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 Role:- ❖ Kirtesh Verma(kirteshverma12345@gmail.com) Contribution: o Data understanding o Handling null or missing values o Performing EDA o Removing Outliers o Linear Regression Model o Random Forest Model o Hyperparameter Tuning on random forest ❖ Pravin Bejjo(praveen.bejo.pb@gmail.com) Contribution: o Data understanding o Data visualization o Multivariate analysis o Ridge Regression Model o Gradient Boosting Model o Hyperparameter Tuning on Gradient Boosting ❖ Sahil Pardeshi(8623879021.sp@gmail.com) Contribution: o Data understanding o Data visualization o Multivariate analysis o Lasso Regression Model o Decision Tree Model o Hyperparameter Tuning on decision tree
Please paste the GitHub Repo link.
GitHub Link:- https://github.com/praveenbejo95/Bike-sharing-demand-prediction-ML-regression-project-

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

The contents of the data came from a city called Seoul. A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short-term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system. The data had variables such as date, hour, temperature, humidity, wind-speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day and rented bike count.

The problem statement was to build a machine learning model that could predict the rented bikes count required for an hour, given other variables. The first step in the exercise involved exploratory data analysis where we tried to dig insights from the data in hand. It included univariate and multivariate analysis in which we identified certain trends, relationships, correlation and found out the features that had some impact on our dependent variable. The second step was to clean the data and perform modifications. We checked for missing values and outliers and removed irrelevant features. We also create dummy variables for categorical features. The third step was to try various machine learning algorithms on our split and standardized data. We tried different algorithms namely; Linear regression, Lasso and Ridge Model, Decision Tree, Random Forest and Gradient Boosting. We did hyperparameter tuning and evaluated the performance of each model using various metrics. The best performance was given by the Gradient boosting and Random Forest model where the R2_score for training and test set was 0.87 and 0.83 for Random Forest and 0.92 and 0.88 for Gradient Boosting respectively.

The most important features who had a major impact on the model predictions were; hour, temperature, Humidity, solar-radiation, and Winter. Demand for bikes got higher when the temperature and hour values were more. Demand was high for low values of Humidity and solar radiation. Demand was high during springs and summer and very low during winters.

The model performed well in this case but as the data is time dependent, values of temperature, wind-speed, solar radiation etc. will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time