Predicting Bike Rental counts

Smit Savjiyani

20/12/19

Contents

3.2 Model Selection 13

Chapter 1

Introduction

1.1 Problem Statement

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

1.2 Data

Our task is to build regression models which will predict the bike rental counts on daily bases.
Counts of bike is depending of multiple factors like Season, weathercast, temperature etc.
Given below is a sample of the data set that we are using to predict the Bike rental counts

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	2011-01-08	1	0	1	0	6	0	2	0.165000	0.162254	0.535833	0.266804	68	891	959
9	2011-01-09	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.361950	54	768	822
10	2011-01-10	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321

As we can see in the table below we have the following 16 variables, using which
we have to correctly predict the counts of bike:

Predictor Variables :
instant, dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, temp, atem, hum, windspeed, casual, registered, cnt

Chapter 2

Methodology

2.1 Pre Processing

Any predictive modeling requires that we look at the data before we start modeling.
However, in data mining terms looking at data refers to so much more than just looking.
Looking at data refers to exploring the data, cleaning the data as well as visualising
the data through graphs and plots.
This is often called as Exploratory Data Analysis.
To start this process we will first try and look at all the
probability distributions of the variables.
Most analysis like regression, require the data to be normally distributed.
We can visualise that in a glance by looking at the probability distributions or
probability density functions of the variable.

In Figure 2.1 we have plotted the histogram with Kernel density Estimations (KDE)
all the continuous data columns.
The blue lines indicate Kernel Density Estimations (KDE)1 of the variable.
So as you can see in the figure most variables either very closely,
or somewhat imitate the normal distribution.

2.1.1 Missing value Analysis

I have analyzed missing value through R and Python code. But there is no missing value found in given dataset.

	Missing_percentage
dteday	0.0
season	0.0
yr	0.0
mnth	0.0
holiday	0.0
weekday	0.0
workingday	0.0
weathersit	0.0
temp	0.0
atemp	0.0
hum	0.0
windspeed	0.0
casual	0.0
registered	0.0
cnt	0.0

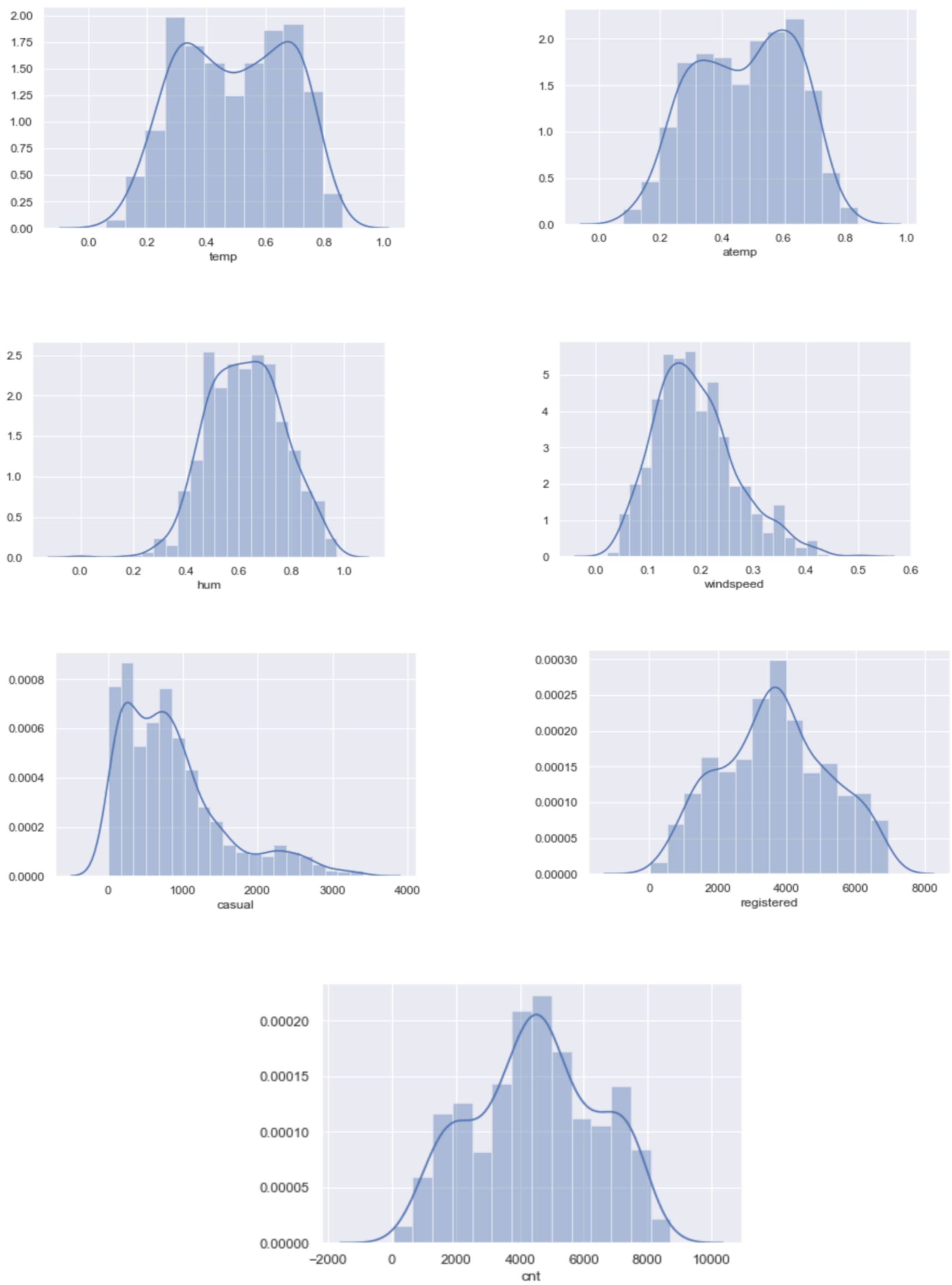


Fig 2.1

2.1.2 Outlier Analysis

Sometimes outliers can mess up an analysis
We usually don't want a handful of data points to skew the overall results.
It's important to dig into what is causing our outliers, and understand where they are coming from.
You also need to think about whether removing them is a valid thing to do, given the spirit of what it is we're trying to analyze.

One of the other steps of pre-processing apart from checking for normality is the presence of outlie
In this case we use a classic approach of removing outliers
We visualize the outliers using *boxplots*.

As we can see in figure 2.1.2 that variables "hum" and "windspeed" have outliers
So, we have to remove that

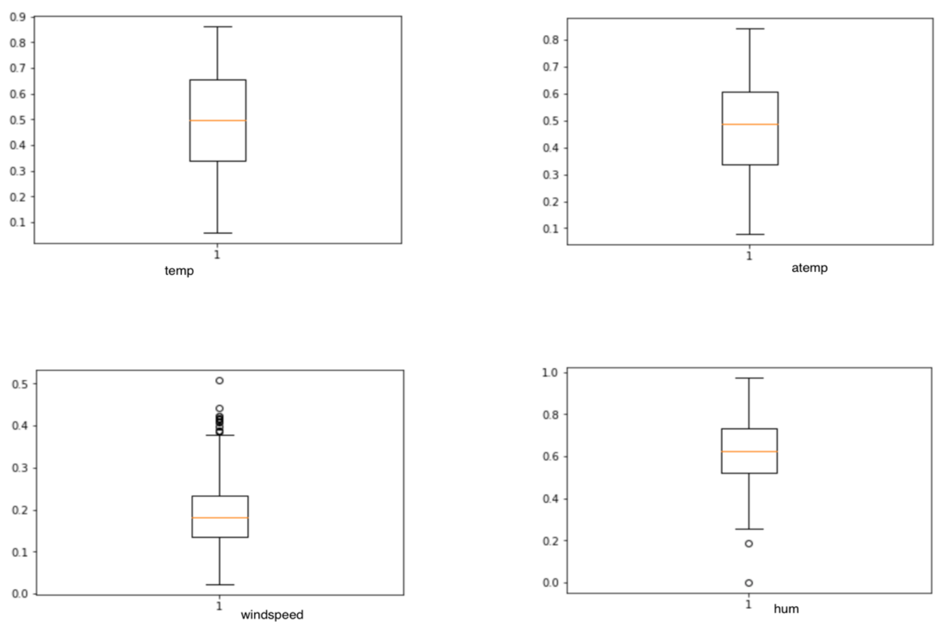


Fig 2.1.2

I've handle outliers in two ways :
1) Remove obeservation which contains outliers
2) Try to impute outliers with Mean, Median, and KNN

I've remove observations in Python which contains outliers
In R I've impute with mean/median
As records are low we can not impute with KNN method

2.1.3 Feature Engineering

In Feature Engineering I've checked data types
And set accourding for model to suitable

Before changing datatypes vs after changing datatypes
In Python initially data types are int,float, object
We need to define which variable is categorical and which variable is numeric/any other type
So we can train our model according to it.
If categorical variable used as numeric in Model development, it might reduce accuracy of our Mod

Below are the variables before changing it's datatype

Variable	Type
dteday	object
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

Feature Engineering

```
bike_data.dteday = pd.to_datetime(bike_data.dteday, yearfirst=True)
bike_data.season = bike_data.season.astype('category')
bike_data.yr = bike_data.yr.astype('category')
bike_data.mnth = bike_data.mnth.astype('category')
bike_data.holiday = bike_data.holiday.astype('category')
bike_data.weekday = bike_data.weekday.astype('category')
bike_data.workingday = bike_data.workingday.astype('category')
bike_data.weathersit = bike_data.weathersit.astype('category')

bike_data.temp = bike_data.temp.astype('float')
bike_data.atemp = bike_data.atemp.astype('float')
bike_data.hum = bike_data.hum.astype('float')
bike_data.windspeed = bike_data.windspeed.astype('float')
bike_data.casual = bike_data.casual.astype('float')
bike_data.registered = bike_data.registered.astype('float')
bike_data.cnt = bike_data.cnt.astype('float')
```

Below are the variables after changing it's datatype

Variable	Type
dteday	datetime64[ns]
season	category
yr	category
mnth	category
holiday	category
weekday	category
workingday	category
weathersit	category
temp	float64
atemp	float64
hum	float64
windspeed	float64
casual	float64
registered	float64
cnt	float64
dtype: object	

2.1.4 Feature Selection

Feature selection is Selecting a subset of relevent features (Variables, predictors) for use in model construction.
Subset of a learning algorithm's input variables upon which it should focus attention , while ignorir the rest
Before performing any type of modelling 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 bike counts prediction.

We remove that variables which are highly correlated to each other in Feature Selection.

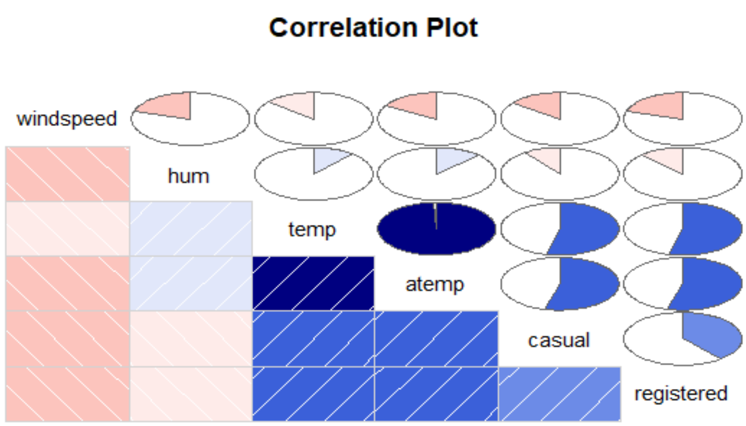


Fig 2.1.4

As we can see that temp & atemp are highly correlated
We will remove atemp

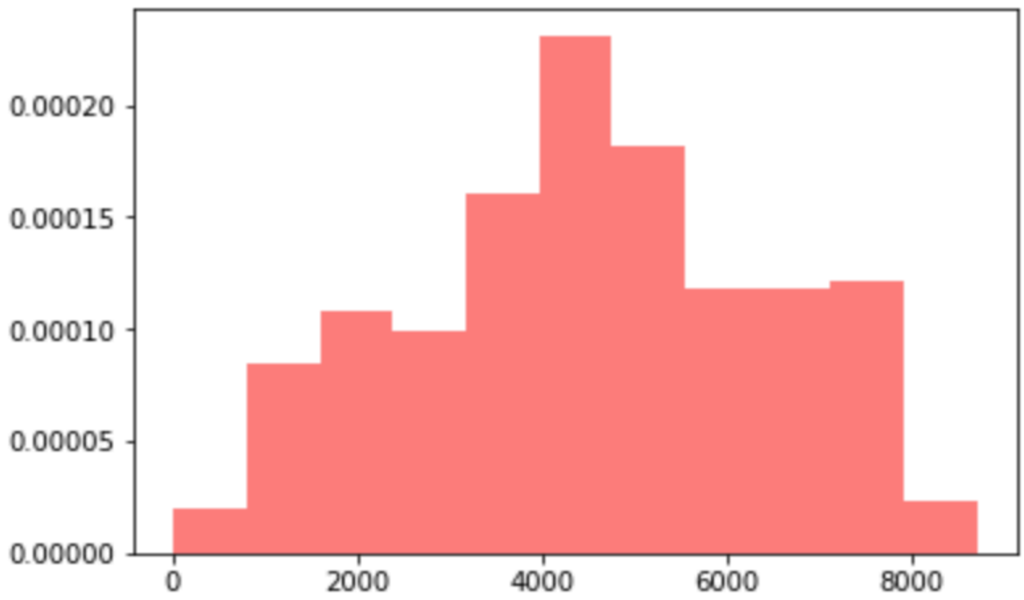
I 've done chi-square test of indepenent variables

We will remove weekday, holiday because they don't provide much to the target variable
We will remove Casual and registered because tha's what we need to predict

2.1.5 Exploratory Data Analysis

Exploratory Data Analysis refers to a set of techniques originally developed by John Tukey. To display data in such a way that interesting features will become apparent.
Unlike classical methods which usually begin with an assumed model for the data, EDA techniques are used to encourage the data to suggest models that might be appropriate.

I've done some EDA through pyrhon to visualise analyze the behaviour of target varible



Distribution of target varable(cnt)

Fig 2.1.5.1

Fig 2.5.1.2

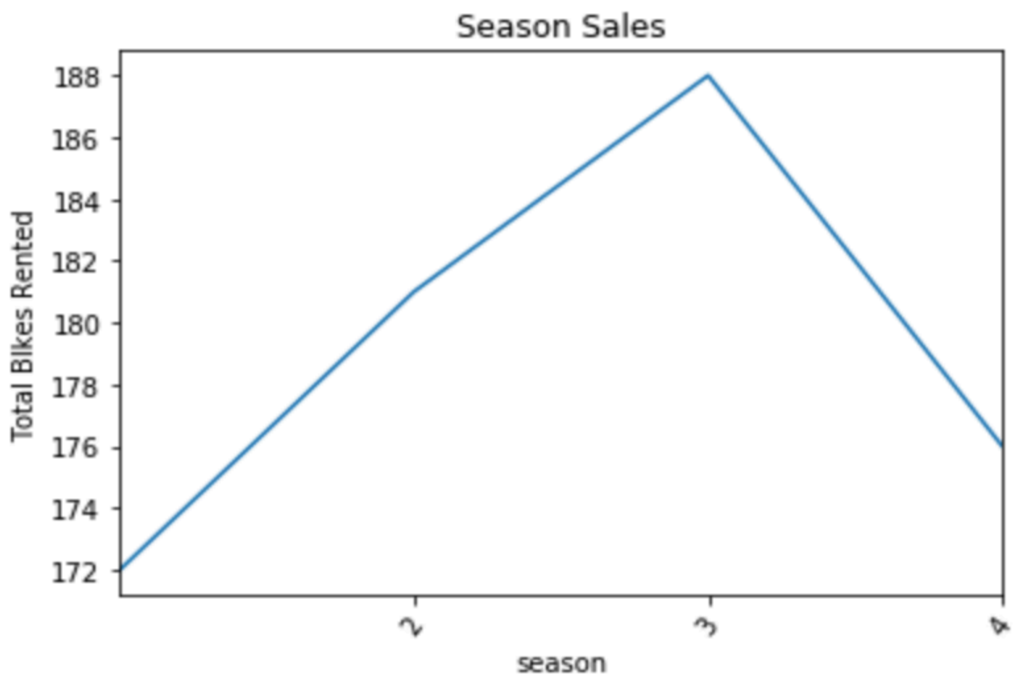


Fig 2.5.1.3

2.2 Modeling

2.2.1 Model Development

In model development we should have clean data
As we have done data cleaning and pre-processing that data is used in model development

I have try below three model to predict bike counts
As these problem is regression analysis problem we can not use classiffication ML algorythm

1) Decision Tree
2) Linear regression
3) Random Forest

1) Decision Tree

Using Decision Tree Algorythm on Train data and apply it to test data
We can achieve 84.14% accuracy

2) Linear regression

Using Decision Tree on Train data and apply it to test data
We can achieve 83.25% accuracy

3) Random Forest

Using Random forest Algorythm on Train data and apply it to test data
We can achieve 86.38% accuracy

Chapter 3

Conclusion

3.1 Model Evaluation

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

3.1.1 Mean Absolute Percentage Error (MAPE)

MAPE is one of the error measures used to calculate the predictive performance of the model.
We will apply this measure to our models that
we have generated in the previous section.

3.2 Model Selection

We can see that three models perform comparatively on average and
therefore we can select either of the one model

As we have seen that Random Forest is more accarte than other models
We are going to select Random Forest

As Random forest model have much better value of R squared
And adjustend R squared