Capstone Project-2

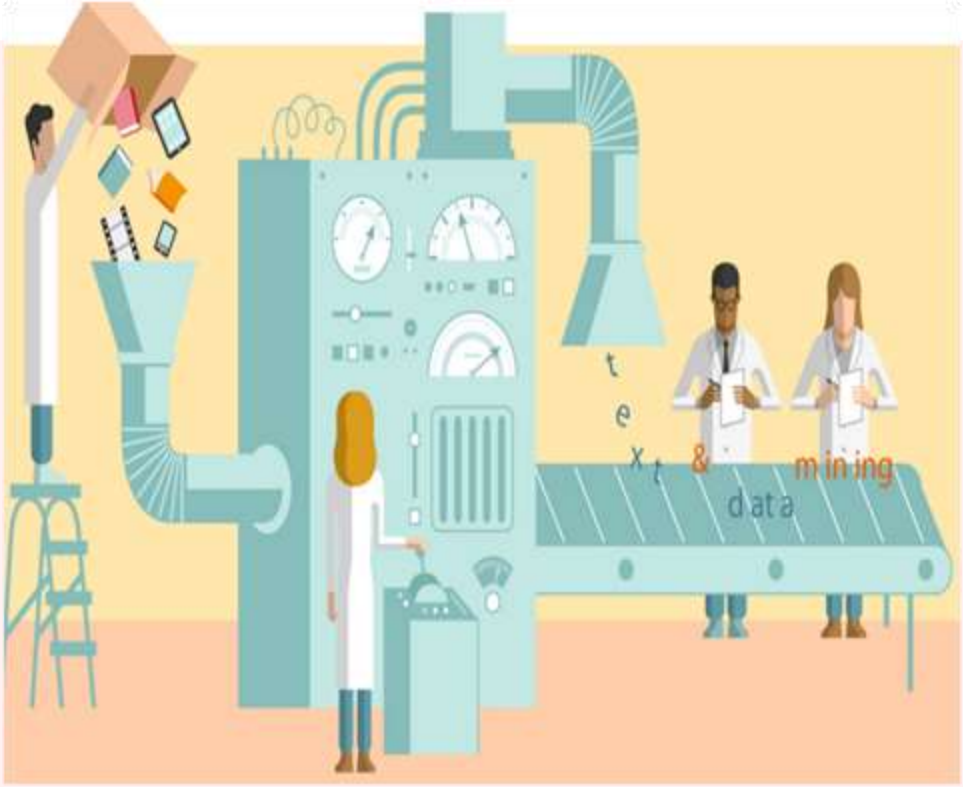
Presentation on

Bike Sharing Demand Prediction

Presented By: Anis Bagwan
Ankit Rai

Agenda

- Defining Problem Statement
- Data Pipeline
- Data Summary
- EDA
- Applying Model
- Model Validation and Selection
- Conclusion



Problem Statement

- In Rented Bike sharing system, the demand keeps fluctuating which means it varies over time depending on various factors. This fluctuating demand may cause the uneven distribution of availability and could possibly lead to the problem of insufficient bikes in certain area for a certain period.

- Insufficient availability of bike for such time could cause user dissatisfaction followed by the unpopularity of the service.

- The main objective is to build a predictive model, which could help the system in predicting shared Bike Demand proactively. This would in turn help the system in matching the user's demand quickly and efficiently.

Data Pipeline

- **Data Pre-processing :** After checking for null ,duplicate values and outliers, we derived new variables (month, year, weekend) from existing variable “Date”. We also dropped irrelevant variables like – “Functioning Day”. We also eliminated redundant variable like - “Dew Point temperature”
- **EDA :** We did exploratory analysis of dependent and independent variables
- **One Hot Encoding** : We applied hot encoding to categorical columns
- **Creating a model :** Finally we tried creating different regression models and applied them. We started with simple Linear regression model and then increases the complexity.

Data summary

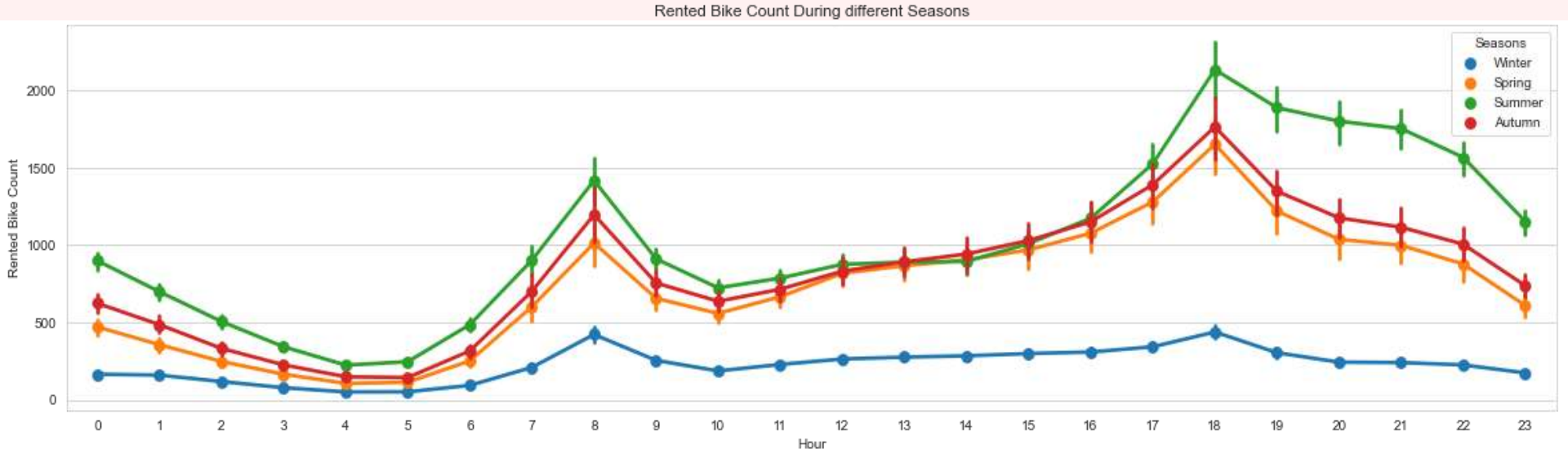
- “**Date:** Date on which bike was rented
- **Rented Bike count :** It tells number of Bikes rented ,for a given date and for a given one hour duration
- **Hour :** Hour of the day for which data is given
- **Temperature :** Temperature of city in Celsius for the given time duration
- **Humidity - % :** Humidity in the city for the given time duration
- **Windspeed - m/s :** Windspeed in the city for the given time duration
- **Visibility - 10m:** This feature gives the information about the visibility condition of the city

Data summary

- **Dew point temperature – Celsius :** Dew point temperature of city in Celsius for the given time duration
- **Solar radiation - MJ/m2 :** Tells the Solar Radiation Intensity for a given time
- **Rainfall – mm:** Talks about the Rainfall intensity
- **Snowfall – cm:** Talks about the Snowfall intensity
- **Seasons : Winter, Spring, Summer, Autumn:** Type of Season for which data is taken
- **Holiday - Holiday/No holiday:** Whether the given day was a holiday or not
- **Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours) :** For the given day, whether the Bike Sharing System was functioning/working or not

Exploratory Data Analysis

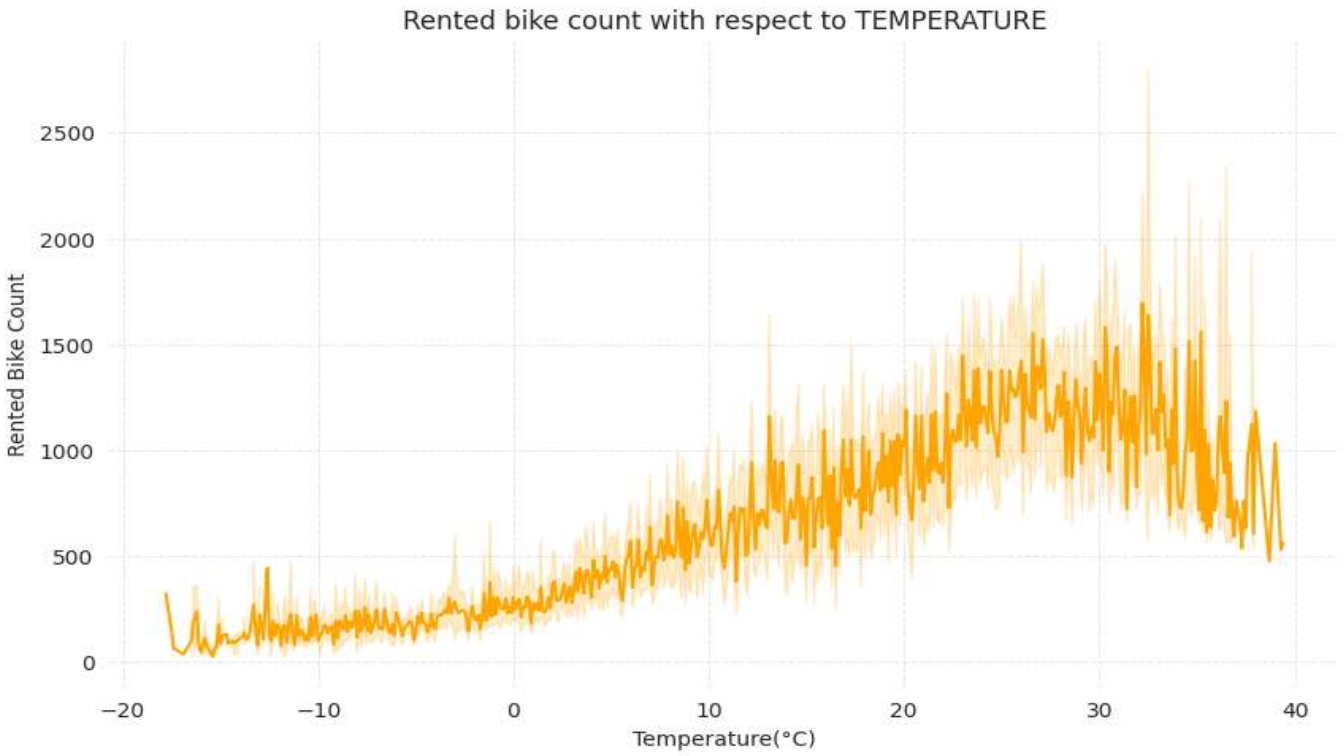
- The bike sharing demand shoots up between 5pm to 7pm and 7 am to 9 am.
- The demand is lowest between 4am to 5 am.
- The hourly based demand pattern is almost same for all seasons.



Exploratory Data Analysis

Rented bike count with respect to TEMPERATURE

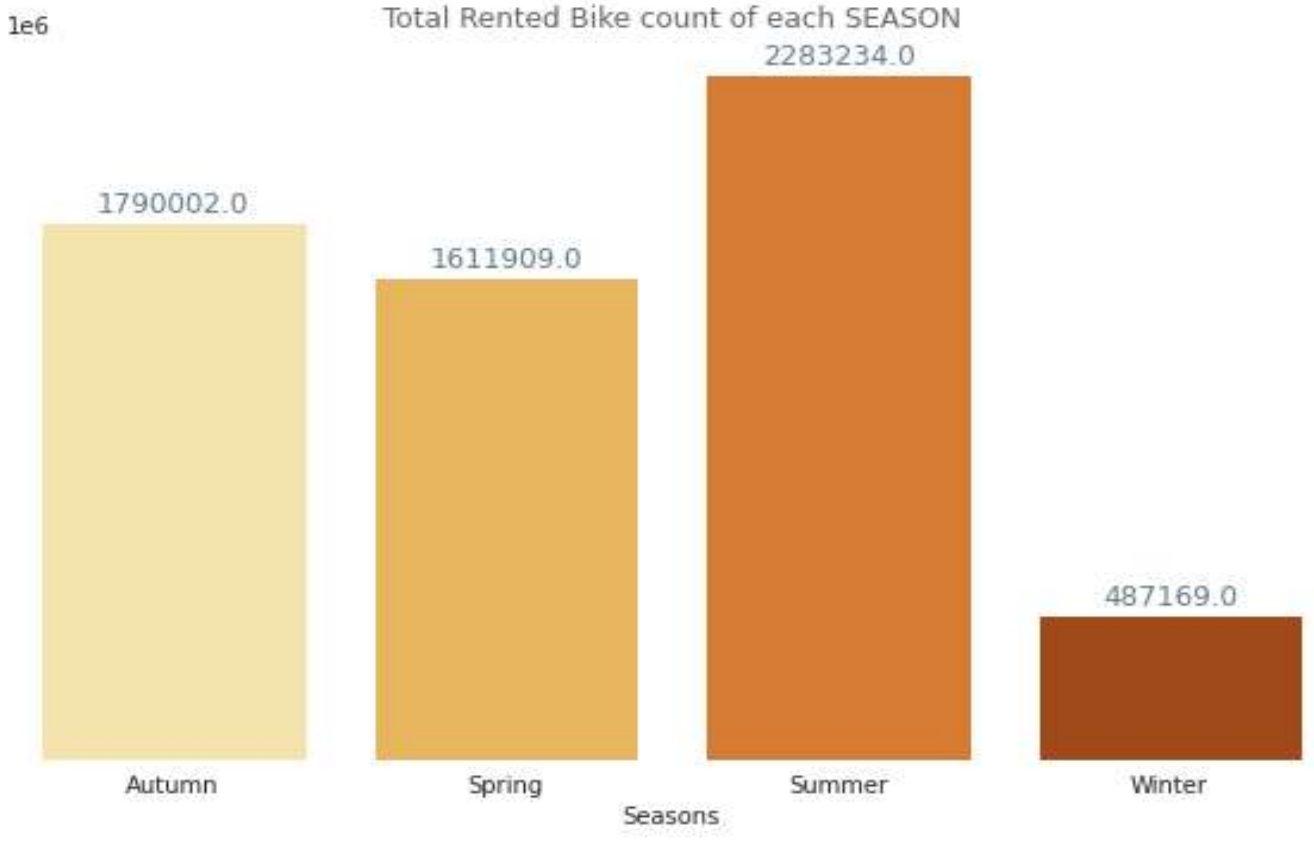
- The bike demand shows linear relation with temperature up to some extend.
- The demand is maximum when temp is between 25 to 35 deg. Celsius
- The demand falls when temp is below 10 Deg Celsius.



Exploratory Data Analysis

Total Rented Bike count of each SEASON

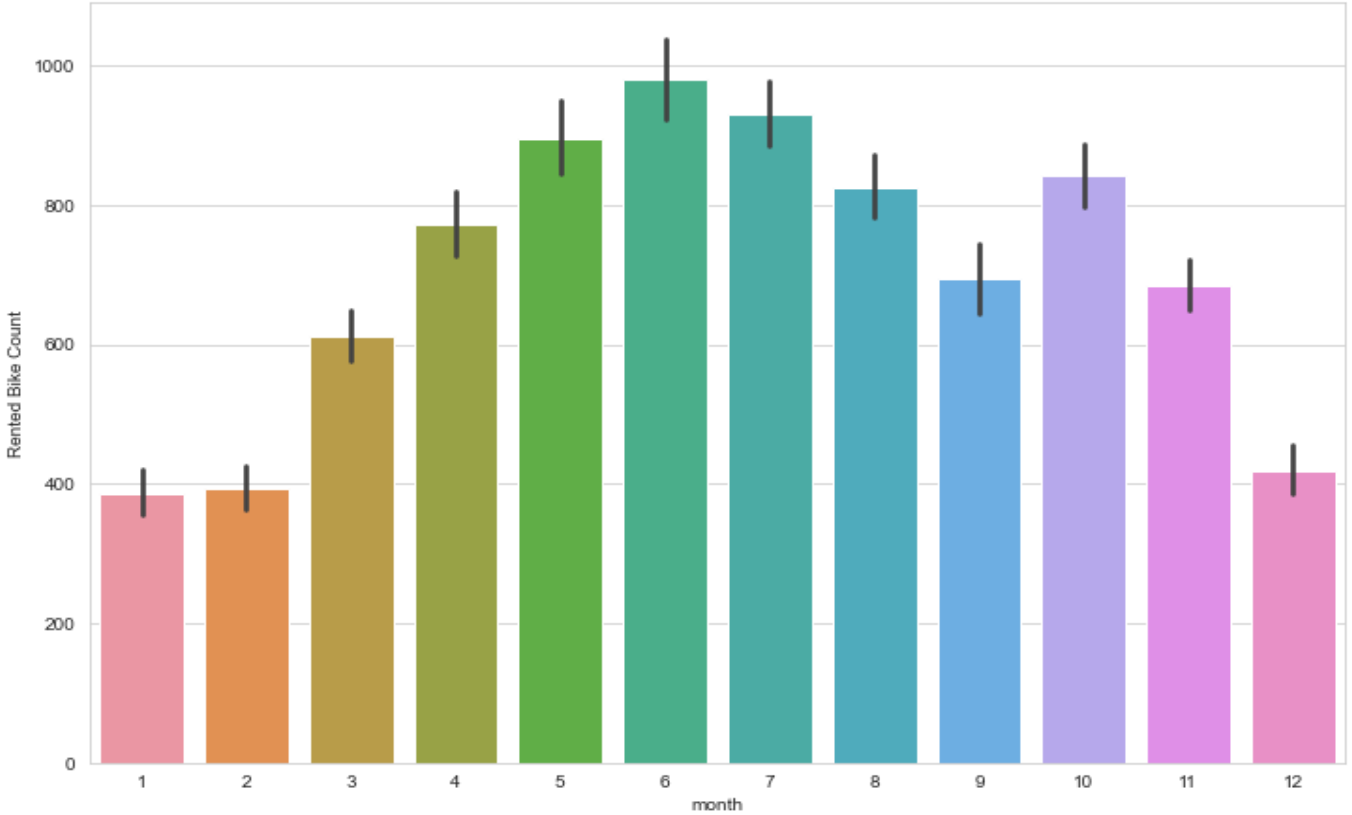
- As we can observe for each season, people took bikes on rent.
- In Summer the demand for rented bike is maximum. On the other hand, in winter the demand falls and in Spring and Autumn it is moderate.
- From this we can say that people react differently in each season.



Exploratory Data Analysis

- The bike demand is maximum for June month, followed by July , may and October.
- The Demand falls in January and Feb month

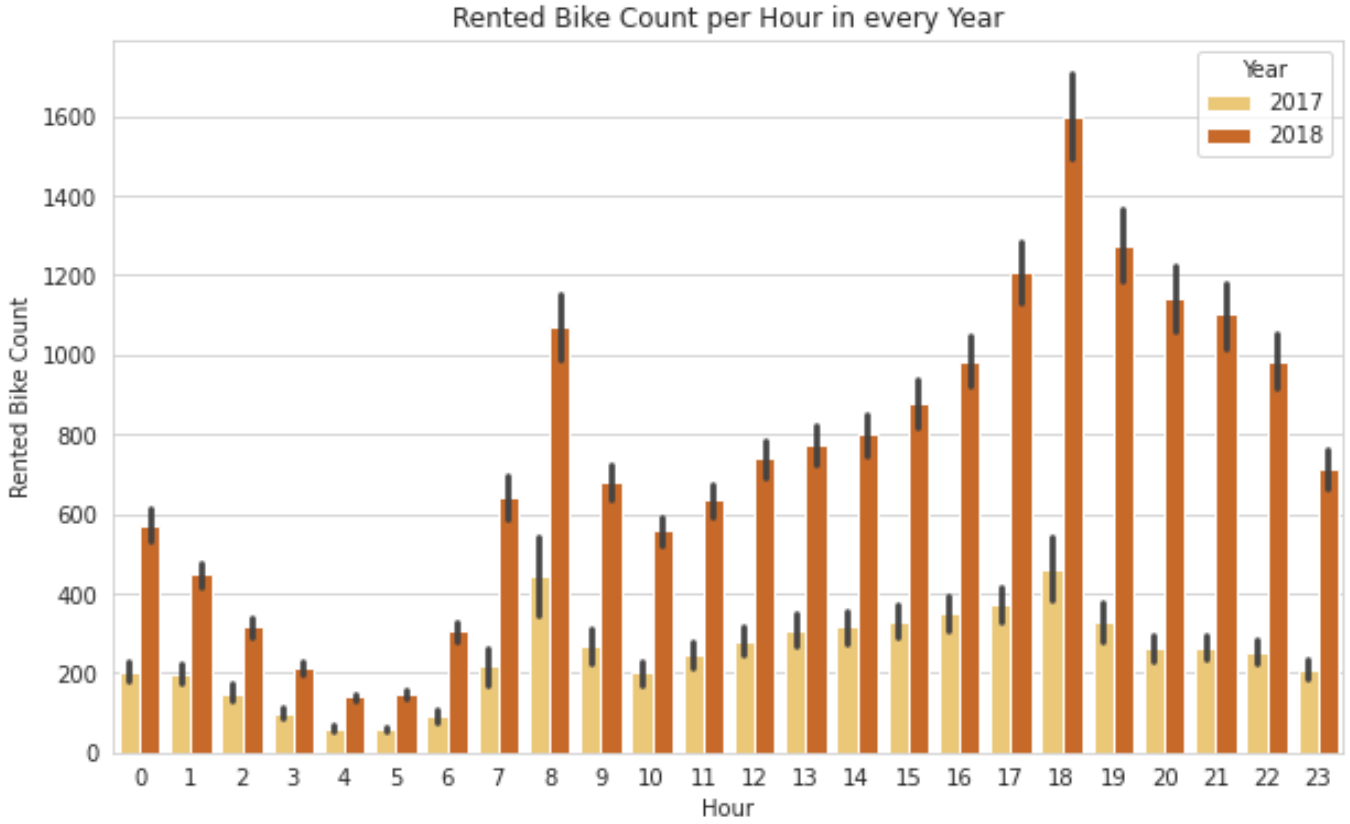
Total Rented Bike count for each month



Exploratory Data Analysis

Rented Bike Count per Hour in every Year

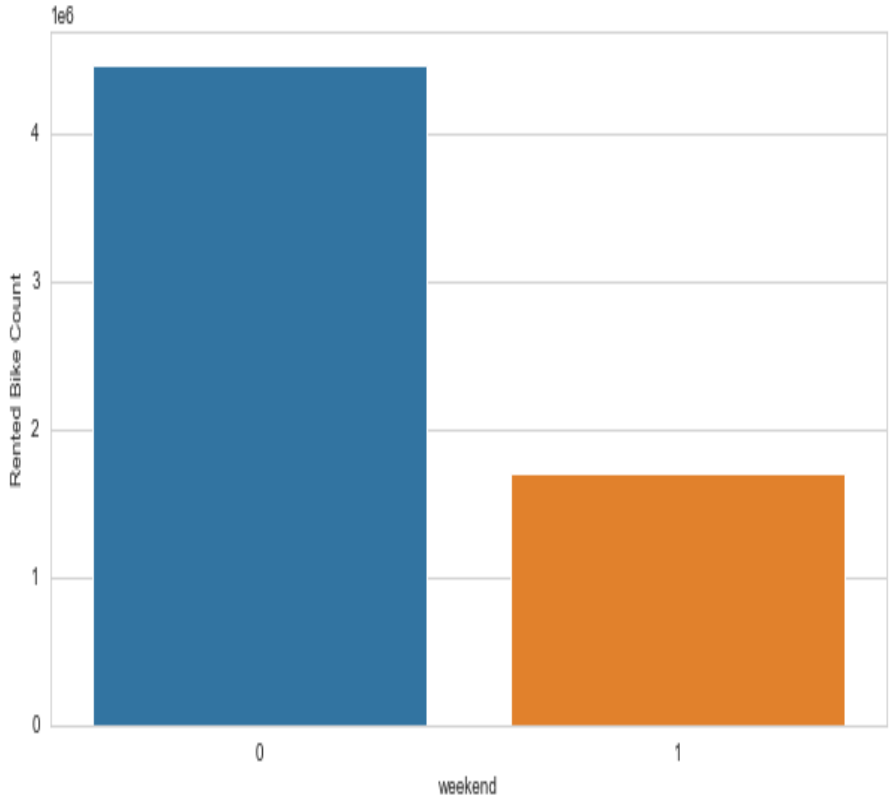
- From plot it seems that the bike demand has increased from previous year (2017).



Exploratory Data Analysis

Effect of Weekend on Bike Demand

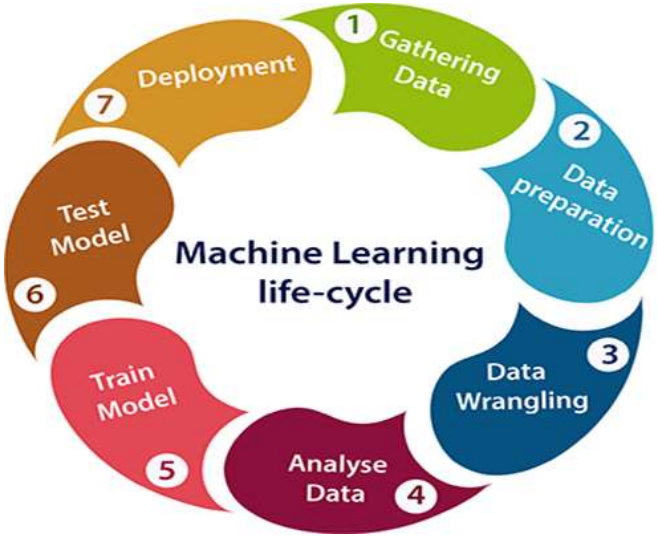
- The bike demand falls drastically in weekends.
- Since on weekdays the demand is double to that of in weekend, we can say that most people are using bikes for commuting their workstations.



Data Preparation for Modelling

Here we are processing the dataset for modelling purpose.

- First of all we created the two lists containing the dependent_variable and independent_variable.
- After that we stored the values of these variables in X and y variables respectively.
- Then we split the dataset into train and test set like “X_train”, “X_test”, “y_train” and “y_test”.



Applying Model

1. Linear Regression

- The first model we are fitting is the linear regression.
- After applying this model we will get the following scores of evaluation metrics:

RMSE: 397.65

R2 score: 0.619

Applying Model

2. Gradient Boosting Algorithm

- The second model we are fitting is the gradient boosting algorithm.
- After applying this model we will get the following scores of evaluation metrics:

RMSE: 316.129

R2 score: 0.778

Applying Model

3. Random Forest Algorithm

- The third model we are fitting is the random forest algorithm.
- After applying this model we will get the following scores of evaluation metrics:

RMSE:288.479

R2 score: :0.79

Applying Model

4. Elastic Net Algorithm

- The fourth model we are fitting is the Elastic Net algorithm.
- After applying this model we will get the following scores of evaluation metrics:

RMSE: 530.034

R2 score: 0.32

Applying Model

5. Hubber Regressor

- The fifth model we are fitting is the Hubber Regressor
- After applying this model we will get the following scores of evaluation metrics:

RMSE: 461.731

R2 score:0.479

Model Validation and Selection

Among all the models, Random Forest algorithm is giving the best r2_score of 0.79 and hence we are taking this algorithm into consideration.

Model	R2 Score	RMSE	MAE
Linear Regressor	0.610872	399.446207	294.821347
Elastic Net	0.405654	493.663886	365.547692
Gradient Booster	0.745530	323.020262	213.574808
Random Forest	0.797628	288.063048	164.844368
HuberRegressor	0.607123	401.365872	290.843800

Hyperparameter Tuning

- By now we had already found that Random Forest model working well for our data. Our next job was to do Hyperparameter tuning to get more refined results.
- We used Randomized Search CV for hyperparameter tuning for improved results.
- Randomized Search CV combines a selection of hyperparameters

BEFORE HYPERTUNING	AFTER HYPERTUNING
RMSE: 288.06	RMSE: 275.99
MSE: 82980.31	MSE: 76175.90
MAE: 164.84	MAE: 168.57
R2 score: 0.79	R2 score: 0.81

Conclusion

- After loading our data, we performed feature engineering on it by removing irrelevant variables and deriving the new ones. Further, we performed EDA and tried drawing insights from the data, like :
  - Rented biked sharing demand goes highest in "summer" while it falls to least in "winter".
  - the bike sharing demand shoots up to peak between 5pm to 7pm and 7 am to 9 am. The demand is lowest between 4am to 5 am
  - The bike counts starts increasing the afternoon (from 3pm to 8pm) where temperature is the highest, with the most visibility, windspeed, and least humidity
- Then we hot encoded categorical variables and rescaled the Numerical Variables.
- After splitting our data into 1:4 ratio for testing and training, we applied five regression models. We also used metrics to find the performance of all the applied models.
- From the metrics result we found that the "**Random Forest**" model gave the best result , with 81% R2 score. Finally, We did hyperparameter tuning for better results

THANK YOU