Project report - Bike Prediction

Darshan Bhansali

Aim

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

The Dataset

Structure of the given dataset:

- ID variables :
  - “instant” : Numeric (unique record instance)
  - “dteday” : Date (Calendar date corresponding to each day)

The Dataset

Structure of the given dataset:

- Independent features
  - “Season” : Categorical (4 levels - Spring, Summer, Fall, Winter)
  - “yr” : Categorical (2 levels - 2011, 2012)
  - “mnth” : Categorical (12 levels - January to December)
  - “holiday” : Categorical (2 levels - Holiday/Non-Holiday)
  - “weekday” : Categorical (7 levels - Sunday - Saturday)
  - “workingday” : Categorical (2 levels - Weekday/Weekend)
  - “weathersit” : Categorical (4 levels - Clear, Mist/Fog/Cloud, Light Rain/Snow, Heavy Rain/Snow)
  - “temp” : Numeric (temperature in celcius)
  - “atemp” : Numeric (temperature as it “feels like”)
  - “hum” : Numeric (humidity)
  - “windspeed” : Numeric (windspeed)

The Dataset

Structure of the given dataset:

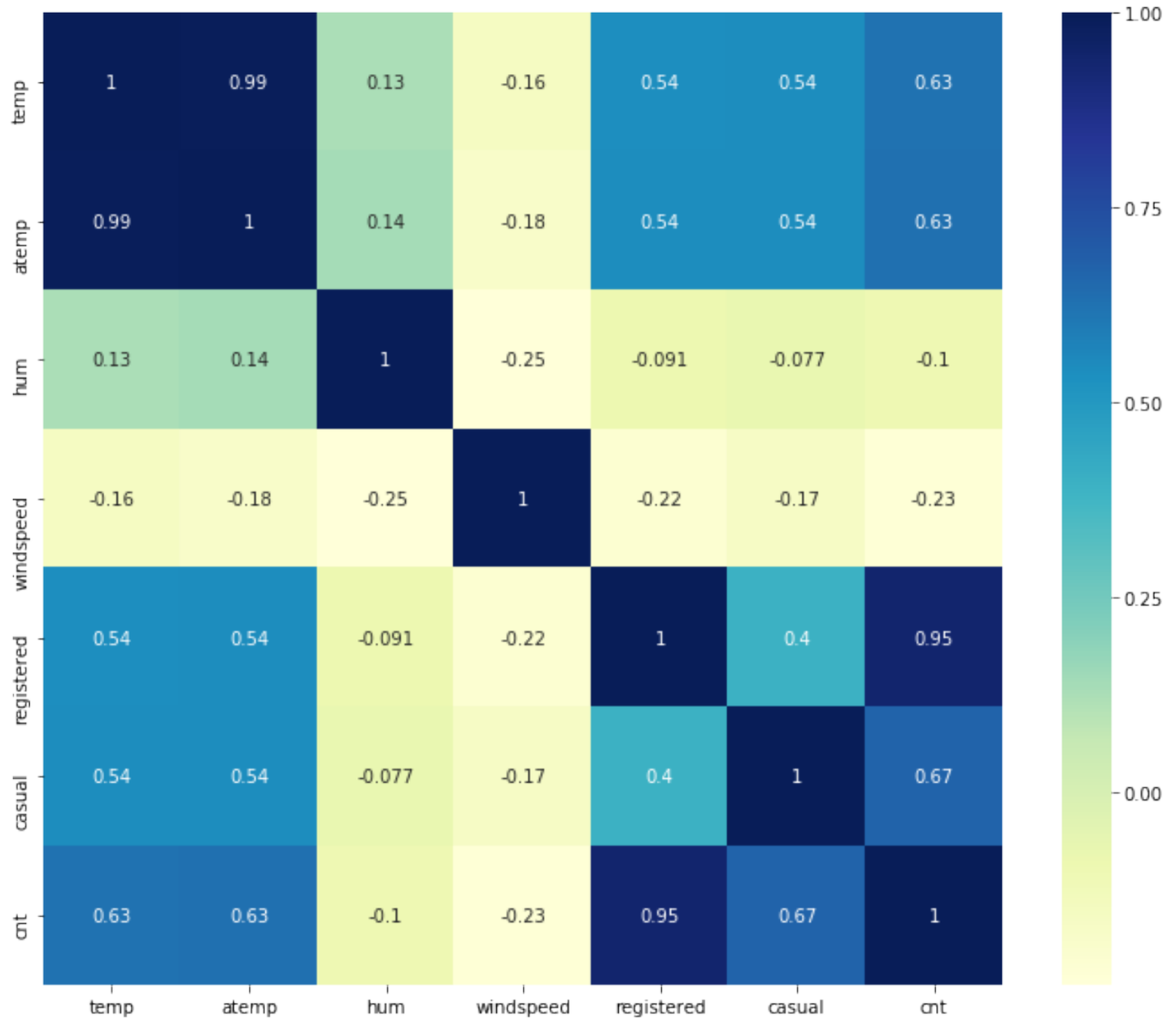
- Dependent features
  - “casual” : Numeric (Count of casual users)
  - “registered” : Numeric (Count of registered users)
  - “cnt” : Numeric (total count of users “casual + registered”)

Overview of the data

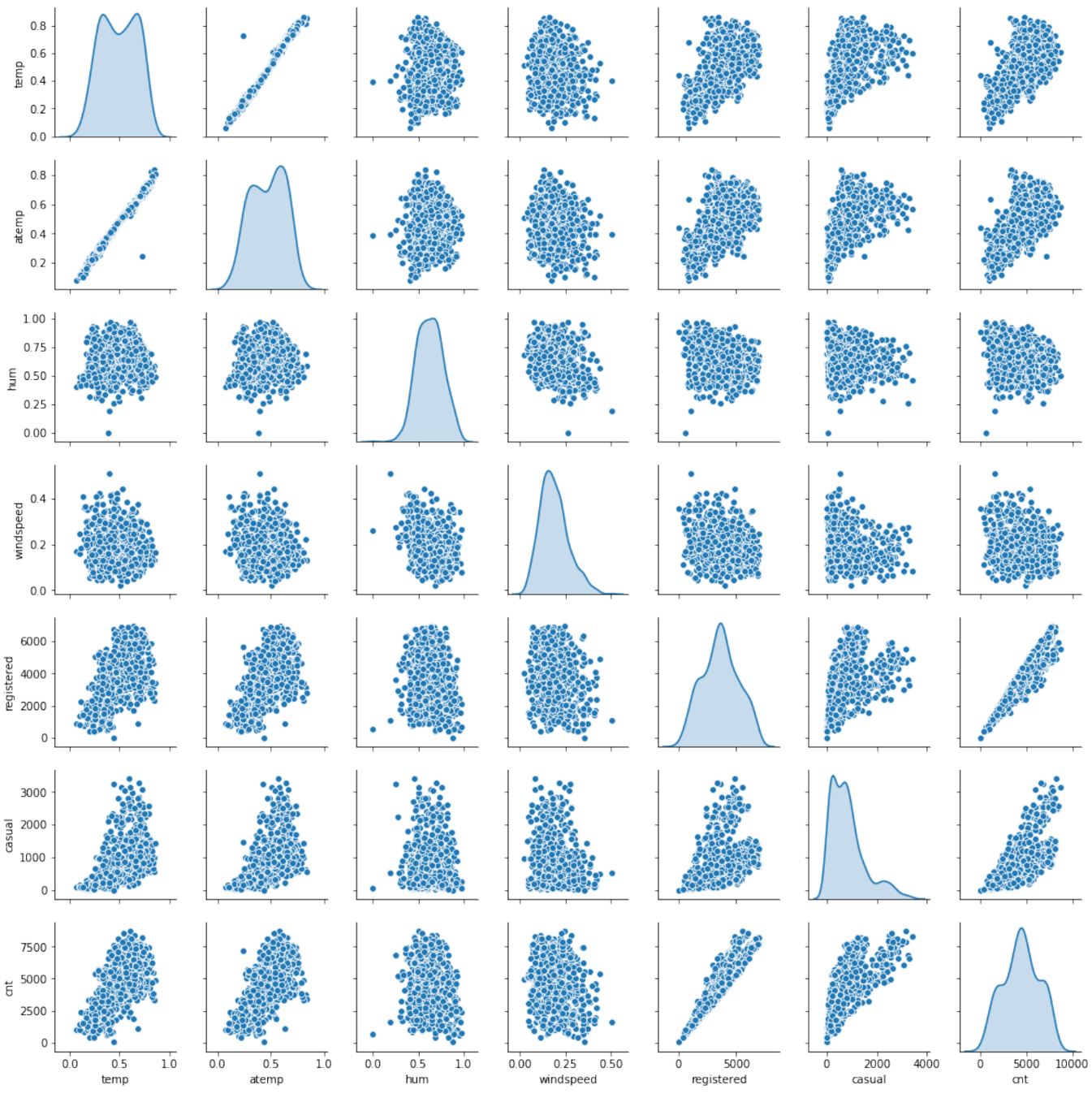
We'll see the following:

- Correlation heat map of the continuous features of the dataset
- Pair plots of continuous features for distribution
- Bar plots of categorical features for classification
- Box plots of continuous variables for outlier analysis

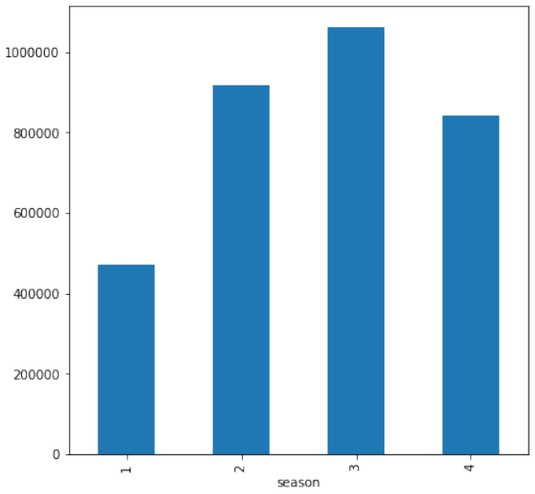
Heat map of continuous features

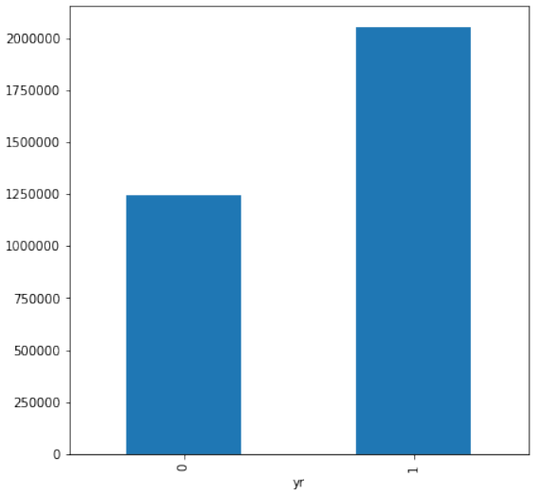


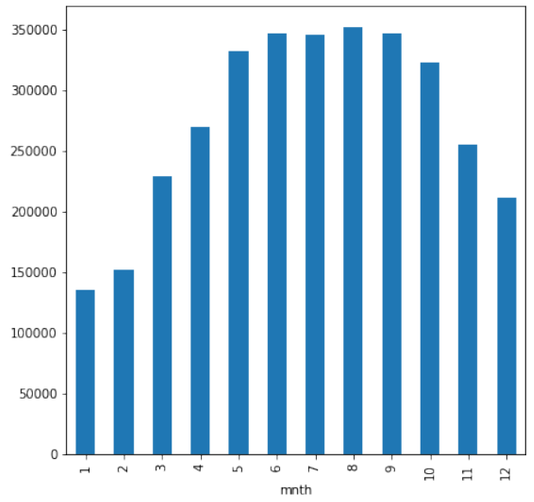
Pair plots of continuous features

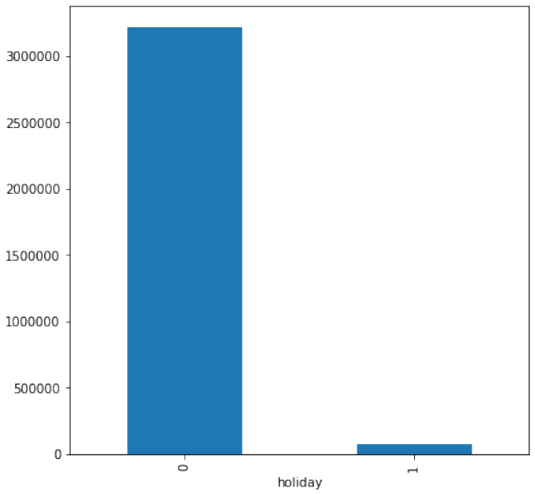


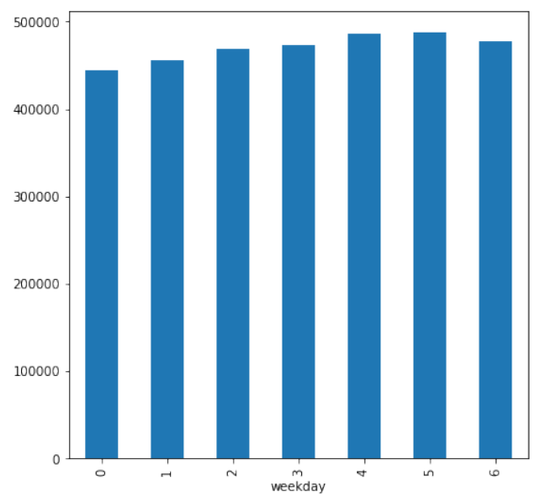
Bar plots of categorical features

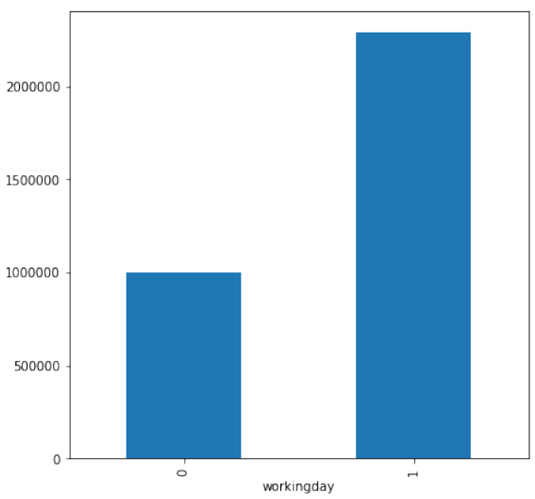


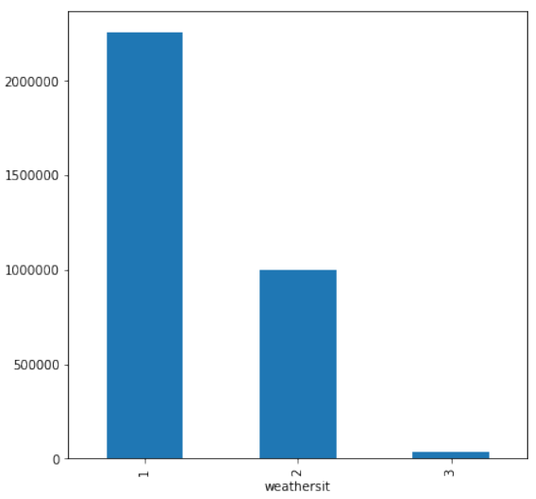




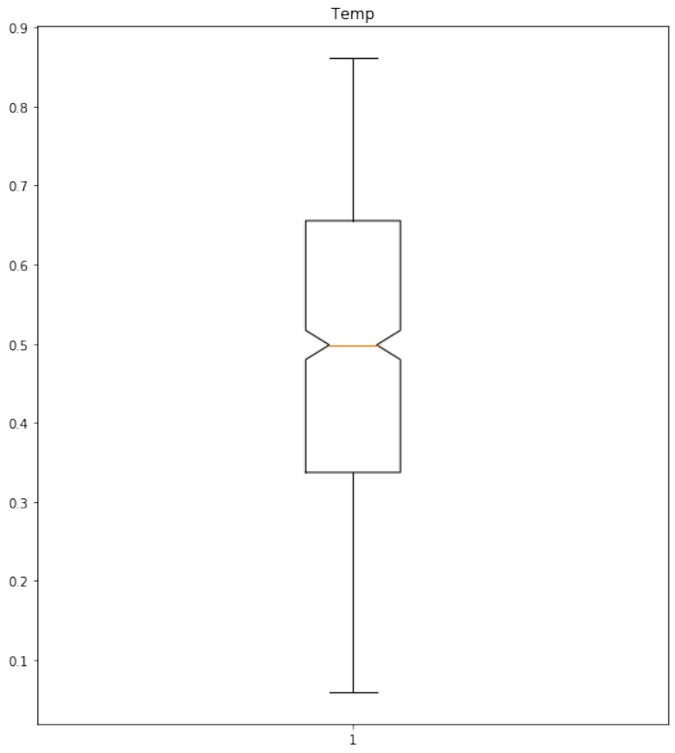


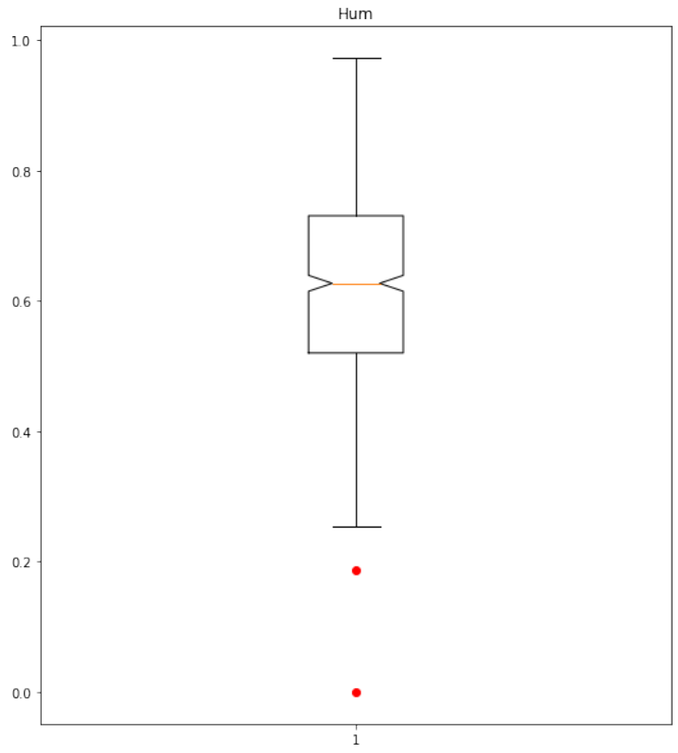


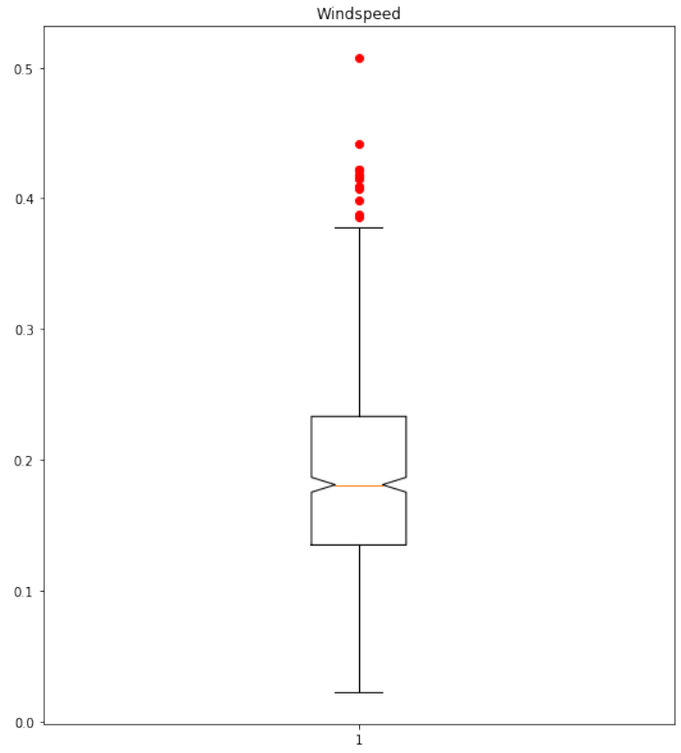




Box plots of continuous features







Observations & Preprocessing

- From the study of the various plots of the features in the data, we can made the following inferences:

  - ‘atemp’ is highly positively correlated to ‘temp’ and can be omitted from the model

  - ‘casual’ and ‘registered’ are highly correlated to ‘cnt’ (since cnt is the total) and so we can focus on predicting count alone as that is the aim of the project

  - Outliers in windspeed and humidity features need to be handled before constructing a model

Building models - R

The following algorithms are selected for the trials with R

- Decision Tree
- RandomForest
- Multivariate adaptive regression splines

Building models - R

Using MAPE as the primary evaluation metric

Decision Tree	RandomForest	MARS
MAPE : 18.65%	MAPE : 12.02%	MAPE : 15.51%

Building models - R

RandomForest model provides the least MAPE and the highest accuracy score

Decision Tree	RandomForest	MARS
MAPE : 18.65%	MAPE : 12.02%	MAPE : 15.51%

Building models - Python

The following algorithms are selected for the trials with Python

- Decision Tree
- RandomForest
- Multivariate adaptive regression splines

Building models - Python

Using MAPE as the primary evaluation metric

Decision Tree	RandomForest	Linear Regression
MAPE : 18.65%	MAPE : 12.02%	MAPE : 15.51%

Building models - Python

RandomForest model provides the least MAPE and the highest accuracy score

Decision Tree	RandomForest	Linear Regression
MAPE : 18.66%	MAPE : 15.54%	MAPE : 18.67%