Machine Learning

Capstone Regression Project

BIKE SHARING

DEMAND PREDICTION

PROJECT

FULLY EXPLAINED

By

Diksha Shejao

CONTENT

Defining Problem Statement

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually , providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Data Summary

The dataset contains weather information (Temperature , Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

(1) Date : year –month-day

(2) Rented Bike count – Count of bikes rented at each hour

(3) Hour – Hour of the day

(4) Temperature – Temperature in Celsius

(5) Humidity - %

(6) Wind speed - m/s

(7) Visibility – 10m

(8) Dew point temperature – Celsius

(9) Solar radiation – MJ/m2

(10) Rainfall - mm

(11) Snowfall – cm

(12) Seasons – Winter , Spring, Summer, Autumn

(13) Holiday – Holiday/ No holiday

(14) Functional Day – No(Non Functional Hours), Fun(Functional hours)

‘Rented Bike count is dependent variable’

· This dataset contains 8760 lines and 14 columns.

· Three categorical features ‘ Seasons’, ‘Holiday’, & ‘Functioning Day’.

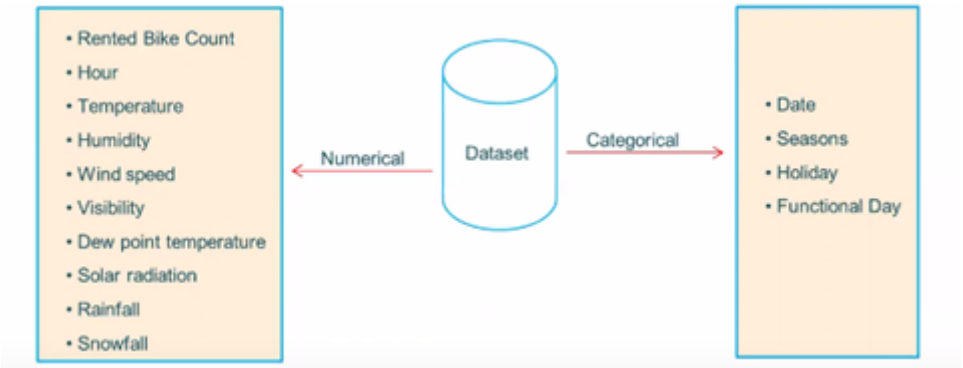
· One datetime features ‘Date’.

· We have some numerical type variables such as Temperature, Humidity , Wind , Vsibility , Dew point temp, Solar radiation , Rainfall, Snowfall which tells the environment conditions at that particular hour of the day.

```
[ ] df.head(5)
```

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

Data Summary



Insights From Our Dataset

- There are No Missing Values present.
- There are No Duplicate values present.
- There are No null values.

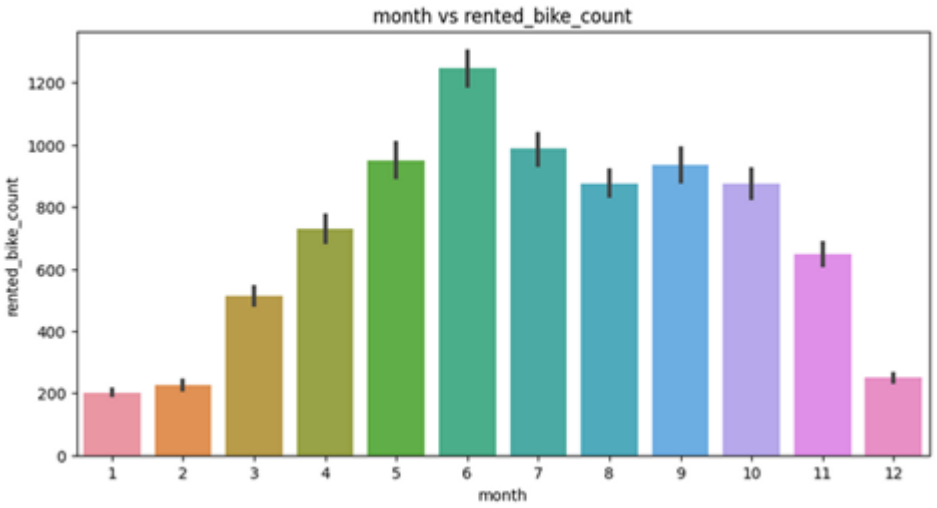
- And finally we have ‘rented bike count’ variable which we need to predict for new observations
- The dataset shows hourly rental data for one year

(1 December 2017 to 30 November 2018) (365days).We consider this as a single year data)

- So we convert the ‘date’ column into 3 different column i.e ‘year’, ‘month’, ‘day’.
- We change the name of some features for our convenience, they are as below ‘Rented_Bike_Count’, ‘Hour’, ‘Temperature’, ‘Humidity’, ‘Wind_speed’, ‘Visibility’, ‘Dew_point_temperature’, ‘Solar_Radiation’, ‘Rainfall’, ‘Snowfall’, ‘Seasons’, ‘Holiday’, ‘Functioning_Day’, ‘Month’, ‘Weekdays_weekend’

EDA

Let us see how the values of ‘Rental Bike Count’ are distributed in given dataset. Distribution of values is highly positively skewed.



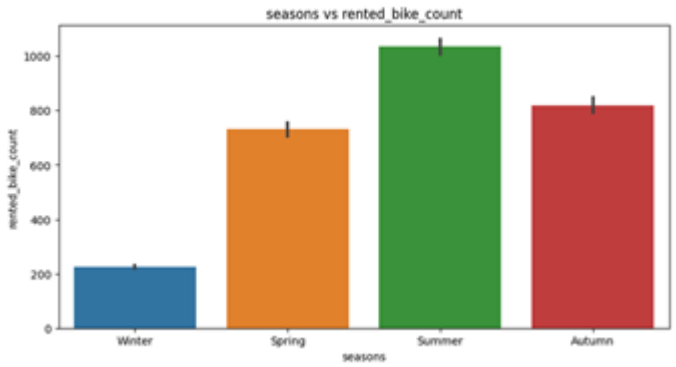
· From the above point polt and bar plot we can say that in the weekdays which represent in blur colour show that the demand of the bike higher because of the office.

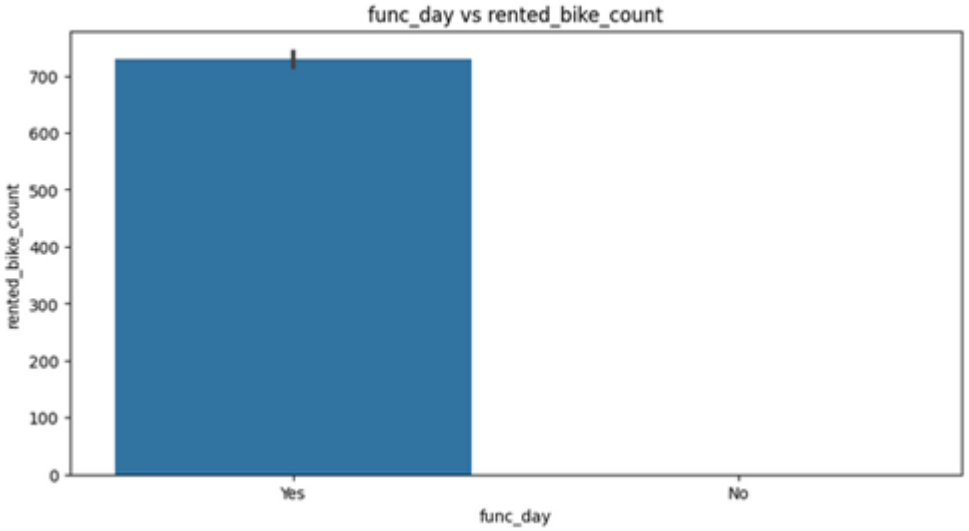
· The orange colour represent the weekend days, and it show that the demand of rented bikes are very low.

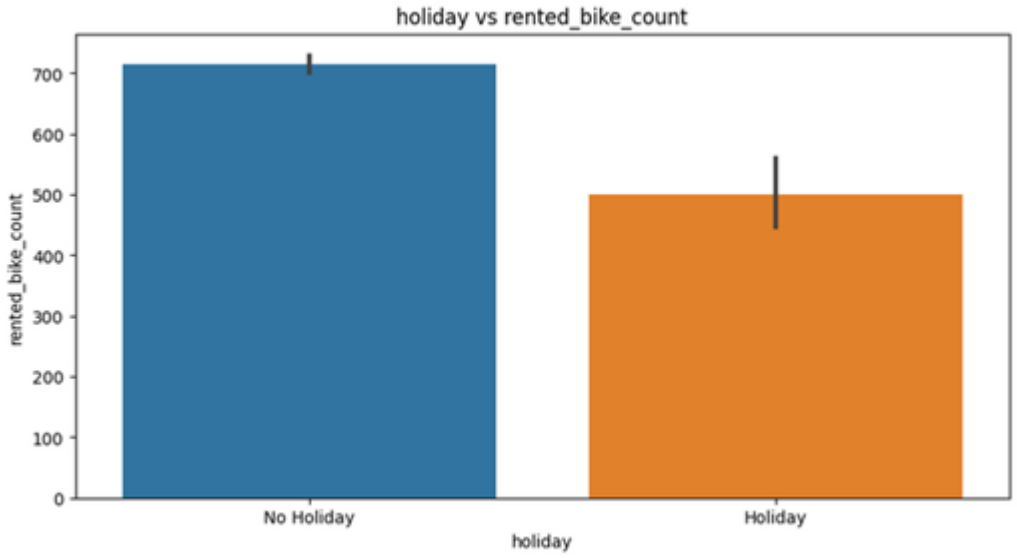
· From the month 5 to 10 the demand of the rented bike is high as compare to other months , these months are comes inside the summer season.

Count of values of categorical features.

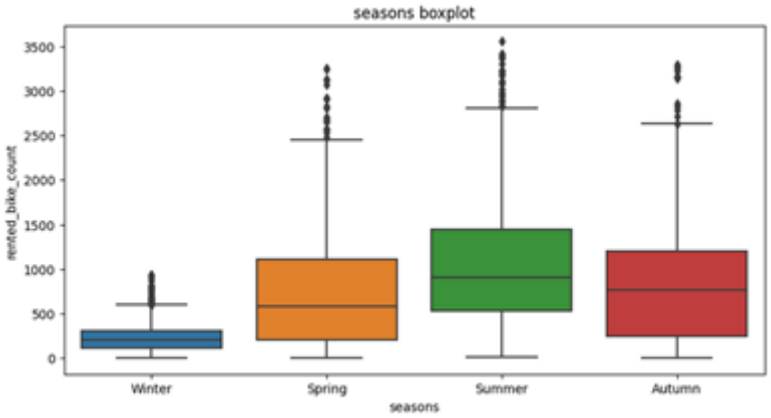
Functioning Day and Holiday have highly imbalanced count of values.

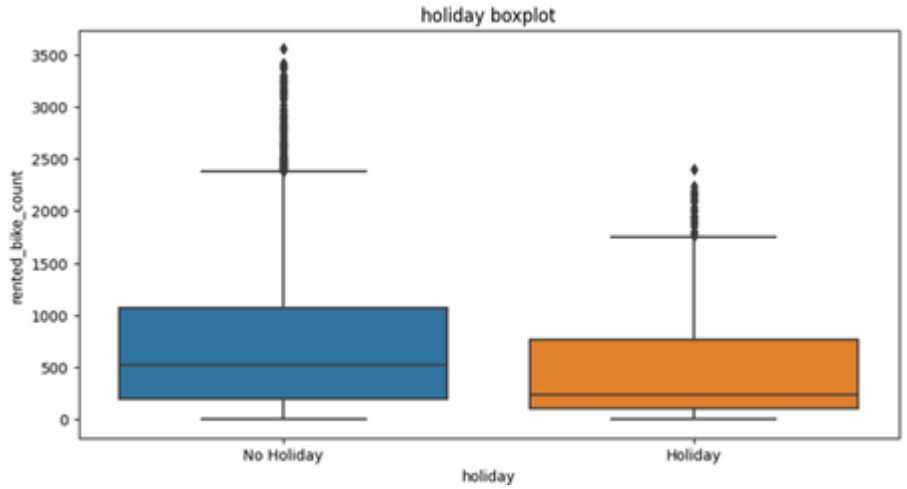


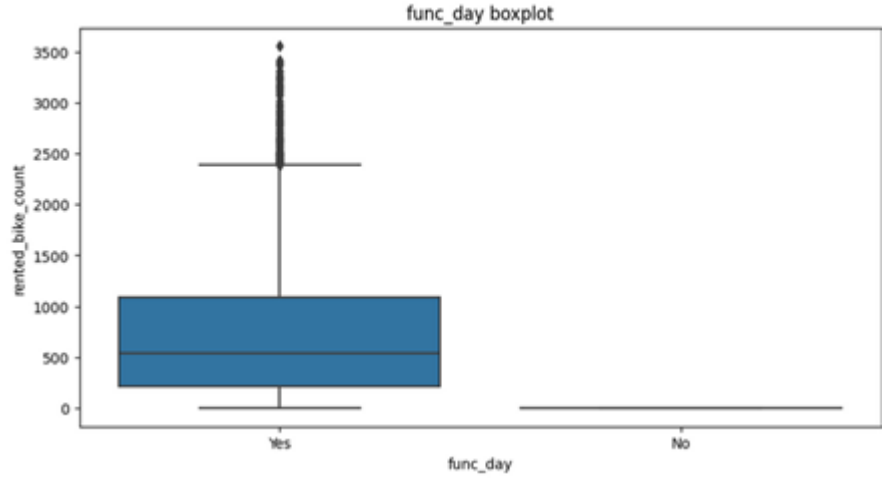


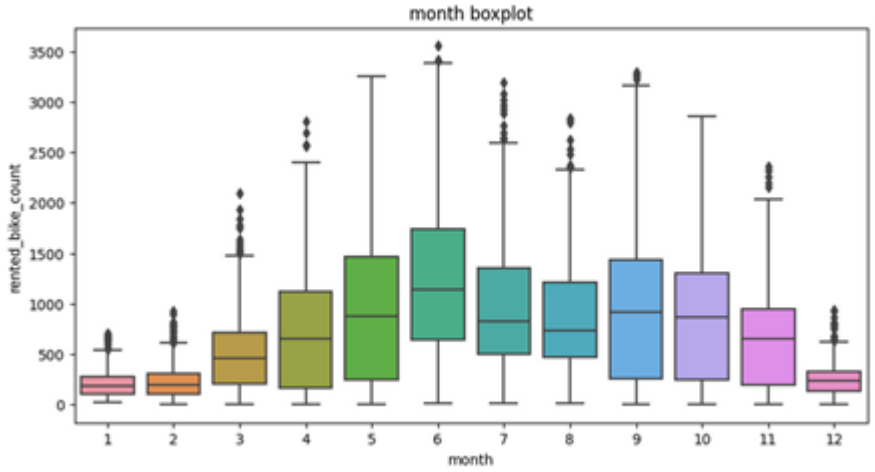


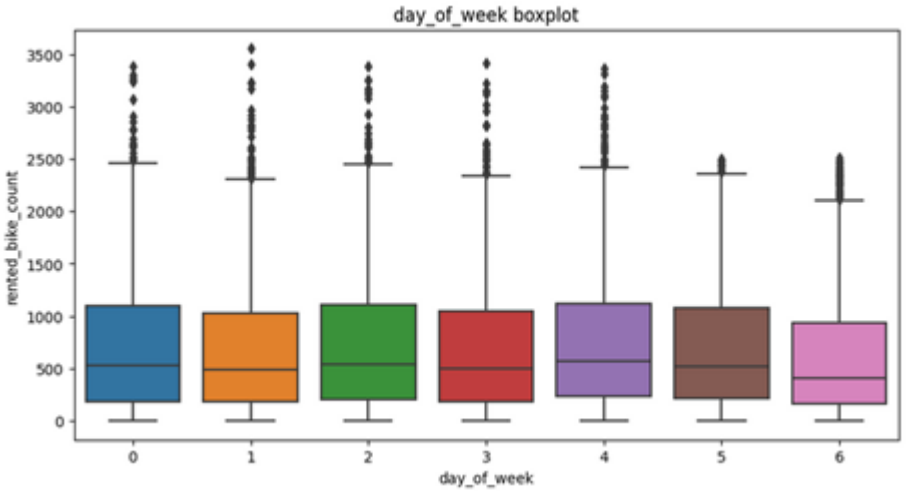
Detect outliers in Rented Bike Count column.


seasons boxplot
rented_bike_count
0
500
1000
1500
2000
2500
3000
3500
Winter
Spring
Summer
Autumn
seasons


holiday boxplot
rented_bike_count
0
500
1000
1500
2000
2500
3000
3500
No Holiday
Holiday
holiday


func_day boxplot
rented_bike_count
0
500
1000
1500
2000
2500
3000
3500
Yes
No
func_day



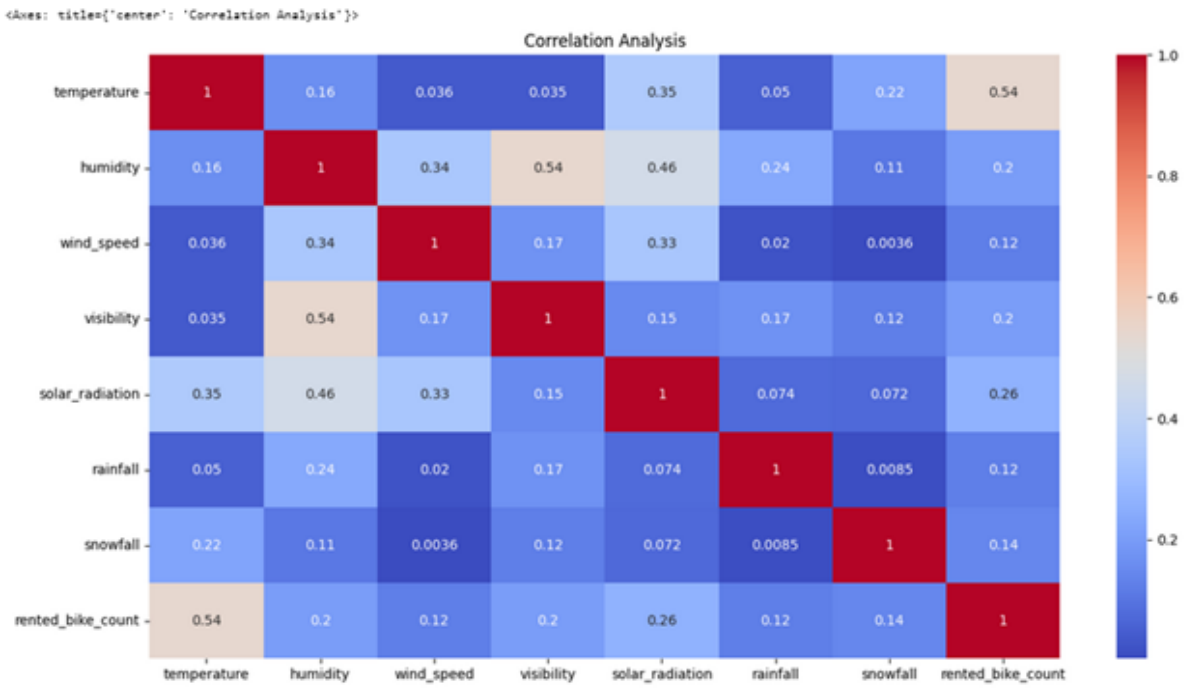


ik

.

- The above boxplot shows that we have detect outliers in rented bike count column.
- There are outliers in the data this must be taken into consideration in the model building pahse.

Correlation



Model building

Choice of model

The prediction model should be chosen in such a way that it:

- Is able to predict the dependent variable with high accuracy (Accuracy).
- Is easy to interpret (interpretability)
- Is easy to explain the model (Explainability).

Since we are working with data that:

- Contains outliers
- Contains many categorical independent variables
- Is not lineraly related to the dependent variable

· It is not advisable to use linear models to make the predictions. We can use tree models instead.

Modeling approach:

· Use a decision tree regressor model to come up with initial set of predictions (baseline model)

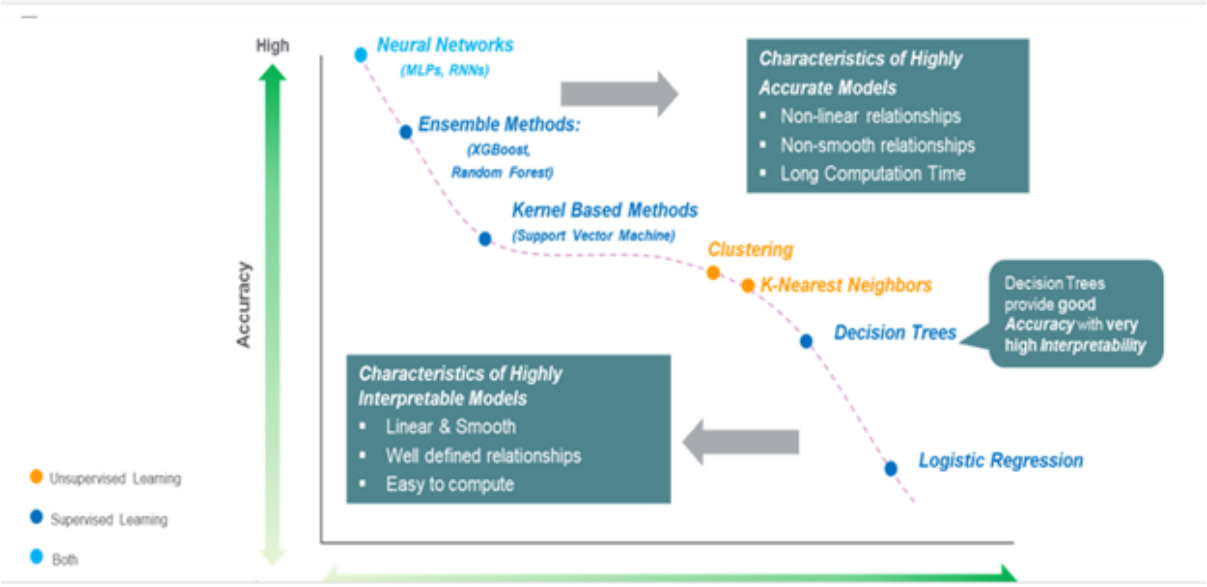
· Calculate the model accuracy using the regression evaluation metrics

· Improve the model accuracy by hyperparameter tuning

· Find feature importances from the baseline model

· Use bagging / boosting algorithms to increase the model accuracy .(PS : explainability reduce as the model complexity increases)

Choose the model with the highest accuracy for deployment



Why is accuracy in prediction of dependent variable important in businesses?

There are many reasons , some of them in this context are:

- Ensuring enough supply of bikes
- Scheduling the right amount of staff
- Making attainable financial plans
- Strategically planning marketing campaigns

These are some of the reasons why it is necessary to achieve higher

prediction accuracies with the help of complex ML algorithms.

Evaluation metrcs:

- We know that the data we are working with contains outliers ,we did not drop them because if we do so, we may loose out important trends/patterns in the data.
- Decision Trees or any tree based algorithms that we will use here are known to handle outliers. Hence we can use RMSE as the evaluation metric.
- Since RMSE we can use R2 score to make the results more explainable to a larger audience.

ML MODEL IMPLEMENTATION

Decision Tree

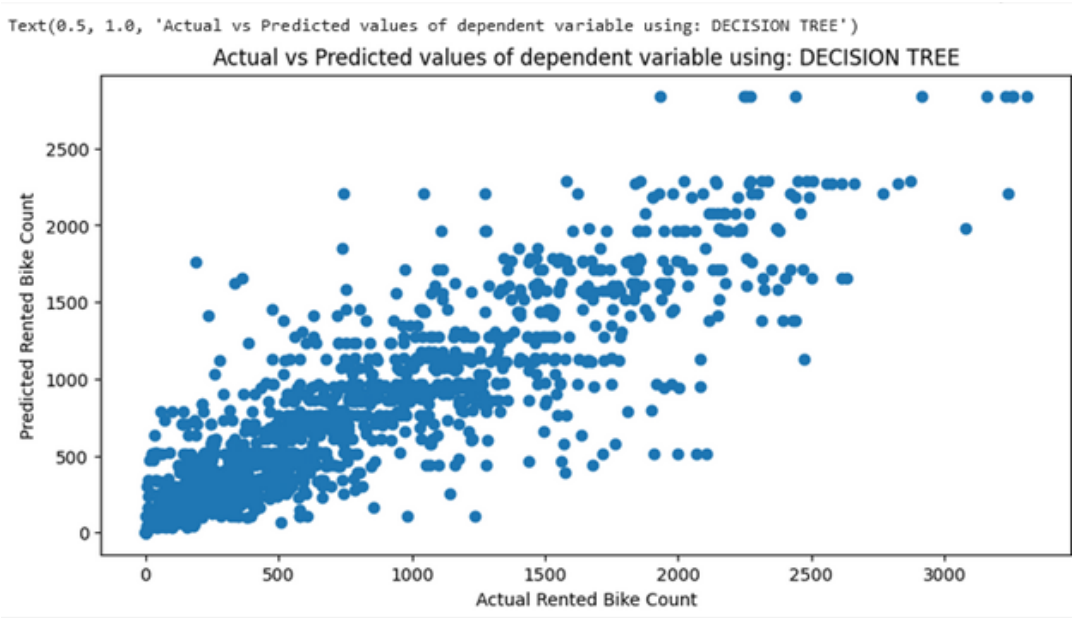
dt_train_r2_score=0.83

dt_test_r2_score=0.79

dt_train_rmse=263.273

dt-test_rmse=294.397

- Decision tree is low bias, high variance model. If we fit a decision tree model on a dataset without tuning the hyperparameters, we get zero RMSE for training data and high RMSE for test data.
- Also the R2 score is 1 for train data, and is significantly low when that model is fit on test data.
- Our aim is to build a generalized model, that is able to predict the dependent variable for unseen data with less error.
- To achieve this, we can tune the decision tree hyperparameters, thereby reducing the model complexity, which in turn improve predictions for the test data.



Scatter plot of the actual and predicted values of the dependent variable on test data using decision tree.

Random for

Train score

rf_train_r2_score = 0.8432435167367067

test score

rf_test_r2_score = 0.8136314152720869

train rmse

rf_train_rmse = 255.13140852412621

test rmse

rf_test_rmse = 279.2838385709083

Applying ML algorithms

Gradient Boosting

XG Boost:

XG Boost model explainability using Shapley values

Challenges

- Large dataset to handle.
- Needs to plot lot of graph to analyse.
- Feature engineering
- Feature selection
- Model implmentation

Conclusion

- We trained 4 unique Machine Learning models using the training dataset, and the its respective performance was improved through hyperparameter tuning.
- We initially started with the decision tree model, mainly because it is easily explainable to the stakeholders, and its low training time.
- Once we were successfully able to fit a decision tree, it was necessary to improve the prediction accuracy, and reduce errors in the predictions.
- To achieve this, we fit a random forest model on the training data, and the final predictions showed less errors compared to that of decision tree model.

- To further improve the predictions of the model, we fit 2 boosting models namely; Gradient boosting machine (GBM) and Extreme gradient boost (XG Boost). The predictions obtained from these models showed errors in the same range, but the errors were lower than that of decision tree model.

The XG Boost model has the lowest RMSE, and the highest R2 score.

Final choice of model depends on:

- If it is absolutely necessary to have a model with the best accuracy, then XG boost will be the best choice, since it has the lowest RMSE than other models built.
- But as discussed above, higher the model complexity, lower is the model explainability. Hence if the predictions must be explained to stakeholers, then XG Boost is not an ideal choice.
- In this case decision tree can be used, since they are easier to explain. By choosing a simpler model, we will be compromising with the model accuracy (Accuracy vs Interpretability tradeoff).