Capstone Project Submission

Team Member's Name, Email and Contribution:

Name: Sri harish A
Email: sriharishanand@gmail.com
Contribution:

- Data cleaning
- Box plot to detect outliers
- Outlier Treatment
- Count plot
- Dependent variable count plot
- Variance Inflation Factor
- Linear Regression
- Lasso Regression
- Decision Tree
- Cross Validation and Hyperparameter Tuning
- Gradient Boosting Regressor
- Feature Importance from Gradient Boosting Regressor
- Model Performance Analysis and Selection

Name: Sethupathy M
Email: sethuqr@gmail.com
Contribution:

- Format conversion to access the date, month, year features
- Distribution plot
- Mean Median plot
- Scatter and correlation plot
- Pie Chart
- Multicollinearity
- OneHotEncoding
- Feature Engineering
- Polynomial Regression
- Ridge Regression
- Random Forest Regressor
- Feature Importance from Random Forest Regressor
- Model Performance Visualization

Github Link:- https://github.com/Sriharish19/Bike-Demand-Prediction/tree/main

Supervised Machine Learning – Bike Sharing Demand Prediction

Summary

This project contains a real-world data record of Seoul Rented Bike Demand from 2017 to 2018, including details such as weather, holiday, working day, date, and so on. There are 14 variables that describe 8800 observations. Each observation represents Rented Bike Demand per hour.

The goal of this research is to experiment with machine learning models for anticipating bike sharing demand in Seoul by providing the hour, day, and weather information. Data exploration, feature engineering, choosing an appropriate scoring metric, cross algorithms, cross validation, tuning the algorithms, feature significance analysis, residual analysis, and performance evaluation are all part of the research.

We started by importing the data set from Google Drive into our colab notebook, which was in.csv format, and executing simple operations like shape, description and info, is null, and so on to get a basic idea of the dataset's contents.

Data cleaning was the next process to remove unwanted variables and values from our dataset and get rid of any irregularities in it. Since these values disproportionately skew the data and hence adversely affect the results.

Our dataset had several outliers, which might have an impact on the model's efficacy. Outliers were found using the bar plot, and all outliers were imputed with the median using the Z- score approach.

Exploratory Data Analysis was performed on the dataset to understand the relationship between the target variable and independent variables. The hypothesis was generated using the EDA, and the target variable is impacted by hourly trend, temperature, rainfall, snowfall, humidity, holidays, working day, visibility, wind speed, and solar radiation.

One Hot Encoding transformation was performed for machine learning algorithms to access information such as Seasons, Working Day, and Holiday.

The values were standardized to center them on the mean with a unit standard deviation. Splitting arrays into train and test subsets was accomplished using train test split.

A Linear regression model was fitted to the modified dataset. The test set's r2 score was determined to be 67 percent after model validation.

The r2 score of the test set was discovered to be 82.5 percent employing the Polynomial regression model. Even with Lasso and Ridge Regressors, the needed r2 score was not obtained. As a result, we intuitively assumed that Decision Tree Regressor would be the best fit model for the given dataset.

Although the Decision Tree Regressor is robust to outliers and multicollinearity, it is prone to over fitting. We obtained a 100% r2 score for the train dataset and a 75% r2 score for the test dataset. As a result, the Ensemble of Decision Trees was chosen. Random Forest Regressor generates n trees and takes the mean of those trees into consideration to reduce over fitting. We received a r2 score of 90%.

After completing grid search CV and parameter tuning, the Gradient Boosting regressor proved to be the best model for the data set, with the best r2 score of 94 percent.