Capstone Project

Bike Sharing Demand Prediction

(Regression)

Presented

by

Sudhanshu Kumar

AlmaBetter Trainee, Bangalore

AlmaBetter

Contents

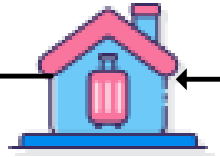
Problem statement

- ✓ For smooth functioning of rented bike system, required number of bike must be available on right time is essential.
- ✓ The main aim of this project is to build a predictive model which can predict number bike required at each hour for the stable supply of rental bikes.
- ✓ This project explores the correlation between the rental bike usage per hour and the numerous predictors, including meteorological and time information.

8760 Rows and 14 Features

Rented Bike Count

- Holiday
- Functioning day



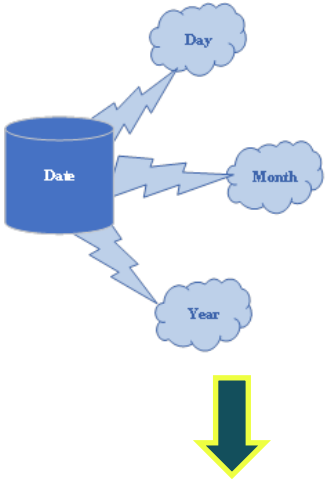
Data description

- Temperature
- Humidity
- Wind speed (m/s)
- Visibility (10m)
- Dew point temperature
- Solar Radiation (MJ/m2)
- Rainfall(mm)
- Snowfall (cm)
- Seasons

- Hour
- Date

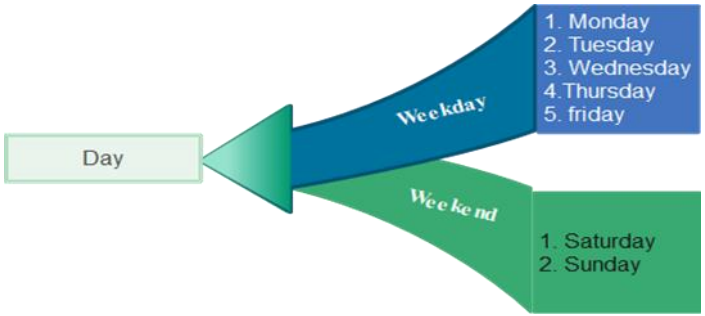
Feature Engineering

Extracting Day, Month and Year from date columns




```
bike_df['Year']=bike_df['Date'].dt.year
bike_df['Month']=bike_df['Date'].dt.month_name()
bike_df['Day']=bike_df['Date'].dt.day_name()
```

Type of Day



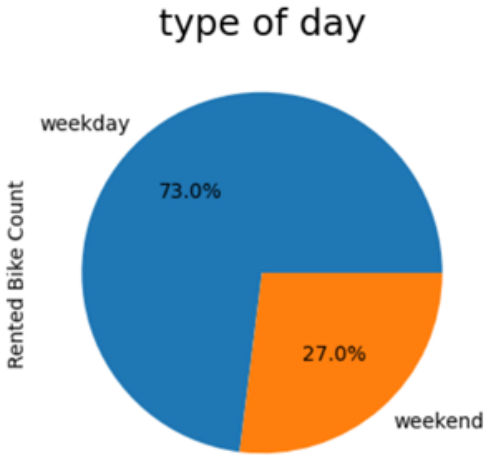

```
bike_df['type of day']=bike_df['Day'].apply(lambda x:
                    "weekend" if x=='Saturday' or
                    x=='Sunday' else "weekday")
```

Exploratory data Analysis

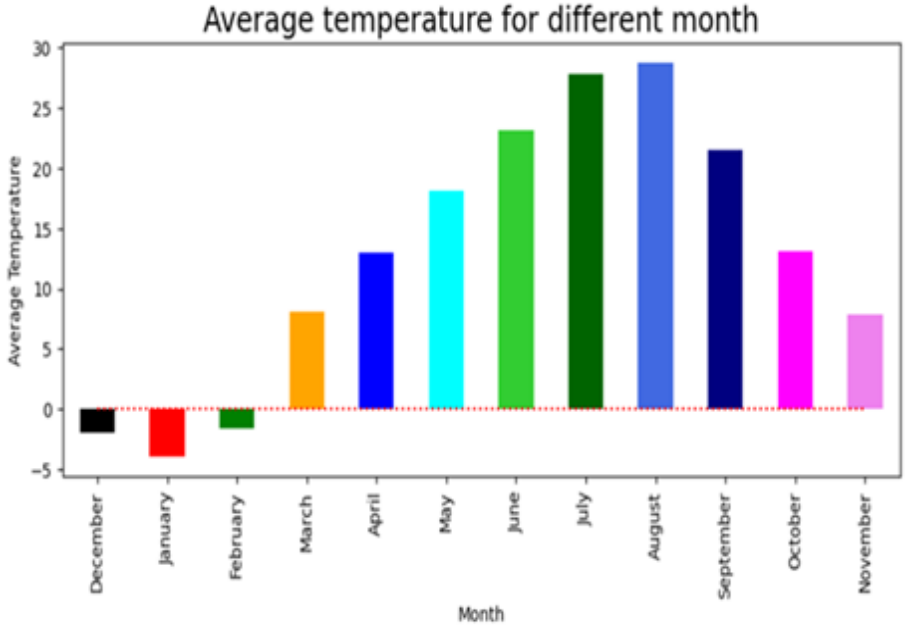
Distribution of rented bike on weekday and weekend

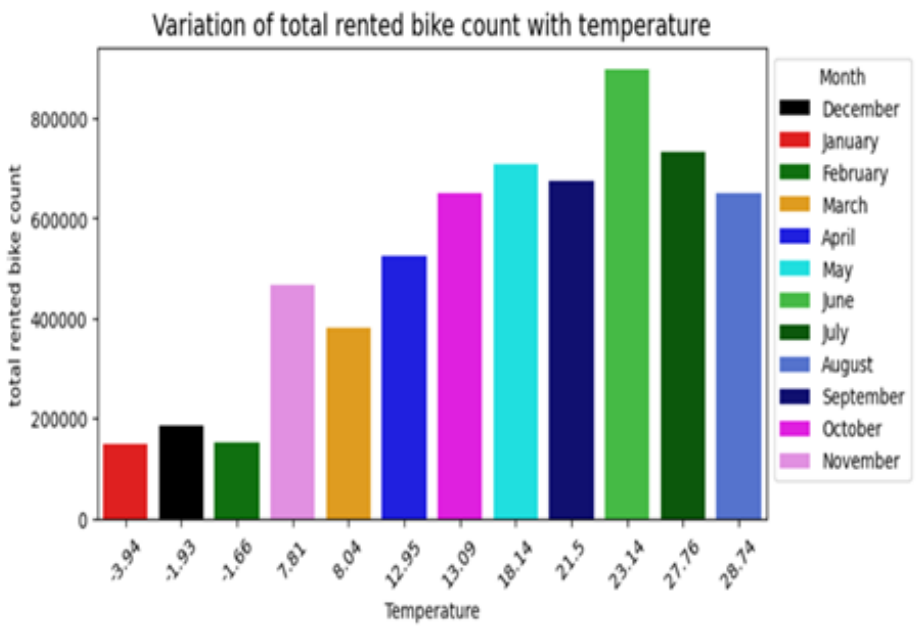
- ✓ 73% of bike booked on weekday and remaining 27% are book on weekend
- ✓ the possible reason for this is people have compulsion to go their office, hence they book bike so that they reach office on time.
- ✓ On weekend people booked bike only for their pending work.

	type of day	Rented Bike Count
0	weekday	4506628
1	weekend	1665686



Exploratory data Analysis

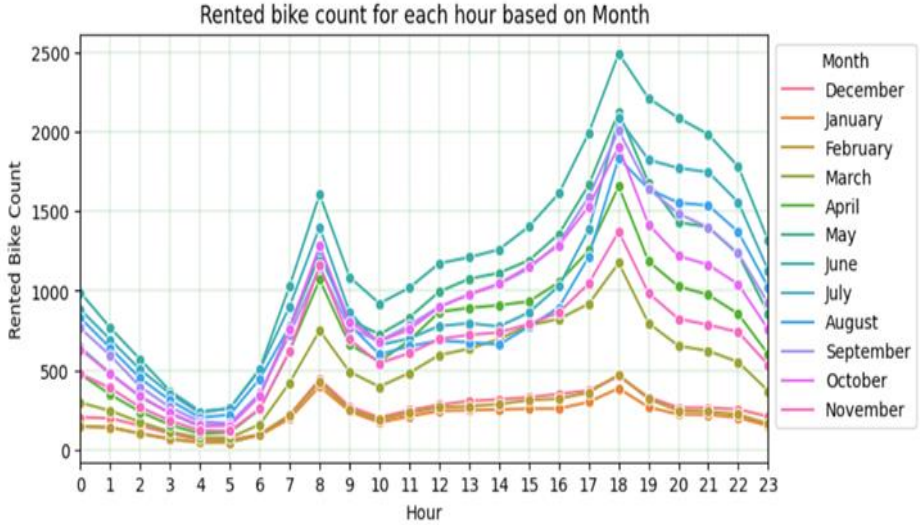


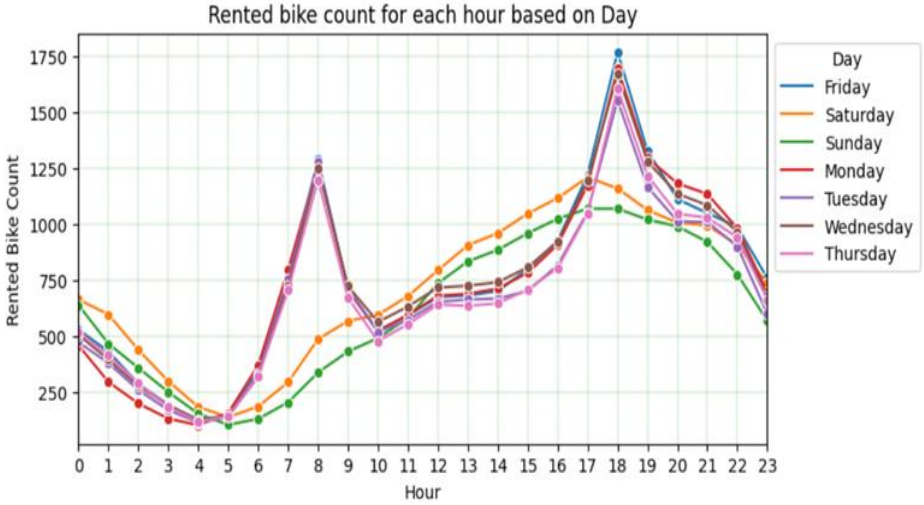


- ✓ December, January and February having negative mean temperature, this may hamper the rented bike count
- ✓ It is observe that, month which have negative mean temperature having less number of booking.

☛ hence, it can conclude that due to lower temperature people prefer to travel with other medium of transportation instead of using bike.

Exploratory data Analysis

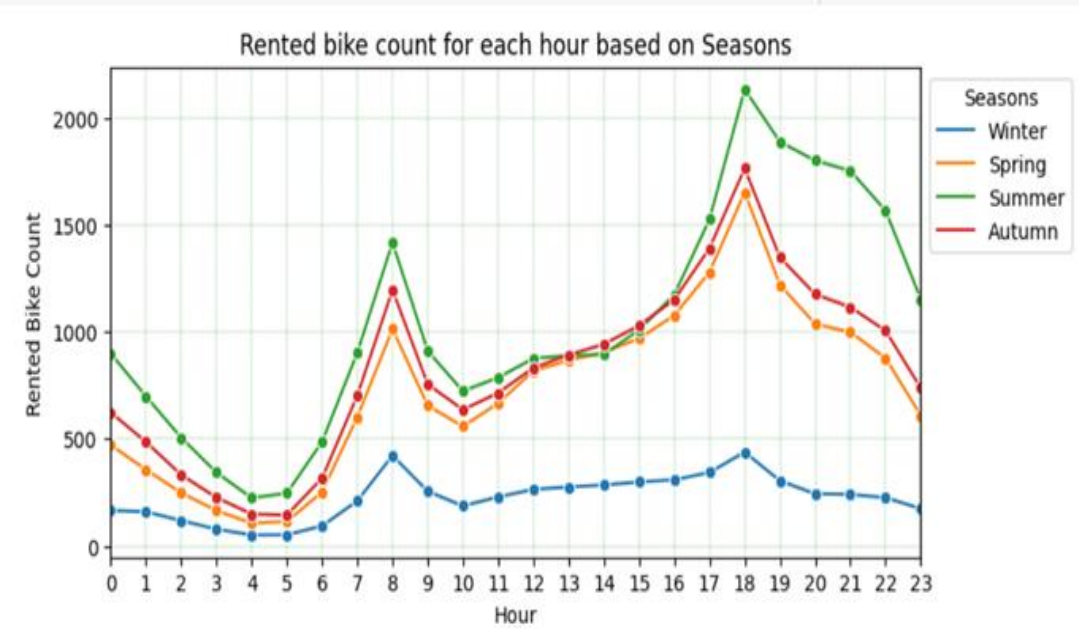


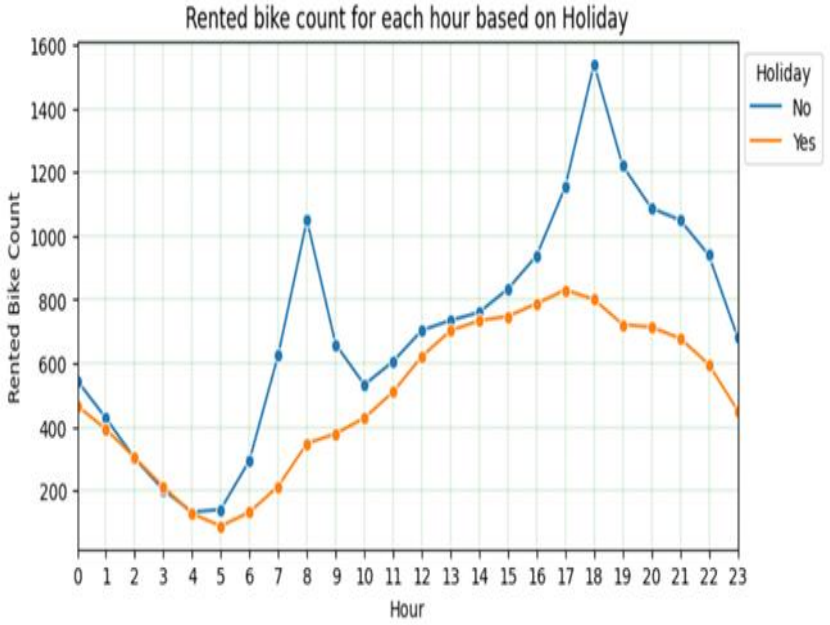


- ✓ December, January and February have a smaller number of bike bookings as compare to other month, possible reason for this temperature is very less in these months.
- ✓ Saturday and Sunday (weekend) do not show peak on 08:00 hours and 18:00 hours, possible reason for this is holiday.

- On 08: 00 hours and 18:00 hours maximum number of bikes are booked.
- It may be office start timing at 08:00 hours and leaving time at 18:00 hours.
- 05:00 shows least number of bookings

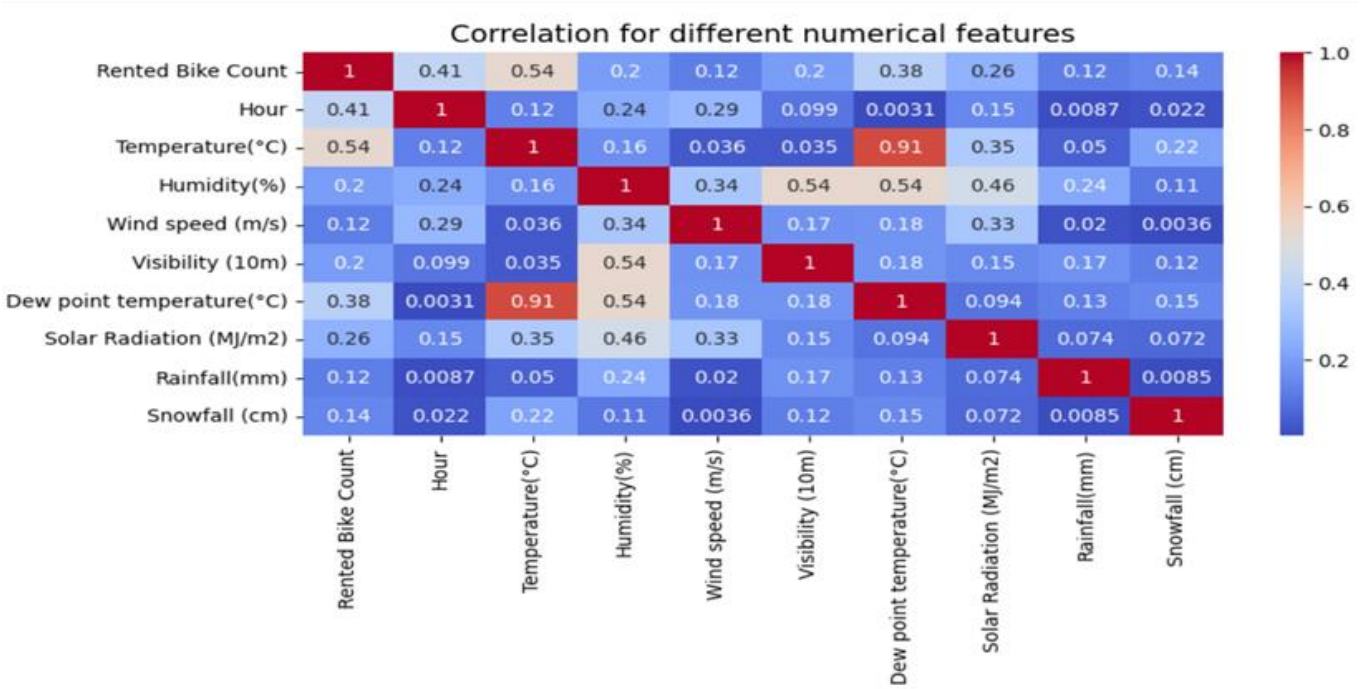
Exploratory data Analysis





- ✓ Rented bike count follow a similar pattern for each of the four seasons.
- ✓ Winter season have less number of booking while summer have maximum.
- ✓ Booking count on working day is more than holiday.
- ✓ There is no peak time on holiday as compare to working day.

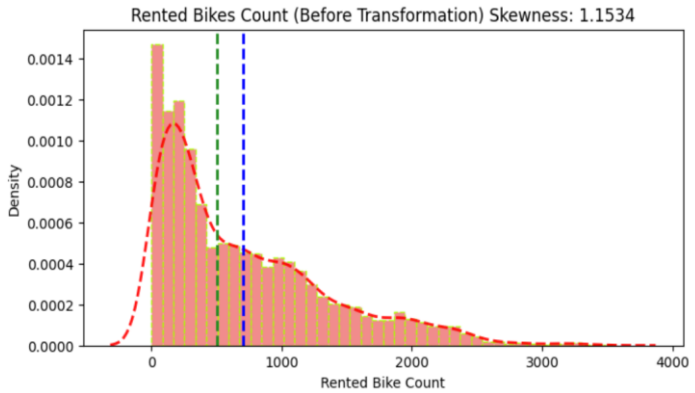
Exploratory data Analysis

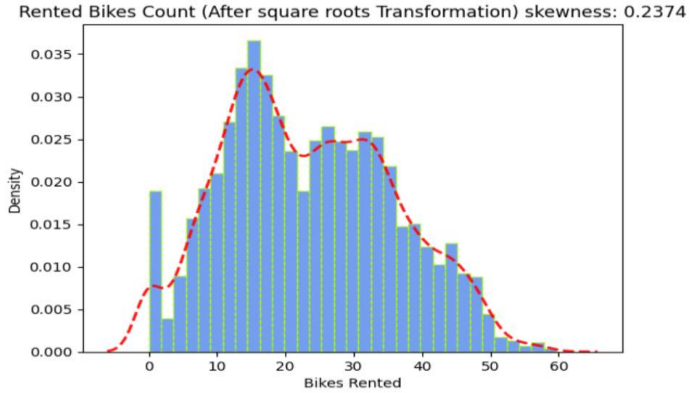


- ✓ Temperature shows high collinearity with dew point temperature.
- ✓ Humidity and visibility are moderately correlated.
- ✓ temperature and hour show positive correlation with rented bike count
- ✓ As per domain knowledge, there dew point temperature column is irrelevant to find the rented bike count

Model building

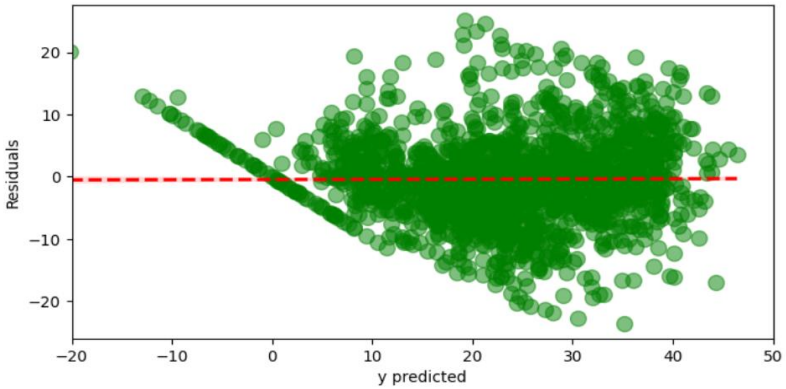
1. Checking for skewness in target variable.

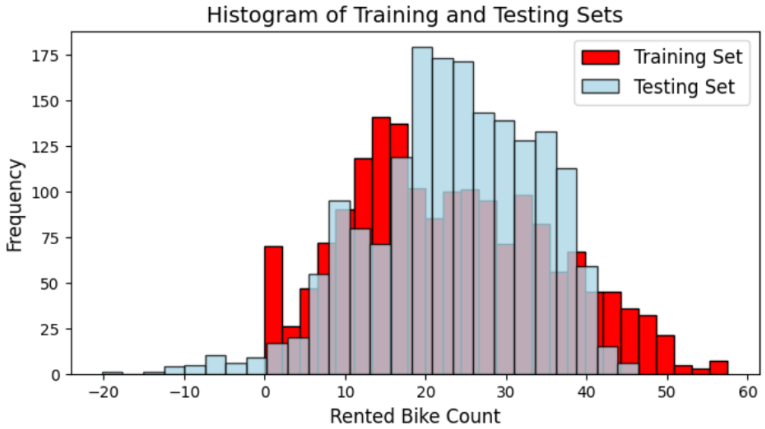




2. Using VIF we remove the multicollinearity between the independent columns.
3. Using one hot encoding we convert the categorical feature into numerical form.
4. Split the data into train data and test data, we considered 20% of whole data as test data and remaining 80% as train data.
5. Scale the data using MinMaxScaler

Linear Regression

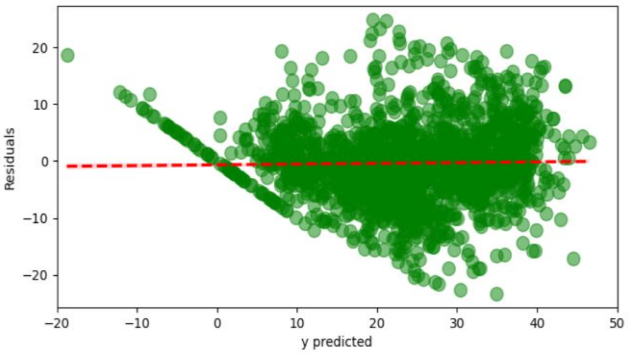


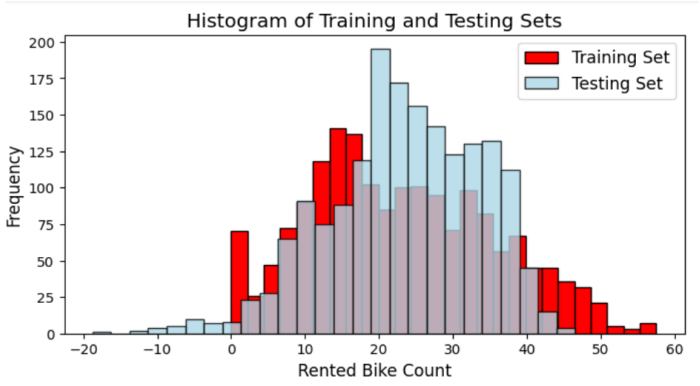


data	Model	R2_score
Train data	Linear Regression	0.683435
Test data	Linear Regression	0.686053

- ✓ Lower R2_score on train data indicate that model is not able to understand the data.
- ✓ From the above two plot it is observe that is not train with the data properly.
- ✓ To avoid this problem, we have to go with some complex model.

Ridge regression (α=0.1)

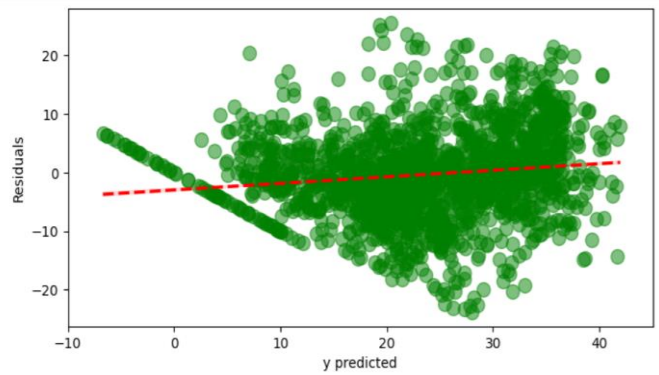


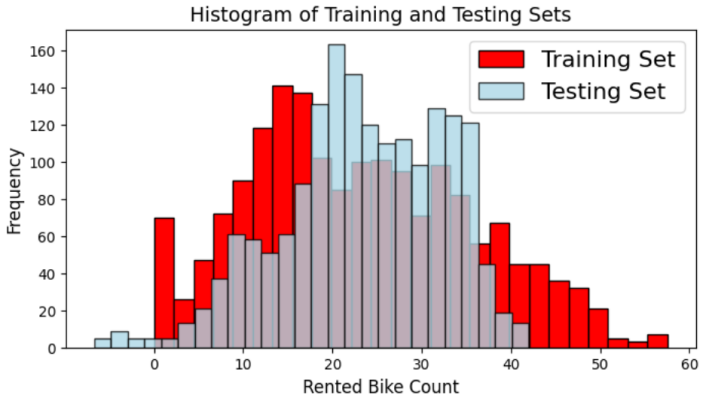


data	Model	R2_score
Train data	Ridge Regression	0.683809
Test data	Ridge Regression	0.686100

✓ Model performance is not improved, it is almost same as linear regression.

Lasso regression (α=0.1)

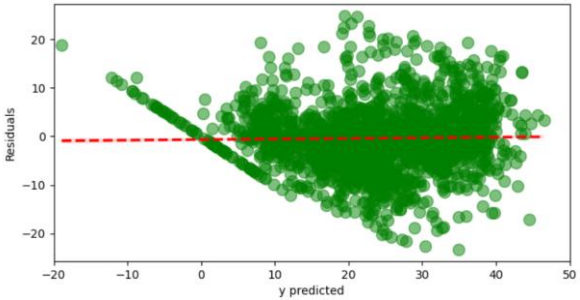


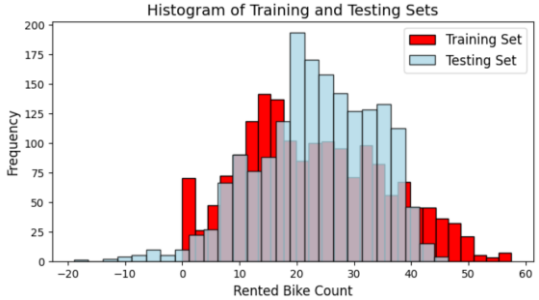


data	Model	R2_score
Train data	Lasso Regression	0.642200
Test data	Lasso Regression	0.636000

✓ Model performance reduces.

Ridge with GridSearchCV



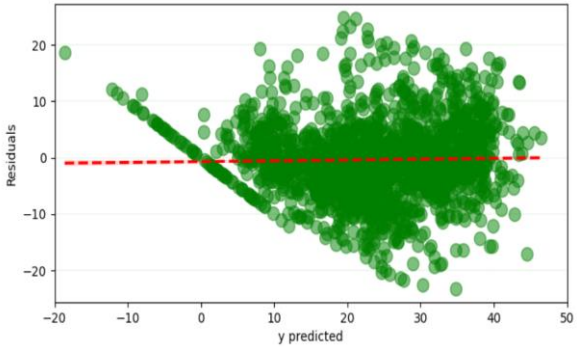


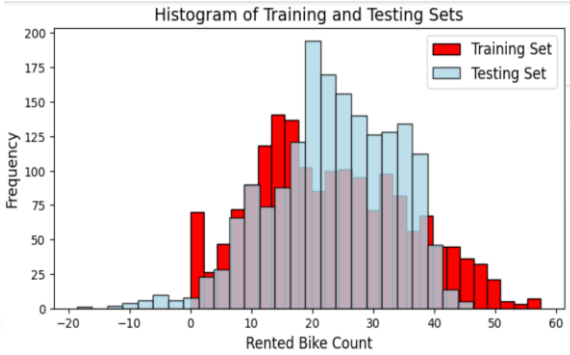
Best parameter
'alpha': 0.01

data	Model	R2_score
Train data	Ridge with GridSearchCV	0.683813
Test data	Ridge with GridSearchCV	0.686204

✓ There is no improvement in performance of model

Lasso with GridSearchCV



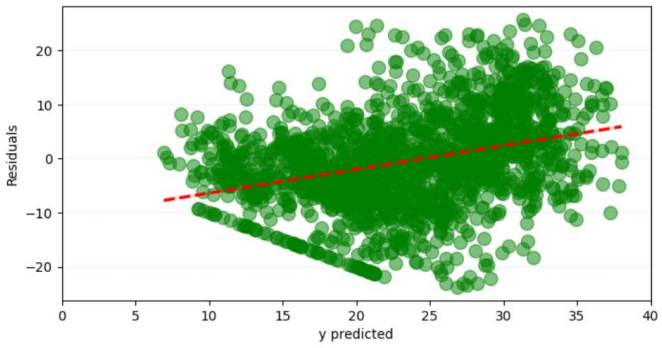


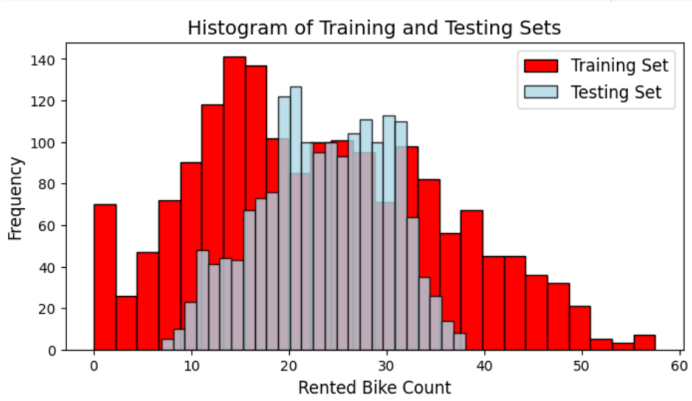
Best parameter
'alpha': 0.0014

data	Model	R2_score
Train data	Lasso with GridSearchCV	0.683790
Test data	Lasso with GridSearchCV	0.685954

✓ There is no improvement in performance of model

Elastic net

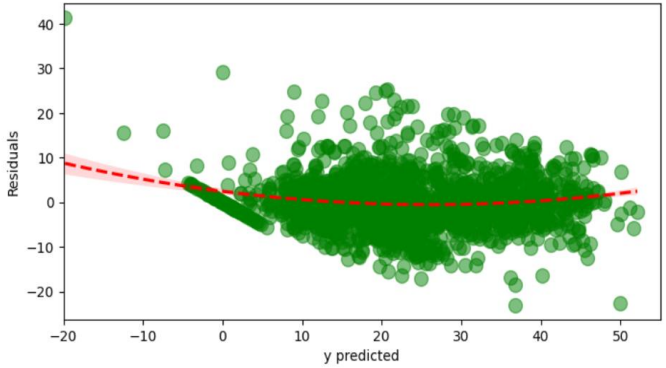


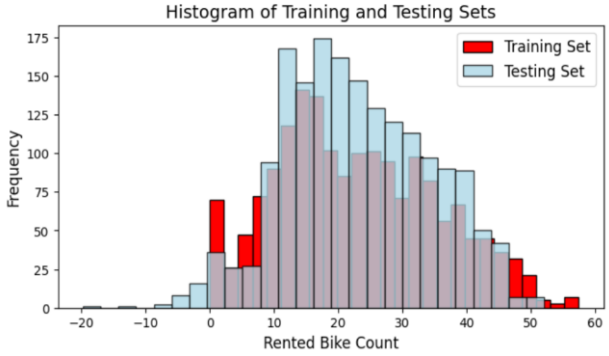


data	Model	R2 score
Train data	ElasticNet	0.529043
Test data	ElasticNet	0.507558

- ✓ Very low R2_score indicate that model is not train well with the given data.

Polynomial Regression





data	Model	R2 score
Train data	Polynomial Regression	0.807985
Test data	Polynomial Regression	0.774544

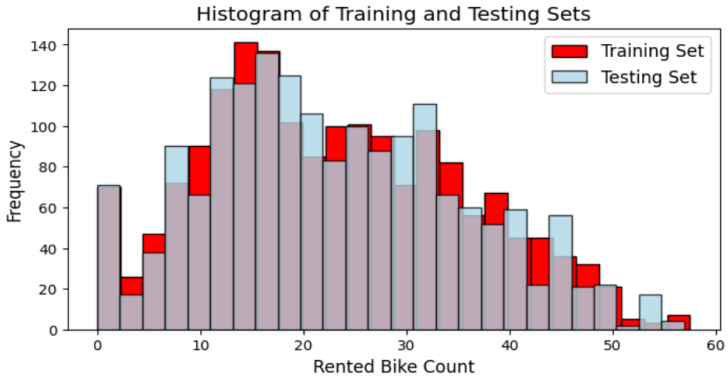
- ✓ Model R2_score is improved to 0.8 for train data, and 0.77 for test data.

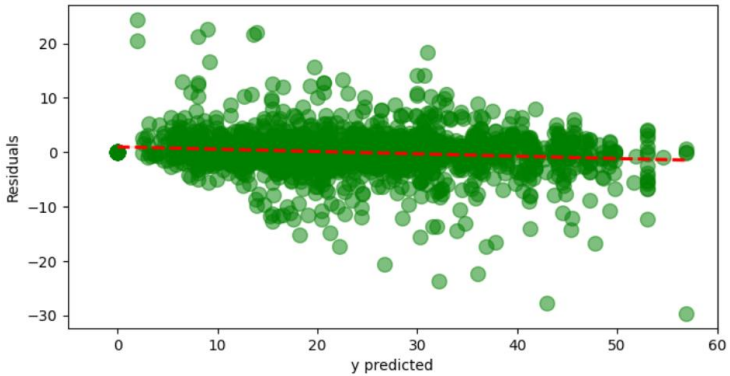
Decision tree with GridSearchCV

Best parameter:
'max_depth'= 15

data	Model	R2_score
Train data	DecisionTreeRegressor with GridSearchCV	0.990103
Test data	DecisionTreeRegressor with GridSearchCV	0.894431

- ✓ Model R2_score is improved to 0.99 for train data and 0.89 for test data, difference between R2_score of test and train indicate that there is small percentage of overfitting.





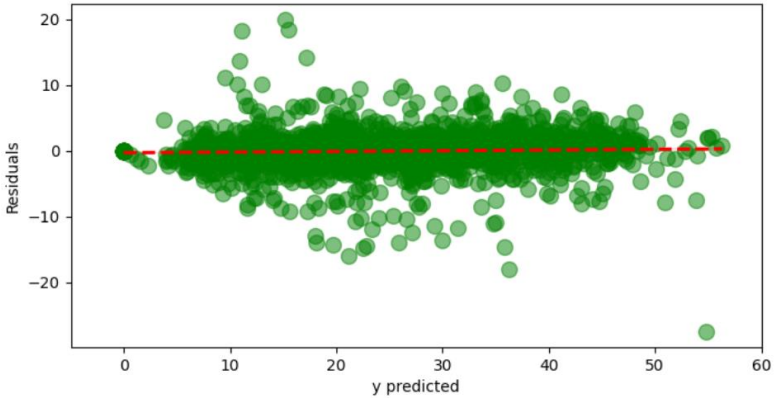
Random forest GridSearchCV

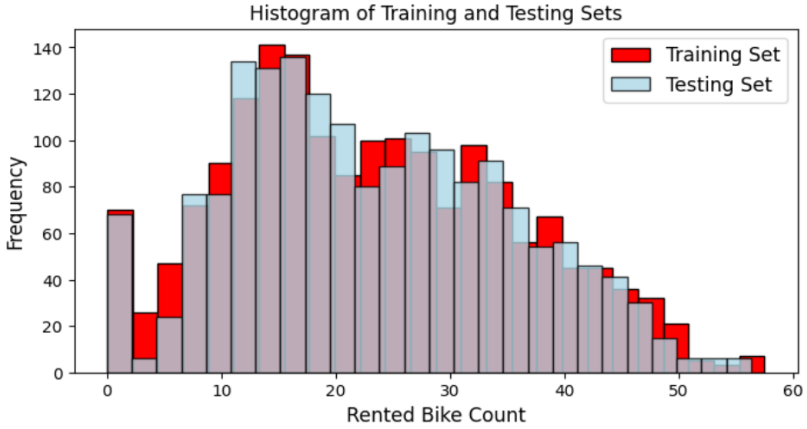
Best parameters:

'bootstrap': True,
'max_depth': 40,
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 36

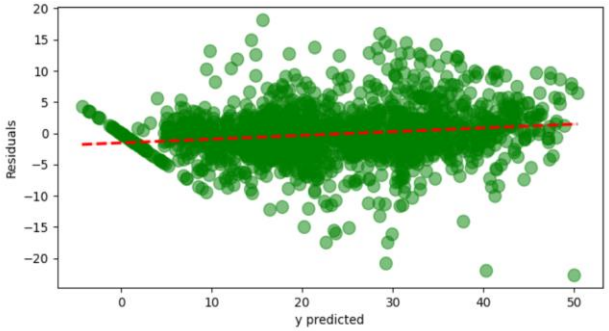
data	Model	R2_score
Train data	RandomForest with GridSearchCV	0.991365
Test data	RandomForest with GridSearchCV	0.939602

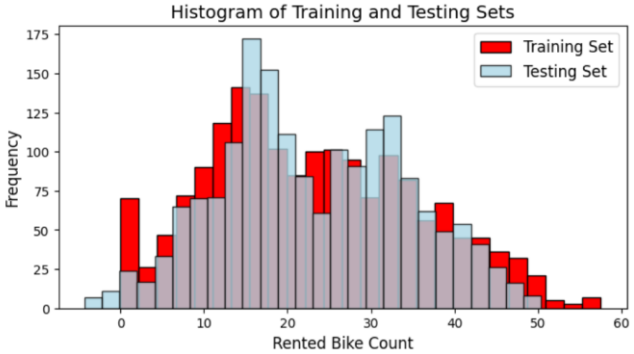
✓ Here model overfitting reduces





Gradient boost

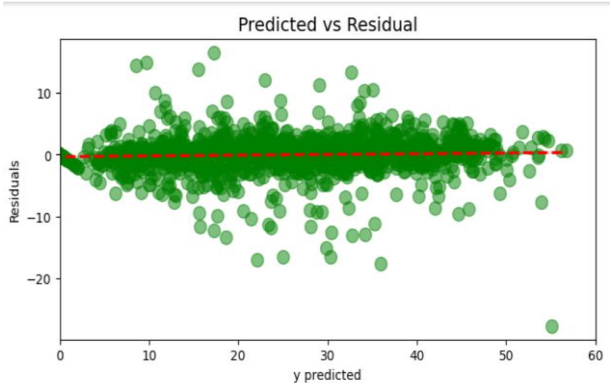


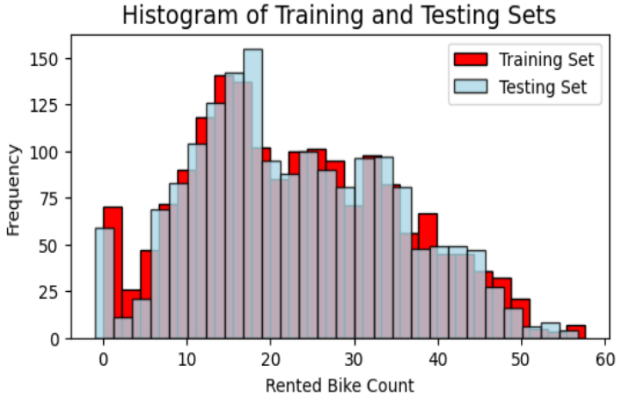


data	Model	R2_score
Train data	Gradient Boosting	0.901926
Test data	Gradient Boosting	0.890287

✓ Here R2_score reduces but model generalized very well.

Gradient Boosting with GridSearchCV





Best parameters:

'learning_rate'=0.06,
'max_depth'= 8,
'n_estimators'= 150,
'subsample'= 0.9

data	Model	R2_score
Train data	Gradient Boosting with GridSearchCV	0.991249
Test data	Gradient Boosting with GridSearchCV	0.950793

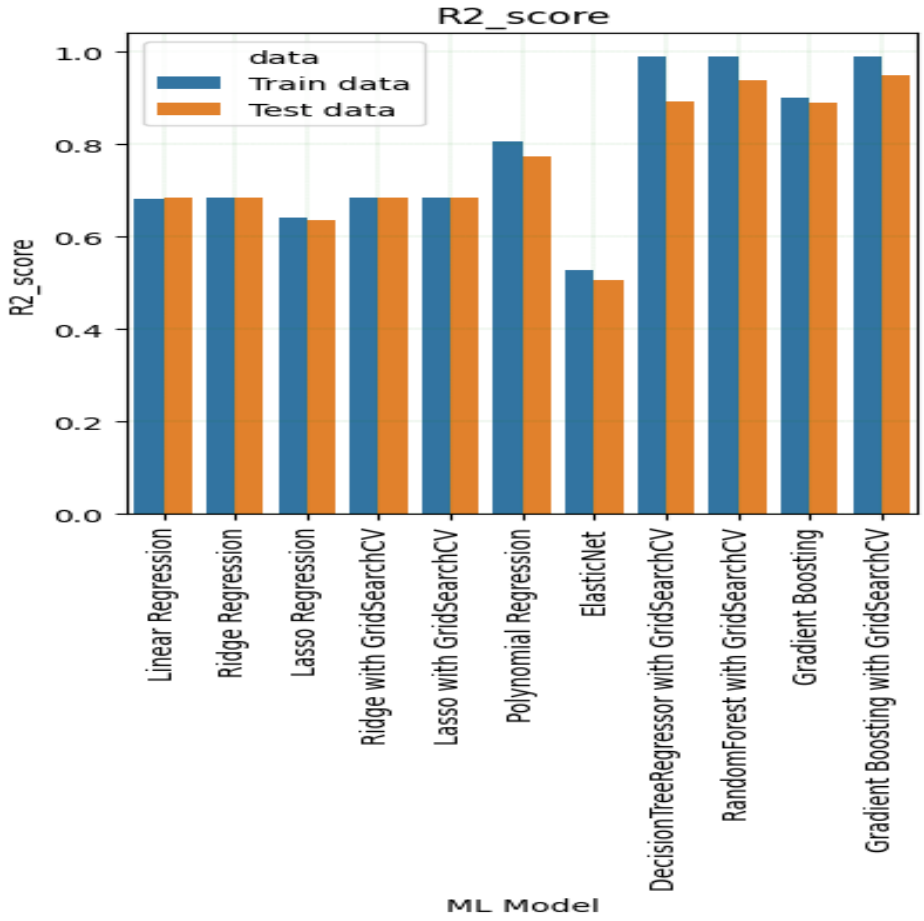
✓ R2_score is improved, and we can deploy this model.

Challenges

- ✓ Choosing the appropriate plot along with their plotting parameter for best representation of data into pictorial form.
- ✓ Removing multicollinearity between the columns.
- ✓ Choosing the appropriate ML model for predictions.
- ✓ To get best parameter from Random forest and gradient boost with hyperparameter tuning (GridSearchCV) takes more computational time.

Conclusions

- ✓ Gradient boosting without hyper parameter tuning have almost same train and test R2 score, it means that model is generalized very well with the given data.
- ✓ Gradient boosting with GridSearchCV having train data R2_score is 0.99 and test data R2_score is 0.95.
- ✓ Difference between R2_score of test data and train data is so small that it can be neglected and we can deploy this model.



Thank You