# Capstone Project Submission

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

| | |
|---|---|
| **Team Member's Name** | **Email** |
| Ajinkya Shingote | shingoteajinkya65@gmail.com |

**Contribution:**

**Ajinkya Shingote:**

- Data Wrangling
- Loading and Preprocessing
- Structuring data
- Enriching data
- Data Mining
- Data Analysis
- Building Machine Learning Algorithms
- Visualizations
- Observations
- Summarization
- Conclusions
- Technical Document
- Power Point Presentation

**Please paste the GitHub Repo link.**

Github Link: https://github.com/Ajinkya6597/Bike_sharing_demand_prediction

**Project Name:** Bike Sharing Demand Prediction.

**Problem Statement:**

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information. Based on the given data we have to build a machine learning model which will be helping us to predict the number of bikes that must be made available by predicting the demand for bikes rented per day..

**Approach:**

I first performed an EDA on given dataset.
Building machine learning algorithm.
After applying linear regression model, I got R2 score of 0.779 for training data and R2 score of 0.774 for test data, which signifies that model is optimally fit on both training and test data i.e. no overfitting is seen. • Therefore, for even better fit, I applied polynomial regression model with degree = 2, we got R2 score of 0.933 for training data and 0.90 for test data • I also tried Tree based classifiers for our data, I applied Decision Tree Regressor, since decision tree is prone to overfit, I gave certain parameters like maximum depth of the tree, maximum leaf nodes etc, with that I got R2 score of 0.835 for training data and 0.803 for test data which is less than polynomial regression. • To get better accuracy on tree based model, I applied Random forest with n_estimator as 180 and with maximum depth as 13, with that I got R2 score of 0.888 for training data and 0.875 for test data. • Finally, I applied Gradient boost with parameters selected after grid search which resulted in highest R2 score of 0.958 for training data and 0.933 for test data with very less mean squared error of 6 and 10 in training as well as in test data. • Also we can see from SHAP summary that high Hour_18 value increasing prediction. Also we can see low Snowfall value increasing prediction and it is a common phenomenon in all the models. • Lastly, In bar graph from SHAP we can see Winter has the highest feature value while Wind Speed has the Lowest shap value.We can conclude that Hour_21,Hour_8 and Wind Speed is not contributing in Decision Tree,Random Forest and Gradient Boost in model prediction.

**Conclusion:**

- In the summer season the highest number of bikes were rented as compared to other seasons
- Higher number of Bikes were rented on a weekday as compared to weekends
- Lowest number of bikes were rented in January and after gradually increasing, the highest number of bikes were rented in May
- Bike Rental is at its peak at 6 PM
- Bikes are rented most on a clear day, i.e. where there is no snowfall or rainfall
- In Hour vs Rented Bike Count we can see that during 18:00 Hrs(i.e 6:00 PM) highest number of bike was rented as
- compared to 5:00 Hrs(i.e 5:00 AM). This means people tends to rent less bikes at early morning.
- In Rainfall vs Rented Bike Count and similarly with Snowfall vs Rented Bike Count we can see that people tend to rent highest number of bikes during 0.00mm of Rainfall or no rainfall and 0.00cm of snowfall or no snowfall as compared to when there is actually rainfall or snowfall. In other words people rent less bikes or no bikes with the increase of rainfall or snowfall.
- In month vs Rented Bike Count we can see that people tends to rent more bike in 6 or june month as compared to less bike during dec or january.From this we can assume that people tends to rent more bikes in summer as compared to winter.

- In weekend vs Rented Bike count we can see that people tends to rent more bike during weekdays as compared to weekends.
- In Average Bike Rented vs Hour we can clearly see that at 6:00 PM average number of bike rented by the people was 1550. While at 00.00 or at midnight average bike rented was lowest with just around 550 bikes.
- In Average Bike Rented vs Month we can clearly see that Average Bike rented in July was highest around 1250 and
- Average Bike Rented during month of February was the Lowest with just 200 average bike.
- After applying linear regression model, we got R2 score of 0.779 for training data and R2 score of 0.774 for test data, which signifies that model is optimally fit on both training and test data i.e. no overfitting is seen.
- Therefore, for even better fit, we applied polynomial regression model with degree = 2, we got R2 score of 0.933 for training
- data and 0.90 for test data
- We also tried Tree based classifiers for our data, we applied Decision Tree Regressor, since decision tree is prone to overfit, we gave certain parameters like maximum depth of the tree, maximum leaf nodes etc, with that we we got R2 score of 0.835 for training data and 0.803 for test data which is less than polynomial regression.
- To get better accuracy on tree based model, we applied Random forest with n_estimator as 180 and with maximum depth as
- 13, with that we got R2 score of 0.888 for training data and 0.875 for test data.
- Finally, we applied Gradient boost with parameters selected after grid search which resulted in highest R2 score of 0.958 for training data and 0.933 for test data with very less mean squared error of 6 and 10 in training as well as in test data.
- Also we can see from SHAP summary that high **Hour_18** value increasing prediction. Also we can see low **Snowfall** value
- increasing prediction and it is a common phenomenon in all the models.
- Lastly, In bar graph from SHAP we can see **Winter** has the highest feature value while **Wind Speed** has the Lowest shap value.We can conclude that Hour_21,Hour_8 and Wind Speed is not contributing in Decision Tree,Random Forest and Gradient Boost in model prediction