Bike Rental Linear Regression

Contributor :Ishita Das

Introduction

Problem Statement

A bike sharing company has experienced considerable dips in their revenues due to Covid.

The company aims to revive in post pandemic market situation

Business Objective

The company management would like to understand the key variables that influence their bike rentals in positive / negative way.

There from they will adapt necessary actions to increase their revenues.

Implementation Methodologies

1. Data understanding
2. Visualize data
3. Test-Train Split
4. Scaling of train data
5. Linear model building
6. Evaluate model on train data
7. Apply model on test data
8. Predict and Evaluate

Data Understanding

- Data is loaded in pandas data frame from .csv file
- We investigate the shape and size of the data frame
- We note the columns and significance of each
- Data set has 730 rows
- Columns weathersit, season although are given as int but they are categorical variables in nature.
- We find no null values present in the data set

Exploratory Analysis

We plot scatter plot and box plot between all the numerical variables

Inference:

- We find both positive and negative linearity between target variable cnt and the other numerical variables.

- Variable temp and atemp are strongly colinear

- Season wise we can see median is much higher for 3 (fall) compared to 1 (spring).

- Year wise a higher trend in 2019 compared to 2018
to year growth.

- Month wise we see the pick at around 6,7,8,9 with median is more than 5000. Where as month 1,2,3 and 11,12 saw median less than 4000. This indicates the possibility of month wise booking trend.

- Demand is higher on weekday compared to holidays.

Dummy variables

- From data dictionary we know that few of the variables although given as int they are not continuous variables, rather they are categorical variables. We have to create dummy variables for those categorical variables.

- Such variables are
  - mnth
  - weekday
  - weathersit
  - season

Linear Model data Preparation

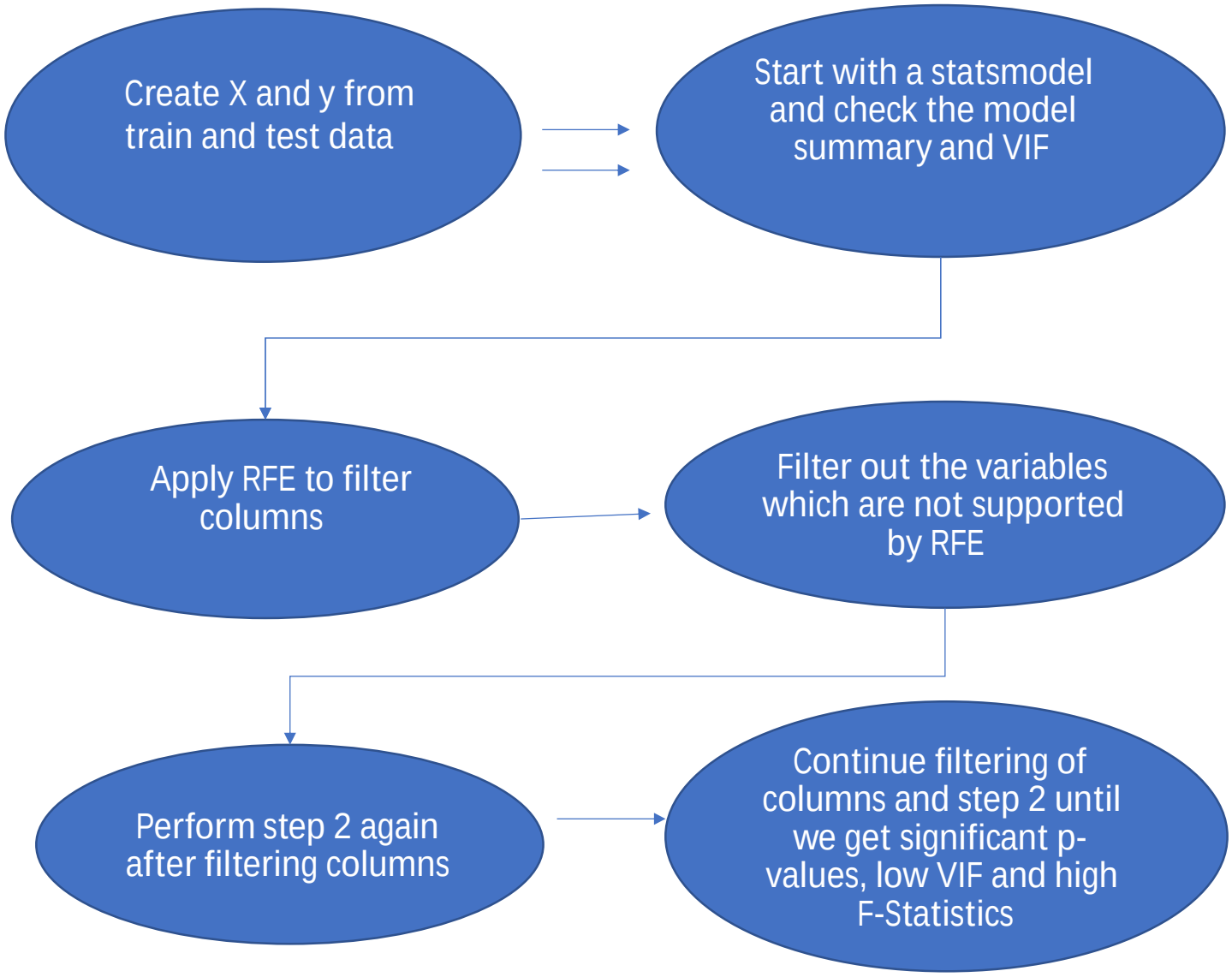
Test-Train Split

We will perform train-test split to a ratio of 70:30

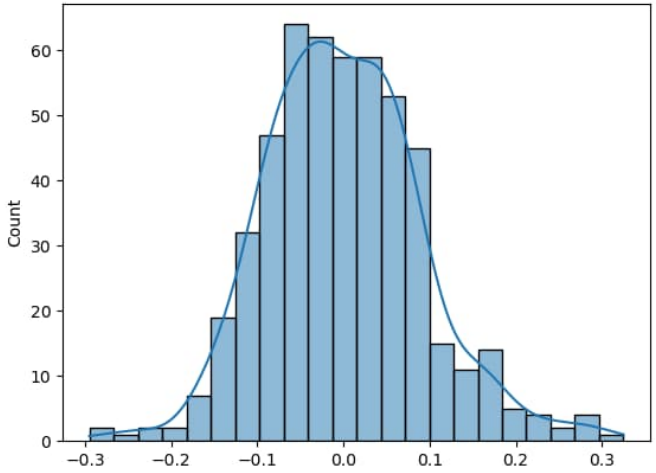
Scaling of train data

We want to perform scaling for the numeric
Variables. We see humidity is much higher in values .So, unless we perform scaling, we will not be able to find the dependencies of the predictor variables.

Linear model building



Step 8 - Evaluate model on Train data



- We will see the residual and plot it and check if the residual is normally distributed.

- We conclude that the residuals are normally distributed from the above plot.

Apply model on test data

Scaling of Test Data

On test data we don't perform fit() as fit() calculates min , max and we are not supposed to know that on the test data set (unknown data). We use the same fit from the train data and transform the test data using the scaler object we created in train data set.

We filter all the columns from the test data those came out as insignificant in our model

Predicting on Test Data

Finally, we predict on test data using the model

We find the R-square value on test data

And we draw a scatter plot between the predicted values and the actual values

Conclusion

- The model shows R-square value of 75 which is close to the train data
- We see a linear the predicted values and the actual values are mostly fitted on a straight line