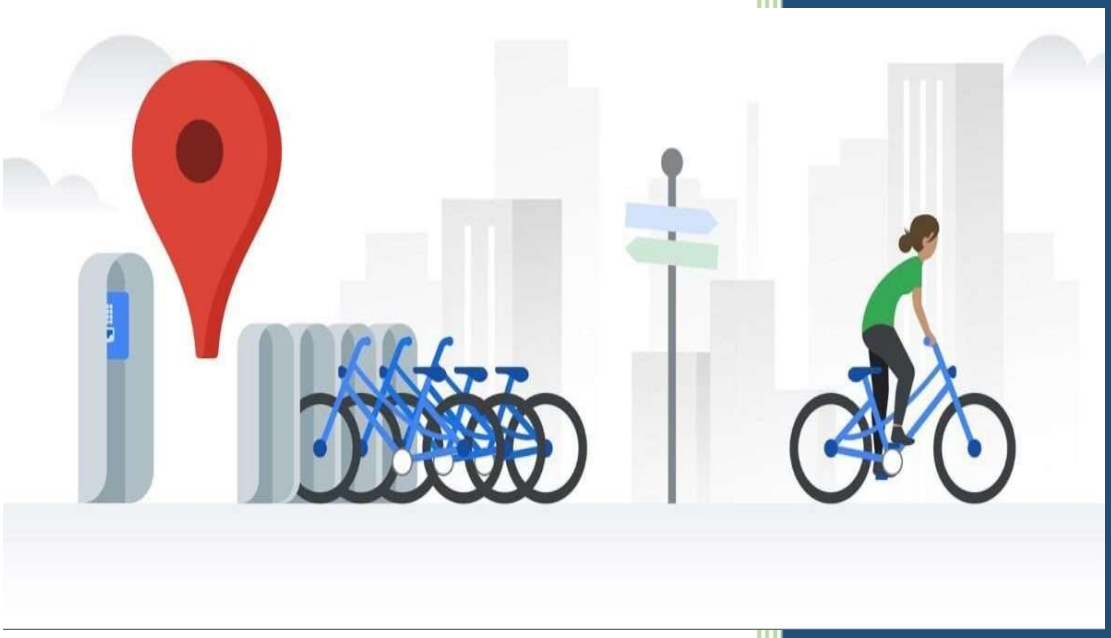
Capstone Project-2

Bike Sharing Demand Prediction



- **Dr. Raj Kumar**
- **Prachi Jadhav**

Abstract- *One of the biggest challenges faced in the bike-sharing system is the unavailability or shortage of bike. This issue has attracted numerous researchers to predict the demand of bike-sharing so that the company is able to redistribute the bikes efficiently and accurately. Correctly predicting the count of the bikes can be challenging especially when the data collected are often imbalance (Sathiskumer and Cho, 2020a). Moreover, despite the several efforts to train models to predict the demand, there is no consensus on which machine learning techniques that will provide the best performance due to the different features applied (Albuquerque et al., 2021). Meanwhile, no standardizedfeatures have proven to be the variables that will significantly improve the models (Albuquerque et al., 2021). Lastly, it is observed that feature engineering is heavily focused in Kaggle kernels, but not in published journal articles. As a result, this report will investigate three main areas such as the best performing machine learning techniques, the feature engineering methods and the features that will significantly enhance the prediction of the bike-sharing demand.*

Table of Contents:

1. Problem Statement:

Bike Rentals have became a popular service in recent year and it seems people are using it more often with relatively cheaper rates and ease of pick up and drop at own convenience is what making this business thrive. 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. 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.

2. Introduction:

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.

3. Overview of data:

We are given the following columns in our data:

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), Yes(Functional hours)

variables.

4. Steps involved:

a. Performing EDA (exploratory data analysis)
b. Drawing conclusions from the data
c. Training the model
d. Evaluating metrics of our model

a. Performing EDA (Exploratory Data Analysis):

A. Exploring head and tail of the data to get insights on the given data.

B. Looking for null values and removing them if it affects the performance of the model.

C. Converting the data into appropriate data types to create a regression model.

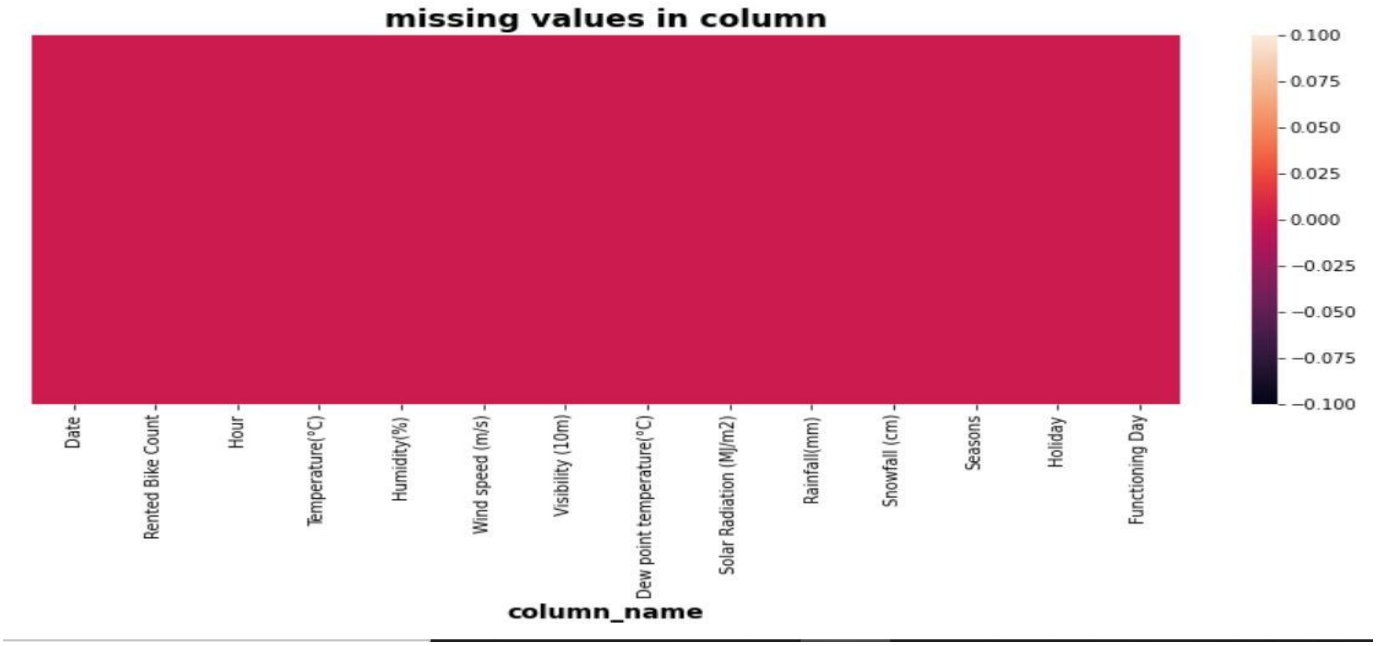
D. Creating data-frames which help in drawing insights from the dataset.

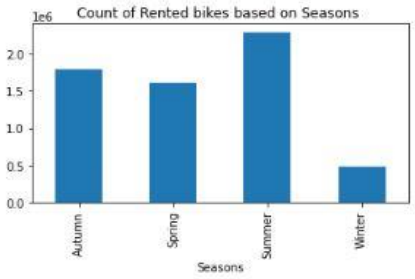
E. Creating more columns in our dataset which would be helpful for creating model.

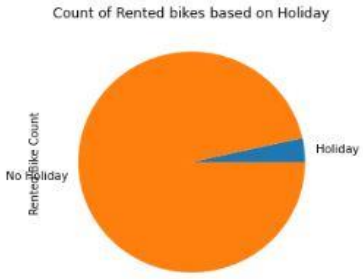
F. Encoding the string type data to better fit our regression model.

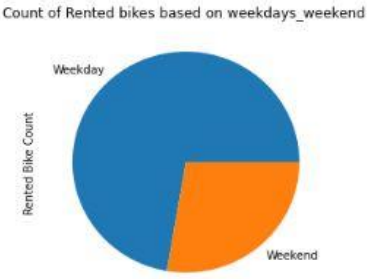
G. Calculating inter-quartile range and filtering our data.

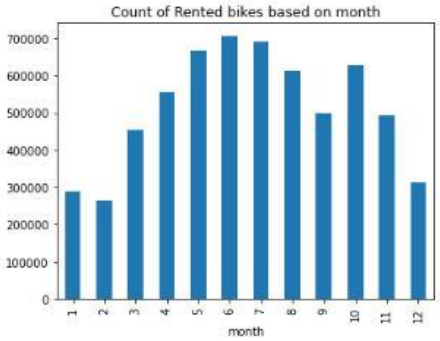
H. Extracting correlation heat-map and calculating VIF to remove correlated and multi-collinear

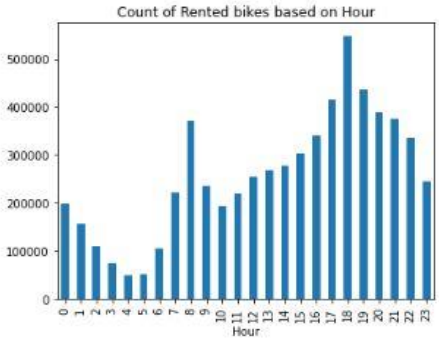


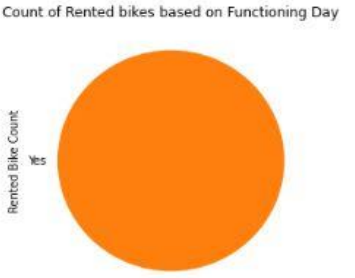

Count of Rented bikes based on Seasons
1e6
2.0
1.5
1.0
0.5
0.0
Autumn
Spring
Summer
Winter
Seasons


Count of Rented bikes based on Holiday
Rented Bike Count
No Holiday
Holiday

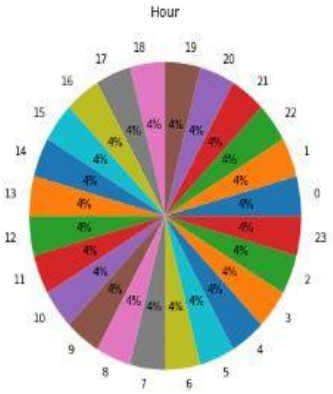

Count of Rented bikes based on weekdays_weekend
Rented Bike Count
Weekday
Weekend

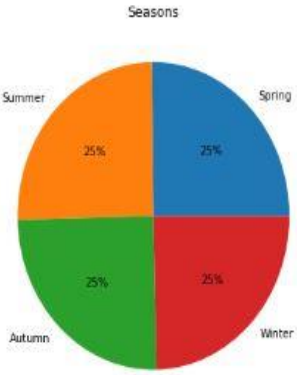

Count of Rented bikes based on month
700000
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
11
12
month

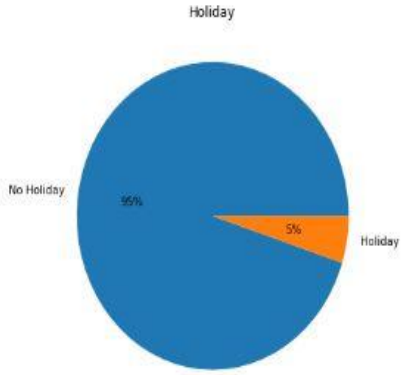

Count of Rented bikes based on Hour
500000
400000
300000
200000
100000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

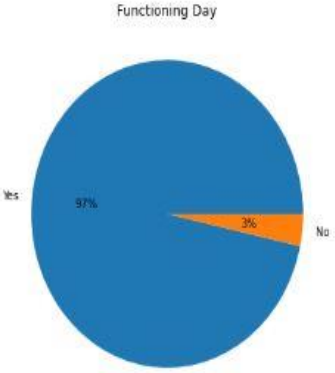

Count of Rented bikes based on Functioning Day
Rented Bike Count
Yes

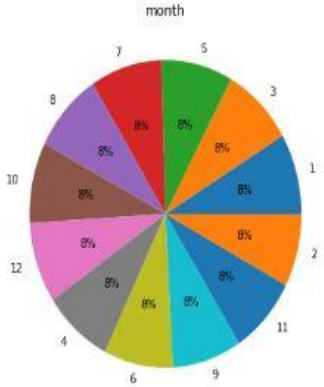

Rented Bike Count
Visibility (10m)
Temperature(°C)
Dew point temperature(°C)
Humidity(%)
Solar Radiation (MJ/m2)
Wind speed (m/s)
Rainfall(mm)
Snowfall (cm)

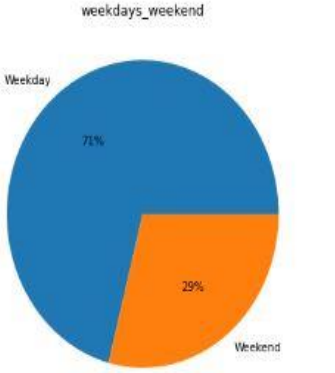

Hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

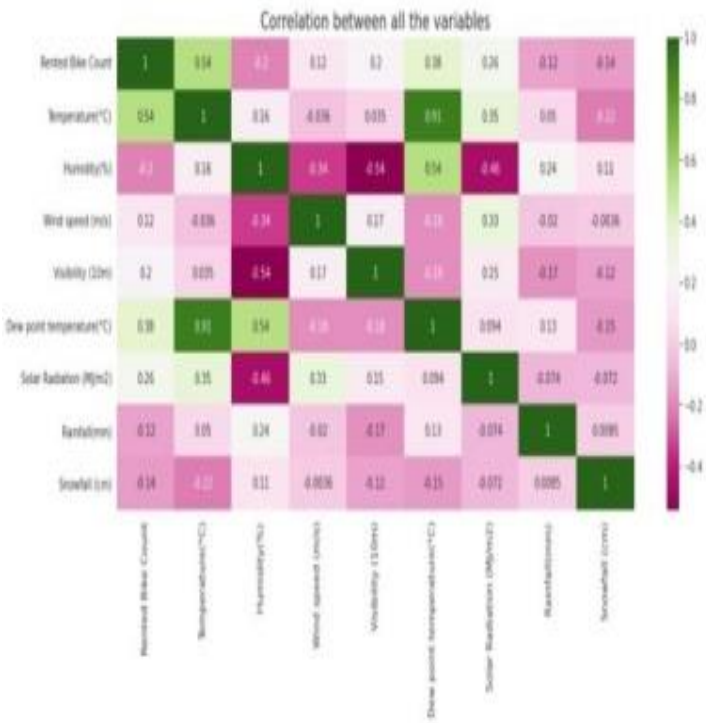

Seasons
Summer
Spring
Autumn
Winter
25%
25%
25%
25%


Holiday
No Holiday
Holiday
95%
5%


Functioning Day
Yes
No
97%
3%


month
1
2
3
4
5
6
7
8
9
10
11
12
8%


weekdays_weekend
Weekday
Weekend
71%
29%


Correlation between all the variables
Rented Bike Count
Temperature(°C)
Humidity(%)
Wind speed (m/s)
Visibility (10m)
Dew point temperature(°C)
Solar Radiation (MJ/m2)
Rainfall(mm)
Snowfall (cm)


Correlation between all the Variables
Rented Bike Count
Temperature(°C)
Humidity(%)
Wind speed (m/s)
Visibility (10m)
Solar Radiation (MJ/m2)
Rainfall(mm)
Snowfall (cm)

b. Drawing conclusions from the data:

Plotting necessary graphs which provides relevant information on our data like :

A. Most bikes have been rented in the summer season.

B. Least bike rent count is in the winter season.

C. Autumn and Spring seasons have almost equal amounts of bike rent count.

D. Most of the bikes have been rented in the year 2018.

E. Most of the bikes have been rented on working days.

F. Very few bikes have been rented in December which is winter season.

G. Most bikes have been rented in December in the year 2017 as we don't have data before that.

H. People tend to rent bikes when there is no or less rainfall.

I. People tend to rent bikes when the temperature is between -5 to 25 degrees.

J. People tend to rent bikes when the visibility is between 300 to 1700.

K. The rentals were more in the morning and evening times. This is because people not having personal vehicle, commuting to offices and schools tend to rent bikes.

c. Training the model:

A. Assigning the dependent and independent variables.

B. Splitting the model into train and test sets.

C. Transforming data using Min-Max Scaler.

D. Fitting linear regression on train set.

E. Getting the predicted dependent variable values from the model.

d. Evaluating metrics of our model:

A. Getting MSE, RMSE, R2-SCORE, ADJUSTED-R2 SCORE for different models used.

a. MSE - the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors.

b. RMSE - Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.

c. R2-SCORE - R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

d. ADJUSTED-R2 SCORE - Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when

predictor improves the model by less than expected.

B. Comparing the r2 score of all models used, to get the desired prediction.

5. Model Used:

1. Linear Regression:

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis

2. Lasso Regression:

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso solutions are quadratic programming problems, which are best solved with software (like Matlab). The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

3. Ridge Regression:

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values. The cost function for ridge regression:

Min(||Y – X(theta)||^2 + λ||theta||^2)

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients is reduced.

- It shrinks the parameters. Therefore, it is used to prevent multi-collinearity.
- It reduces the model complexity by coefficient shrinkage.

4. Decision Tree regression model:

Linear model trees combine linear models and decision trees to create a hybrid model that produces better predictions and leads to better insights than either model alone. A linear model tree is simply a decision tree with linear models at its nodes. This can be seen as a piecewise linear model with knots learned via a decision tree algorithm. LMTs can be used for regression problems (e.g. with linear regression models instead of population means) or classification problems (e.g. with logistic regression instead of population modes).

5. Random Forest regression model:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

6. Gradient Boosting:

"Boosting" in machine learning is a way of combining multiple simple models into a single composite model. This is also why boosting is known as an additive model, since simple models (also known as weak learners) are added one at a time, while keeping existing trees in the model unchanged. As we combine more and more simple models, the complete final model becomes a stronger predictor. The term "gradient" in "gradient boosting" comes from the fact that the algorithm uses gradient descent to minimize the loss. When gradient boost is used to predict a continuous value. We're using gradient boost for regression.

7. Extreme Gradient Boosting:

XGBoost is an implementation of Gradient Boosted decision trees. It was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on **regression, classification**, ranking, and user-defined prediction problems.

6. 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.

7. Conclusion:

- We have trained seven unique Machine Learning models using the training dataset, and its respective performance was improved through hyper parameter tuning.
- Thus, we have successfully built predictive models that can predict the demand for rental bikes based on different weather conditions and other.
- The XG Boost prediction model had the lowest RMSE.
- The final choice of model for deployment depends on the business need; if high accuracy in results is necessary, we can deploy XG Boost model.
- If the model interpretability is important to the stakeholders, we can choose deploy the decision tree model.