Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Team Member's Name: Sanchit Misra
Email: sanchit.misra.a13@gmail.com
Contribution:

1. **Imported dataset and visualize first glimpse of data.**
2. **Data Wrangling**
   - **Bike Sharing Demand Prediction**
   - **Null Values, Duplicates, Unique Values, Date Time feature**
   - **Encoding the features**
3. **Descriptive analysis of numerical features in dataset.**
4. **Numbers when the facility for sharing bicycle rentals is not functioning day**
5. **Exploratory Data Analysis**
6. **Identified Multicollinearity**
7. **Outlier handling**
8. **Feature Engineering**
9. **Model Building**
   - **Divided data into dependent and independent variables.**
   - **Train-Test split**
   - **Standard Scaling of variables**
   - **Implemented various models**
   - **Evaluation metric for model performance**
   - **Plotted feature importance graph.**
10. **Model comparison**
    - **Compared hyper tuned models.**
    - **Compared plain vanilla models.**

Team Member's Name: Tushar Hande
Email: handetushar3@gmail.com
Contribution:

1. **Imported libraries for data cleaning and visualization.**
2. **Mounted the drive and uploaded the dataset.**
3. **Checked shape and size of the data and count of the variables.**
4. **Checked null values counts and duplicates.**
5. **Collected information about each variable i.e. type, count**
6. **Statistical analysis- (mean, median, quartiles and the distribution of the data).**
7. **Checked basic assumptions for regression models.**
8. **Outlier handling.**
9. **Multicollinearity detection and reduction of multicollinearity.**
10. **Basic modeling**

Team Member's Name: Mohit Jain
Email: jainmohit02.mj@gmail.com
Contribution: -

1. **Imported all the libraries for data exploration, Sorting, Cleaning and Visualization and Modelling.**
2. **Imported and mounted data set require for analysis from google drive.**

3. **Exploring data set such as number of columns and row with feature name and what is data type of each feature using python libraries like Info(), shape, describe(),Head(),Tail().**
4. **Checked Null Values, Duplicates, Unique Values, Date Time feature.**
5. **Exploratory Data Analysis: - Performed Uni variant analysis and bi variant analysis to check dependency of each independent variable with dependent variable and checked multicollinearity by heat MAP and dropped feature that is mostly affected.**
6. **Identified and Handled outlier with box plot and IQR method and retain all the data as provided.**
7. **Divided data into dependent and independent variables.**
8. **Performed Train-Test split on clean data.**
9. **Implemented various models and Checked Evaluation metric for model performance to identify best model as per the business requirement.**
10. **With help of all group member prepared presentation.**
11. **Helped group member to prepare technical documentations.**
12. **Made Conclusion on analysis and model evaluation metric.**

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/jainmohit02/Bike-Sharing-Demand-Prediction-Solution

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

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. In our project Seoul bike sharing demand prediction used different regression models to predict the count of the bike using the given variables with a good accuracy.

Initially we imported the data and did basic exploratory data analysis (EDA) steps to understand data very well. We checked the relation of the independent variable with dependent variable using bivariate analysis.

Outliers affect's the accuracy of the regression model we identified the outliers using boxplot's and replaced them using IQR (Inner quartile range) method.

Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model. We identified the multicollinearity using heatmap and VIF (Variance inflation factor). We dropped one variable to reduce multicollinearity.

To handle categorical features, we used label encoding and One Hot encoding. For label encoding we used binary mapping.

We started modeling from basic linear regression. In linear regression we used ridge, lasso and elastic net regression models. R2 score for these regression models was between 0.60 – 0.65. To get better accuracy we moved to tree based algorithms. In that we used decision tree regression, random forest regressor, XG Boost and Cat Boost algorithms. R2 score for these algorithms was between 0.80-0.90.

To get better performance from the models we did hyperparameter tuning. In linear

regression algorithms we used GridSearchCV. For tree-based algorithms we controlled the depth of a tree.

To calculate the model accuracy, we used serval metrics like R2 score, root mean squared error (RMSE), mean absolute percentage error (MAPE).

Finally, we reached to the conclusion:

- The results clearly suggest that Cat Boost is the best model for predicting bike sh aring demand, as the performance measure (MAPE, RMSE) is lower and (R2, adjus ted_R2) value is greater for XG Boost and Cat Boost.
- We may say that the XG Boost model helps us to anticipate the number of rental bikes more accurately.
- Temperature is regarded as the most important variable for the linear regression model.
- The Cat Boosting algorithm's hour feature is considered to be its most significant factor.
- We discovered through EDA that people ride bikes more frequently in the summ er season and also when the winds are strong.
- The months of May, June, and July can be considered to have a larger demand for rented bikes.
- There is a considerable demand for the rented bikes during the peak office hours. Therefore, base stations can be setup near office buildings.
- In order to reduce public waiting times, the number of bikes should be raised dur ing the summer. As a result, they can easily rent bikes whenever they need to or a t any time.