Capstone Project -2

Bike sharing Demand Prediction

Submitted By:
Somya Jain

Problem Statement

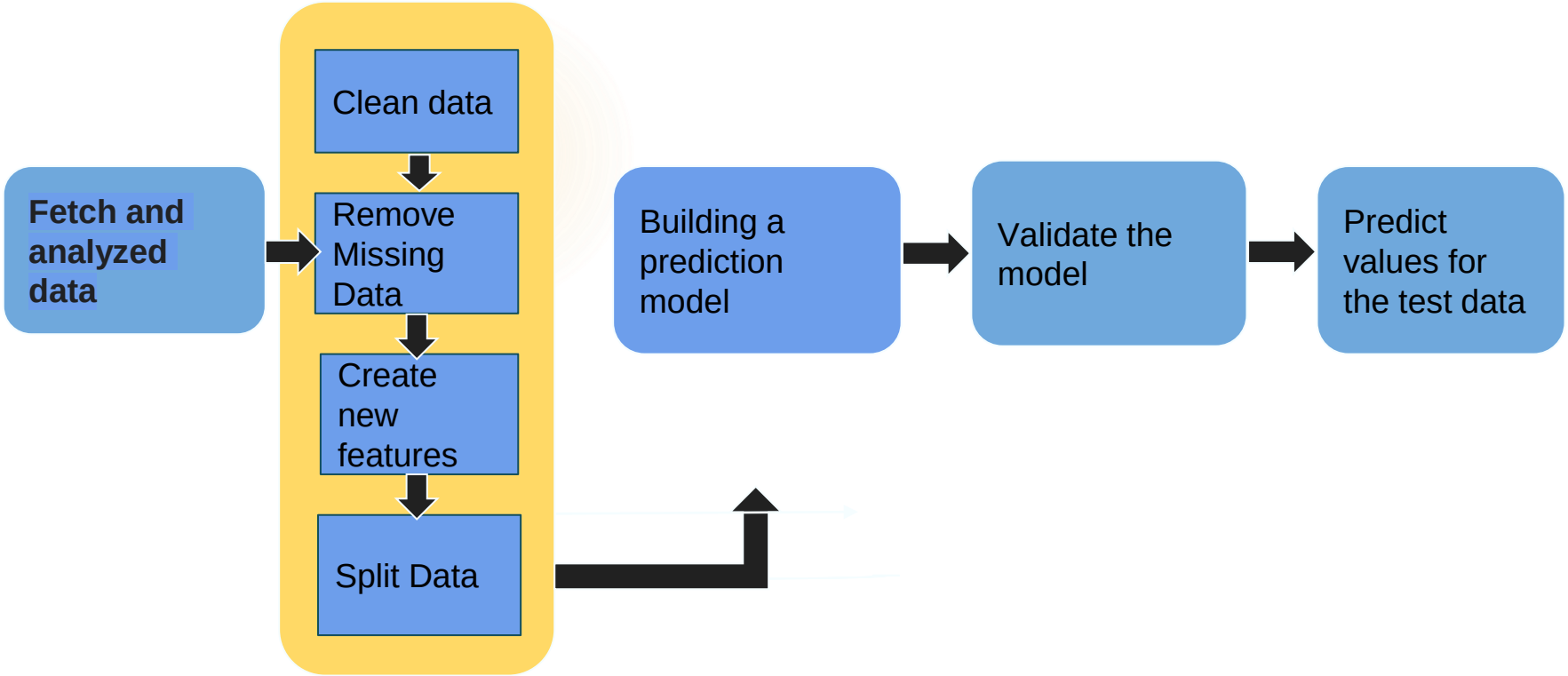
- **Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time.**
- **It is important to make rental bikes available and accessible to customers at the right time as it lessens the waiting time, eventually, providing the city with a stable supply of rental bikes.**
- **Predicting the important features related to the availability of rental bikes.**
- **The goal of this project is to build a ML model that is able to predict the demand of rental bikes in the city of Seoul.**

Objectives

- ➢ Primary Objective: To Build a superior statistical model to predict the number of bicycle that can be rented with availability of data
- ➢ Secondary objectives:

1) To learn how real time data is represented in dataset
2) To understand how to preprocess such data
3) To fit various models such as Regression, Regularized regression(Lasso, ridge and Elastic Net), Decision Tree, Random Forest ,XGBoost, Light GBM

Proposed Methodology



Data Description

The dataset contains 8760 rows and 14 columns,below are the features details:

Data	Data Type	Description
datetime	Object	From Dec 2017 to Nov 2018
Rented Bike count	Integer	The total number of bikes share for the hour
Hour	Integer	Per hour information
Temperature	Float	Temperature in degree celsius
Humidity	Integer	Relative humidity percentage
Wind speed	Float	The speed that air is moving in Meter/sec
Visibility	Integer	Visibility upto 10m

Data Description (Contd.)

Data	Data Type	Description
Dew Point Temperature	Float	Dew point temperature in degree celsius
Solar Radiation	Float	Solar radiation in Megajoules per square meter
Rainfall	Float	Rainfall in mm
Snowfall	Float	Snowfall in cm
Seasons	Object	Categorical (spring, Summer, autumn,winter)
Holiday	Object	Holiday /no holiday
Functional Day	Object	yes/no

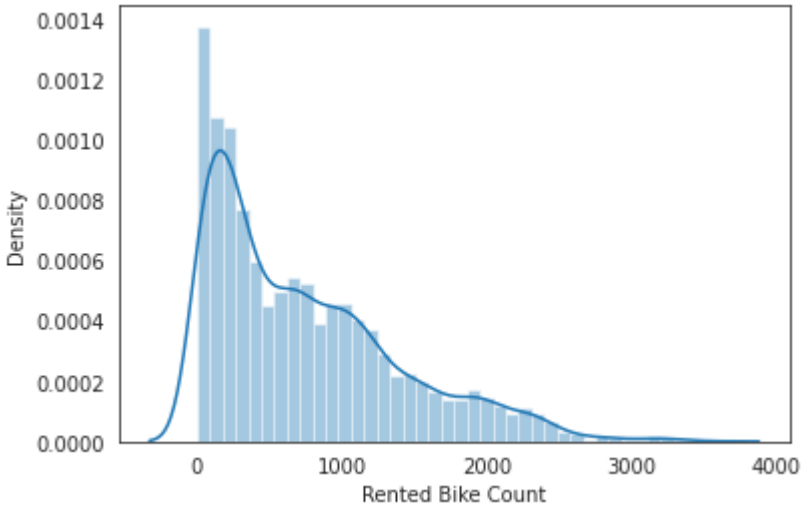
Pre Processing and Data Cleaning

1. **Duplicates Value:** We searched for duplicates value but there were none duplicates in dataset.
2. **Null Values:** There were no null values in dataset
3. **Convert column to appropriate data type:** converted date column to datetime object
4. **Extract new columns:** Extracted day,month, year from date column
5. **Drop Column:** dropped unnecessary column
6. **One hot encoding:** Created dummy variables for season,
   Replace the object values into 0,1 in the columns holiday, functioning day.

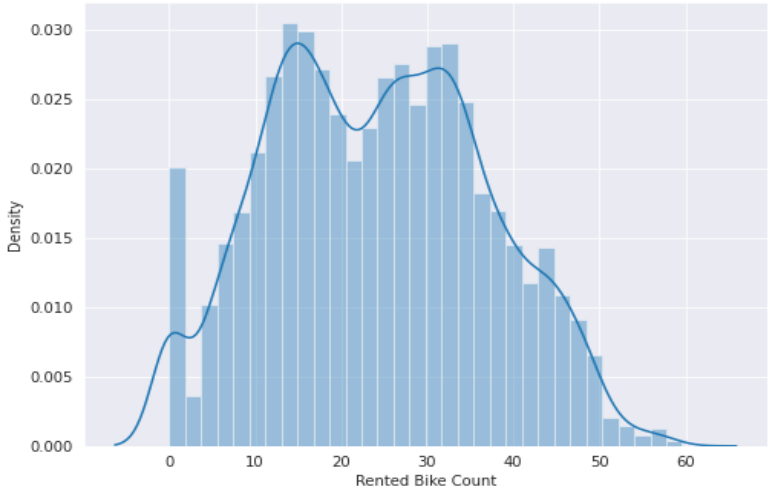
Exploratory Data Analysis

Skewness: Target variable (Rented bike count) is positively skewed

Before Transformation

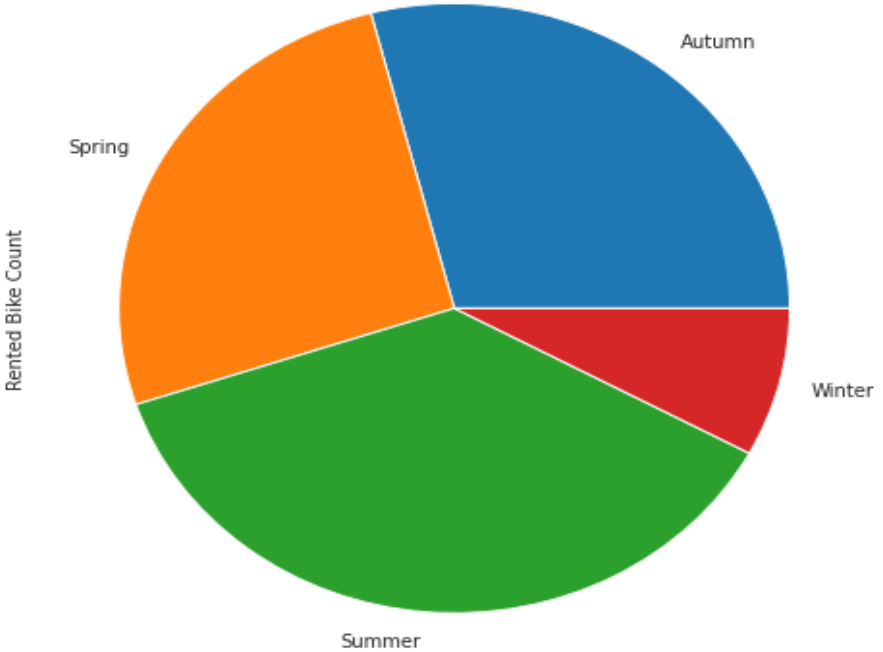


After Square Root Transformation



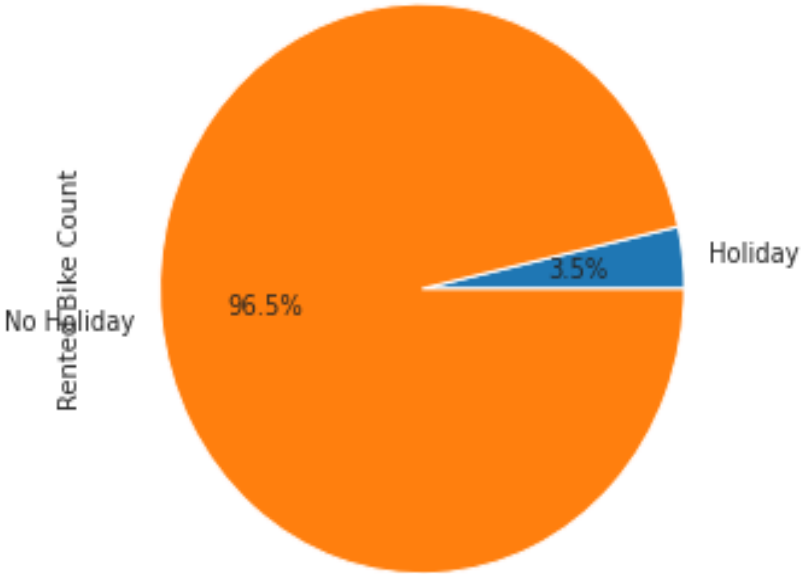
Rented bike count as per seasons

- As we can see most of the bikes rented on the summer season and least number of bikes rented in the winter season
- Business should keep in mind that there is enough availability of bikes in the summer season



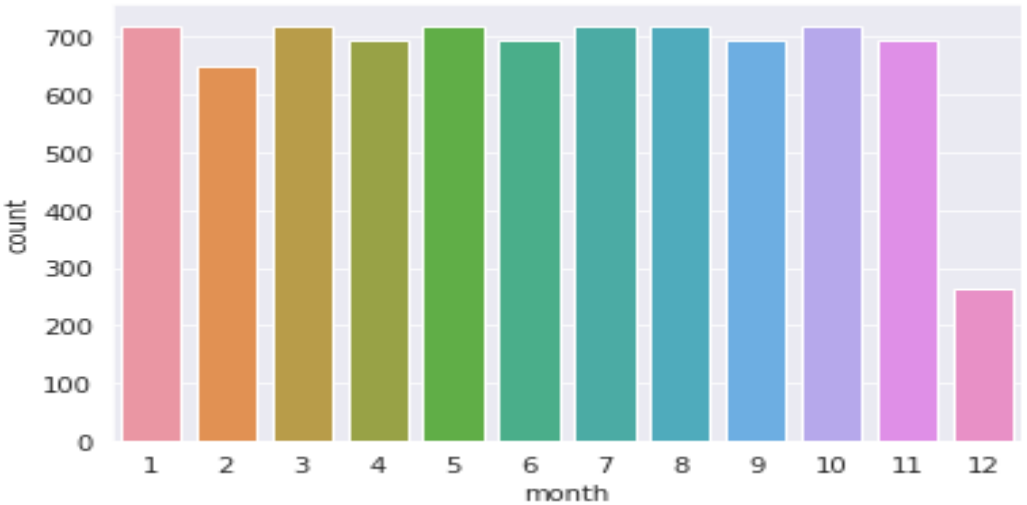
Rented bike count as per Holiday

- As we can see most of the bikes rented on the working days.
- Business should keep in mind that there is enough availability of bikes in working days.



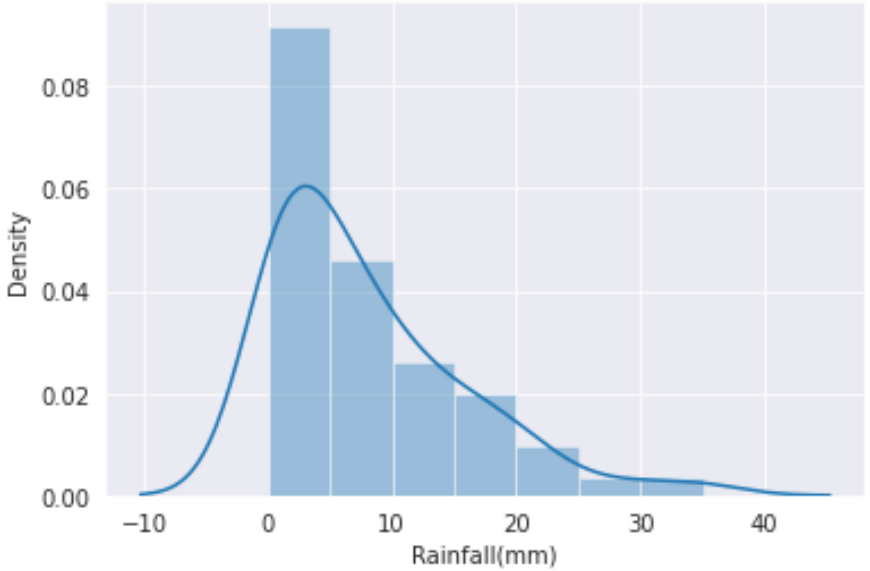
Rented bikes in different months in the year 2018

- The plot shows that very less bikes have been rented in december which is winter season.



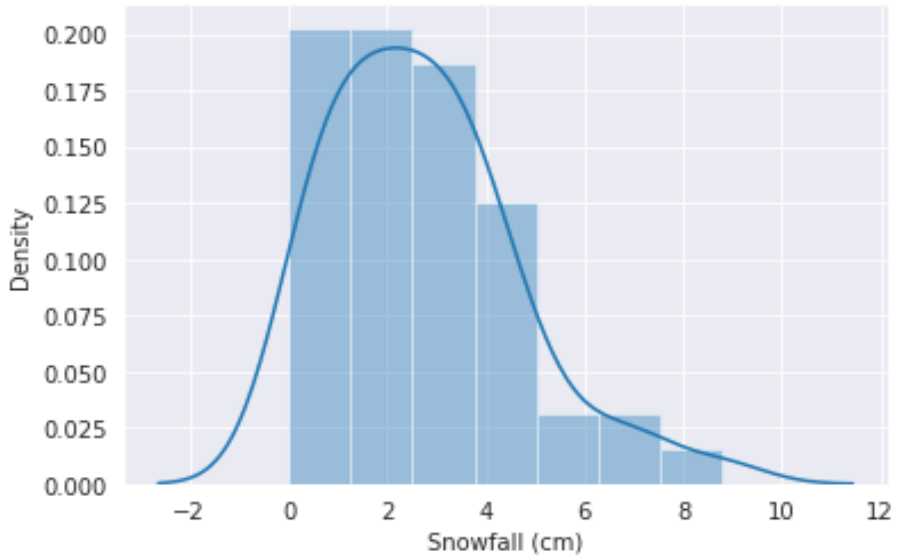
Bikes rented in different intensities rainfall

- Plot shows that people tend to rent bikes when there is no or less rainfall.



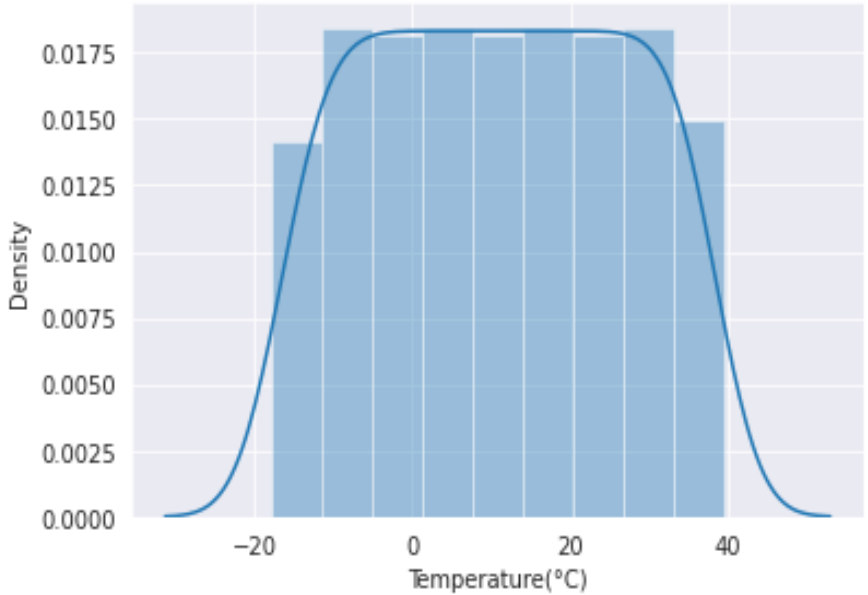
Bikes rented in different intensities snowfall

- Plot shows that people tend to rent bikes when there is no or less snowfall.

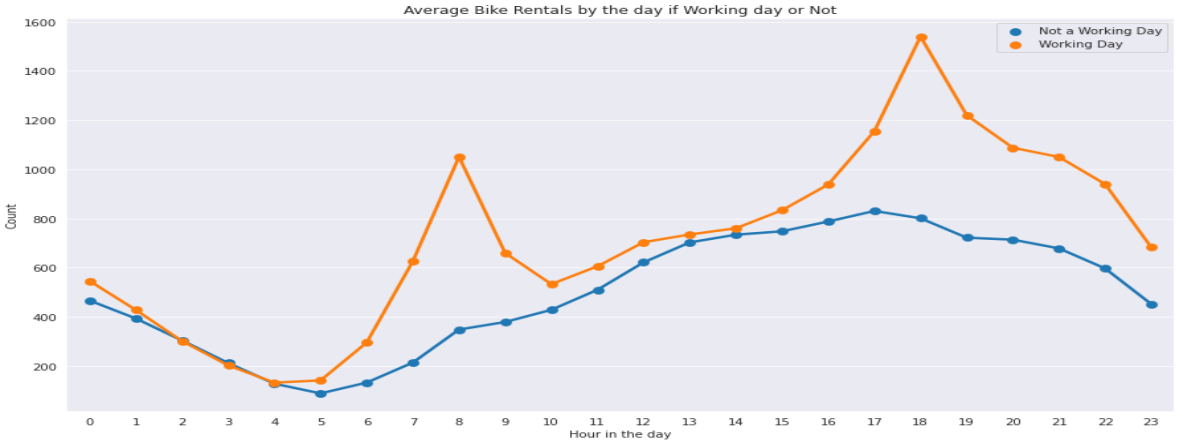


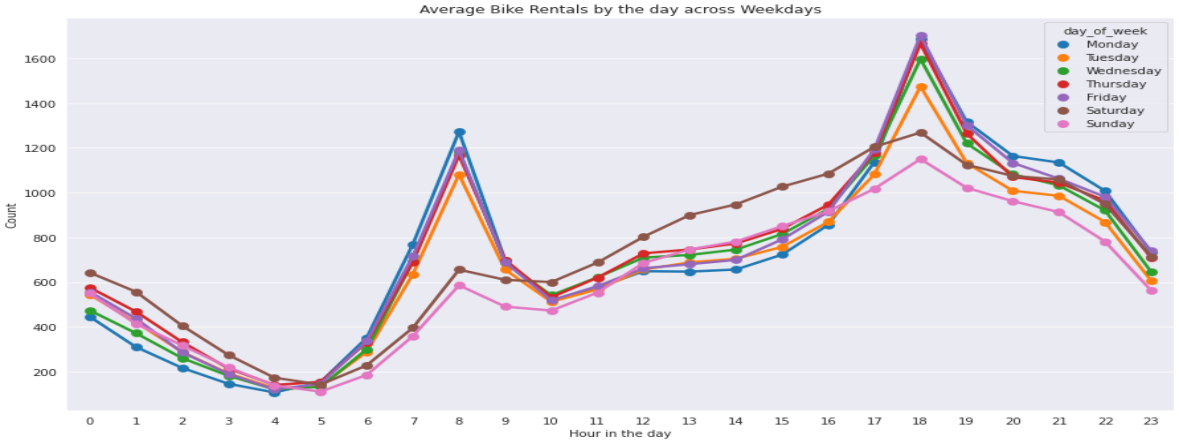
Bike rentals according to temperature intensity

- Plot shows that people tend to rent bikes when the temperature is between -5 to 25 degrees.



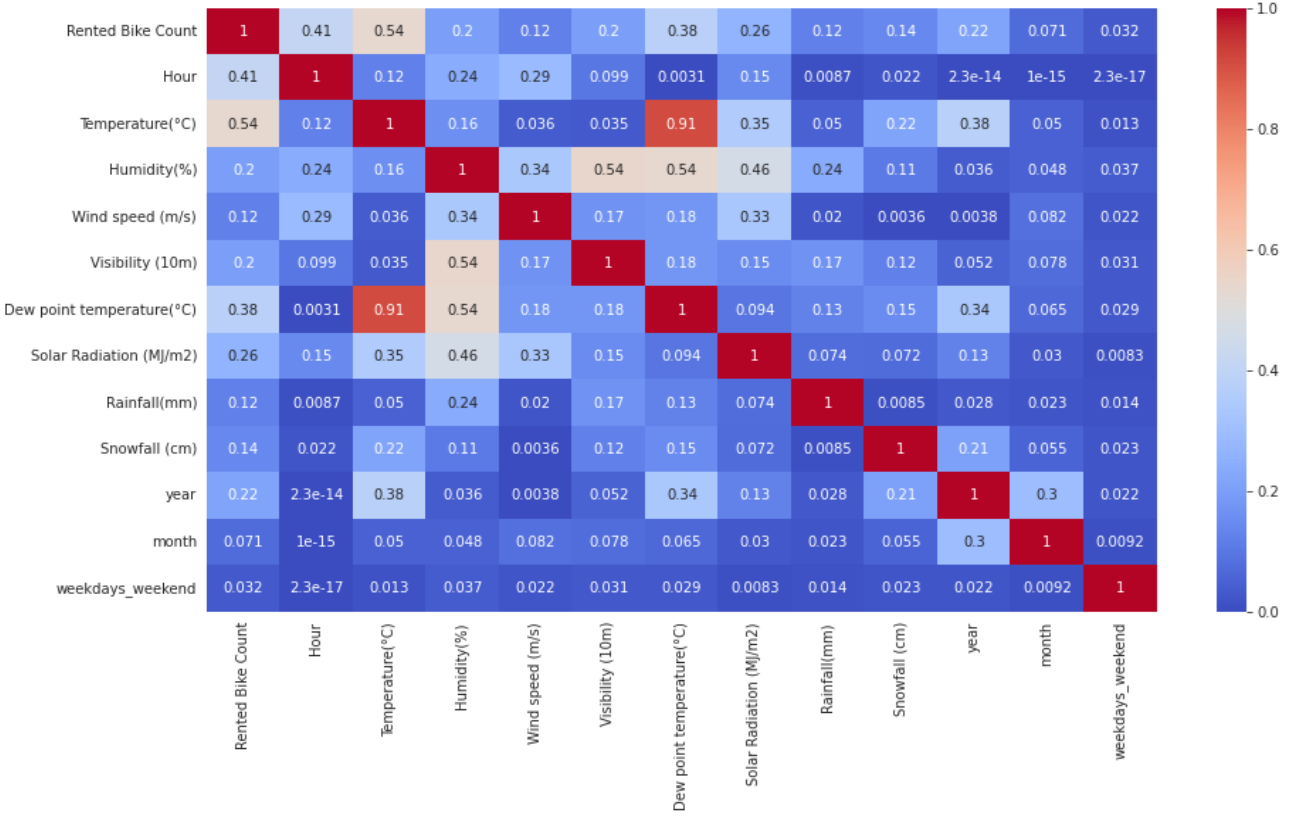
- We see 2 rental patterns across the day in bike rentals count - first for a Working Day where the rental count high at peak office hours (8am and 5pm) and the second for a Non-working day where rental count is more or less uniform across the day with a peak at around noon.
- Bike rental count is mostly correlated with the time of the day. As indicated above, the count reaches a high point during peak hours on a working day and is mostly uniform during the day on a non-working day.



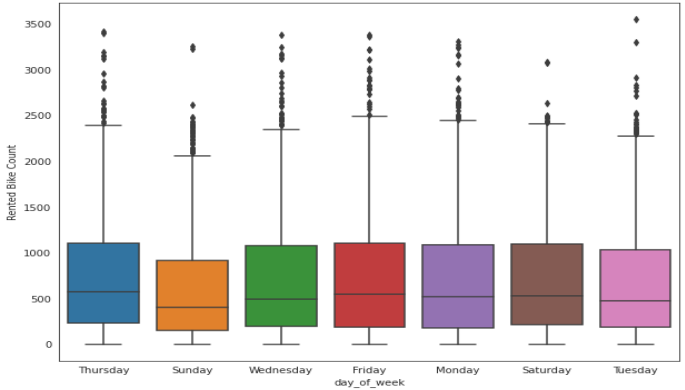


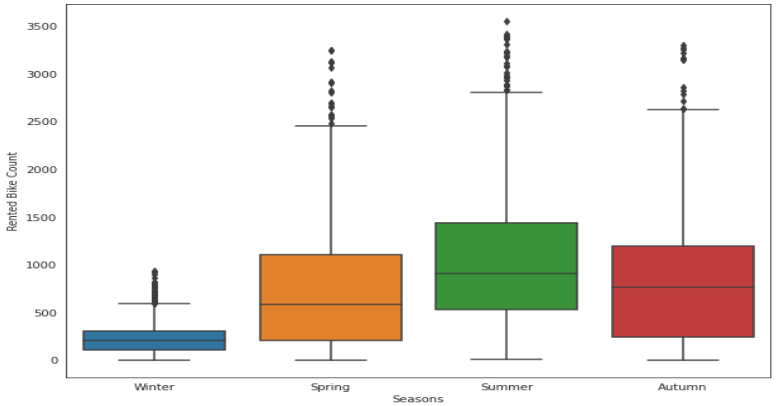
Correlation among the features

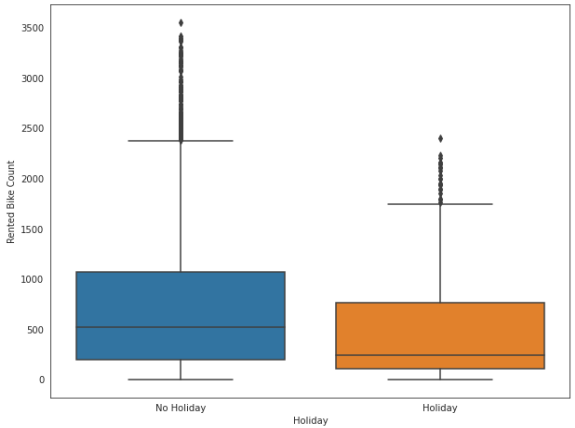
- There is high correlation between temperature and dew point temperature.
- To remove collinearity, used VIF.

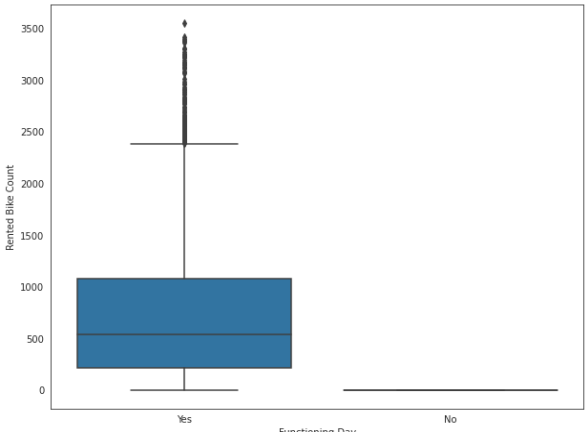


Relation between Categorical and Target Variable









- There are outliers in the data so we have removed them using IQR

Modelling Approach

• Since the data contains outliers, and many categorical attributes. It won't be wise to fit linear models, but we will fit the linear one and check errors.

• We will also use tree models, since they can handle outliers and categorical attributes better than linear models.

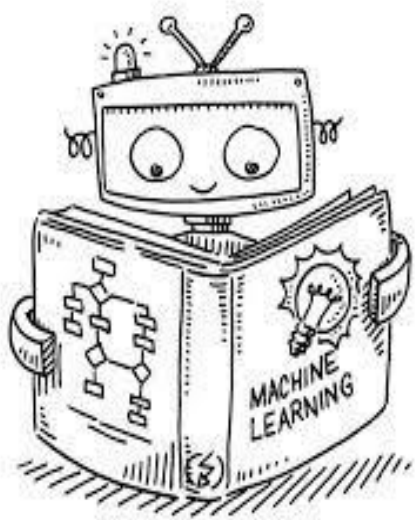
• We will use decision tree as a baseline model.

• Subsequently, to get better predictions, we will use ensemble models: Random forests, XGBoost, Light GBM.

• Final choice of model will depend on whether interpretability or accuracy is important to the stakeholders.

Fitting various model

- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic Net Regression
- Decision Tree
- Random Forest
- XGBoost
- Light GBM



Model Performance Comparison

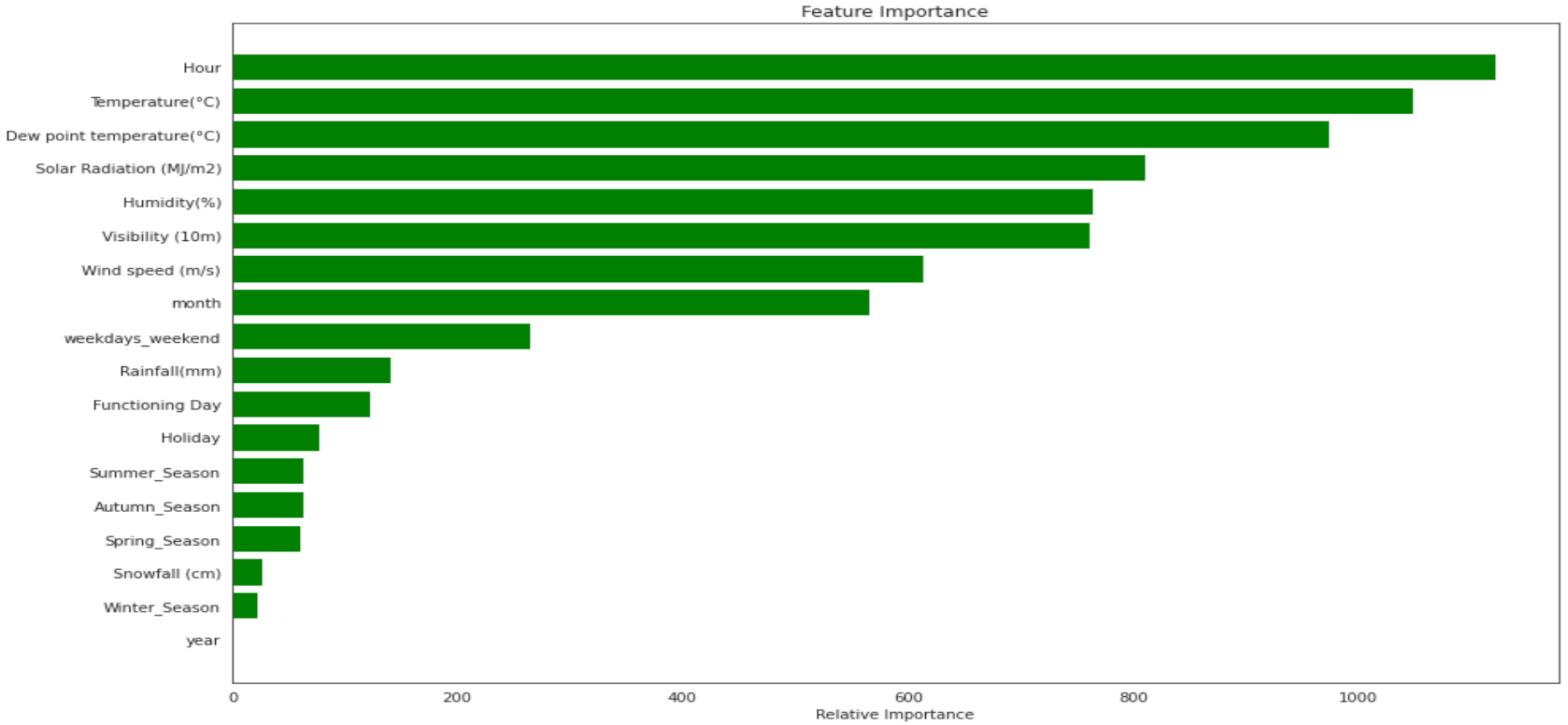
	Linear	Lasso	Ridge	Elasticnet	Decision_Tree	Random_Forest	xgboost	LightGBM
Mean_square_error	193577.938673	194075.891177	194107.613165	193938.455635	99052.302761	48238.569930	67891.326748	39467.309253
Root_Mean_square_error	439.974930	440.540454	440.576456	440.384441	314.725758	219.632807	260.559641	198.663810
R2	0.545853	0.544685	0.544610	0.545007	0.767617	0.886829	0.840722	0.907407
Adjusted_R2	0.540676	0.539494	0.539419	0.539820	0.764967	0.885539	0.838907	0.906351

- The RMSE of Light GBM is least among all the models.
- The R2 score is also the highest in Light GBM.

It indicates that Light GBM is performing best among all the other models.

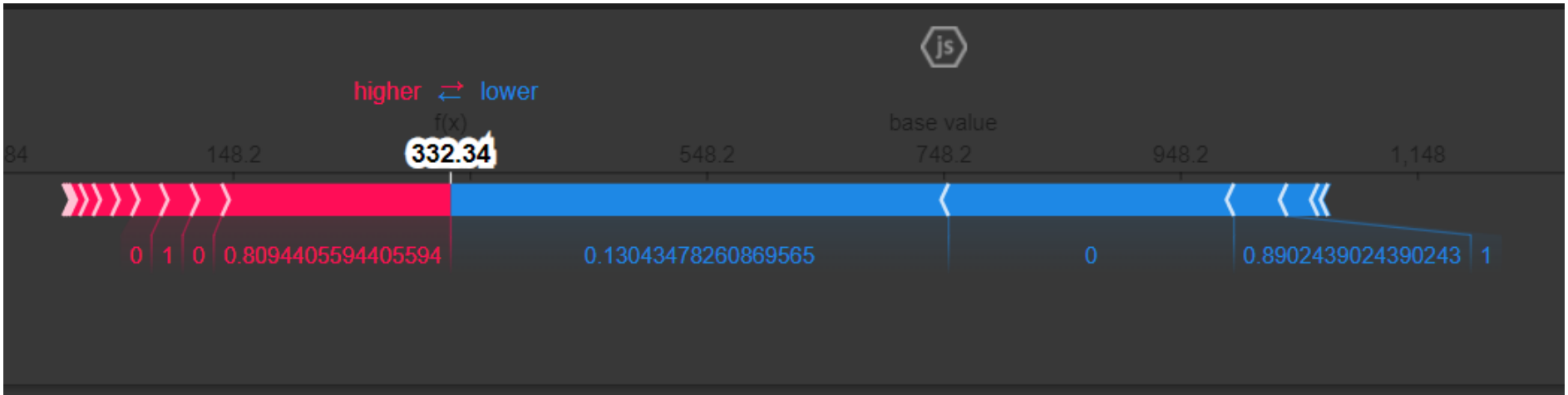
Feature Importance

- Hour of the day is most important features which helping to make our prediction.



Model Explainability

- The force plot is another way to see the effect each feature has on the prediction, for a given observation.
- In this plot the positive SHAP values are displayed on the left side and the negative on the right side, as if competing against each other.
- The highlighted value is the prediction for that observation.



Challenges Faced

- Comprehending the problem statement, and understanding the business implications.
- Feature engineering – deciding on which features to be dropped / kept / transformed
- Choosing the best visualization to show the trends among different features clearly in the EDA phase
- Deciding on how to handle outliers
- Choosing the ML models to make predictions
- Deciding the evaluation metric to evaluate the models
- Choosing the best hyperparameters, which prevents overfitting

We used 6 Regression Models to predict the bike rental count at any hour of the day - Linear Regression, Ridge, Lasso, Random Forest,XG Boost and LightGBM Model.

Below is a summary of the model performances :

- Of all the models, we found LightGBM Model providing the best/lowest RMSE score and highest R^2 score.
- Hour of the day is the most important feature in the respect of all independent feature which provide highest bike rented count.
- Thus, we have successfully built predictive models that can predict the demand for rental bikes based on different weather conditions and all other features.
- If the model interpretability is important to the stakeholders, we can choose deploy the Light GBM model.


Thank You