Capstone Project Submission

Team Member's Name, Email and Contribution:
Name : Ajit Sharad Mane Email : ajitmane36@gmail.com
Please paste the GitHub Repo link.
Github Link:- https://github.com/ajitmane36/Bike-Sharing-Demand-Prediction-ML-Regression.git
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)

This Seoul Bike Sharing Demand Prediction machine learning project aims to predict demand for bike-sharing based on historical data. The dataset used contains data from a bike-sharing system including the date, hour, weather conditions, temperature, and the number of bikes rented.

The aim of the project is to use this data to build a predictive model that can accurately estimate bike rental demand for any given hour. The model will be evaluated on its accuracy in predicting the number of bikes rented during a given hour.

The data set included Rental bike information. As part of the analysis, descriptive statistics were calculated for each variable, and visualizations were created to explore the relationships between various variables. To get insight from the dataset, we built a variety of charts, including a distplot, count plot, bar plot, line plot, heatmap, and boxplot.

Dataset for rented bikes with 8760 observations and 14 variables. These include Date, Rented Bike Count, Hour, Temperature (°C), Humidity (%), Wind speed (m/s, Visibility (10m), Dew point temperature, Solar Radiation, Rainfall (mm), Snowfall (cm), Seasons, Holiday, and Functioning Day. The parse_date() method was used to transform the feature date from its original object form to datetime64.

There are no duplicate values in the dataset. Additionally, there are no null or missing values in the dataset. 3 variables in the dataset Seasons, Holidays, and Functional Day are categorical variables, however, the remaining 11 variables are all of a numerical character. The Outliers found for Rented Bike Count, Wind speed (m/s), Solar Radiation (MJ/m2), Rainfall (mm), and Snowfall (cm). It is advisable to omit the columns labeled "Rainfall (mm)" and "Snowfall (cm)" because of their flat interquartile ranges. We eliminated outliers from Rented Bike Count, Wind speed (m/s), and Solar Radiation (MJ/m2) by using the interquartile range.

We delete the original Date variable from the dataset and replace it with new variables that include the day, month, and year that we extracted from the Date variable. The dataset is now prepared for analysis.

After doing univariate, bivariate, and multivariate analyses, we discovered insights which are following :

- Customers favor rental motorcycles equally in all seasons.
- When there are no holidays, customers choose to rent motorcycles. Customers hardly ever use the bikes they rent while traveling on holiday.
- Nearly all consumers preferred to rent bikes during functional hours.
- Bicycle rentals are popular all month long.
- Renting bicycles was not very popular in 2017, but it increased by 83.02 percent in 2018.
- At night, customers do not prefer to use rented bikes.
- Customers do not prefer rented bikes in the mornings 4 and 5, but from 7, 8, and 9, the use of rented bikes increases, possibly due to working people going to the office, and it is the same in the evenings 5, 6, and 7, because people are travelling from the office to home. Overall, the rented bike was the most frequently used during office in and out times.
- Customers mostly use rented bikes for transportation in the evening.
- Customers who travel most commonly use rented bikes in the morning at 8 a.m. and in the evening at 6 p.m.
- When the humidity level is between 10% and 18%, people prefer to rent bikes.

- wind speed is between 2 m/s and 3.5 m/s, people consistently use rented bikes, and it is at its peak when wind speed is normal, which is 3.2 m/s.
- Renting a bike is the best option for customers in dew point temperatures ranging from 12°C to 18°C. The use of a rented bike increases with increasing dew point temperatures, but it still reaches normal dew point temperatures.
- According to the graph, solar radiation has no effect on customer use of rented bikes.
- When it's not raining, people prefer rental bikes the most.
- When there is no snowfall, most people opt to rent bikes. However, the majority of customers prefer to rent bikes when it snows up to 4 cm.
- In the first 10 days of the month, most rented bikes are used by customers. Customers consistently use rented bikes in the last 15 days of the month.
- In June, most rented bikes are used through the year, followed by October. Customers' use of rent bikes is at its peak from April to September.
- The count of rented bikes on that day is unaffected by the day's visibility, but when visibility exceeds 1750, use of rented bikes increases more than usual.
- During the summer and autumn seasons, most people rent bikes. During the winter, fewer people choose to rent bikes.
- Even when there is no holiday other than a holiday, people rent bikes. The use of rented bikes on holidays is lower than on non-holiday days.
- Almost every rented bike is used during its functional hours.
- The use of rented bikes increased by three times in 2018 compared to 2017.

We also verified the presumptions of the regression machine learning model. We restore the distribution of the features Rented Bike Count, Wind speed (m/s), Solar Radiation (MJ/m2), Visibility (10 m), Rainfall (mm), and Snowfall (cm) to normal using log and square root transformation. We also check to see if there is a linear relationship between the independent and dependent variables. Furthermore, we use a heatmap and the variance inflation factor to examine multicollinearity in independent variables (VIF). Due to their high VIFs, we eliminated the variables year, dew point temperature (°C), and humidity (%). Once categorical variables like seasons, holidays, and working days were encoded, our dataset was ready to be used with a machine-learning model.

To make it easier for a model to learn and understand the problem, we divided the dataset into dependent and independent features and scaled them to the same length. Then we implement machine learning models like linear regression, Lasso (L1), Ridge (L2), ElasticNet, Decision Tree regressors, Random Forest, and XGBoost regression. We obtained a high accuracy of 0.9026 and a low root mean squared error of 3.80 from the XGBoost model, become final optimal model for prediction.

The Functioning Days, Rainfall (mm), and Seasons variables have a significant influence on the XBoost model, according to our plot of feature significance. Additionally, we explain our XGBoost model using SHAP.

Some difficulties faced during the implementation of the model include data accuracy and data integration. Additionally, the model may require extensive tuning and testing to ensure the accuracy and reliability of the predictions. Lastly, the model may need to be regularly updated in order to keep up with changes in the environment or bike-rental patterns.