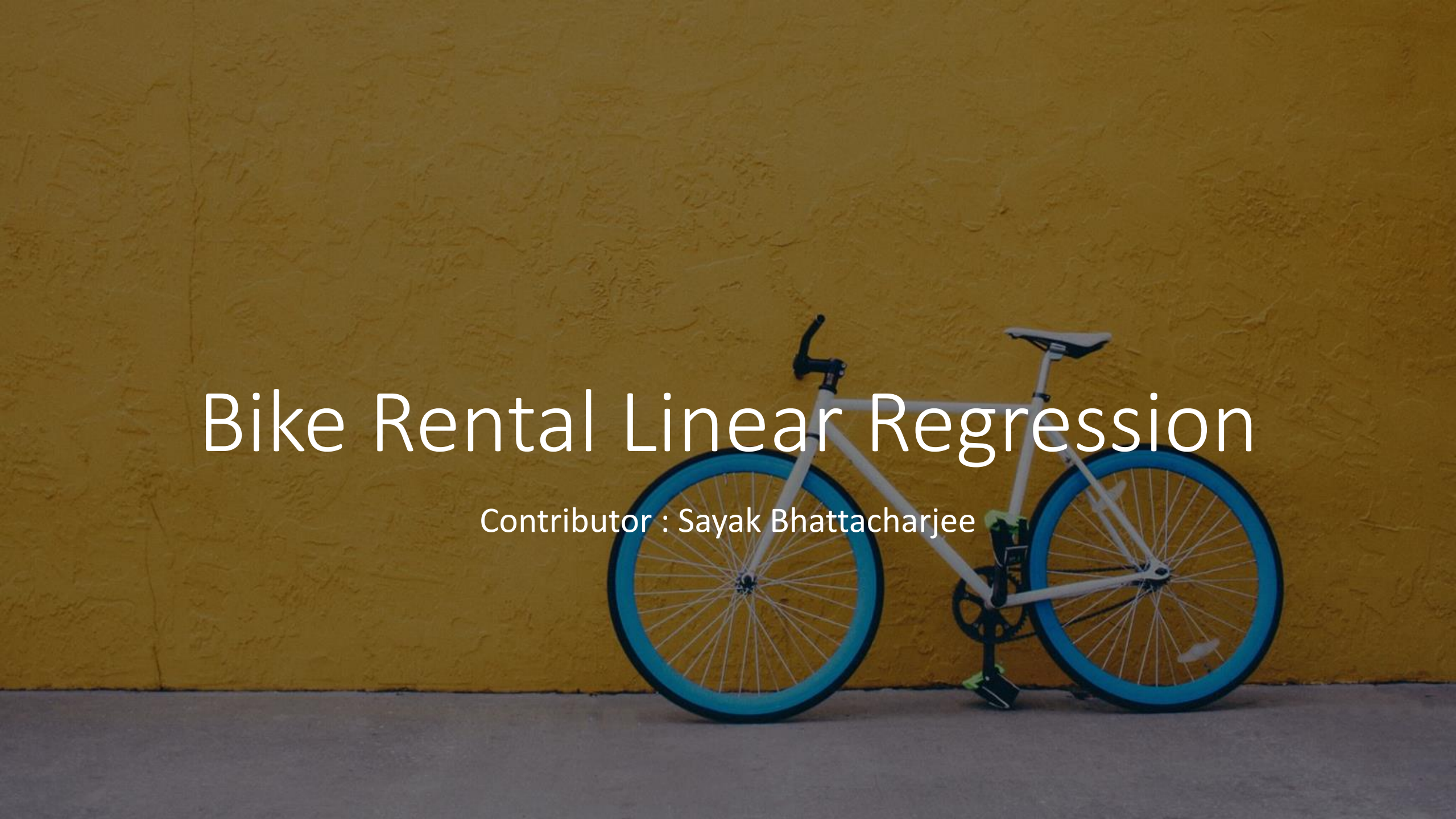


Bike Rental Linear Regression

Contributor : Sayak Bhattacharjee

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.

Solution Approach

We would like to perform data analysis followed by a model set up that can predict the variables with an optimal accuracy.

We will apply linear regression model only to achieve the business objective (we will discuss in detail how we create a linear regression model)

Implementation Methodologies

To achieve the below 2 factors

- Which variables are significant in predicting the demand for shared bikes.
- How well those variables describe the bike demands

We have performed the steps in order mentioned in the right side

1. Data understanding
2. Data quality check
3. Exploratory data analysis on train data
4. Preparation of dummy variables
5. Test-Train Split
6. Scaling of train data
7. Linear model building
8. Evaluate model on train data
9. Apply model on test data
10. Conclude model quality

Step 1 – Data Understanding

- We read the file and load in pandas data frame.
- We investigate the shape and size of the data set
- We observe the data types of the columns

Our findings

1. Dataset is having 730 number of rows which are not enough for a predictive model.
2. We have one date field dteday with datatype as object.
3. We have already learnt from the problem statement that some of the columns like `weathersit,season` although are given as int but they are categorical variables in nature. We will handle them later.

Step 2 - Data Quality Check

We find no null values present in the data set

We also could not trace any junk values in the data set

We conclude the data set does not need any sanity treatment as the quality of data is pretty good

Step 3 - Exploratory data analysis

- We plot scatter plot between all the numerical variables
- We plot box plot between all the categorical variables

- Our Conclusion

- 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). This indicates that season is a potential predictor variable.
- year wise a higher trend in 2019 compared to 2018, i.e. there is an year 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.
- A clear day shows much higher demand compared to rainy or snowy day. So weathersit is a potential predictor.

Step 4 - Preparation of 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

- There are few more as `yr, holiday, workingday` , but we don't have to create dummy var for them as they have only 2 categories

Linear Model data Preparation

Step 5- Test-Train Split

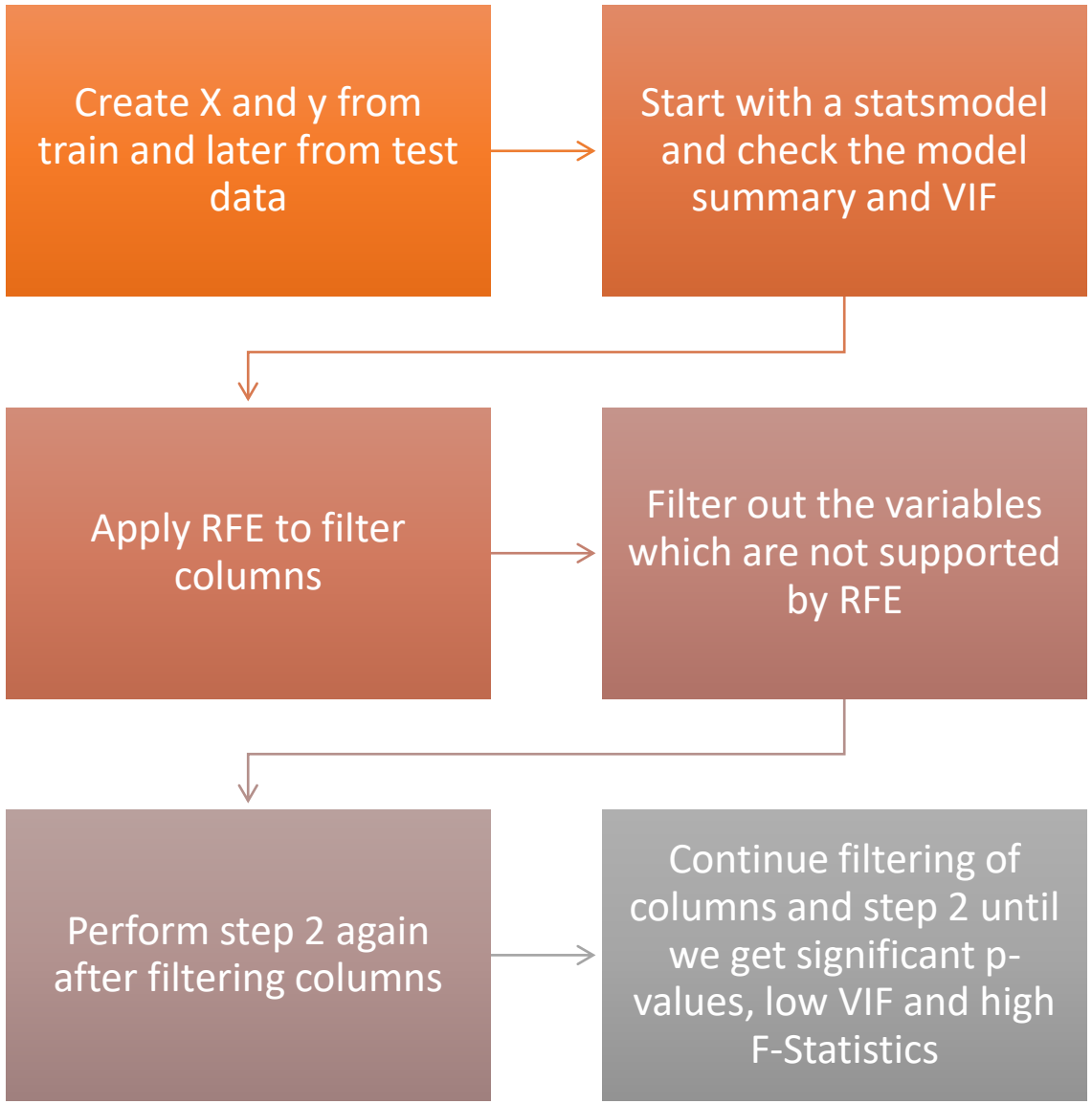
- We will perform train-test split to a ratio of 70:30 using skt-learn library

Step 6 - Scaling of train data

We want to perform scaling for the numeric variables

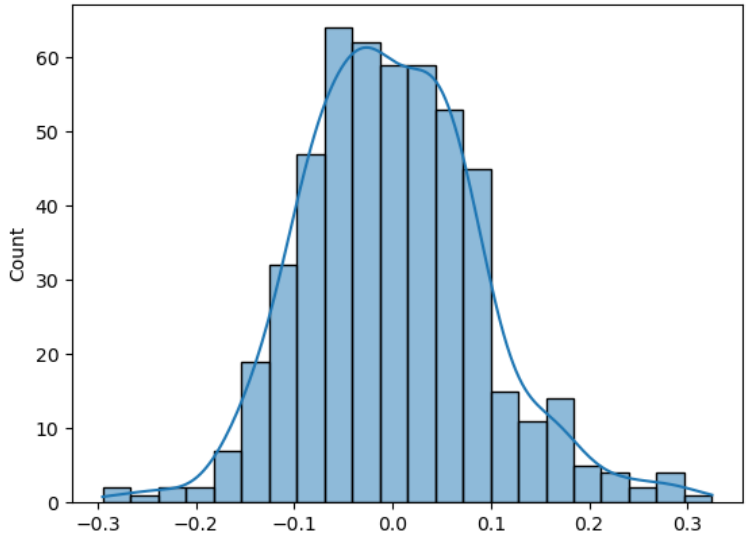
We see humidity is much higher in values than some of the others like - temp,atemp,windspped. So, unless we perform scaling we will not be able to find the dependencies of the predictor variables.

Step 7 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.

Step 9 - 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

Step 10 - Conclude on model quality

- 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