Predicting Bike Rental Count

1. INTRODUCTION

1.1 Problem Statement

The objective of this Case is to predict of bike rental count on daily basis on the environmental and seasonal settings.
The aim of the whole project is to predict the total bike rental count for a day based on environment and seasonal settings.
So that the bike rental company will be able to serve the customer and meet the demand of bikes.

1.2 Data

Data provided with the problem is **day.csv**.
Overview of dataset:

Column Name	**Description**	**Expected Values**	**Value Meaning**
Instant	Record index	Unique numeric identifier	
Dteday	Date	Any valid Date	
Season	Season	1 2 3	1:springer, 2:summer, 3:fall, 4:winter
yr	Year	0 1	0: 2011, 1:2012
Mnth	Month	1 2 3 . . . 12	1: January, 2: Feb, 3: March, . . . 12: Dec
Holiday	Indicates weather day is holiday or not(extracted from Holiday Schedule)	0 1	0: Not a Holiday 1: Holiday
Weekday	Day of the week	0 1 . . 6	0: Saturday 1: Sunday . . 6: Friday
workingday	If day is neither weekend nor holiday is 1, otherwise is 0	0 1	0: Not a workingday(weekend or Holiday) 1: workingday

Weathersit	Weather situation(extracted fromFreemeteo)	1 2 3 4	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 temperature in Celsius.	The values are derived via (t-t_min)/(t_max-t_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		

Here Target variable(dependent variable) **“CNT”**

And predictors(independent variables), on basis of which target needs to be predicted are :
(dteday,season,yr,mnth,holiday,weekday,workingday,weathersit,temp,atemp,hum)

2. PROJECT MAP

To execute this project, i have tested with CRISP- DM process

2.1 CRISP-DM process

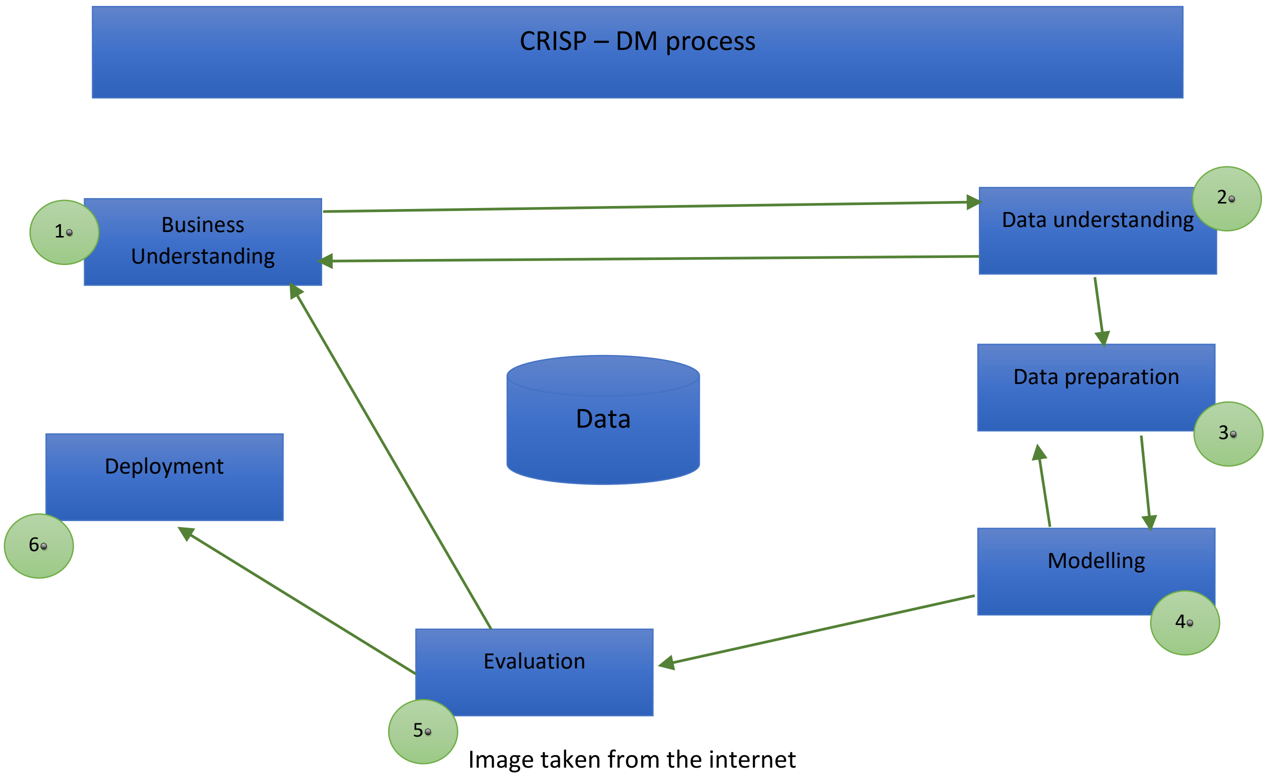


Image taken from the internet

Above process, shows phases of data science projects, each phase having its importance.

3. PROJECT IMPLEMENTATION

Below are the different phases of the implementation of project

3.1 Rough sketch of project

Here is the rough sketch of the project in different phases, using CRISP-DM process. The whole project is divided in 7 phases (and further sub-phases). Below are the phases defined.

- Define and categorize problem statement
- Gather the data

- Prepare data for consumption
- Perform Exploratory Data Analysis
- Modelling
- Evaluate and compare Model performances and choose the best model
- Produce sample output with selected model

3.2 Acutal Implementation of project

exploring each phase in the project

3.2.1 Categorize Problem

The problem statement is "To predict of bike rental count on daily basis on the environmental and seasonal settings"
It is evident from the problem statement above that based on the predictors values (input, both numerical and categorical), the output dependent value (numerical) needs to be predicted.

So, clearly this problem is of category – **Supervised Machine Learning Regression Problem.**

3.2.2 Prepare Data

Next step is to prepare the data for consumption. In this case, the compiled data set is given to us in 1 file. We do not need to join different sources to prepare data for the analysis.

What we need to do here is more of data cleaning activity and make it ready for EDA and modelling. I performed following steps to achieve this.

3.2.2.1 Check the shape/properties of the data

- There are 16 features and 731 observations in the dataset
- float64(4) ,Int64(11) and object(1) datatypes are used in this dataset.
- None of the columns in dataset has nulls

3.2.2.2 Completing

Perform missing value analysis and impute missing values if necessary

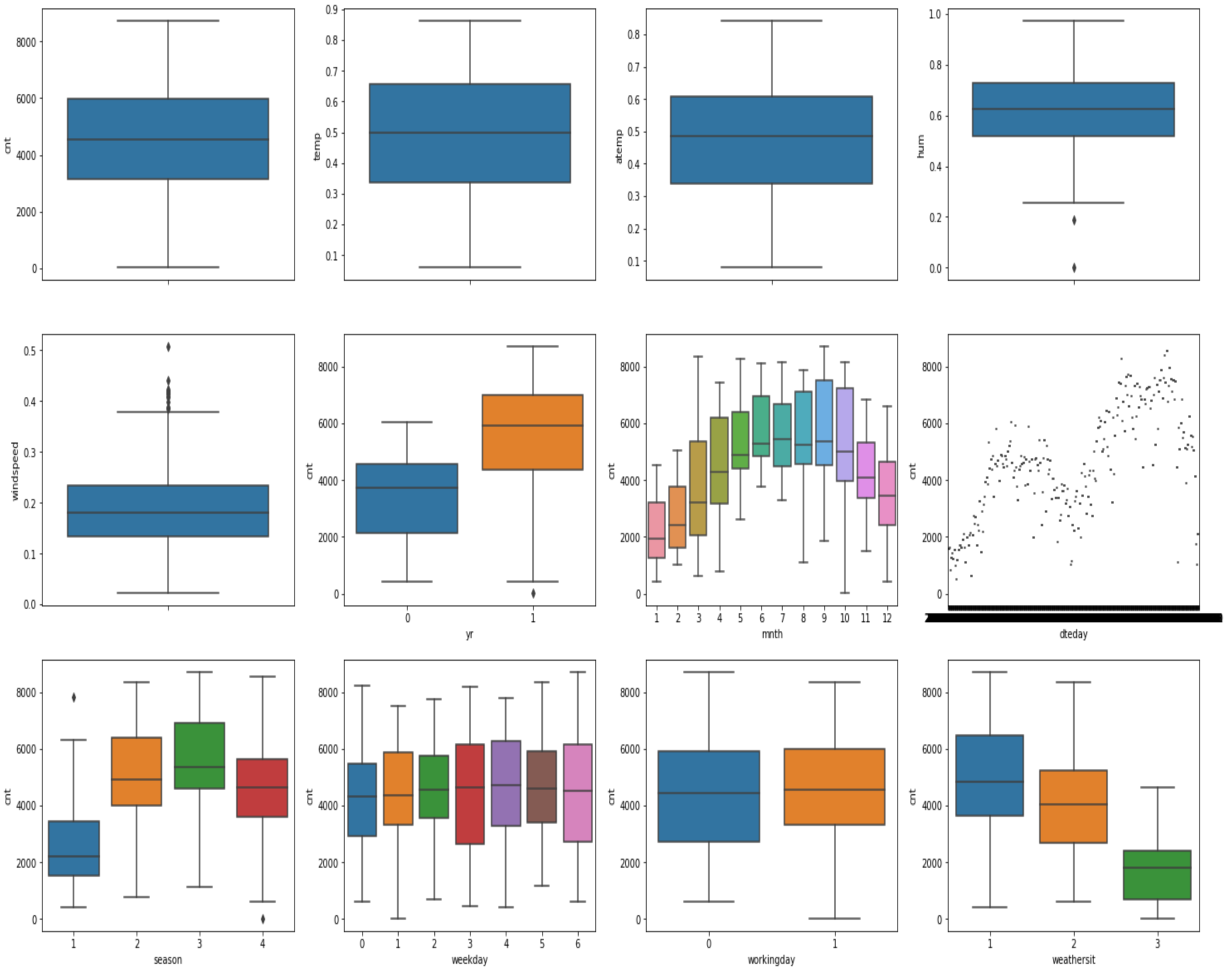
3.2.2.3 Converting

Converting data to proper formats.
Columns like 'yr', 'season' are imported as numeric columns. However, these are categorical in nature. So, these needs to be converted to 'categorical' datatypes.
It is important to convert the categorical values to category as 'numeric' and 'categories' have different features.

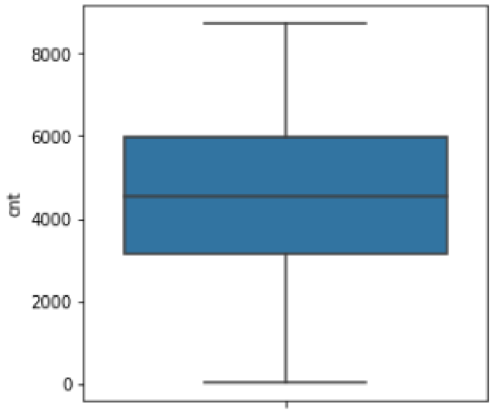
For ex:- Here category feature and numeric feature may seem same but they are not, numeric has order attached to values like 1<2 etc.
But category does not, 1, 2,3,4 does not have any order, they are simply categories. And they can’t be represented quantitively.
Feeding wrong datatypes to model may affect the results.

3.2.3 Perform EDA

3.2.3.1 Outlier Analysis using Boxplots



Boxplot for rented bikes count(‘cnt’)



There are no outliers for target variable ‘cnt’. All values lies with 3IQR range of mean.

Boxplots for numeric predictors():

- There is one outlier for year 2011, where the rental count is extremely low. This could be the result of some extreme weather condition
- Windspeed and humidity has some outliers, which tells about data captured in extreme weather conditions
- From month boxplot, we can see there are no outliers and also can be seen that average rented bikes count is higher between 5th-10th month, which makes sense due to favourable weather conditions
- Season data is pretty much in range, only 2 outliers.
- No other categorical features have any outliers.

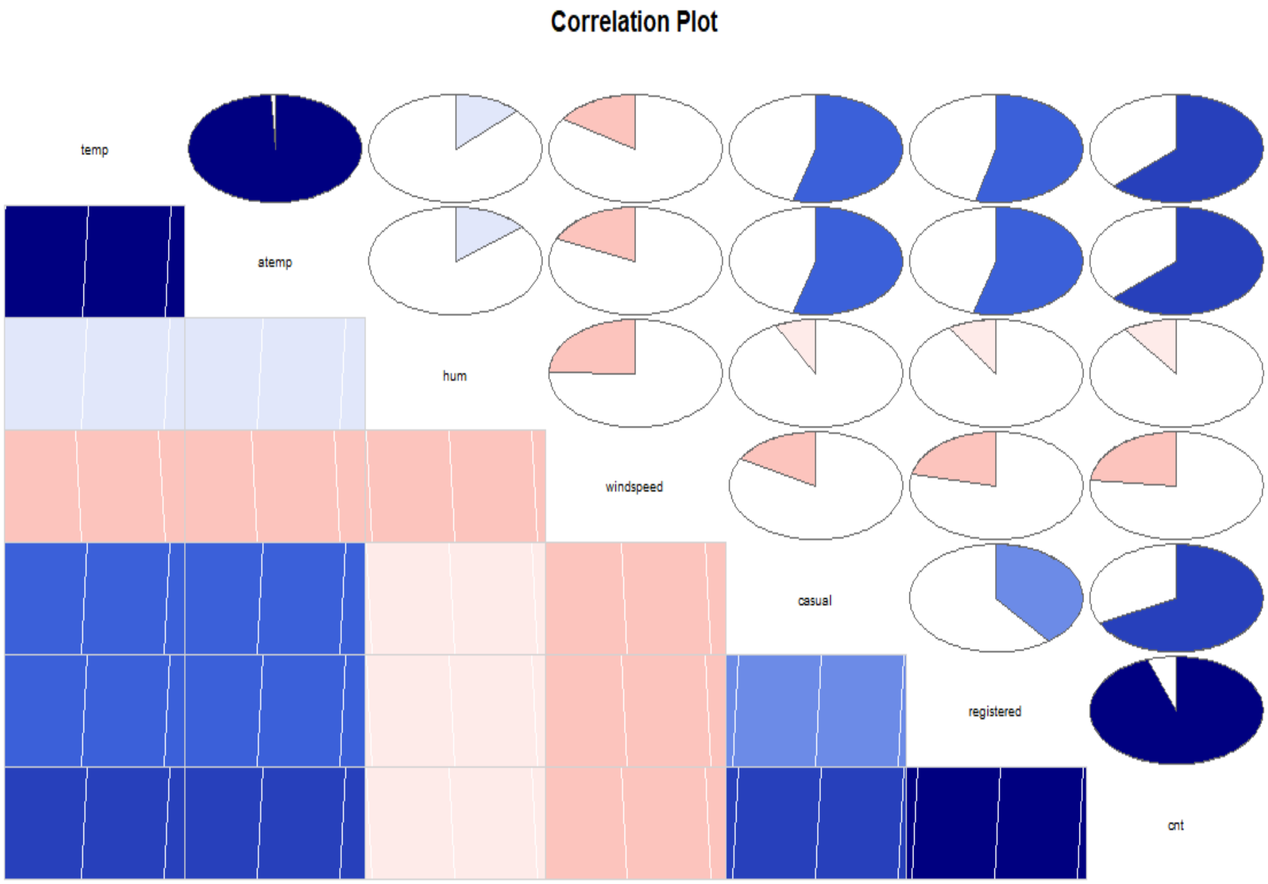
- Most of the data set looks clean. Not many outliers found.
- Only little noise is there in database
- Not removing outliers, as they have valid reasons.

Performing casting of variables types. Converting necessary variables into categorical form,

3.2.3.2 Analysis of Numerical Features

i) Correlation Analysis

analysing the variables using plot



- relationship of all numeric independent variables with target variable(cnt)
- relationship of all numeric variable among themselves(to detect multi collinearity)

For many models, it is important that the target variable follows normal distribution.
If that is not the case, many a times, we apply some techniques(like taking log) to convert the distribution to normal.

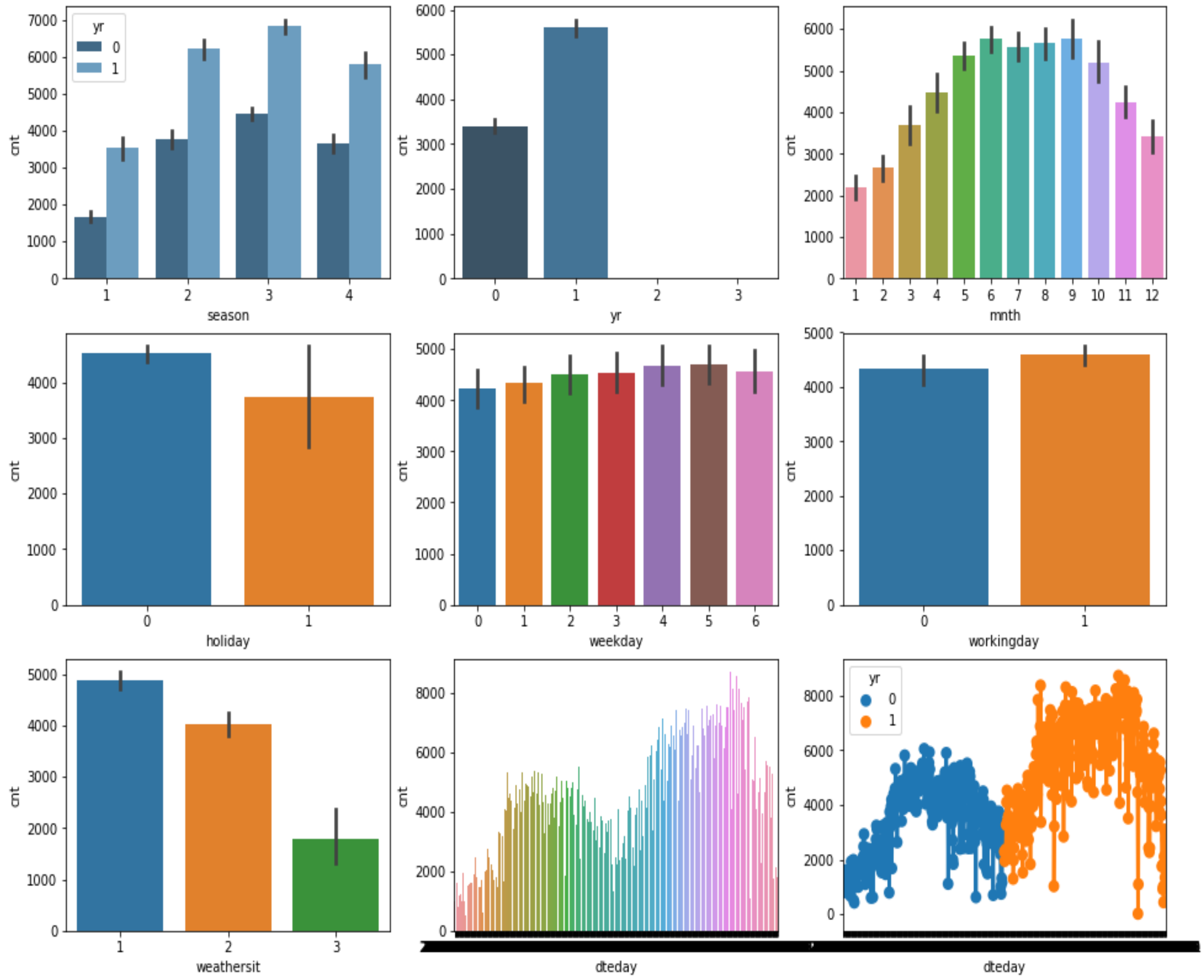
- Clearly, target variable has almost normal distribution, no need to apply any technique here.

- dropping atemp variable as it is highly correlated with temp
- dropping "casual" and "registered" variables as their addition is cnt.

- Treating cnt as target variable
- holidayand workingday seem to explain same thing...So, dropping holiday

3.2.3.3 Exploratory Analysis of Categorical Features

i) Barplot for Count Vs. Categorical features



- 'yr' effect on count of rented bikes
  the count has an upward trend with year
- 'season' effect on count of rented bikes
  seems ppl rent more bikes during season 3 and 2, i.e. highest in fall and summer and less in winter and springs. This makes sense as weather is good to ride during summer and fall.
- 'month' effect on count of rented bikes
  people are likely to rent bikes more between the months May- October and lowest in month of Jan, Feb and Dec(in that order). Because those months have favourable weather conditions

- 'weekday' effect on count of rented bikes
  people seems to rent lesser bikes on Sat/ Sun. ie. over the weekend. Again makes sense as school and offices are closed on weekend.Monday also has lesser count of rented bikes. It may be possible the people visit to other places/cities over weekend and travel back in car on Monday, instead of renting bikes.

- **DUMMY Variables**
  After all the feature engineering, i have made dummy variables of categorical variables.
  That is variable having k categories will be converted into k-1 variables now.
  For example : month variable has 12 levels/categories, so after applying dummy function, 11 variables of month will be formed keeping 1 level as reference.

3.2.4 Modelling

3.2.4.1 Choosing the ML Models

Now, the exploratory analysis is done, we need to decide on the machine learning algorithms we'll use to build the predictive models.
I am using 2 Machine Learning algorithms to build 2 different models and later compare them to decide on the best model

- Linear Regression Model
- Random Forest Model

Before going to the predictive models we are going to use, let's understand few concepts:

- **Ensemble methods:**
  Ensemble methods combine several decision trees classifiers to produce better predictive performance than a single decision tree classifier. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model.
- **Bagging**
  Bagging is a technique that is used when the goal is to reduce the variance of a decision tree classifier. **Here the objective is to create several subsets of data from training sample chosen randomly with replacement.** Each collection of subset data is used to train their decision trees. As a result, we get an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree classifier.

Now, lets get on to the models we are going to build.

i) **<u>Linear Regression Model:</u>**

Linear regression is one of the most commonly used predictive modelling techniques. The aim of linear regression is to find a mathematical equation for a continuous response variable Y as a function of one or more X variable(s). So that you can use this regression model to predict the Y when only the X is known.

Y= a1 X1 + a2X2 + …………………………… + anXn

ii) **<u>Random Forest Model:</u>**

Random Forestis an ensemble machine learning algorithm that uses 'bagging' technique.

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[1][2] Random decision forests correct for decision trees' habit of overfitting to their training set.

3.2.4.2 Choosing the Performance Measures for the Models

We are working on regression problem, so I think the best performance matrix could be

- RMSE
- MSE
- MAPE

We'll measure above error metrics and compare them.

3.2.4.3 Building the ML predictive Models

Building linear regression by using sm.OLS()
And random regression by using RandomForestRegressor()

Basic things while building the model, I have split the data between train and test then used the training set data features to train with training set target variable and then used that model to test on testing data set variable to produce results and then compared those results with target variable of test data set.

After that step i have performed cross validation so that 5 folds of the datasets should be formed and there will 5 different results of the model and average of those 5 datasets will be counted.

3.2.5 Model Evaluation and Comparision

Results are as below: these are mean scores after testing model on cross validation :

Model	**MSE**	**RMSE**	**MAPE**
Linear Regression	871878	933	17.24
Random Forest	393180	627	14.65

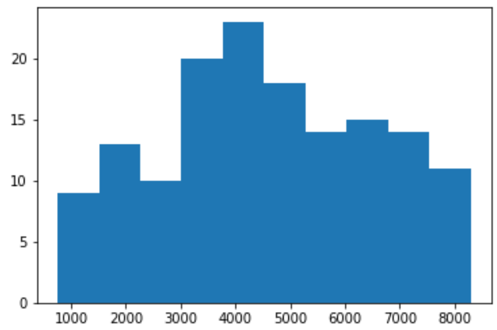
From the above table, we can see, Random Forest has better results than Linear Regression. So, finally Random Forest model is chosen.

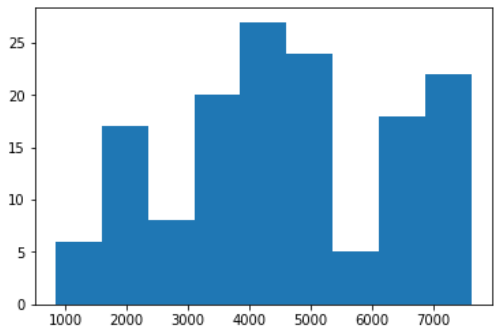
After generating sample output i have plotted the graph of cnt target variable

4. CONCLUSION

Both predicted values and actual counts have almost similar distribution.
This model is ready for deployment now.

Histogram plot of cnt and predicted values


0
5
10
15
20
1000
2000
3000
4000
5000
6000
7000
8000


0
5
10
15
20
25
1000
2000
3000
4000
5000
6000
7000