Capstone Project 2 - Bike Sharing Demand Prediction

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Individual Project. Name : Shubham Chandrakar Email-Id: chandrakar.shubham17@gmail.com Contributions: 1) Data Wrangling a) Checking duplicate row b) Checking Null/Nan Values c) Outlier Removal d) Feature engineering. e) Dependent variable distribution check. 2) EDA on Project a) Univariate analysis b) Bivariate analysis c) Multivariate analysis d) Correlation analysis e) VIF analysis f) Categorical Encoding 3) Building ML model a) Linear regression model b) Lasso regression model c) Lasso regression model with Cross Validation d) Random forest regression model e) Random forest regression model with cross validation f) XGBoost regression model g) XGBoost regression model with cross validation
Please paste the GitHub Repo link.
Github Link :- https://github.com/chandrakar-shubham/Bike-Sharing-Demand-Prediction--ML-regression-Project
Google drive link is posted below.

Google drive link :-
https://drive.google.com/drive/folders/1glmti7DyRdW-PMB_C5wESwZuaCJfDTt_?usp=sharing

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

In Seoul (a city in South Africa), rental bikes are currently 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. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. In this EDA project, following features are provided i.e. date, rented bike count, hour, temperature, humidity, wind speed, visibility, dew point temperature, solar radiation, rainfall, seasons, holiday , functional day. Dependent feature here is rented bike count.

As the first step, data wrangling is performed over raw data in the given database. After that we divided the project into various parts such as Univariate analysis, Bivariate analysis, Multivariate analysis, Correlation in data, VIF analysis, Building ML model.

In Univariate analysis, individual analysis is performed over all features to analyze the distribution and trend of numerical and categorical variables. Various types of chart plot are used for comprehending the enumerative properties as well as a descriptive summary of the particular data variable.

In Bivariate analysis, It is performed to compare two features. This analysis is used to find the relationship between the two variables. Using this analysis we can also find relationships between dependent features and other features. Various plots like box plot, line plot, scatter plot, count plot etc. can be used to identify relationship between features

In Multivariate analysis, It is performed to compare three or more relevant features to establish a relationship between them.

Correlation analysis is used to find the pairwise correlation of all columns. This analysis is used to do feature engineering efficiently.

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables. Feature engineering is used to maintain VIF value of any feature less than 10.

After that, multiple supervised machine learning models to predict dependent variables with accuracy. The multiple ML models built were Linear regression model, Lasso regression with/or without Grid Search CV, Random forest regression model with/or without Grid Search CV, XGBoost regression model with/or without Grid Search CV. It is found that after analyzing various models, random forest after hyper parameter tuning performed best with R2 score of 94.33% accuracy and Adjusted R2 score of 94.14%.

Conclusion : It is found out that demand for bikes rises with rise in temperature. At

night demand for rental bike is most, In summer season the demand for rental bike is most, In monthly period it is seen that rental bike demand is low on January, February and December and high between may to august, It can be seen that bike demand rises after 5 AM and peaks at 8 AM, then again rises after 2 PM and peaks at 5PM then demand remain significantly above average demand 6PM and 11PM.That means in this 11 hours of a day bike demand is most. XG boost regression model can predict rental bike demand with 94.14% accuracy.