the dataset.

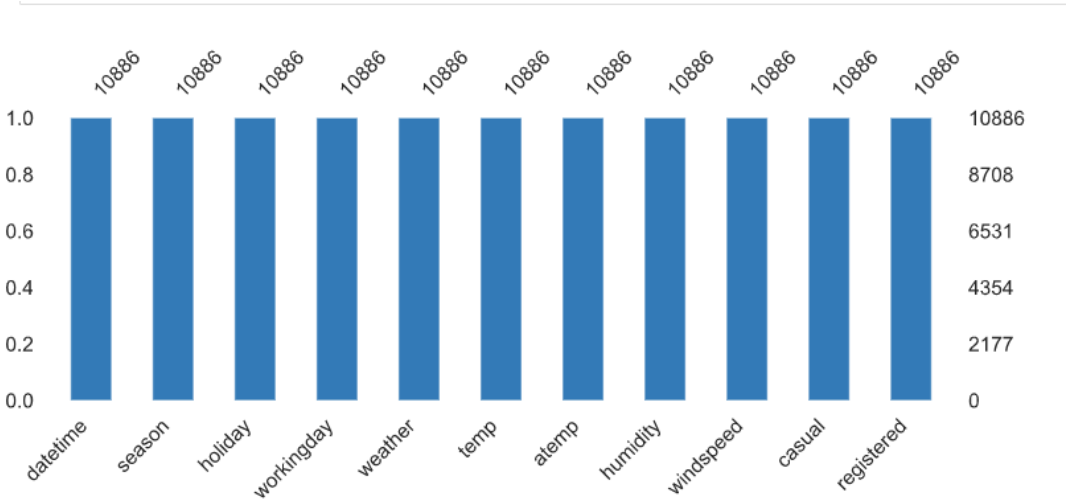
Dataset statistics

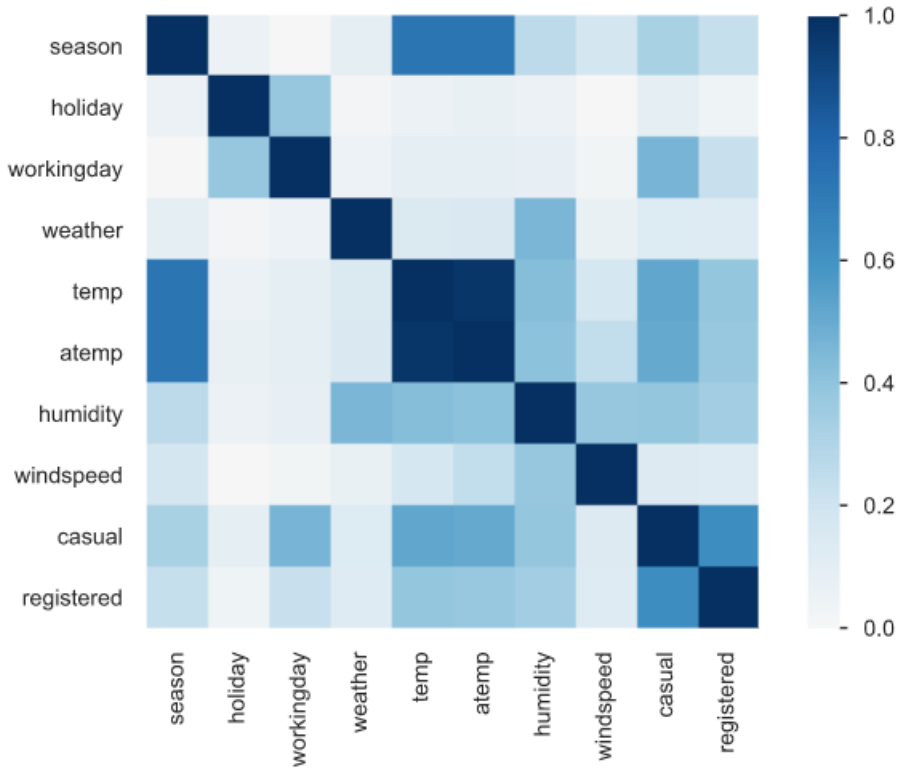
Number of variables	11
Number of observations	10886
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	935.6 KiB
Average record size in memory	88.0 B

Variable types

NUM	6
CAT	2
BOOL	2
DATE	1

Pandas Profiling also shows us that there are no missing values and duplicate rows in the dataset and also shows the correlation heatmap.

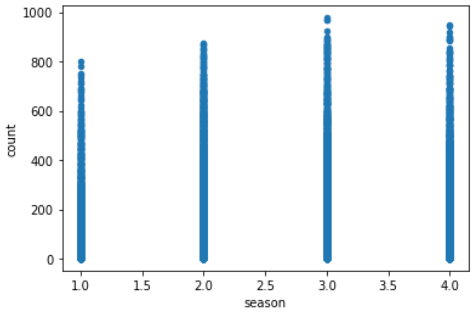




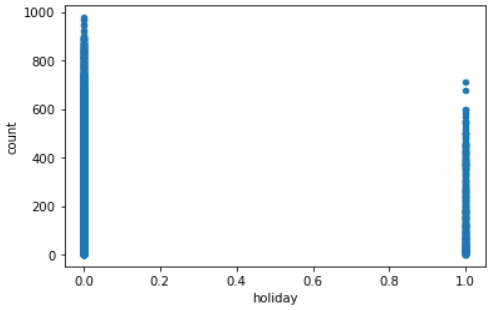
2.2.1 Plotting target variable against other variables:

By plotting the target variable against other variables, we are trying to understand the data and its distribution with respect to the target variable.

Season Versus Counts

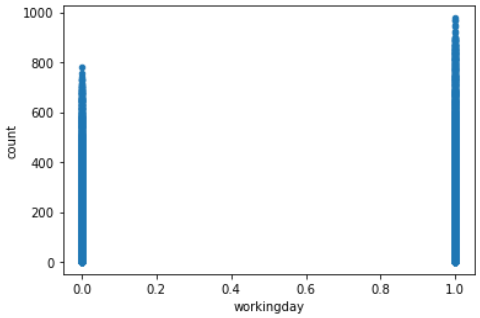


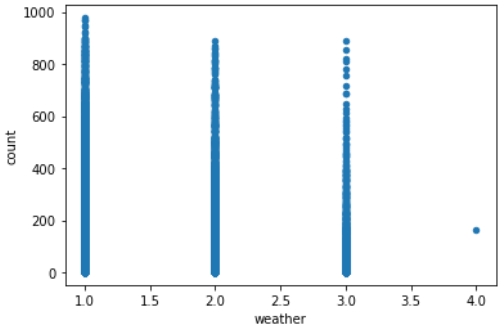
Holiday Versus Counts



Workingday Versus Counts

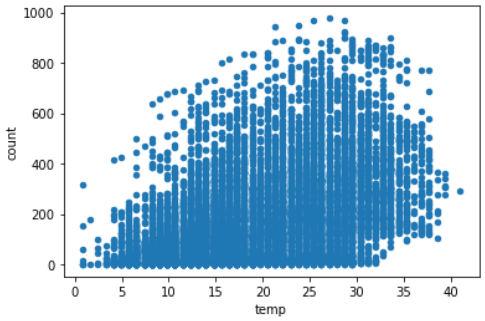
Weather Versus Counts

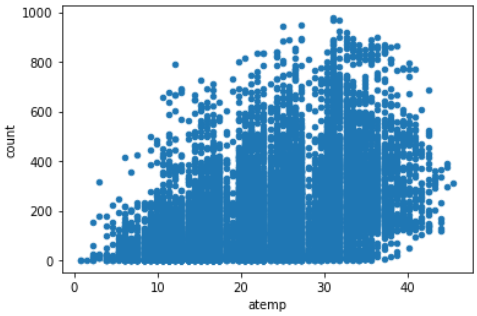




Temperature Versus Counts

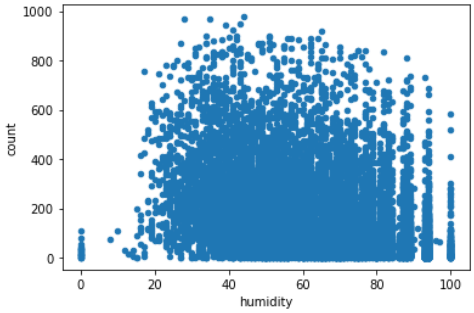
Atemp Versus Counts

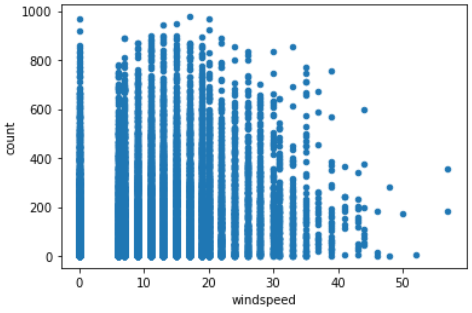




Humidity Versus Count

Windspeed Versus Count

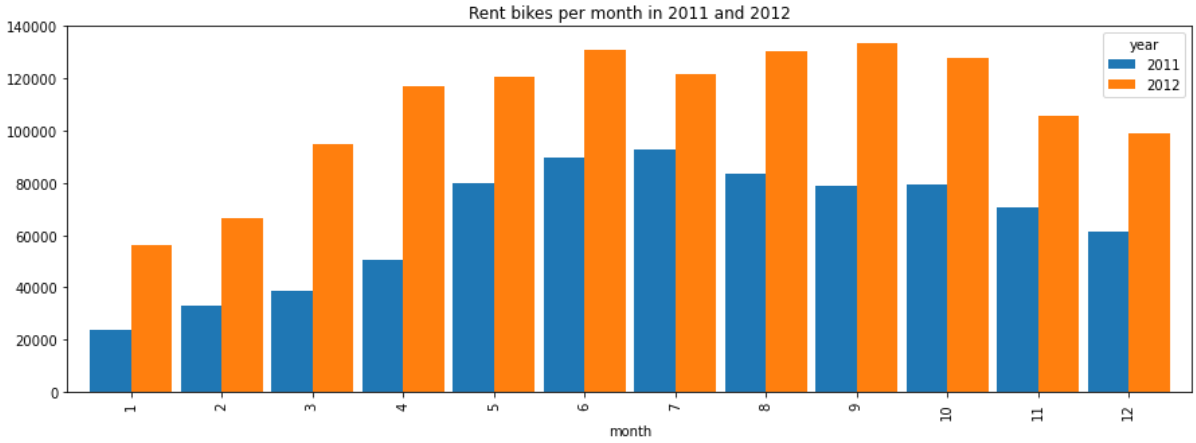




Plotting all the important feature againt count varible we can say the people of UK prefer Fall and Winter season to go out on bike ride. Similarly it seems that bike are mostly used as a transport to work since the of bikes increase on workingday. Also it can be fewer people use the bike than on working day form holiday versu count plot. From weather versus count plot the bikes seems to be useful in range on weather from a clear day to light snowfall/rainfall. And clearly are abandoned in Heavy storm and rainfall. People go out more often using bike when humidity is increases and in low to medium windspeed also temperature between 15^{o}C to 32^{o}C is suitable to go out.

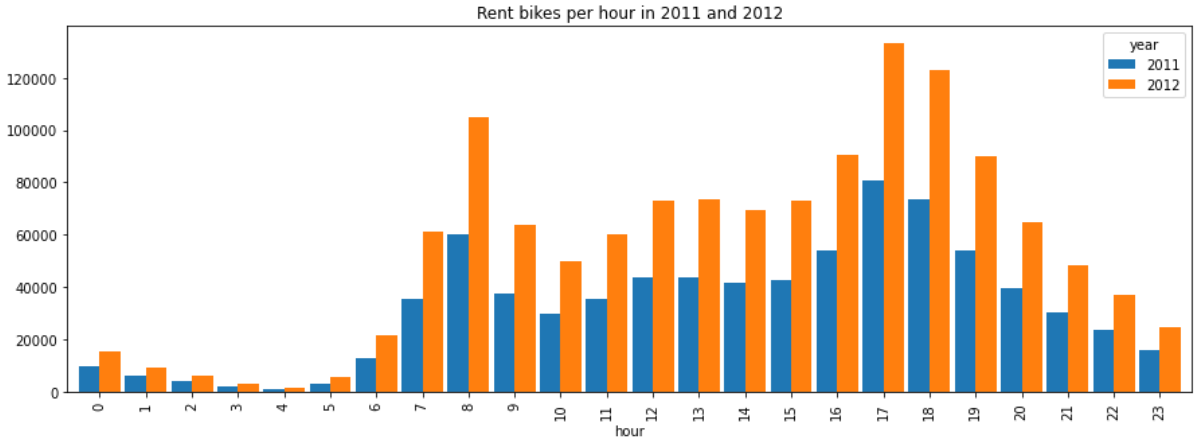
2.2.2 Plotting data against time:

To drive more insights from the data, we will plot it against time i.e. hourly and month-wise.

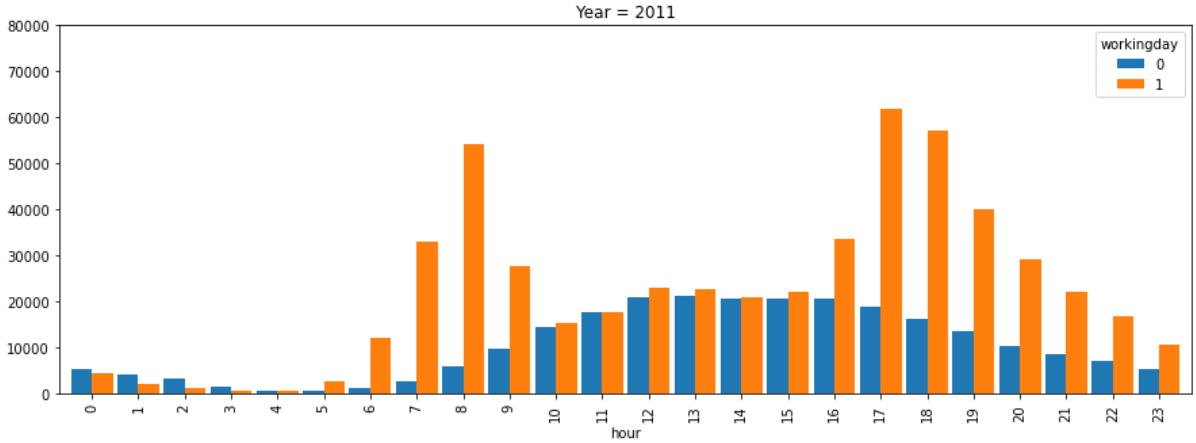


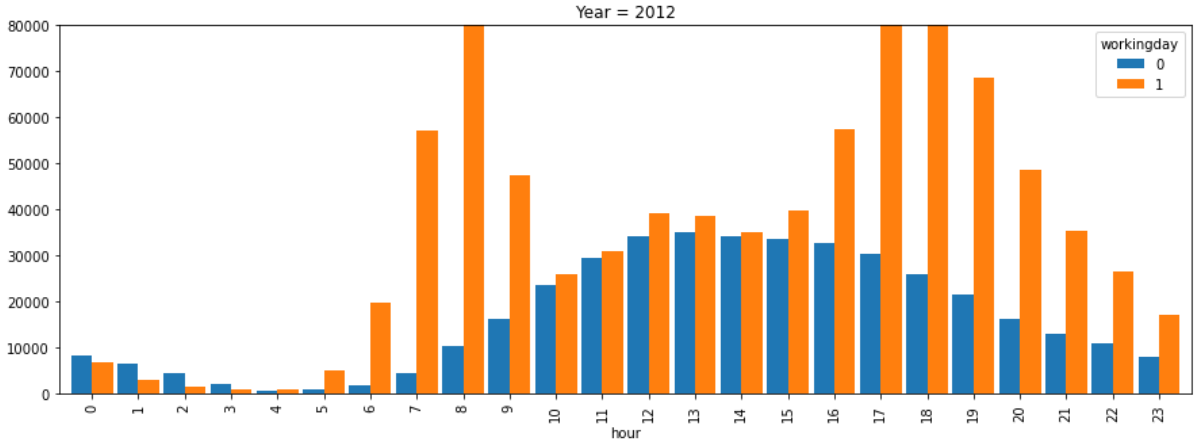
From the above bar plot it is clear that bike users have increased significantly in 2012 from 2011. It also portrays April to October as the months in which bikes have been more used whereas January and February stands as the months with lowest user of bikes.

By plotting the data hourly for 2011 and 2012 we get the following plot.



Plotting yearly data and differentiating by workingday.

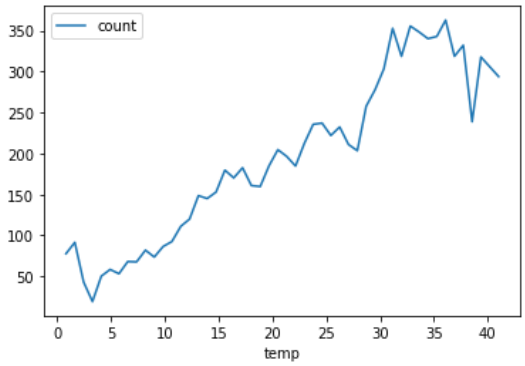




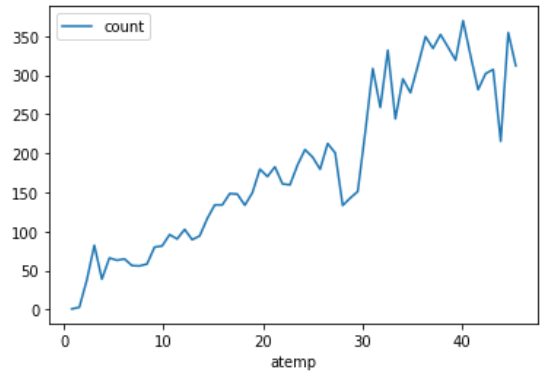
Here we can see that the bikes are rented mostly when people are either going to work i.e. 7 to 9 hrs or coming from work i.e. 16 to 19hrs on working day whereas on holiday bikes are mostly rented during 10 to 18 hours.

2.2.3 Plotting data against Temperature:

Temperature Vs Count



Apparent Temperature Vs Count



The bike are rented more as the temperature increases. People love to go out and enjoy a nice warm weather.

Model Evaluation

3.1 Model Preparation:

After completing pre-processing of data now we go forward with model preparation. For this data we have selected four models Dummy Regressor (mean, median), Random Forest Regressor, XGBoost Regressor.

The Dummy Regressor is a regression algorithm which use simple rules and is used only for comparison with the other algorithms.

Random forest is a supervised learning algorithm. It creates an ensemble of decision trees using bagging method. Bagging method is generally improving the result of decision tree over time.

XGBoost Algorithm is an ensemble of decision tree which uses gradient boosting framework. This algorithm is seen to generally performs better than other algorithm.

For model preparation we have defined a function `simple_modeling`, which will give us the direct output of the mean absolute error. From which we can decide which model to select.

3.2 Mean Absolute Error:

Mean absolute error (**MAE**) is a measure of errors between paired observations expressing the same phenomenon. Examples of *Y* versus *X* include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$ [1]

We got the following result for mean absolute error for the models:

```
[('dummy-mean', 33006.9386813781),
 ('dummy-median', 35172.1331496786),
 ('random-forest', 14313.964669753537),
 ('xgboots-regressor', 14019.816346457774)]
```

3.3 Model Selection:

Based on the above results we can see that XGBoost has the least Mean Squared Error. So XGBoost is the selected model for this dataset.