CAPSTONE PROJECT
SUPERVISED ML REGRESSION

BIKE SHARING DEMAND PREDICTION

by : Sachin Yallapurkar

Contents

- Introduction
- Problem Statement
- Points for discussion
- Data Summary
- Exploratory Data Analysis
- Modeling Overview
- Feature Importance
- Conclusion

Introduction

A bike rental or bike hire business rents out motorcycles for short periods of time, Usually for a few hours. Most rentals are provided by bike shops as a sideline to their main businesses of sales and service, but some shops specialize in rentals.

As with car rental, bicycle rental shops primarily serve people who do not have access to vehicles, typically travelers and particularly tourists.

Bike rental shops rent by the day or week as well as by the hour, and these provide an excellent opportunity for those who would like to avoid shipping their own bikes but would like to do a multi-day bike tour of a particular area.

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 the bike count required at each hour for the stable supply of rental bikes.

Discussion Topics

- Bike booking in each season, on functioning days, holidays, and months.
- Comparing Rented Bike Count against Numerical data columns.
- Checking for Linear relation between the Rented bike count and the Numerical data columns.
- Climate Effect in Different seasons on Bike Sharing.
- Heat Map(OR) Correlation Map.
- Linear Regression analysis, Lasso Regression Analysis, Grid Search CV for Hyperparameter tuning,
- Decision Tree Analysis, XG boost, and Random Forest Analysis
- Feature Importance.

Data Analysis Steps

Imported Libraries

In this part, we imported the required libraries NumPy, Pandas, matplotlib, and seaborn, to perform Exploratory Data Analysis and for prediction, we imported the Scikit learn library.

Descriptive Statistics

In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() this told mean, median, standard deviation

Missing Value Imputation

We will now check for missing values in our dataset. after checking not existed any missing values, In case there are any missing entries, we will impute them with appropriate values.

Graphical Representation

We will start with Univariate Analysis, bivariate Analysis and conclude with various prediction models driving the Demand for bikes.

Attributes of each variable

Date: Date in year-month-day format

Rented Bike Count: Count of bikes rented at each hour

Hour: Hour of the Day

Temperature: Temperature in Celsius

Humidity: Humidity in %

Windspeed: Speed of wind in m/s

Visibility (10m): Visibility

Dew point temperature: Dew Point Temp (Celsius)

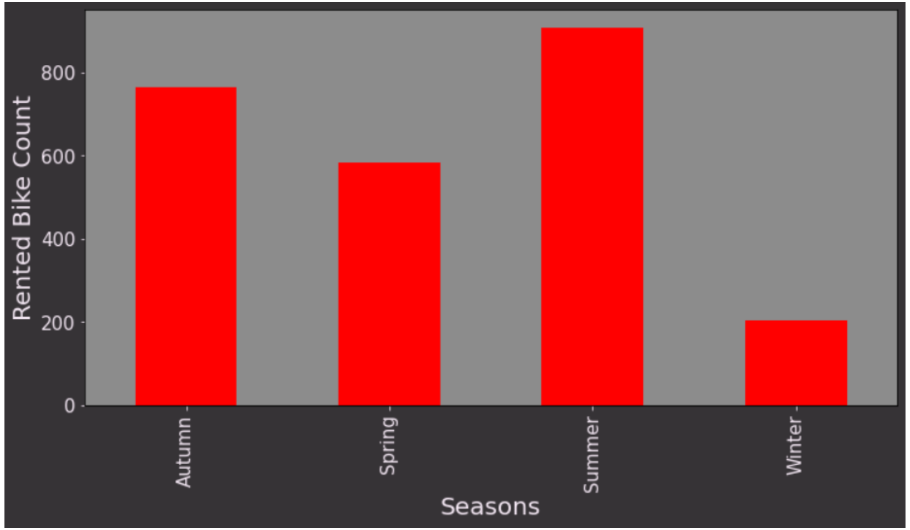
Bikes Rented per Season

- Highest number of bikes were rented in **Summer.** The total count of bikes rented in summer was 2.28 million

- Second highest Bikes were rented in **Autumn** around **1.79 million** followed by **Spring** in which **1.6 million** bikes are rented.

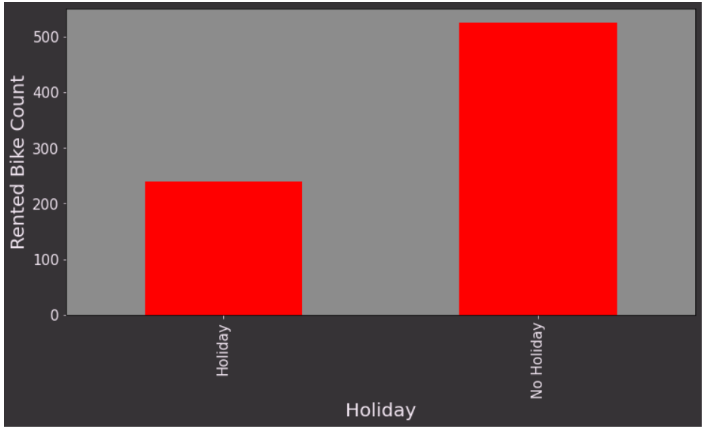
- **Winter** appears to be the least popular season for bike rentals. In the winter, just 487K bikes were rented.

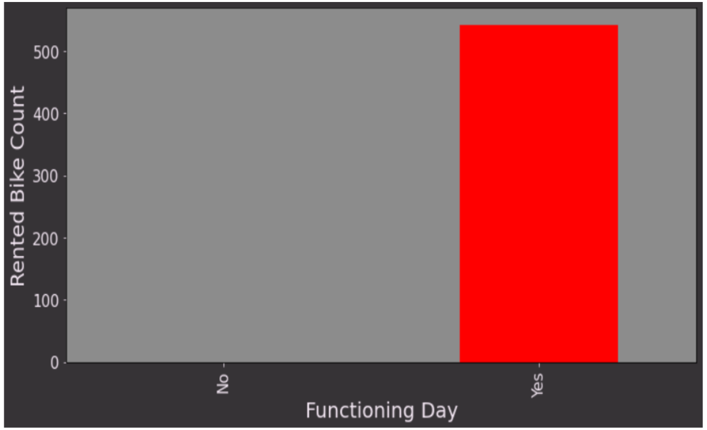
- The **extreme temperatures** in Seoul in the **winter** might be a factor in the **low demand** for bikes in the winter.



Bike Renting trend on holidays, Functioning days

- **People prefer** to use the bike on **Non-holiday more** compared to **Holidays**.
- **5.9 million** bikes are rented on **Non-holidays**, only a meager **215K** bikes were rented on **holidays**.
- It's reasonable to conclude that the **majority of clients** in the **bike rental sector** are from **Seoul's working class**.
- All the bikes rented were on the functioning days.



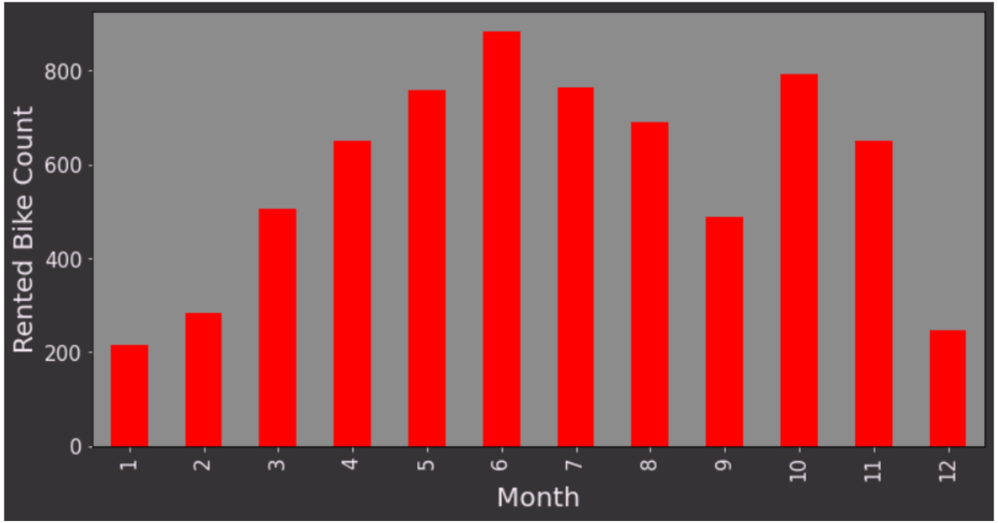


Bike Booking Monthly Trend

- **June** is the most preferred month for bike booking around **896K** bikes were rented in June.

- **July** and **May** are the second and third best.**734K** bikes were booked in **July**, and **707K** were booked in **May**.

- Demand for bikes was **least** in **Jan**, followed by **Feb** and **Dec**. **150K** bikes were rented in **Jan**, **151k** in **Feb**, and **185K** in **Dec.**

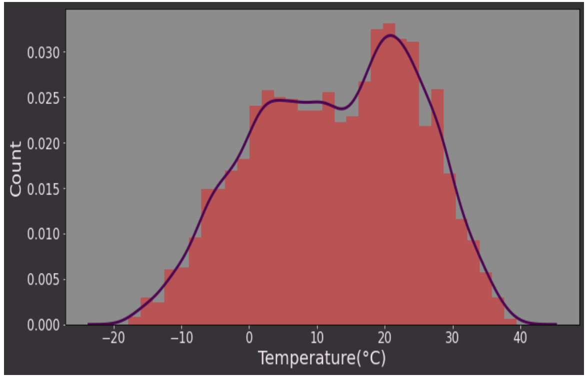


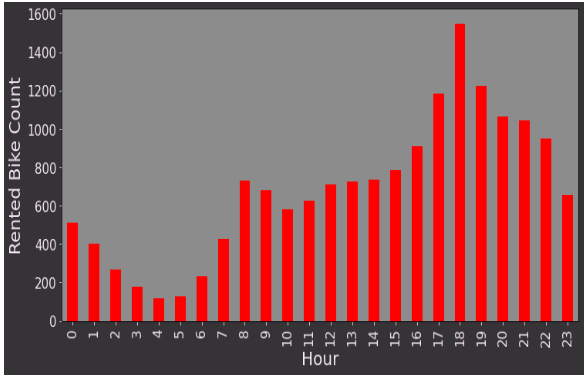
Rented Bike Count Against Numerical Data

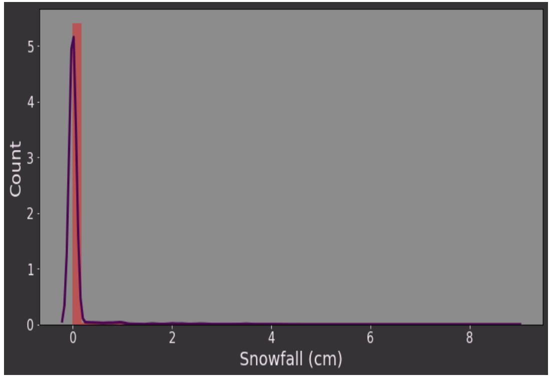
- **Most preferred** bike-sharing **temperature** is **20- 30** degrees Celsius. Bike renting is **minimal** when the **temperature** is **>35 or <5** degrees Celsius.

- Bike sharing is at its **peak between 4 pm to 8 pm**. Bike-sharing is at **least between 2 am to 6 am**, it **increases from 6 am onwards until 8 am**.

- **Snowfall** is **least favorable** for the bike renting Business.





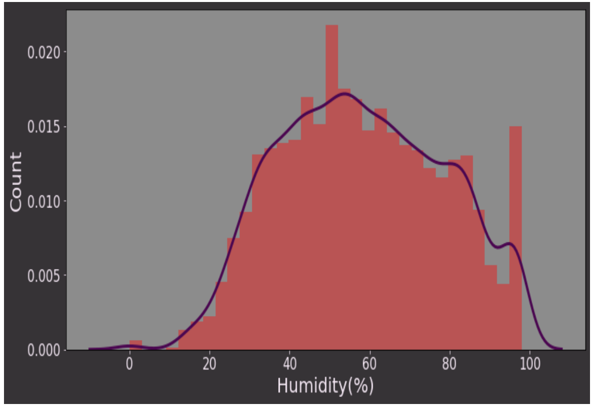


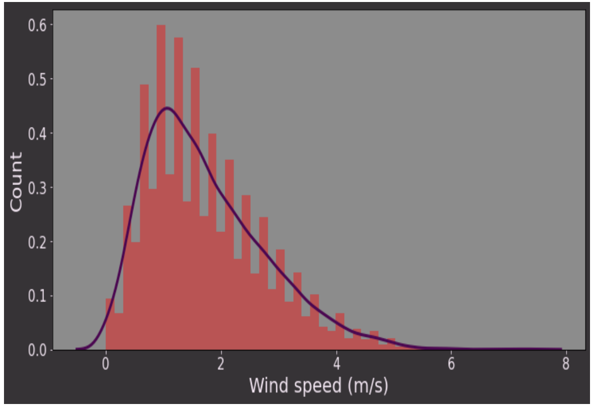
Rented Bike Count Against Numerical Data

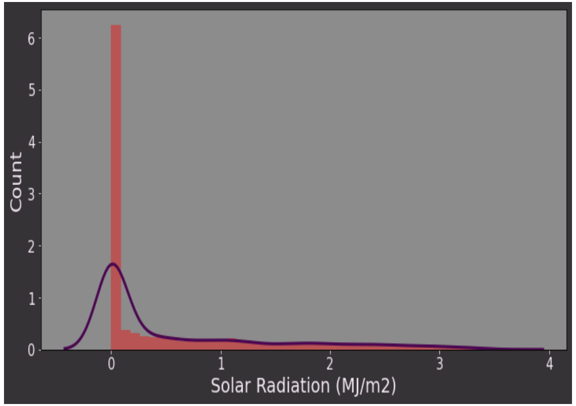
- Bike renting is at its **peak** when the **humidity is 40%- 60%**. People avoid bikes when the climate is too humid or too dry.

- Favorable wind speed for Bike sharing is 1m/s -2 m/s as wind speed goes beyond 2m/s the count of bike-sharing starts dropping reaching minimal when the **speed > 5m/s.**

- Bike sharing is at its **peak** when the **radiation is minimal**.

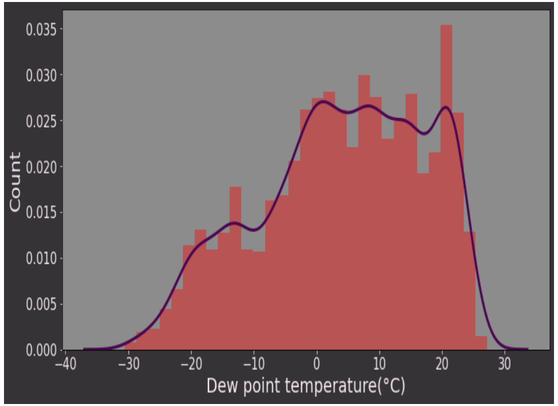


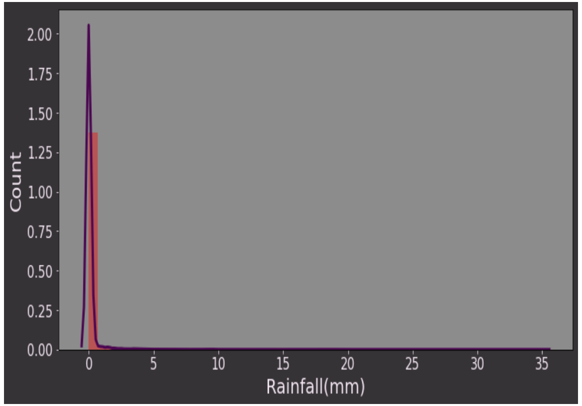


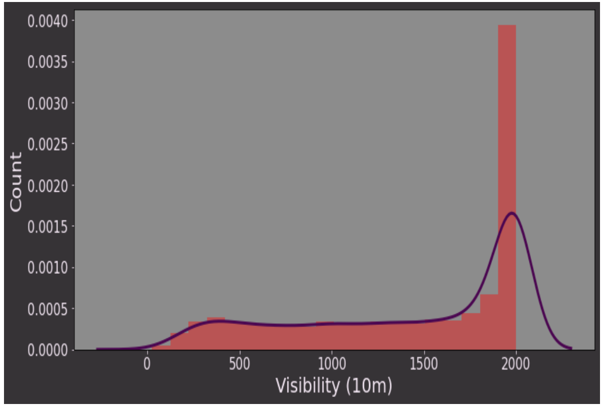


Rented Bike Count Against Numerical Data

- Dew point temperature between **5-25 Degrees** is **most favorable** for Bike sharing.
- Demand for bikes **dwindles** in case of **rainfall**.
- **Visibility** is an important factor for bike riders, bike sharing is at its **peak** when the **visibility is maximum**

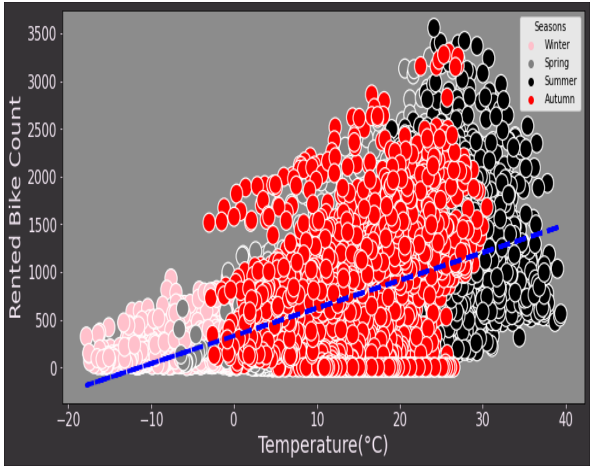


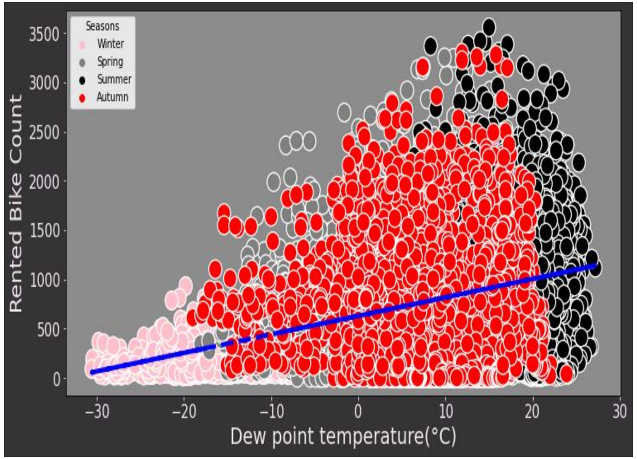


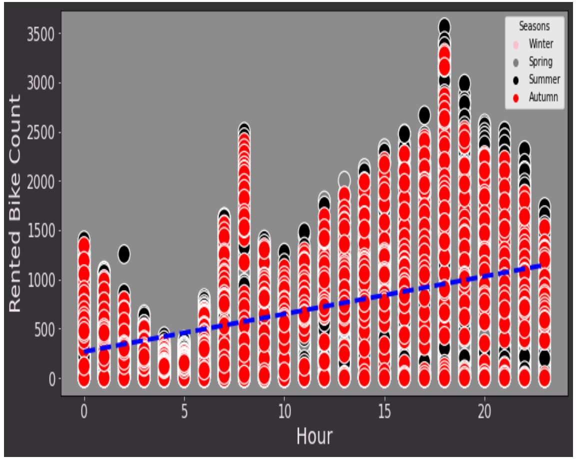


Co-relation: Rented bike count vs Temp, Dew point Temp, Hour

- Bike sharing is positively co-related to temperature and Dew point Temperature as the temperature approaches **30 degrees**.
- Though one thing to notice the positive co-relation is applicable only because the temperature in Seoul rarely crosses **40 Degrees**.
- Bike sharing count is positively co-related to hours as the Hours Progress from 0 (12 am) to 20 (8 pm) the bike-sharing count increases.



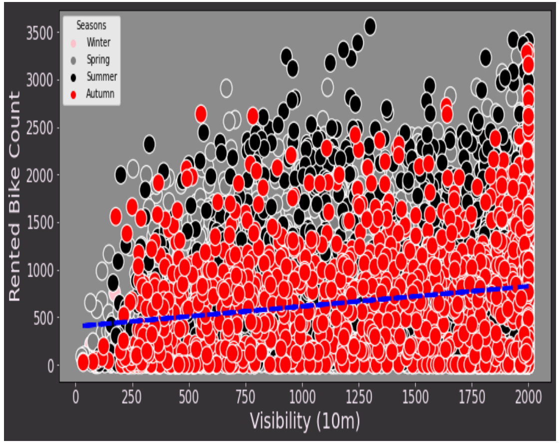


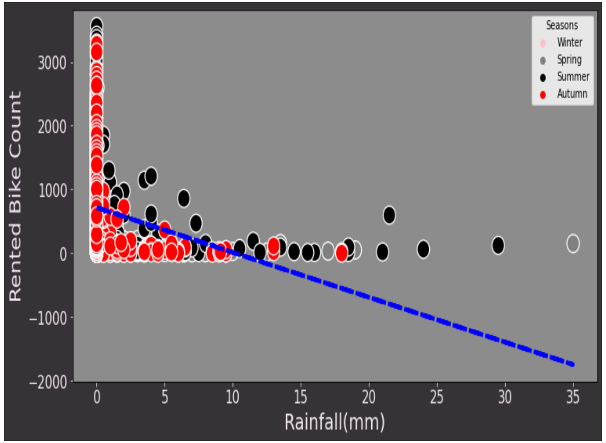


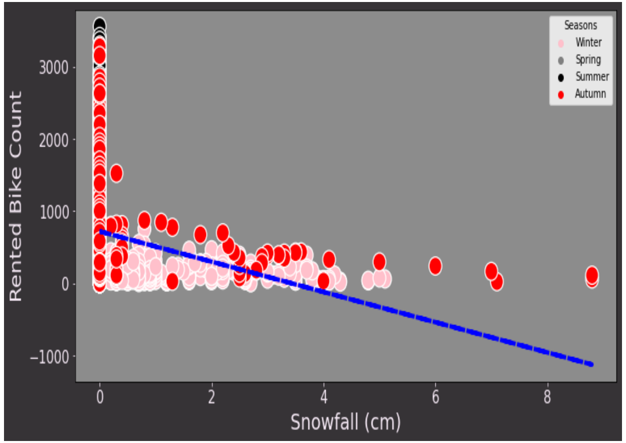
Co-relation: Rented bike count vs Visibility, Rainfall, Snowfall

- Visibility is Also slightly positively co-related with Bike Bookings.

- Snowfall, Rainfall are negatively co-related to Bike rented count.

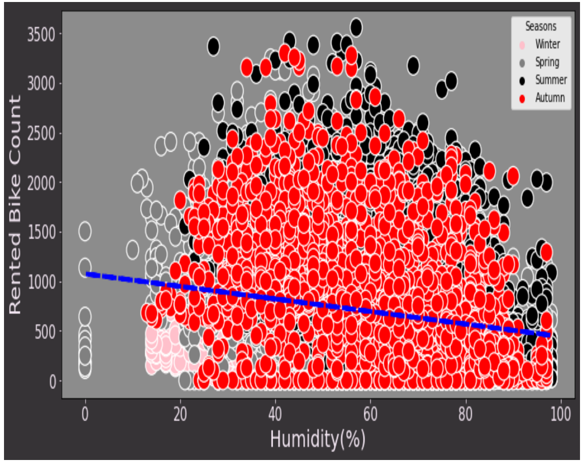


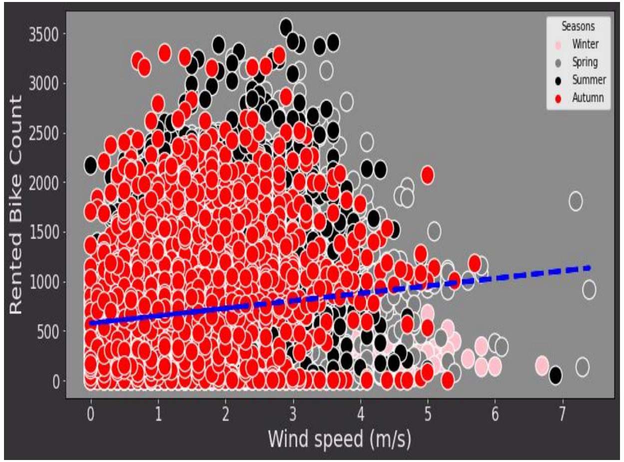


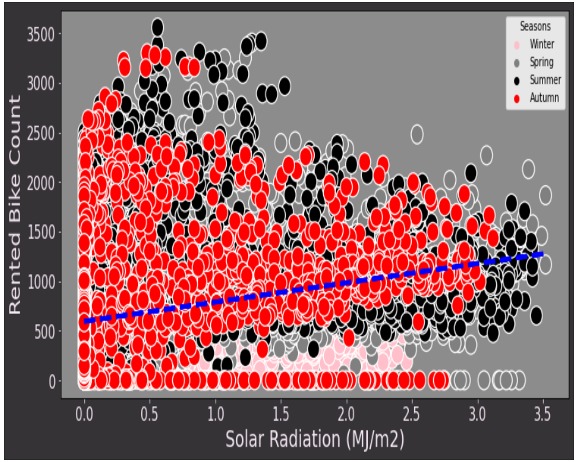


Co-relation: Rented bike count vs Humidity, Wind Speed, Radiation

- The bike-sharing count is slightly negatively correlated to Humidity.
- Wind speed and Solar radiation are slightly positively related to Bike-sharing count.







Correlation map

- Heat map shows slightly positive relation of Rented bike count with **Hour, Temperature, Dew point Temperature, Solar Radiation.**

- Bike sharing count is negatively co-related to **Humidity, Snowfall, Rainfall.**

- Temperature and Dew point temperature are positively co-related.

Models List

In this project we used total twelve models, so that we can compare the final Root mean square error and R2 score of this models.

```
# List of models that we are going to use for this dataset
models = [
            ['LinearRegression: ',              LinearRegression()],
            ['Lasso: ',                         Lasso()],
            ['Ridge: ',                         Ridge()],
            ['KNeighborsRegressor: ',           neighbors.KNeighborsRegressor()],
            ['SVR:' ,                           SVR(kernel='rbf')],
            ['DecisionTree ',                   DecisionTreeRegressor(random_state=42)],
            ['RandomForest ',                   RandomForestRegressor(random_state=42)],
            ['ExtraTreeRegressor :',            ExtraTreesRegressor(random_state=42)],
            ['GradientBoostingRegressor: ',     GradientBoostingRegressor(random_state=42)],
            ['XGBRegressor: ',                  xgb.XGBRegressor(random_state=42)] ,
            ['Light-GBM: ',                     lightgbm.LGBMRegressor(num_leaves=41, n_estimators=200,random_state=42)],
]
```

Result

As we can see clearly out of twelve models Lightgbm, ExtraTreeRegressor and RandomForestRegressor give us max R2 score and Less Root mean square error on test set.

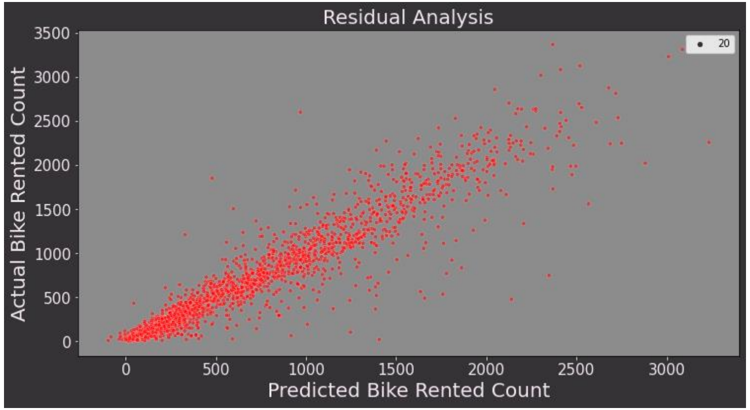
	Name	Train_Time	Train_R2_Score	Train_RMSE_Score	Test_R2_Score	Test_RMSE_Score
0	LinearRegression:	7.152557e-07	0.539158	439.103836	0.555399	419.033651
1	Lasso:	7.152557e-07	0.534718	441.213908	0.552075	420.596945
2	Ridge:	1.430511e-06	0.538927	439.213876	0.555500	418.986169
3	KNeighborsRegressor:	7.152557e-07	0.863473	239.001167	0.800327	280.816708
4	SVR:	1.192093e-06	0.263863	554.971348	0.289494	529.721550
5	DecisionTree	1.192093e-06	1.000000	0.000000	0.743764	318.114797
6	RandomForest	9.536743e-07	0.983017	84.294528	0.873439	223.570369
7	ExtraTreeRegressor :	4.768372e-07	1.000000	0.000000	0.878614	218.951240
8	GradientBoostingRegressor:	4.768372e-07	0.867964	235.037831	0.848602	244.525043
9	XGBRegressor:	1.192093e-06	0.866833	236.042007	0.848415	244.676517
10	Light-GBM:	9.536743e-07	0.974173	103.950667	0.888898	209.471137

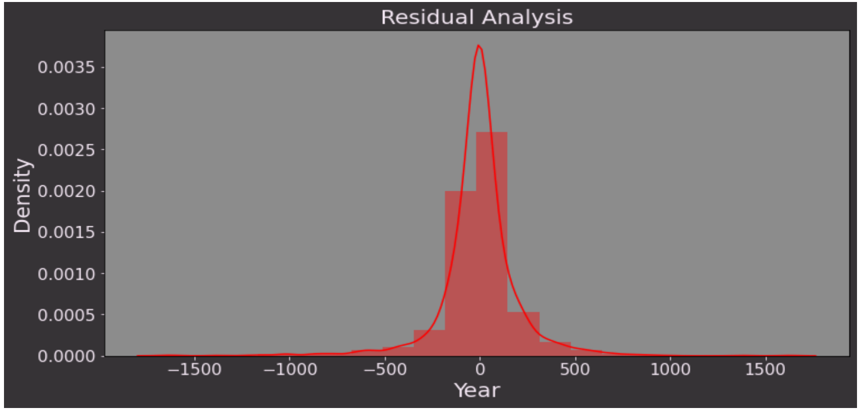
Model-1 ExtraTreeRegressor

- Extratree improves the RMSE significantly on the Test set. its evident from the below plot the Predicted and actual values are much closer compared to other Models.

- R-score of **0.878** and RMSE : **218.95**

- Residual values are reduced remarkably for the Extratree. The KDE plot is much leaner and most of the Residual values are closer to zero.





Hyperparameter Tuning of ExtraTreeRegressor

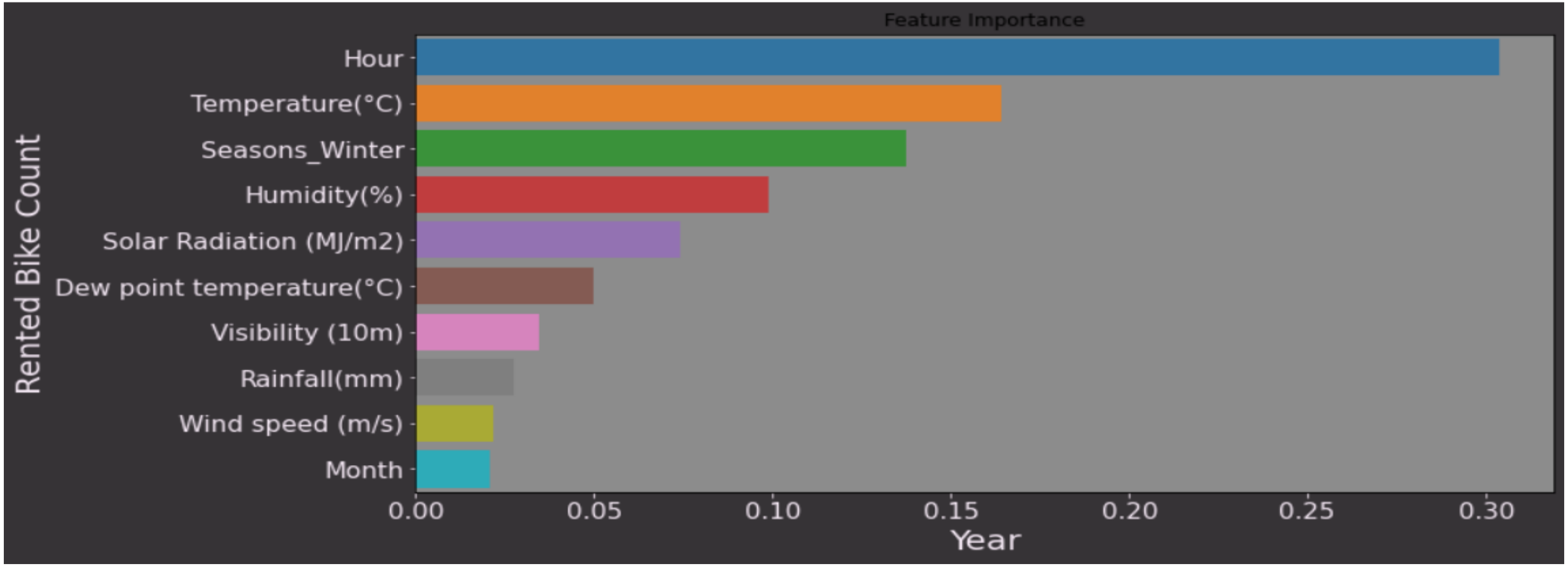
- For Hyperparameter tunning we used random search cv method to find best hyper-paramters for our model.
- R2 Score after Hyperparameter tunning increased by 0.76% only
- R-Score **.885** and RMSE : **212.78 .**

```
# Create the random grid
random_grid ={'bootstrap': [True, False],
              'max_depth': [70, 80, 90, 100, None],
              'max_features': ['auto', 'sqrt'],
              'min_samples_leaf': [1, 2, 4],
              'min_samples_split': [2, 5, 10],
              'n_estimators': [800, 1000]}

RF = ExtraTreesRegressor(n_jobs=-1, random_state=42)

# Random search of parameters, using 3 fold cross validation,
random_search = RandomizedSearchCV(estimator = RF,
                                   param_distributions = random_grid,
                                   n_iter = 100, cv = 3, verbose=2)
# Fit the random search model
random_search.fit(train_inputs, train_targets)
```

```
print('Improvement of {:0.2f}%.'.format( 100 * (0.8853 - 0.87863) / 0.8773))
```

```
Improvement of 0.76%.
```

Feature Importance



- The adjacent Graph shows the importance of the features on our Rented bike count.
- **Temperature** and **Hour** of the day is a major factors **driving** the **demand for bikes**.
- **Solar Radiation, Humidity, Rainfall**, where its working day or not are other variables driving the demand for bikes.

Model 2 - Light GBM

LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

1. Faster training speed and higher efficiency.
2. Lower memory usage.
3. Better accuracy.
4. Support of parallel, distributed, and GPU learning.
5. Capable of handling large-scale data.

LightGBM improves the RMSE significantly on the Test set. its evident from the below plot the Predicted and actual values are much closer compared to other Models.

R-score of **0.878** and RMSE : **218.95**

Hyperparameter Tuning of LightGBM

- For Hyperparameter tunning we used Optuna Library to find best hyper-paramters for our model.
  Optuna works on Bayesian search method.

```
import optuna
from optuna import Trial, visualization
from optuna.samplers import TPESampler
def objective(trial,data=data):


    param = {
        'metric': 'rmse',
        'random_state': 42,
        'n_estimators': 10000,
        'reg_alpha': trial.suggest_loguniform('reg_alpha', 1e-3, 10.0),
        'reg_lambda': trial.suggest_loguniform('reg_lambda', 1e-3, 10.0),
        'colsample_bytree': trial.suggest_categorical('colsample_bytree', [0.3,0.4,0.5,0.6,0.7,0.8,0.9, 1.0]),
        'subsample': trial.suggest_categorical('subsample', [0.4,0.5,0.6,0.7,0.8,1.0]),
        'learning_rate': trial.suggest_categorical('learning_rate', [0.006,0.008,0.01,0.014,0.017,0.02]),
        'max_depth': trial.suggest_categorical('max_depth', [10,20,100]),
        'num_leaves' : trial.suggest_int('num_leaves', 1, 1000),
        'min_child_samples': trial.suggest_int('min_child_samples', 1, 300),
        'cat_smooth' : trial.suggest_int('min_data_per_groups', 1, 100)
    }

    model = lightgbm.LGBMRegressor(**param)
    model.fit(train_inputs,train_targets,eval_set=[(val_inputs,val_targets)],early_stopping_rounds=100,verbose=False)

    preds = model.predict(val_inputs)
    rmse = metrics.mean_squared_error(val_targets, preds,squared=False)

    return rmse
```

```
print('Improvement of {:0.2f}%.'.format( 100 * (0.9037 - 0.8878) / 0.8878))
```

```
Improvement of 1.79%.
```

Conclusion

- Most numbers of Bikes were rented in **Summer**, followed by **Autumn**, **Spring**, and **Winter**. **May-July** is the peak Bike renting Season, and **Dec-Feb** is the least preferred month for bike renting.

- **Majority** of the **client** in the bike rental sector belongs to the **Working class**. This is evident from EDA analysis where bike demand is more on weekdays, working days in Seoul.

- **Temperature** of **20-30 Degrees**, **evening time 4 pm- 8 pm, Humidity** between **40%-60%** are the most favorable parameters where the Bike demand is at its peak.

- **Temperature, Hour** of the day, **Solar radiation**, and **Humidity** are major driving factors for the Bike rent demand.

- Feature and Labels had a weak linear relationship, hence the prediction from the linear model was very low. Best predictions are obtained with a **LightGBM** as r2 score of **0.894** and RMSE of **203.91**

THANK YOU