Final report

Sunday, May 3, 2020 8:47 PM

Objective: The main objective of the project is to produce a model which predicts bike usage.

EDA: Two datasets are present in the data directory namely daily.csv and hourly.csv

For this project we are considering only daily.csv as main objective of the project is to estimate the bike usage provided certain features of the day. The following steps are used for EDA :-

Step 1: Basic Structure

The dataset has 731 rows and 16 columns. Here the rows represent the days.
The Basic structure of the dataset includes:

#	Column	Non-Null Count	Dtype
0	instant	731 non-null	int64
1	dteday	731 non-null	object
2	season	731 non-null	int64
3	yr	731 non-null	int64
4	mnth	731 non-null	int64
5	holiday	731 non-null	int64
6	weekday	731 non-null	int64
7	workingday	731 non-null	int64
8	weathersit	731 non-null	int64
9	temp	731 non-null	float64
10	atemp	731 non-null	float64
11	hum	731 non-null	float64
12	windspeed	731 non-null	float64
13	casual	731 non-null	int64
14	registered	731 non-null	int64
15	cnt	731 non-null	int64

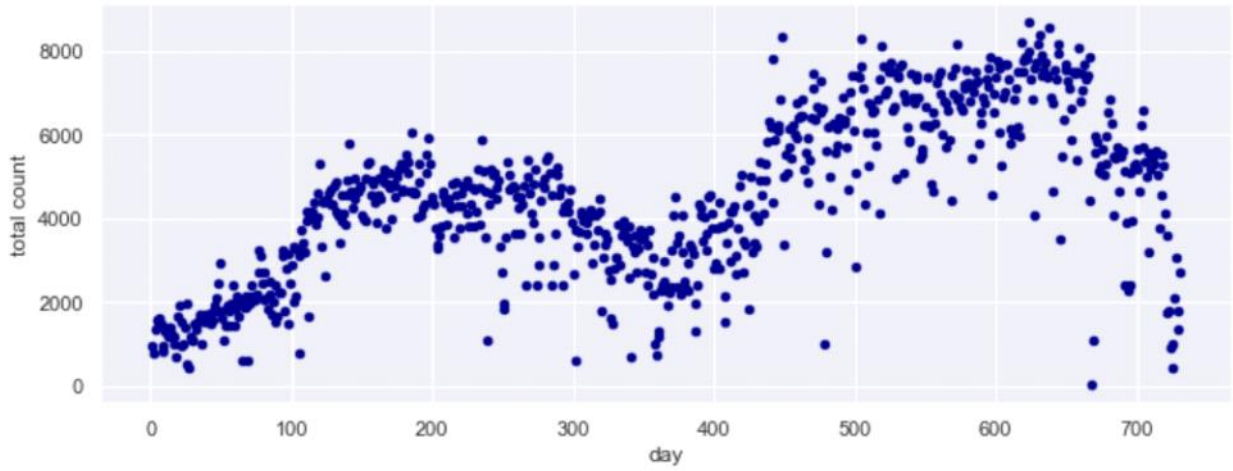
From the above report, we can conclude that we have three datatype features like int, object and float, there are also no categorical features.

Step 2: Data Cleaning

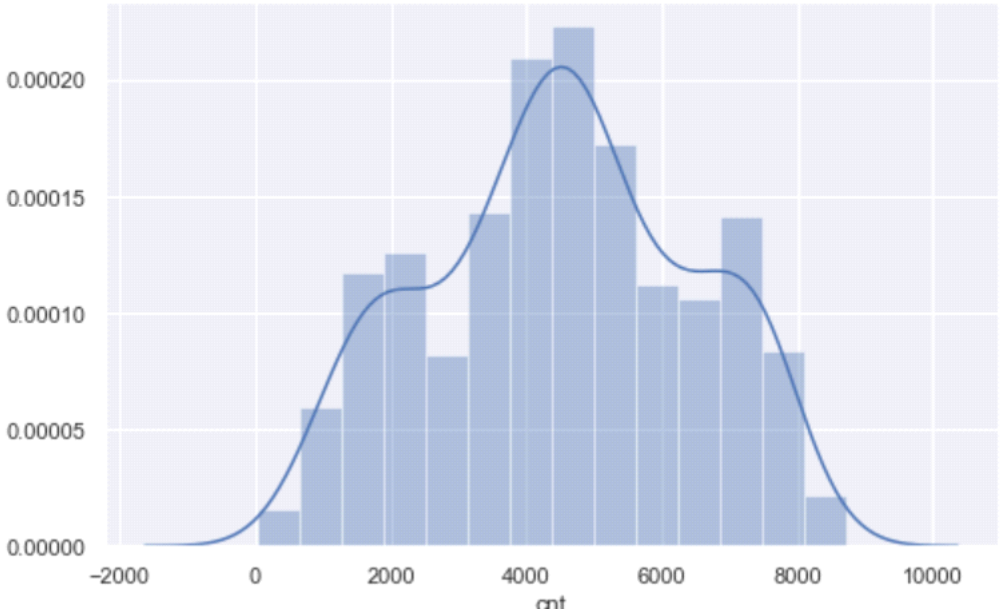
The data is very clean, so there is nothing to clean and data is also normalized. There are no null values and outliers in the data.

Step 3: Exploring the predictive column

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



From the above plot we can conclude that there the total count is has a trend and seasonality with respect time. And there is less linearity in the bike usage. So we will explore with other features.

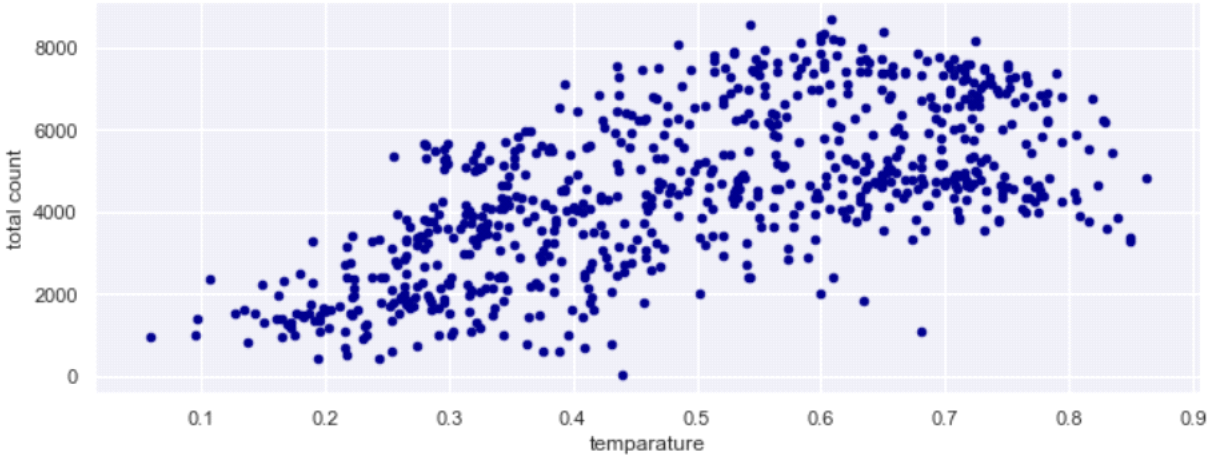


The above plot is used to check the skewness in the total count, the data is normally distributed.

Step 4: Feature selection

There are several methods for feature selection, but since the data is clean and normally distributed, we can simply use correlation matrix and choose columns accordingly.

From the above correlation matrix, we can see that season, year, month, temperature have good correlation with total count.



The above scatter plot shows the correlation between temperature and total count. Some linearity is observed in the data.

MODELING :

Linear Model

Since the object is predication of bike rental count hourly or daily based on the environmental and seasonal settings, a simple linear regression model is modelled. Linear regression requires the relation between the dependent variable and the independent variable to be linear.

After EDA and feature selection, out of 16 columns only season, year, month and temperature are considered for modelling.

The dataset is split into train and test datasets in the ratio of 4:1.

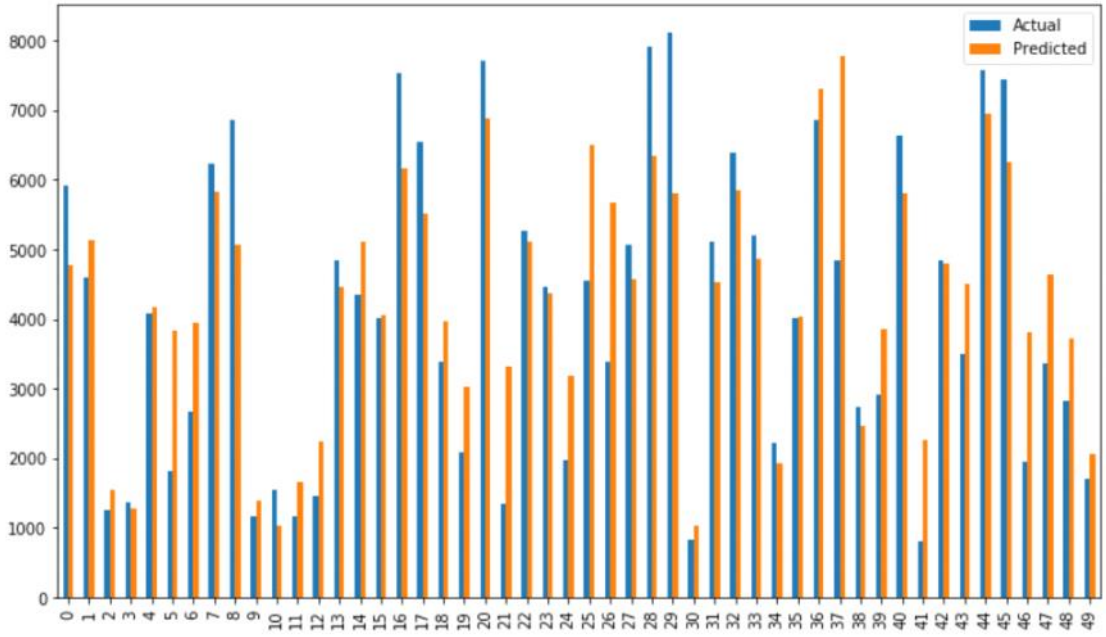
After fitting the train data set with Multi Linear Model in sklearn library, the following equation is formed.

cnt = -174.23 + 471.51 **season** + 2034.78 **yr** -34.88 **mnth** + 5501.55 **temp**

After Predicting the test dataset, the following metrics are observed:

Mean Absolute Error: 812.4685200638834
r-square score : 0.7327183500040215

The R-square of is greater than 0.7, the model is considered good, but MAE is high, so model is unable to capture the patterns in the data. This is an example of **under-fitting**.



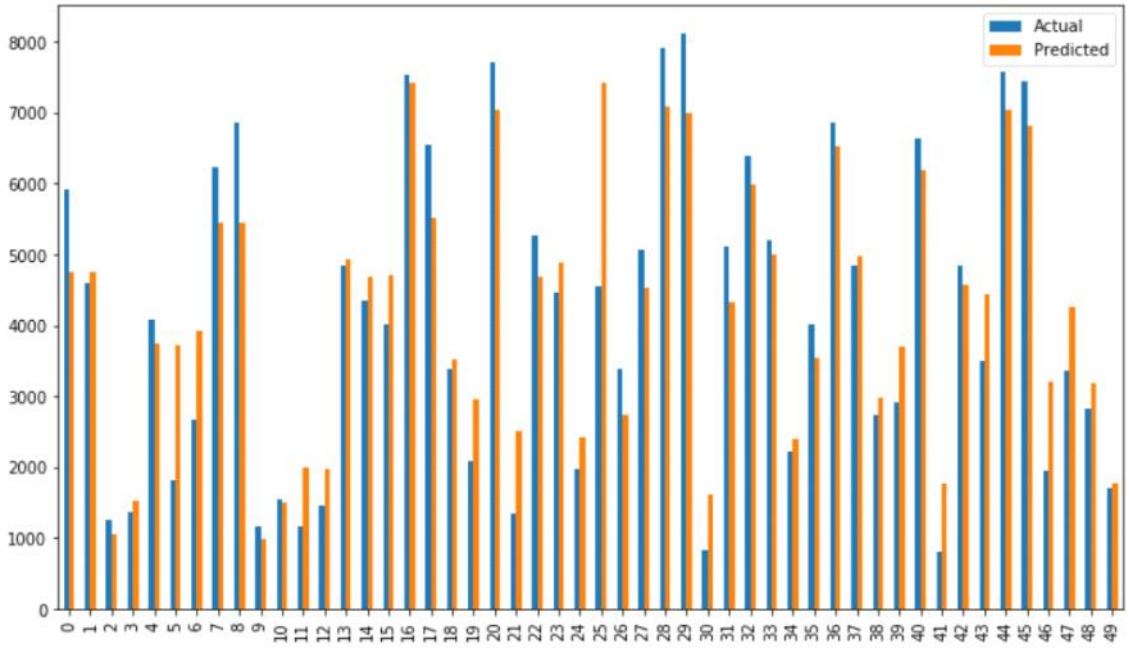
The above plot is plotted on the first 50 rows of the test data and predictions are almost predicting the actual values, with a small error.

Polynomial Model :
In the simple linear model the Mean Absolute error is observed ~812, so in this polynomial model we will try to reduce MAE using different degrees of polynomials. To overcome under-fitting, we need to increase the complexity of the model.

After fitting the training data with polynomial model with degree 3 and predicting the testing data the following metrics are observed.

Mean Absolute Error: 637.4336879427826
r-square score : 0.8228854993567609

We can observe that MEA reduced to 637 from 812, which is considered that the prediction values are much closer to actual values. It is also observed that R-square value is increased.



The above plot show the prediction and actual values of the first 50 rows of the test dataset, when compared with linear model, it is observed that the polynomial model is performing much better than the simple linear model.

Note: Different degrees of polynomial have been tested with the data, but there is high risk of overfitting, so best model is considered with polynomial degree 3 with season, year, month and temperature features.

Git:

The project has been developed using python notebooks and these notes are monitored by git version control software

References:

1 . https://software-carpentry.org/lessons/
2. https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f
3. https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f