Prediction of Bike Rental Count

Aurangzeb Alam

25th May 2019

Contents

Chapter 1: Introduction

1.1 Problem Statement

The aim of this project is to predict the count of bike rentals based on the seasonal and environmental settings. By predicting the count, it would be possible to help accommodate in managing the number of bikes required on a daily basis, and being prepared for high demand of bikes during peak periods.

1.2 Data

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 -
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 fromHoliday 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 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

The goal is to build regression models which will predict the number of bikes used based on the environmental and season behavior. Given below is a sample of the data set that we are using to predict the number of bikes:

Chapter 2: Methodology

2.1 Pre-Processing

A predictive model requires that we look at the data before we start to create a model. However, in data mining, looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is known as Exploratory Data Analysis.

The goal is to build regression models which will predict the number of bikes used based on the environmental and season behavior. Given below is a sample of the data set that we are using to predict the number of bikes:

Table 1.1: Bike Count Sample Data (Columns: 1-9)

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit
1	2011-01-01	1	0	1	0	6	0	2
2	2011-01-02	1	0	1	0	0	0	2
3	2011-01-03	1	0	1	0	1	1	1
4	2011-01-04	1	0	1	0	2	1	1
5	2011-01-05	1	0	1	0	3	1	1

Table 1.2: Bike Count Sample Data (Columns: 10-16)

temp	atemp	hum	windspeed	casual	registered	cnt
0.3441670	0.3636250	0.805833	0.1604460	331	654	985
0.3634780	0.3537390	0.696087	0.2485390	131	670	801
0.1963640	0.1894050	0.437273	0.2483090	120	1229	1349
0.2000000	0.2121220	0.590435	0.1602960	108	1454	1562
0.2269570	0.2292700	0.436957	0.1869000	82	1518	1600

Table 1.3: Data type of variables are given in below Diagram

```
#Data types
Dataset.dtypes
```

```
instant         int64
season          int64
yr              int64
mnth            int64
holiday         int64
weekday         int64
workingday      int64
weathersit      int64
temp          float64
atemp         float64
hum           float64
windspeed     float64
casual          int64
registered      int64
cnt             int64
dtype: object
```

Table 1.4: Statistical description is given in below diagram (Columns: 1-8)

```
#Statistical Summary of the dataset
Dataset.describe()
```

	instant	season	yr	mnth	holiday	weekday	workingday	weathersit
count	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000
mean	366.000000	2.496580	0.500684	6.519836	0.028728	2.997264	0.683995	1.395349
std	211.165812	1.110807	0.500342	3.451913	0.167155	2.004787	0.465233	0.544894
min	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000
25%	183.500000	2.000000	0.000000	4.000000	0.000000	1.000000	0.000000	1.000000
50%	366.000000	3.000000	1.000000	7.000000	0.000000	3.000000	1.000000	1.000000
75%	548.500000	3.000000	1.000000	10.000000	0.000000	5.000000	1.000000	2.000000
max	731.000000	4.000000	1.000000	12.000000	1.000000	6.000000	1.000000	3.000000

Table 1.5: Statistical description is given in below diagram (Columns: 1-9)

temp	atemp	hum	windspeed	casual	registered	cnt
731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000
0.495385	0.474354	0.627894	0.190486	848.176471	3656.172367	4504.348837
0.183051	0.162961	0.142429	0.077498	686.622488	1560.256377	1937.211452
0.059130	0.079070	0.000000	0.022392	2.000000	20.000000	22.000000
0.337083	0.337842	0.520000	0.134950	315.500000	2497.000000	3152.000000
0.498333	0.486733	0.626667	0.180975	713.000000	3662.000000	4548.000000
0.655417	0.608602	0.730209	0.233214	1096.000000	4776.500000	5956.000000
0.861667	0.840896	0.972500	0.507463	3410.000000	6946.000000	8714.000000

2.2 Missing Value Analysis

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a column has more than 30% of data as missing value either we ignore the entire column or we ignore those observations. In the given data there is no missing value.

Table 1.5: Dataframe containing the number of missing value in each column is given below in the diagram.

```
#Check Missing values
missing_value = pd.DataFrame(Dataset.isnull().sum())
missing_value
```

	0
instant	0
season	0
yr	0
mnth	0
holiday	0
weekday	0
workingday	0
weathersit	0
temp	0
atemp	0
hum	0
windspeed	0
casual	0
registered	0
cnt	0

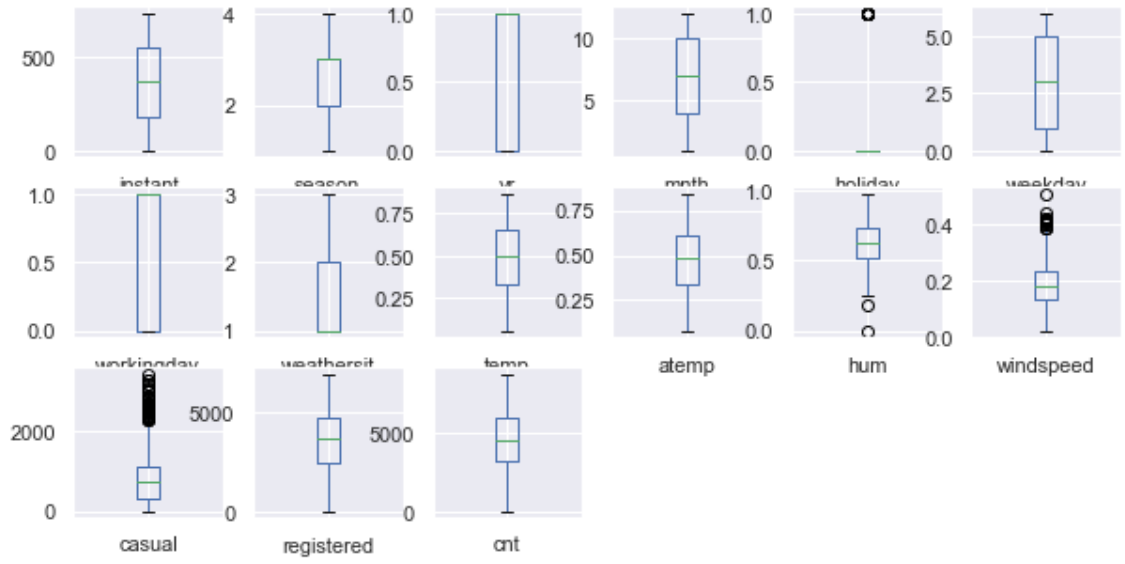
2.3 Detection of outliers and its Removal

We can clearly observe from these probability distributions that most of the variables are skewed. The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots or directly we plot by matplotlib.
In figure we have plotted see if there are outliers. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.

Table 1.6: This shows the plotting of Outliers

```
#Outliers
plt.figure(figsize=(20,20))
Dataset.plot(kind="box",subplots=True,layout=(3,6), figsize=(10,5))
plt.show()
```

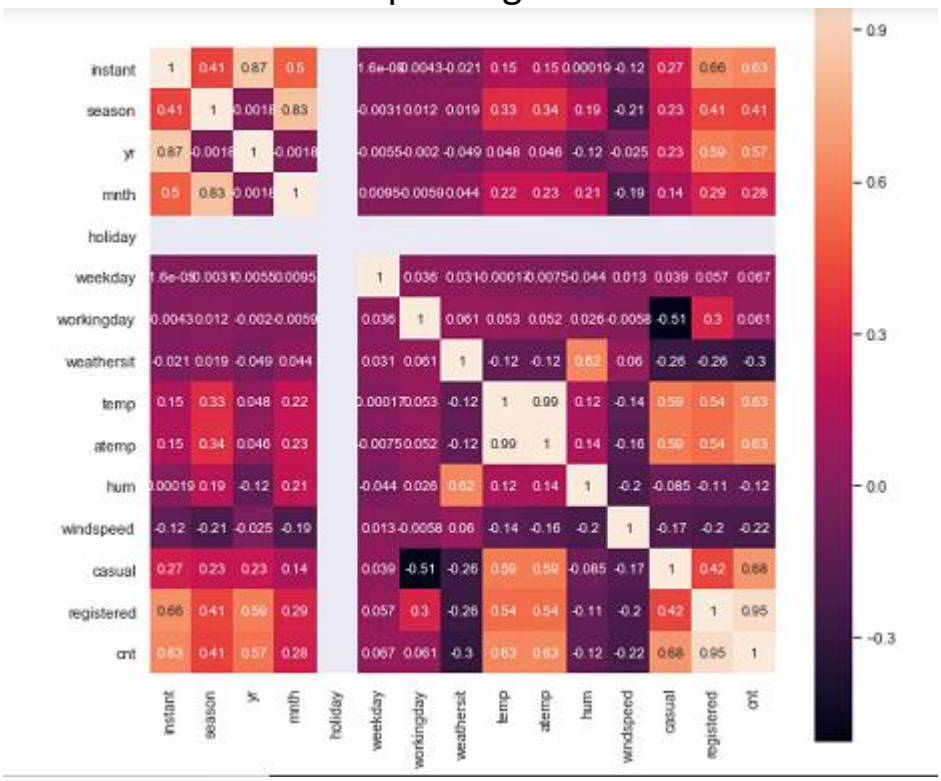
```
<Figure size 1440x1440 with 0 Axes>
```



2.4 Feature Selection

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation Analysis** for numerical variable.

Table 1.7: Correlation plot is given below:



2.5 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies
widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

Chapter 3: Modelling

3.1 Model Selection

The dependent variable in our model is a continuous variable i.e., Count of bike rentals. Hence the models that we choose are Linear Regression, Decision Tree and Random Forest. The error metric chosen for the problem statement is Mean Absolute Error (MAE).

3.2 Linear Regression

Linear regression is the most common form of linear regression analysis. Multiple linear Regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

Here in the Random Forest we have achieved the accuracy of our model as 89%.

3.3 Decision Tree

A decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

Here in the Decision Tree we have achieved the accuracy of our model as 88%.

3.4 Random Forest

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data.

Here in the Random Forest we have achieved the accuracy of our model as 92%.

Chapter 4: Conclusion

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

In our case of Bike count prediction Data, Interpretability and Computation Efficiency, do not hold much significance. Therefore, we will use Predictive performance as the criteria to compare and evaluate models.

Predictive performance can be measured by comparing Predictions of the models with real values of the target variables, and calculating some average error measure.

4.1 Reference

For Data Cleaning and Model Development
https://edwisor.com/career-data-scientist