# Capstone Project Submission

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member's Name, Email and Contribution:** |
|---|
| 1) Kajal Dhun<br>Email ID: kajaldhun@gmail.com<br><br>2) Navinkumar Sambari<br>Email ID: nnnghvs143@gmail.com<br><br>3) Tanu Rajput<br>Email ID: tanuurajput689@gmail.com<br><br>**Contribution:**<br><br>**The entire project was performed as a Team. We daily set meetings and did projects and contributed equally.** |
| **Please paste the GitHub Repo link.** |
| Github Link: https://github.com/TanuRajput110/Seol-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)** |

This Seoul bike sharing demand prediction dataset contains 14 features and 8760 observations of a complete year, i.e., from 1-12-2017 to 31-11-2018. We have a regression problem because our target is the number of rented bikes per hour. This Regression analysis helps to achieve the goal of the company Seoul Bike in providing the city with a stable supply of rental bikes. It becomes a major concern to keep users satisfied. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

After loading the dataset, firstly, we performed data preprocessing, we did some data exploration by checking types, missing values, duplicate values and data description. In this dataset there are neither null values nor duplicate values. We also changed the date type to Date Time which was initially a str object. From the date, we also created three columns with the day of the week and the month and the year corresponding. We also changed the datatype of the hour feature from int to object.

After that, we performed Exploratory Data Analysis to obtain the insights of our dependent variable Rented Bike Counts. Various graphs were constructed comparing the Rented Bike Count column with other columns. List of insights were obtained. We observed that bike rental count is higher during weekdays than weekend days. The rental bike counts are at its peak at 8 AM in the morning and 6pm in the evening. Highest rental bike count is during Autumn and summer seasons and the lowest in winter season.

Next step was feature engineering, in which we detected and took care of multicollinearity. We used the square root method to normalize the target variable. For scaling independent features, we used Yeo Johnson transformation technique. Lastly, we used Pandas dummies for encoding the categorical features.

Now the modeling part begins, here we used 8 regression algorithms, viz., Linear Regression, Ridge, Lasso, Polynomial, Decision Tree, Random Forest Regressor, Gradient Boosting Regressor and Extra Trees Regressor. So, after fitting the models and evaluating metrics (MSE, RMSE, R Square, Adjusted R Square) and also hyperparameter tuning we came to the result and conclusion. We got the Adjusted R2 among all the models**,** Extra Trees Regressor gives the highest Score where Adjusted R2 score is 0.908699 and Training score is 0.987167. Therefore, this model is the best for predicting the bike rental count on an hourly basis.