Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Rahul Shah: shahrahool010@gmail.com 1. Exploratory data analysis – univariate and multivariate analysis. 2. Data Wrangling – checking missing values, outliers, features modification. 3. Fitting Models – splitting the data, applying algorithms, hyper-parameter tuning, model evaluating, model explanation. 4. Presentation, Technical documentation.
Please paste the GitHub Repo link.
Github Link:- https://github.com/rahool010/Bike-Sharing-Demand-Prediction
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. It is the capital city of South Korea and has a population of around 9.7 million people. It has humid continental climate influenced by monsoons. Seoul City has a rental bike sharing program. It is a method of renting bicycles; bike return is automated via a network of kiosk located across the city. The dataset contains the variables such as date, hour, temperature, humidity, windspeed, 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. Imported necessary libraries and dataset for the experiment. Did a quick overview of the dataset. 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 who had some impact on our dependent variable. The second step was to clean the data and perform modifications. Here, we checked for missing values and outliersand removed irrelevant features. We transformed our dependent variable by applying standardization, extracted 'Day', 'Month' and 'Year' from the 'Date' column and also encoded the categorical variable. The third step was to try various machine learning algorithms on our splitted and standardized data. We tried 4 different algorithms namely; Linear regression, Decision Tree, Random Forest and XGBoost. We also did hyper-parameter tuning and evaluated the performance of each model using various metrics. The best performance was givenby the XGBoost model where the R2_score for training and test set was 0.88 and 0.84 respectively. Next we implemented shap technique to understand the working of our model. The most important features that had a major impact on the model predictions were; temperature, hour, humidity, seasons_winter and solar radiation. Demand for bikes got higher when the temperature and hour values were more. After a long exercise we concluded that ensemble learning approach such as XGBoost improves the model performance considerably.