**Capstone Project Submission**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Name: Sethupathy M  Email: [sethuqr@gmail.com](mailto: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 * Gradient Boosting Regressor * Feature Importance from Gradient Boosting Regressor * Model Performance Visualization   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 * Random Forest Regressor * Feature Importance from Random Forest Regressor * Model Performance Analysis and Selection |
|  |
| Github Link:- <https://github.com/SethupathyM/Supervised-ML---Regression> |
| **Supervised Machine Learning – Bike Sharing Demand Prediction**  **Summary** |
| This project contains the real-world data record of Rented Bike Demand of Seoul containing details like weather, holiday, functioning day, date etc. from 2017 to 2018. There are a total of 14 variables describing 8800 observations. Each observation represents a Rented Bike Demand per hour.  The purpose of this project is to try a machine learning approach for predicting bike sharing demand in Seoul by given the hour, day, and information about the weather. The research contains: Data exploration, feature engineering, choosing appropriate scoring metric, cross algorithms, cross validation, tuning the algorithms, analysis of feature importance, analysis of residuals and performance evaluation.  We began by loading the data set from the google drive into our colab notebook, which was in.csv format, and performing basic operations such as shape, describe and info, is null, and so on to gain a basic understanding of the dataset's contents.  [Data cleaning](https://www.simplilearn.com/data-cleaning-why-and-how-to-get-started-article) 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 many outliers which could affect the effectiveness of the model. So, by using bar plot outliers had been detected and by using Z- score technique all the outliers had been imputed with median.  Exploratory Data Analysis was performed on the dataset to understand the relationship between the target variable and independent variables.  From the EDA the hypothesis generation was done which was the Target variable is influenced by Hourly trend, Temperature, Rainfall, Snowfall, Humidity, Holidays, Functioning day, Visibility, Windspeed, Solar radiation.  OneHotEncoding transformation has been done for machine learning algorithm to access the features like Seasons, Functioning day, Holiday.  Standardization transformation was done to centers the values around mean with unit standard deviation.  Train\_test\_split was used to Split arrays or matrices into random train and test subsets.  The transformed dataset was fitted into Linear regression model. After the model validation the r2\_score of test set was found to be 67%.  By using Polynomial regression model, the r2\_score of test set was found to be 82.5%. And even after using Lasso and Ridge Regressors the desired r2\_score was not achieved. So, for the given dataset we intuitively thought Decision Tree Regressor would be the best fit model.  Decision Tree Regressor is robust to outliers and multicollinearity but it is prone to overfitting. We got 100% r2\_score for train dataset and 75% r2\_score for train dataset.  So, Ensemble of Decision Tree was chosen. In RandomForestRegressor n number of tress are generated and mean of those trees are taken into account so overfitting is reduced. Desired r2\_score of 88% is achieved by using RandomForestRegressor. |