Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
• Karan Rawat (rawatkarankr99@gmail.com) • Data Reading • Exploratory Data Analysis • Metrics Selection • Modelling • Model Performance Selection • Final Model Other than this I prepared summary, PPT report and Technical document for this study.
Please paste the GitHub Repo link.
Github Link:- https://github.com/rawatkarankr99/Capstone-Project-2.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) **Summary** ● The job is to predict the number of rented bike counts each hour. ● The dataset contains a total 14 features in which "Rented Bike Count" is response variable and others are predictor variables. ● The dataset contains no missing values. ● The dataset contains a "Date" feature which has an object dtype, we changed the dtype of this feature to correct one. ● The "Date" feature is splitted to "Year", "Month", and "Day" features for better understanding. ● The "Hour" feature has dtype as int , we changed the dtype of this feature to correct one. ● After modification the dataset has a total eight numerical features and seven categorical features. ● The feature "Visibility" is negatively skewed and the features "Wind Speed", "Solar Radiation" ,"Rainfall" and "Snowfall" are positively skewed. The different

transformations used to remove skewness but only for the feature "Wind Speed", skewness can be removed using square root transformation.

- The features "Wind Speed","Solar Radiation","Rainfall","Snowfall" have outliers.
- The features "Dew point temperature" and "Temperature" are highly correlated. So we used the combination of both to remove correlated features from the dataset.
- Each numerical feature is less correlated with the dependent variable and follows a non-linear relationship.
- For categorical features , the demand is seen similar for each values of he features "Hour","Season","Month","Day" and the demand is seen more for the features "Holidays" as yes , "Functioning Day" as yes and the "Year" value as 2018.
- The response variable is positively skewed, so to remove the skewness different transformations are used. The most effective is square root transformation.
- The numerical features are scaled and categorical features are one hot encoded before passing to the model.
- The key metrics to be noticed are R2 and adjusted R2.
- Both Linear and Non Linear Models are used in the task.
- The Linear Regression, tuned Lasso,Ridge and ELasticNet gives R2 score as 0.779 and adjusted R2 score as 0.768.
- The Decision Tree Regressor gives R2 score as 0.835 and adjusted R2 score as 0.827.
- The Random Forest, GradientBoosting, XGBoost Regressors give around R2 as score 0.89 and adjusted R2 score as 0.88.
- The highest scores can be seen for ensemble models.
- Taking model explainability and feature importance into account the Decision Tree is selected as final model.The Decision Tree gives R2 score as 0.835.
- The "Season" feature as winter is the most important factor for the predictions. The "functioning_day" as yes , and Humidity are the other most important factors. The "Solar Radiation","Temperature" ,"Hour" as a whole feature and "RainFall"are the next important factors. So these factors are the most important for the predictions, therefore we need to focus on them to improve the business model.

Challenges Faced

- The dataset contains skewed predictor variables.
- The response variable was also skewed.
- Choosing the right metric was a challenge.