BoomBikes

BIKE SHARING PREDICTION

SIVARAMAN SAPTHARISHI (SIVA)

Problem Statement

Challenges

A US bike-sharing provider BoomBikes has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario.

Expected Outcome

To generate a model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations.

DataSet Analysis

- Number of Columns in the dataset : 16
- Number of Rows in the dataset : 730
- Removing following variables
  - **Dteday** :
    - This has the date, Since we already have seperate columns for 'year' & 'month',hence, we could proceed without this column
  - **Instant** :
    - Its only an index value.
  - **casual & registered** :
    - Both these columns contains the count of bike by different categories of customers. Since our goal is to find the total count of bikes and not by specific category, we will ignore these two columns

DataSet Analysis

- **Categorical Value**

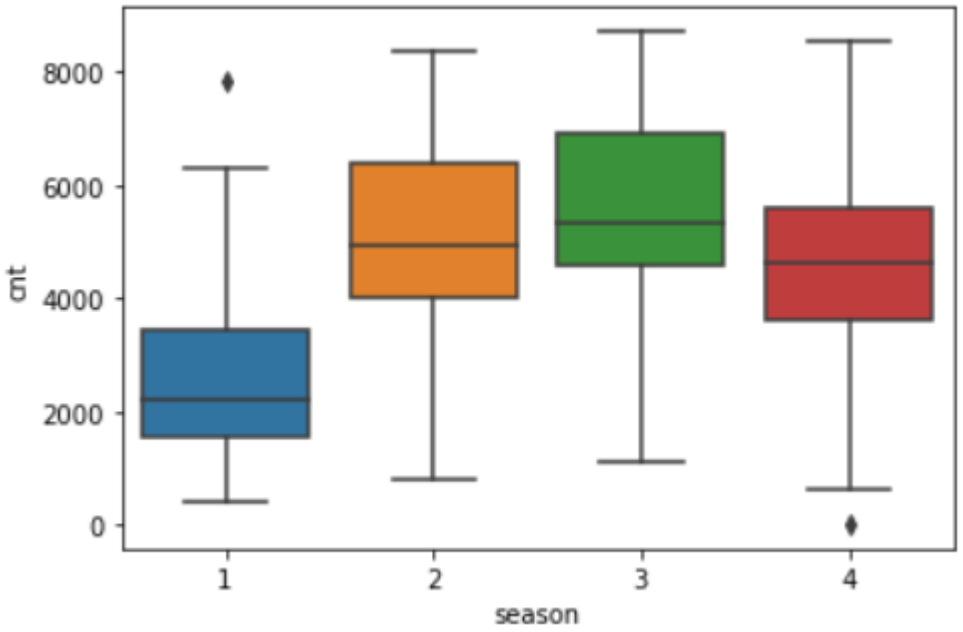
In this dataset below are some of the variable taken for analysis

- mnth
- weekday
- season
- weathersit
- yr
- holiday
- workingday
- temp
- atemp
- hum
- cnt

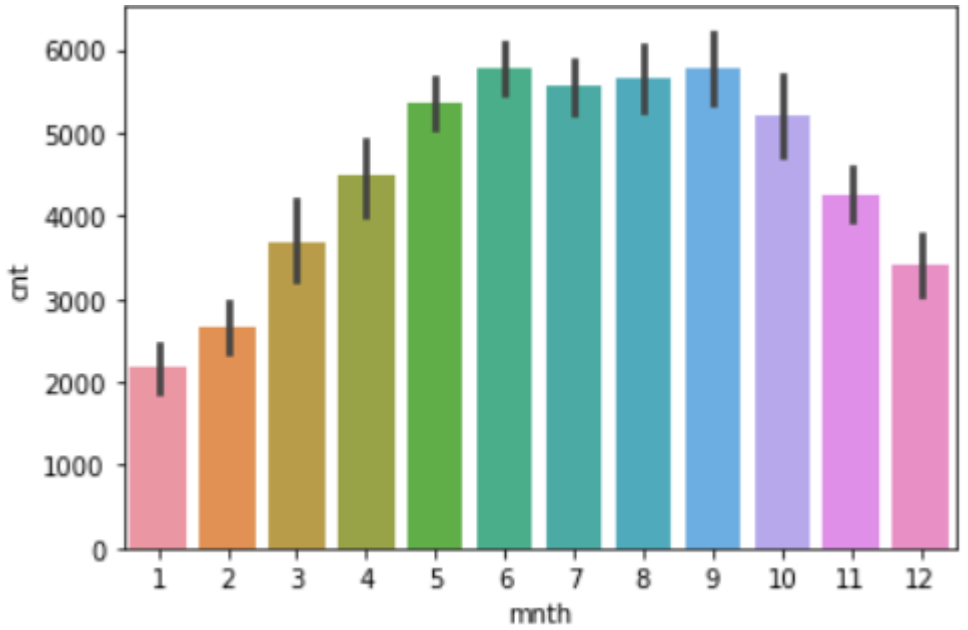
DataSet Analysis

- **DUPLICATE VALUE ANALYSIS**
  - There are no duplicate values in the dataset.
- **NULL VALUE ANALYSIS**
  - There is no null values in the entire dataset

Categorical Data Analysis

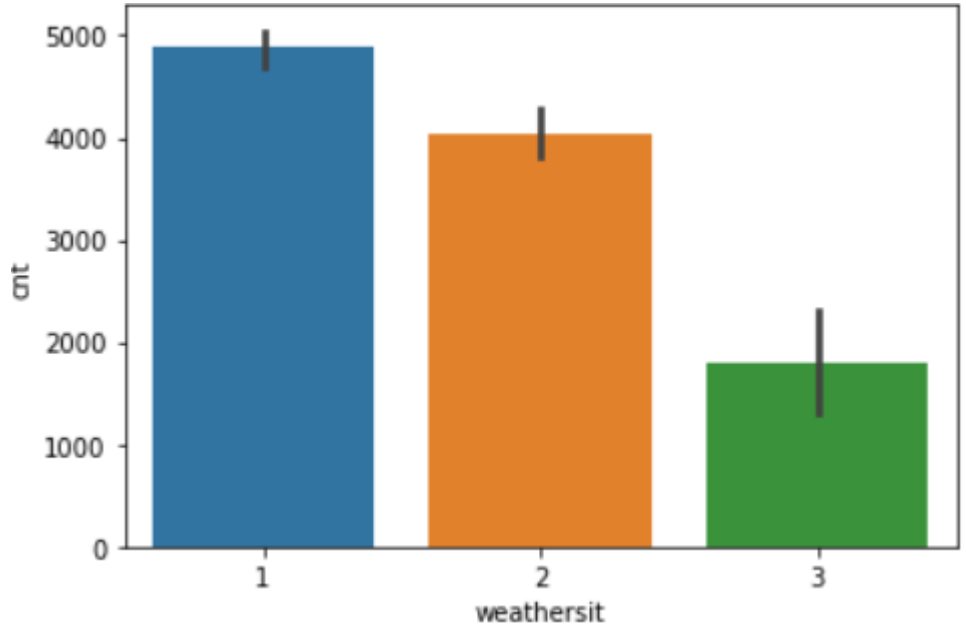


The categorical variable season can be a good predictor where more bike booking(more than 5000 happens) in season 3, following by season 2 , season 4 and then season 1

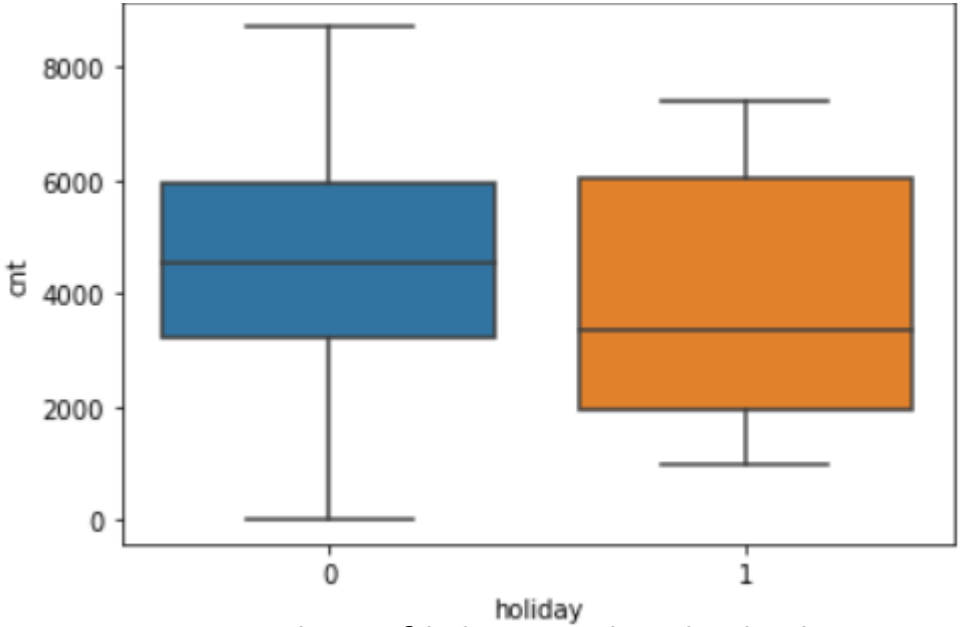


More bike users are in the month 6 to 9. The bike riders keeps increasing from month 1 to 5.
From month 6 to 9, it doesnt vary much and decreases vastly from month 9 to 12.

Categorical Data Analysis



Nearly 70% of the bike booking were happening during ‘weathersit1. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.



More number of bikes are booked when it is not a holiday from median. But not sure if it is a weekday or working day. Hence this variable can be ruled out and variable weekday and workind day can be taken into account.

Categorical Data Analysis

From the categorical variable analysis following variables contribute a lot for prediction of the bike users

1) season 2) mnth 3) weathersit 4) workingday

Holiday and weekday are ruled out since, more bike users are seen when it is not a holiday.

Hence it can be considered as either working day or weekday.

There are possibilities where weekdays can be off and weekends can also be a working day hence, workingday will represent the holiday and weekday data.

Rescaling & Correlation

Correlation is applied after

Rescaling

Before scaling we saw that the Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'. After spliting and scaling the same corrleation exist with train data set

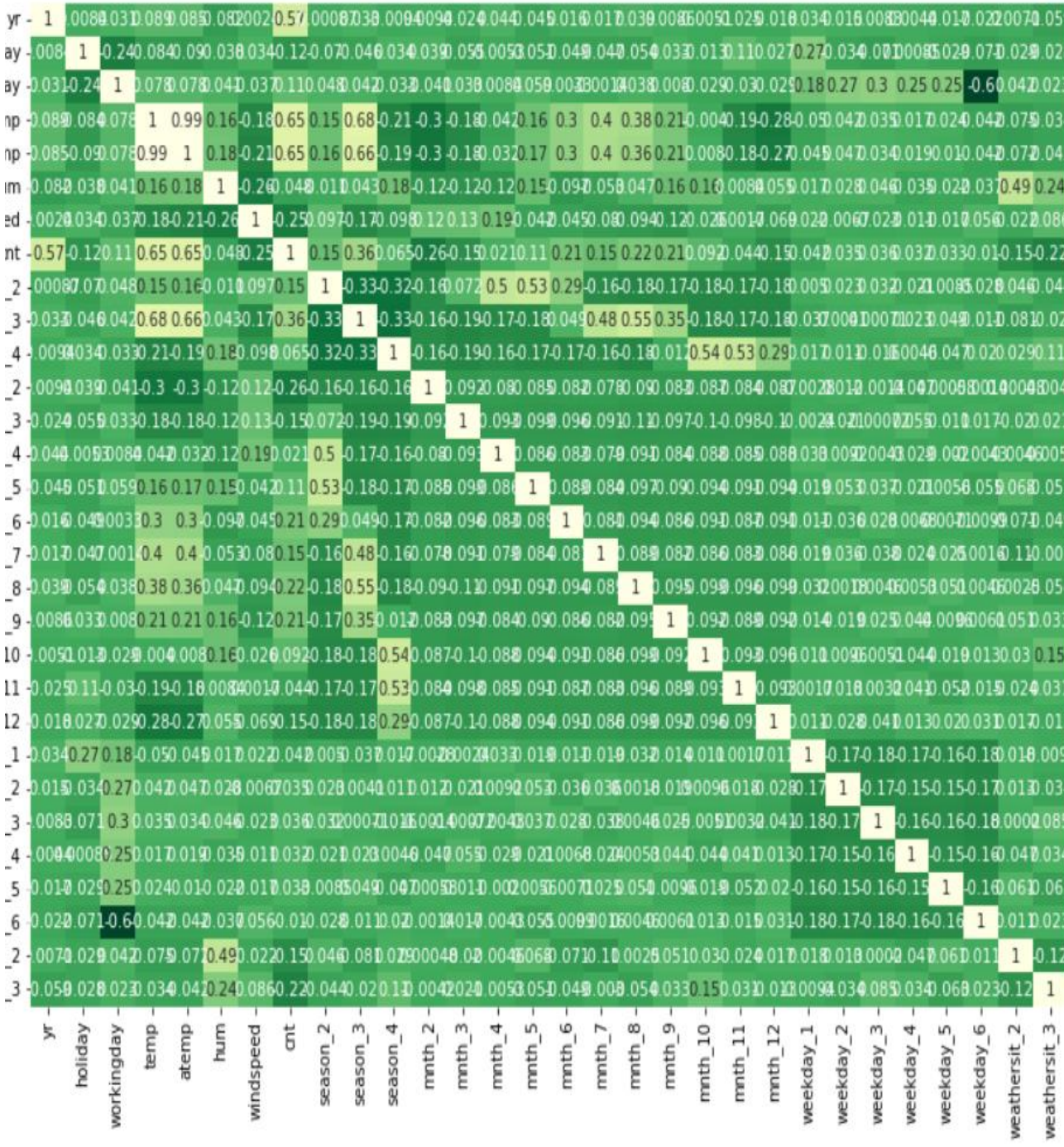
Most correlated values

A) numerical_vars

1) temp
2) atemp
3) hum
4) windspeed

B) categorical_vars

1) season
2) mnth
3) weathersit
4) workingday



Model Generation

- After first train with variable 'temp’ and then with ‘windspeed’
- Significance is deicided with P value, which is 0.
- The result is statistically significant
- R-squared is 0.424, which means that nearly 42% of the variance in bike users count is explained by atemp feature.
- Still model can be improved

Drop Variables based of VIF and p values

From Model_1 -> Removed mnth_11, mnth_12

From Model_2 -> Removed mnth_7

From Model_3 -> Removed 'season_3','weekday_5','weekday_3','weekday_4','mnth_2','weekday_2','weekday_1’

From Model_4 -> Removed mnth_4

From Model_5 -> Removed mnth_5 & mnth_6

From Model_6 -> Removed mnth_3

From Model_7 -> Removed mnth_10

From Model_8 -> Removed mnth_8

Inferences

- ✓From the lates LR_model which is generated out of X_Drop_7
- ✓it is evident that all our coefficients are not equal to zerowhich means We REJECT the NULL HYPOTHESIS
- ✓F-Statistics is used for testing the overall significance of the Model: Higher the F-Statistics, more significant the Model is.
- ✓ F-statistic: 276.0
- ✓Prob (F-statistic): 4.26e-204
- ✓The F-Statistics value of 276 (> 1) and the p-value of '~0.0000' states that the overall model is significant

Multiple Linear Regression Equation

cnt = 0.065078 + (yr × 0.230599) + (workingday × 0.059123) + (temp × 0.560123) − (windspeed × 0.152146) + (season2 × 0.091156) + (season4 × 0.137470) + (mnth9 × 0.095924) + (weekday6 × 0.069696) − (weathersit2 × 0.080801) − (weathersit3 × 0.288993)

Prediction with the model

- **Train R2 = 0.837**
- **Train Adjusted R2 = 0.834**
- **Test R2 = 0.781**
- **Test Adjusted R2 = 0.771**

Observation

Temperature:

A coefficient value of ‘0.560123’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.560123 units.

Weather Situation 3:

A coefficient value of ‘-0.288993’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.288993 units.

Year :

A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

The above three variables have more influence in achieveing maximum bike booking. Since its

coefficient values are more compared other values followed by other values.