# Capstone Project Submission

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

**Team Member's Name, Email and Contribution:**

Team Name : Team Datavengers

Team Members:

1. Kunal Mahadik

Email id : kunalmahadik0811@gmail.com

Contribution:

1. Data Wrangling
2. Data Preparation
3. Data Cleaning
4. Data Preprocessing
5. Implementation of Linear Regression model
6. Implementation of Lasso and Ridge Regression model

2. Aashruti Agarwal

Email id : aaashruti@gmail.com

Contribution:

1. Data Wrangling
2. Data Preparation
3. Data Cleaning
4. Data Preprocessing
5. Implementation of Decision Tree model
6. Implementation of Random Forest Regressor

3. Raneev K

Email id : raneevk36@gmail.com

Contribution:

1. Data Wrangling
2. Data Preparation
3. Data Cleaning
4. Data Preprocessing
5. Implementation of Gradient Boost Regressor
6. Implementation of Gradient Boost Regressor with Grid Search CV

**Please paste the GitHub Repo link.**

https://github.com/AashrutiA/ML_Regression_analysis

Github Link :- https://github.com/AashrutiA/ML_Regression_analysis

Drive link : https://drive.google.com/drive/folders/1eA-AfwjgFWI1ieT8ltS9-GU0S5H1AQgz?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)**

Rental Bike are slowly getting its acclaim in the urban cities for better mobility comfort to the public. In order to maintain the smooth operation, availability and accessibility of the rental bike with lesser waiting time period is the most crucial concern. Thus, it is very necessary to understand the features driving the demand of rental bikes in order to build a model which fulfill the rental bike demand each hour to ensure the stable supply of rental bikes.

In order to do so, we will first deep dive into the dataset to understand the correlation of different feature with bike rented and pull-out insight out of that. Followed by that, we will implement different machine learning algorithm to build a model which could predict the hourly bike sharing demand.

We have come up with following approach to solve the problem statement:

1. Firstly, we will deep dive into the dataset after loading it to understand each of the features. After carefully inspecting the features, we framed out hypothesis questions for EDA such as most active hours, season, month, weekday for rented bike and many more to understand the patterns and relation between different variables.
2. **Data Preprocessing & feature engineering** of the dataset involving removing duplicates, identifying and handling missing values, outliers, dropping the irrelevant columns to enhance the quality of our dataset.
3. **Analyzing and Visualization** of the dataset utilizing relevant plots, charts and correlation heatmap to infer the insights. For example, most active season was summer, bike were mostly rented at morning 7 and evening 6pm, people rented more bike on weekdays and so forth.
4. **Model Building** involving implementation of different regression machine learning algorithms like linear regression, regularization, tree-based regression models to build the model followed by training the dataset, hyper-tuning the parameter via GridSearchCV, and predict the test result.
5. **Comparison of all models** based on r2 score and find the best performing model to predict the rental bike sharing demand.

The dataset was pretty clean without any null values, duplicates. However, the dataset has many outliers, which were treated by applying square root transformation. Feature engineering was performed where some features were dropped while some new features were added. Out of all the built model, random forest and gradient boosting hypertuned via GridSearchCV model performed really well for this dataset with r2 score of 98% and 94% respectively.