Capstone Project - 2

Bike Sharing Demand Prediction

Submitted by

Shaloy Elshan Lewis

Data science trainee, Almabetter

Agenda

- Problem Statement
- Data Summary
- Feature Engineering
- Exploratory Data Analysis (EDA)
- Modelling Approach
- Predictive Modelling
- Model comparison
- XG boost model explanations
- Challenges faced and Conclusions

Problem Statement

- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort
- It is important to make the rental bike available and accessible to the public at the right time as it lessens the **waiting time**, eventually, providing the city with a **stable supply** of rental bikes
- The goal of this project is to build a ML model that is able to predict the demand of rental bikes in the city of Seoul.

Data Summary

- Date
- Rented Bike count
- Hour - Hour of the day
- Temperature - Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature – Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons
- Holiday
- Functional Day
- **Day of week**
- **Month**
- **Weekend**

Feature Engineering

- **Td = T - ((100 - RH)/5)**

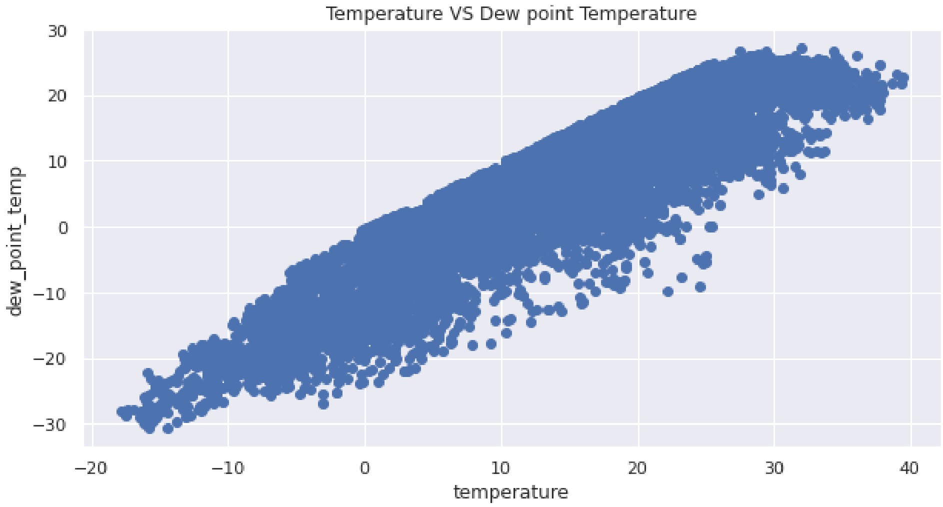
➔ Td = dew point temperature

➔ T = Temperature

➔ RH = Relative humidity (%)

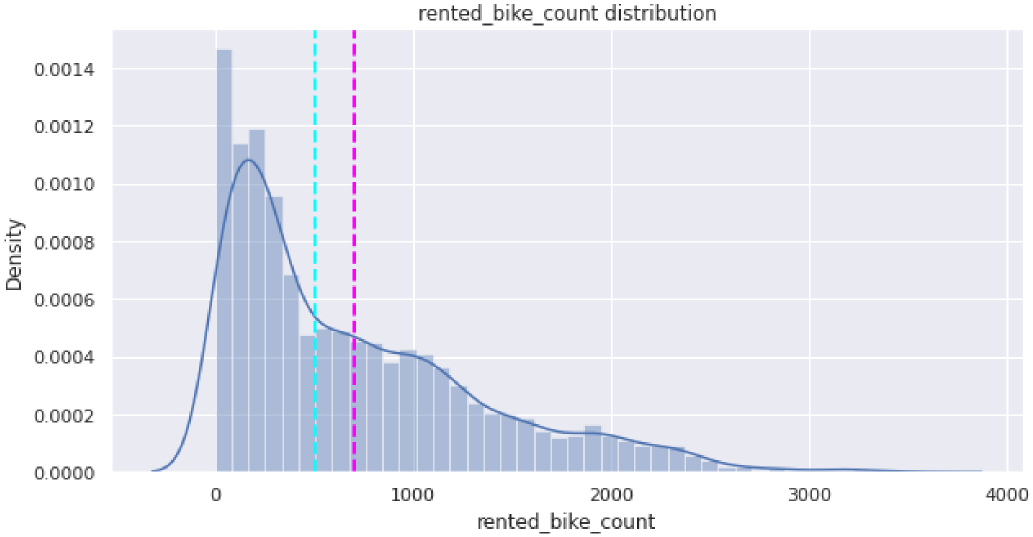
- Also these variables are **highly correlated** (0.912798)
- Hence we can drop dew point temperature

- There are **no missing values** in the dataset



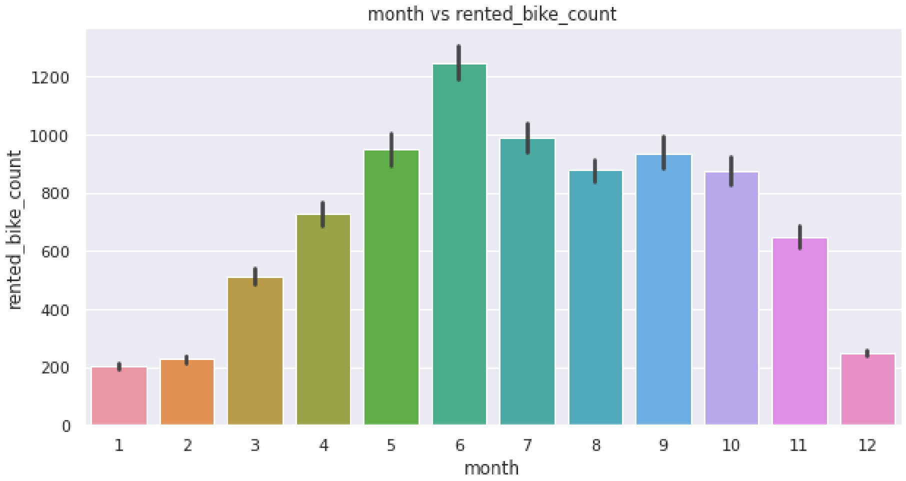
Exploratory Data Analysis (EDA)

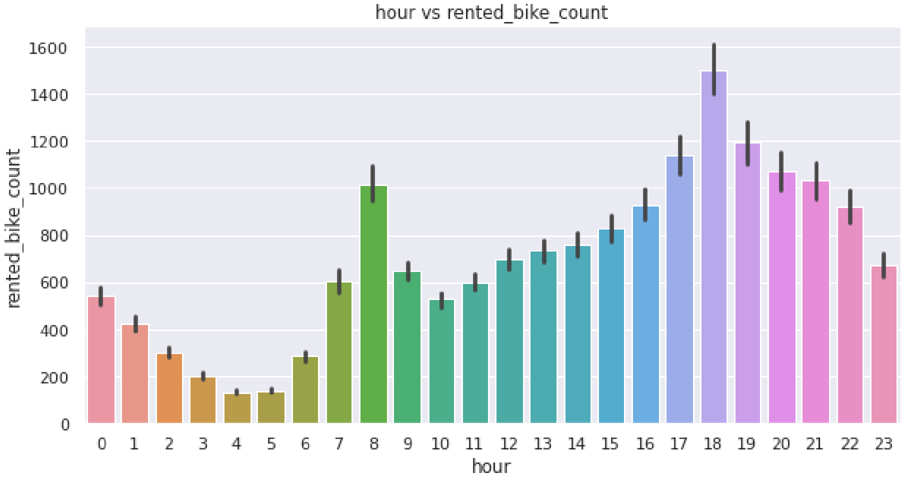
- The dependent variable - rented bike counts is **positively skewed**
- **Normally distributed attributes:** temperature, humidity.
- **Positively skewed attributes:** wind, solar radiation, snowfall, rainfall.
- **Negatively skewed attributes:** visibility.

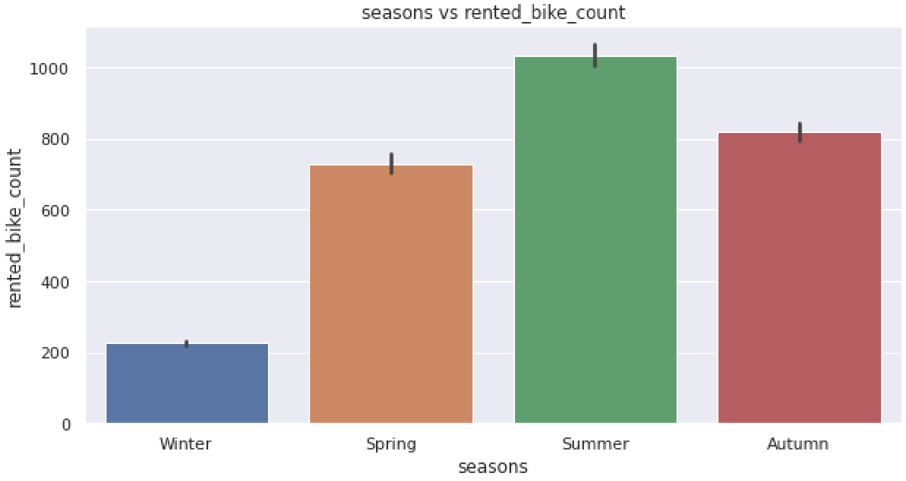


EDA (Contd.)

- Highest demand - **June**
- Lowest demand - **January**
- On a typical day, there is a **surge** in demand for rental bikes during the **rush hours**

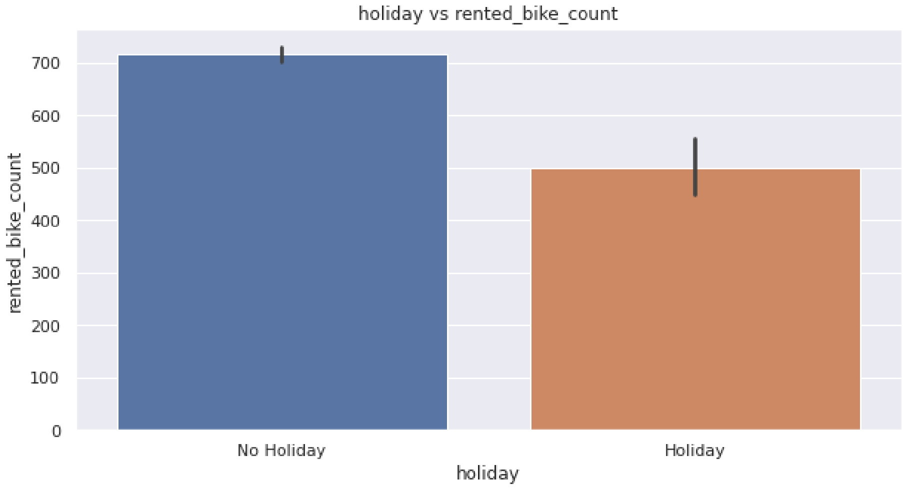


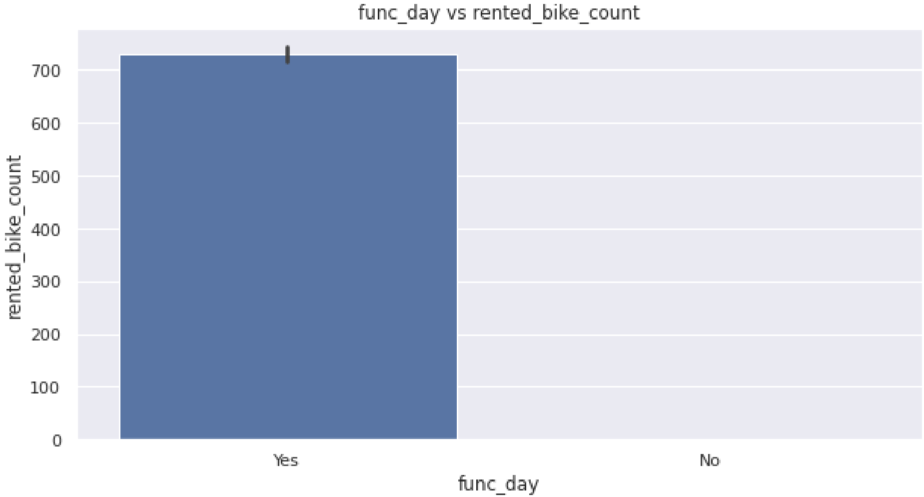


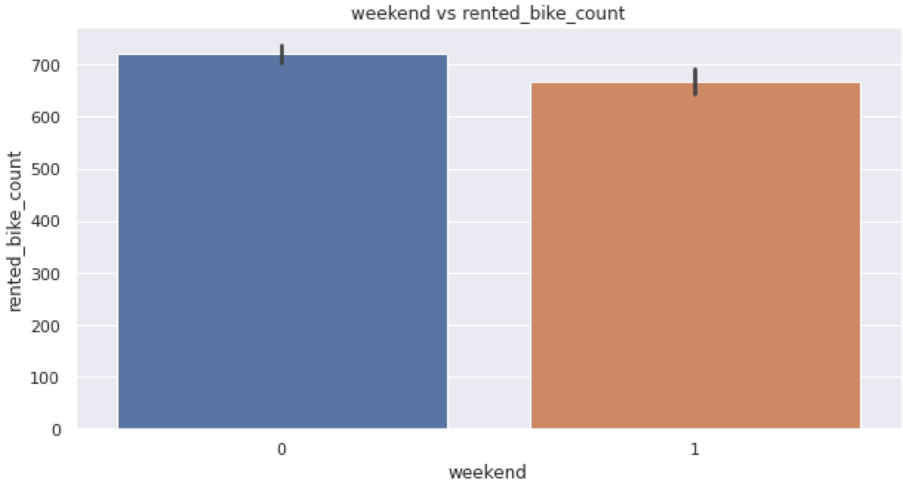


EDA (Contd.)

- Demand for rental bikes is **lower** on **holidays** and **weekends**
- On a non functional day, no bikes were rented in **all** instances

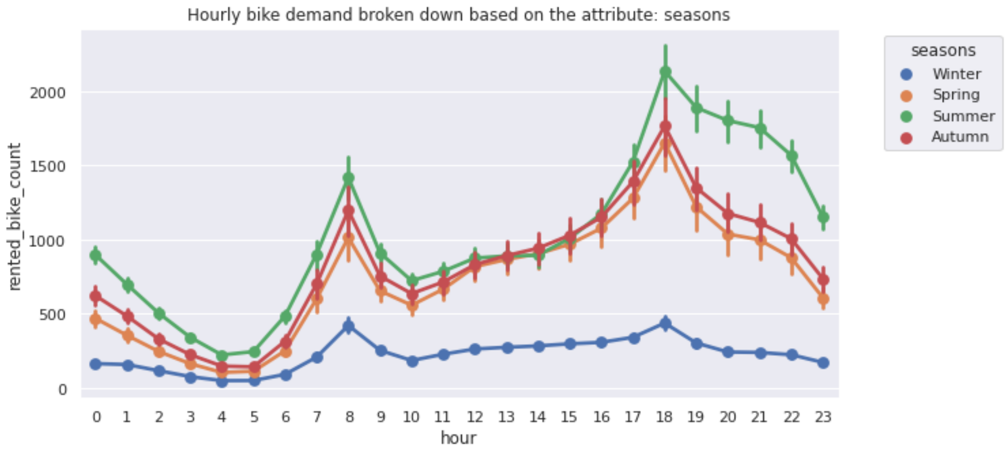


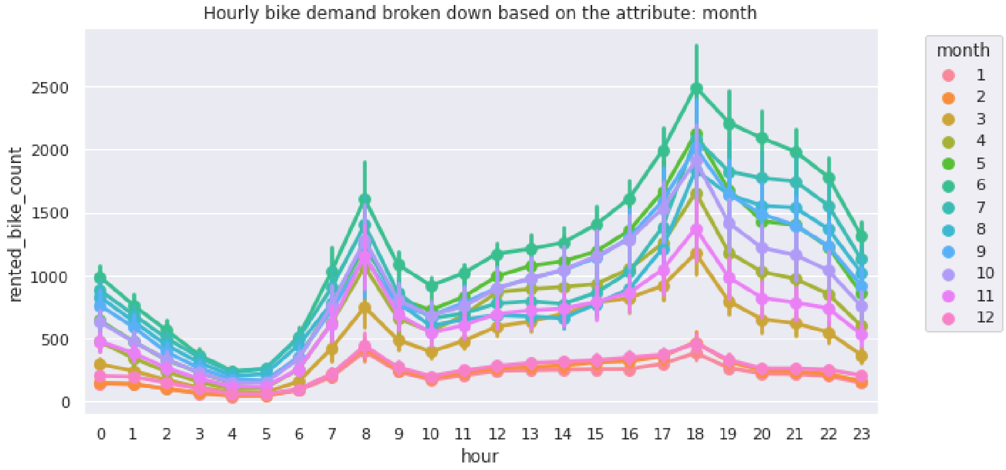




EDA (Contd.)

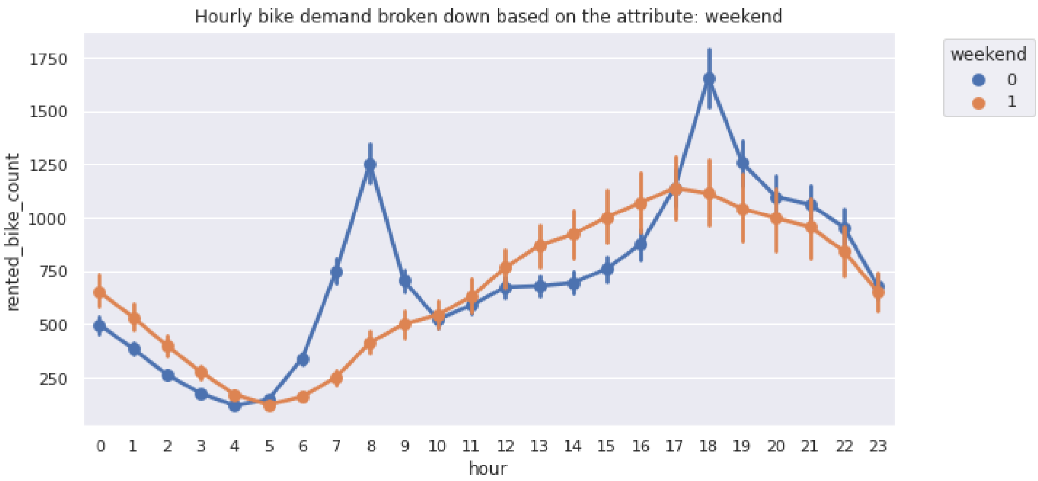
- Lowest demand - **Winter**
- Highest demand - **Summer**
- In autumn and spring, the demand on average is similar throughout the day

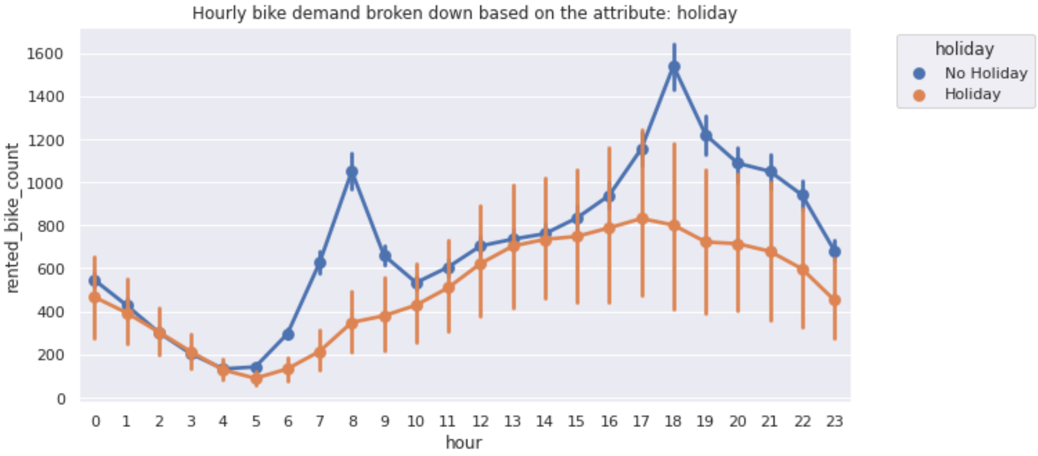




EDA (Contd.)

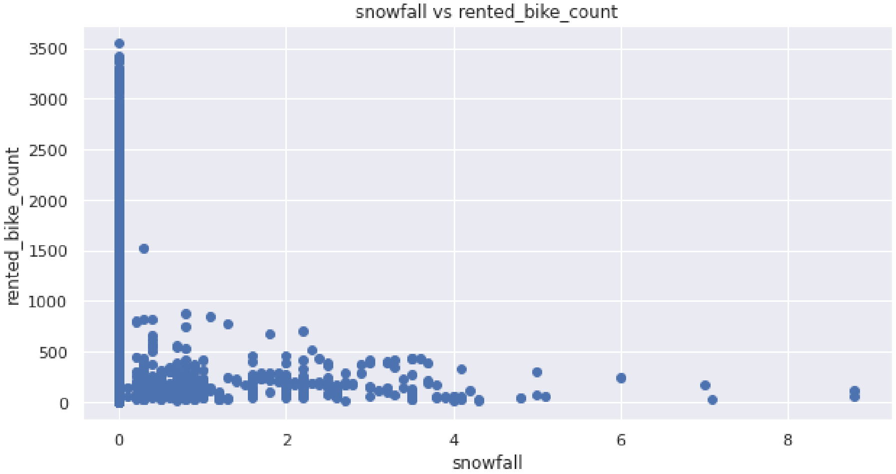
- On a regular day, there is a **surge** in demand for rental bikes during **rush** hours
- On holidays and weekends, the demand for rental bikes **increases** gradually throughout the day

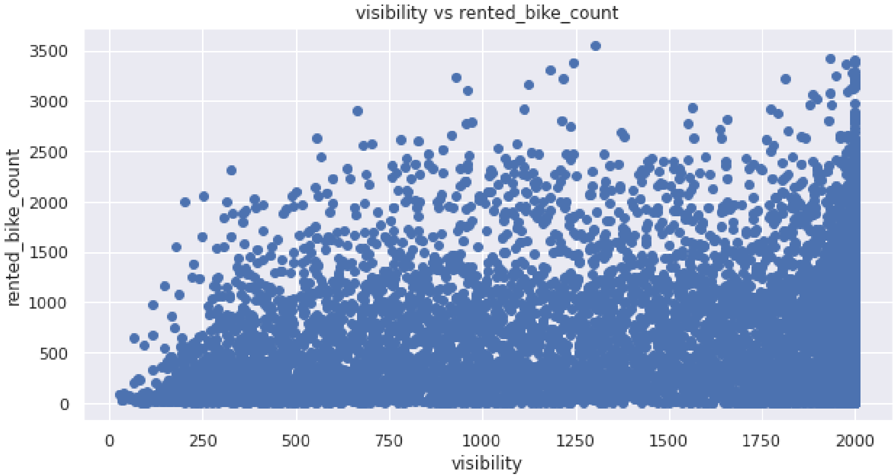


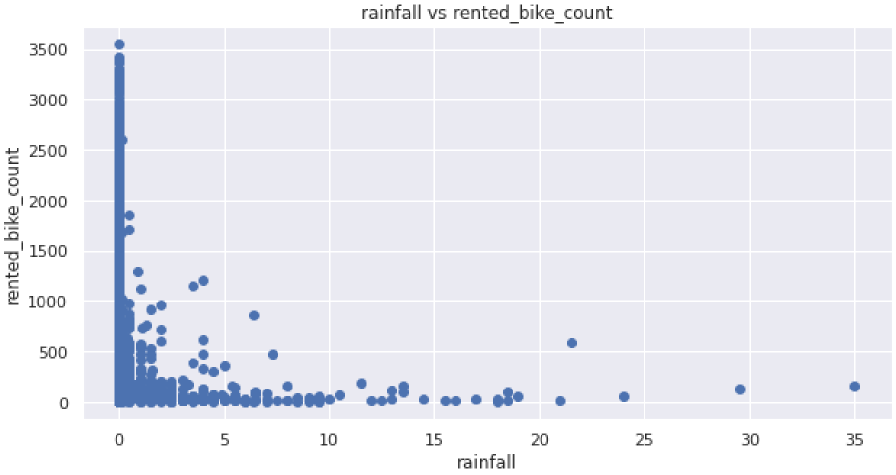


EDA (Contd.)

- The demand for rental bikes is typically **lower** when there is **rainfall / snowfall**, and on days with **low visibility**

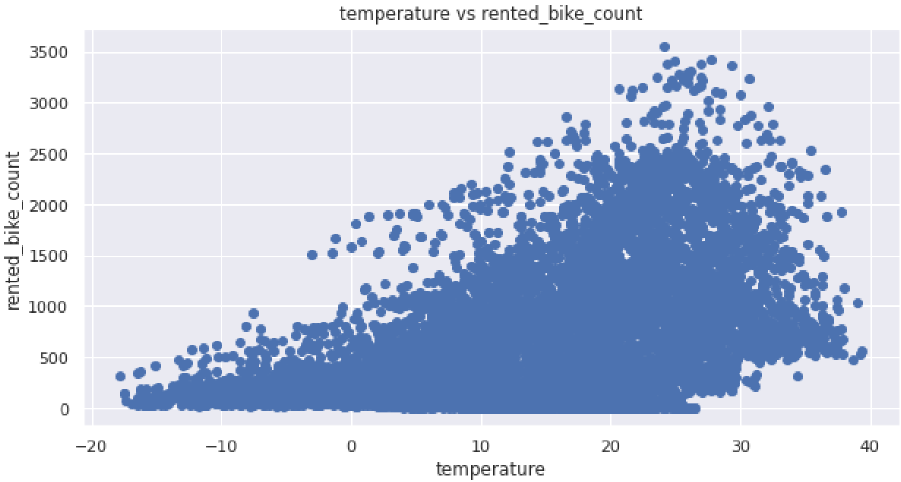


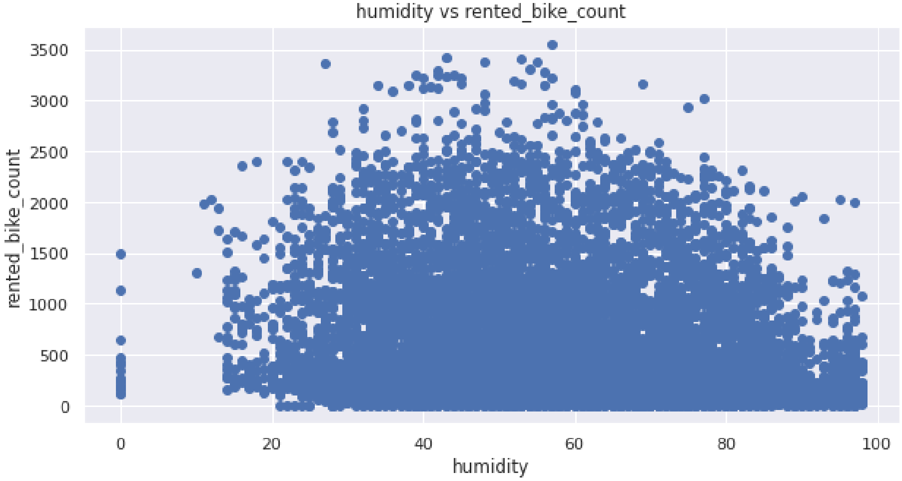


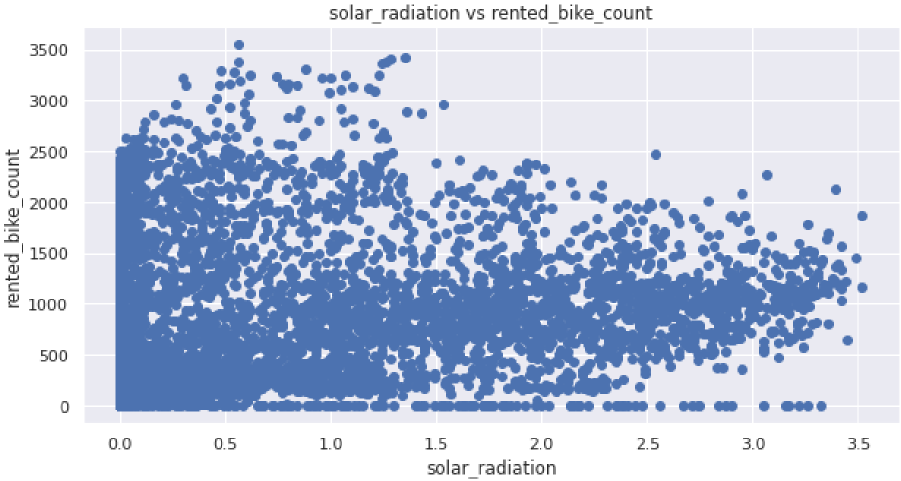


EDA (Contd.)

- The demand for rental bikes remains **low** for days with very **low temperatures**, and on days with high intensity of **solar radiation**

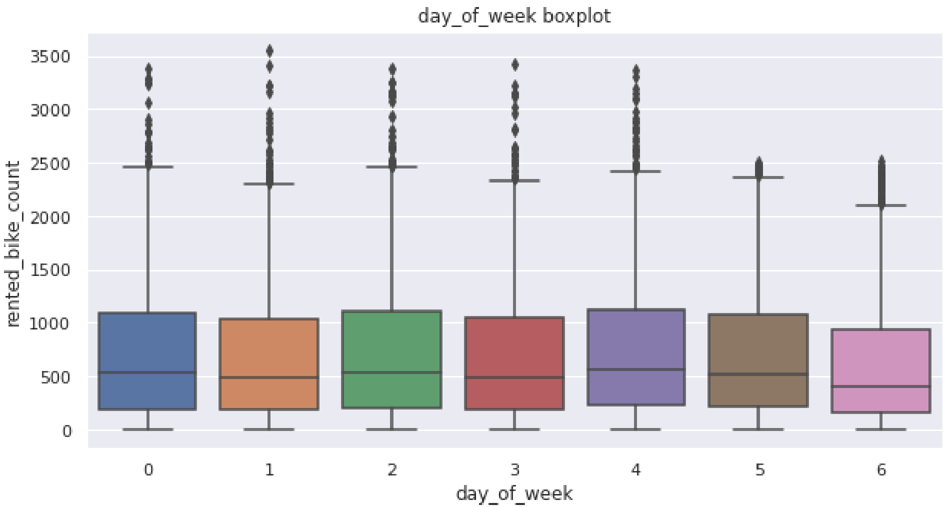


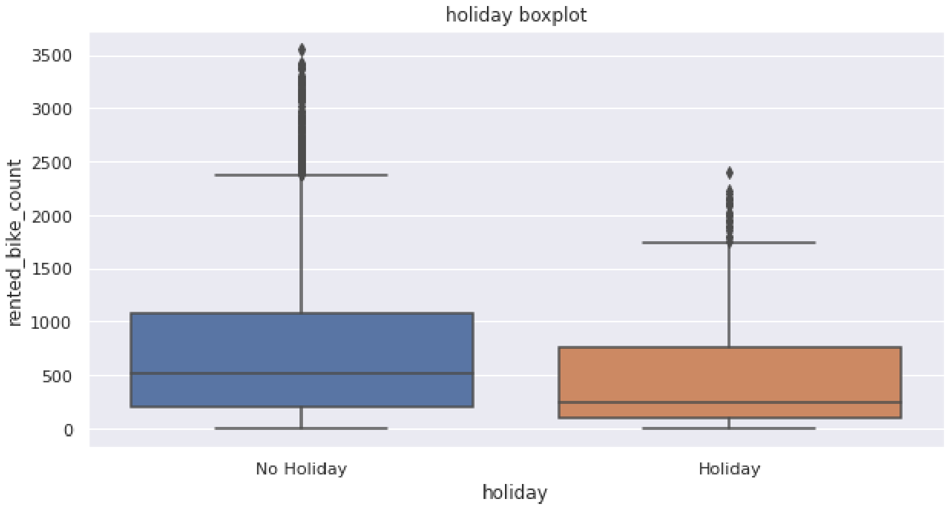


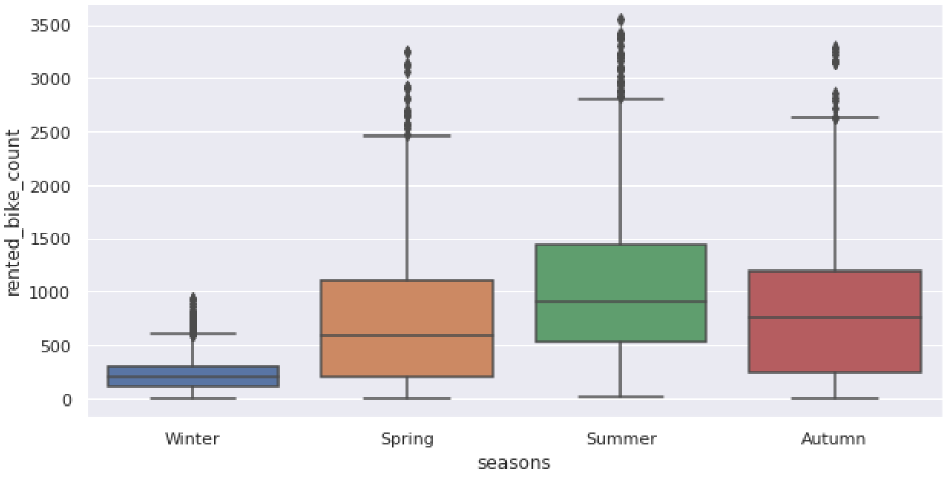


EDA (Contd.)

- There are outliers in the data
- We cannot handle them since we may **eliminate patterns** we had discovered earlier

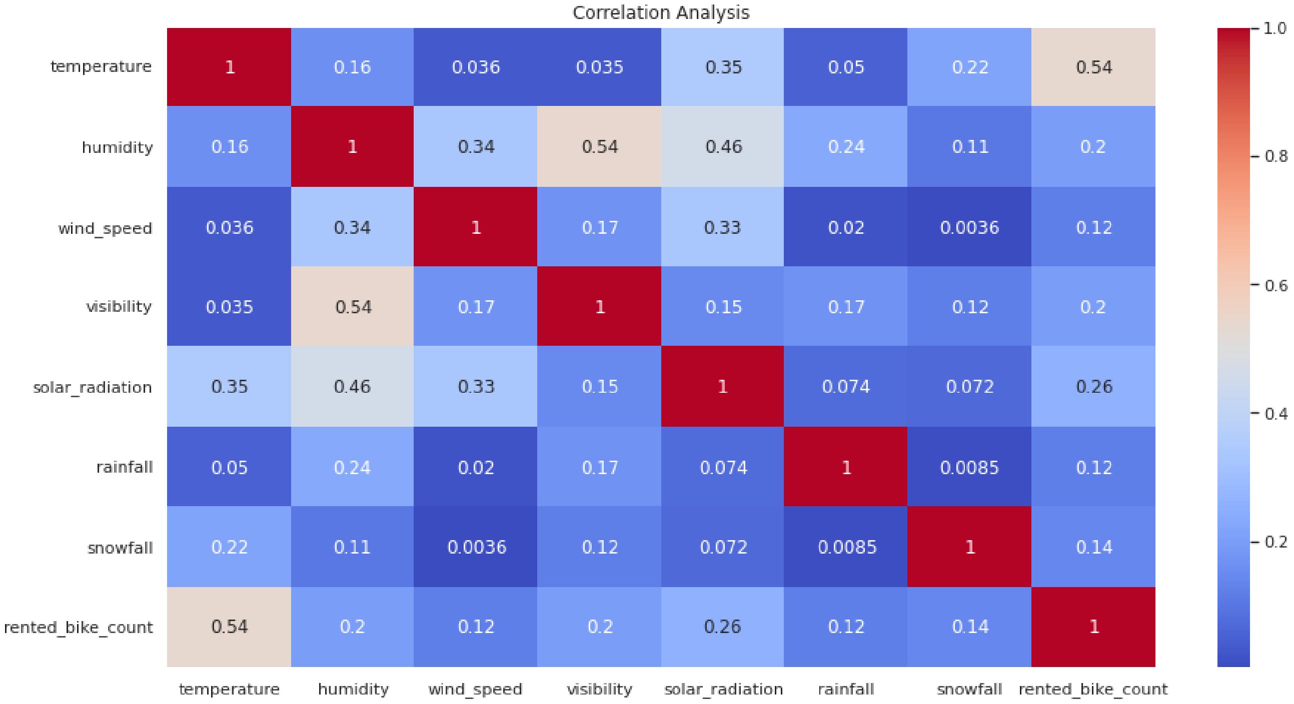






EDA (Contd.)

- Correlation magnitude
- There is no **multicollinearity** in the attributes
- Temperature has the **highest** correlation with the dependent variable



EDA Summary

- The dependent variable - rented bike counts is **positively skewed**
- Demand for rental bikes is lowest in the winters; highest in summers
- On regular days, there is a **surge** in demand for rental bikes during **rush** hours, this was absent during **holidays** and **weekends**
- The demand for rental bikes remains **low** when there is **snowfall / rainfall**, and on days with **low visibility**
- The demand for rental bikes remains **low** for days with very **low temperatures**, and on days with high intensity of **solar radiation**
- The data contains **outliers**, all the numeric variables were log transformed to handle skew, and all datapoints beyond 3 standard deviations from the mean were replaced with the median value
- Temperature has the highest correlation with dependent variable

Modelling Approach

- Numeric features: rainfall, snowfall, and visibility were converted to categorical variables
- Since there are many **categorical** attributes, It won’t be wise to fit linear models, as they will give high errors.
- We can use **tree** models instead, since they can handle outliers and categorical attributes better than linear models.
- We can use **decision tree** as a baseline model.
- Subsequently, to get better predictions, we can use **ensemble models**: Random forests, GBM, XG Boost.
- Final choice of model will depend on whether interpretability or accuracy is important to the stakeholders.

Modelling Approach (Contd.)

- Choice of split is taken as **K-fold cross validation**, with k=6, because of the computational power available and to reduce overfitting
- Model evaluation metric is taken as **RMSE** to punish outliers.

$$RMSE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N}}$$

- Apart from RMSE, **R2 score** was also calculated to explain the model performance to the general audience.

$$R^2 = \frac{Sum\ of\ Squares\ of\ Residuals}{Total\ Sum\ of\ Squares}$$

- Hyperparameter tuning is done using **Grid Search**

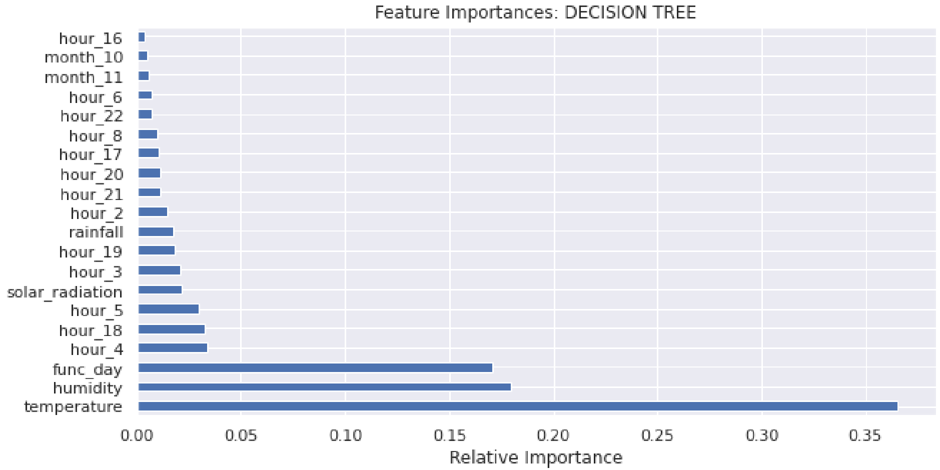
Decision Tree

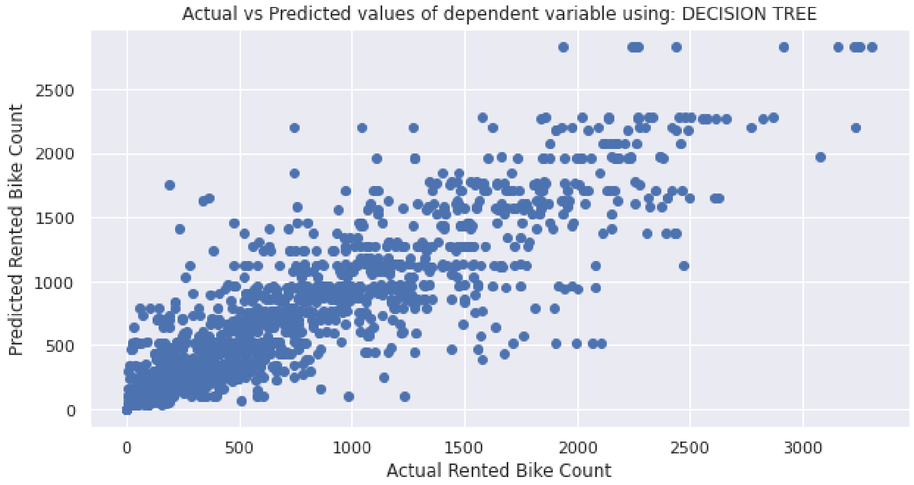
Parameters:

- Max_depth = 24
- Min_samples_leaf = 30

Evaluation metrics:

- Train RMSE = 263.27
- Test RMSE = 294.39
- Train R2 Score = 0.833
- Test R2 Score = 0.7929





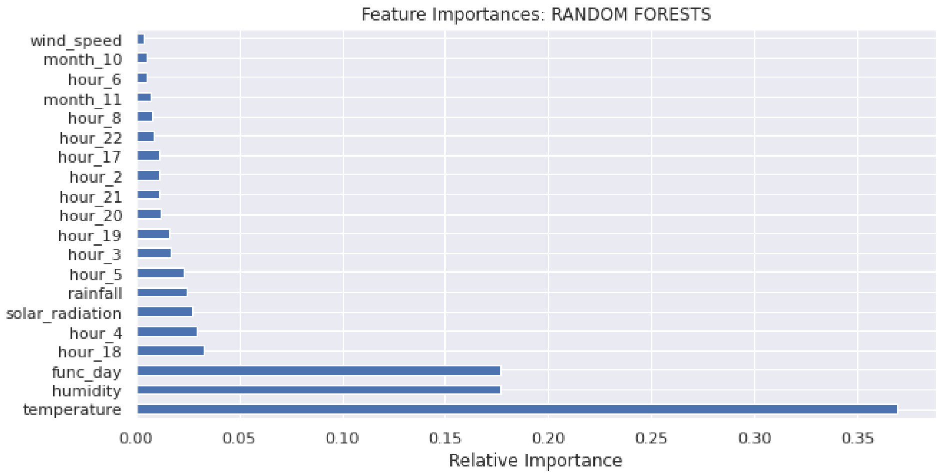
Random Forests

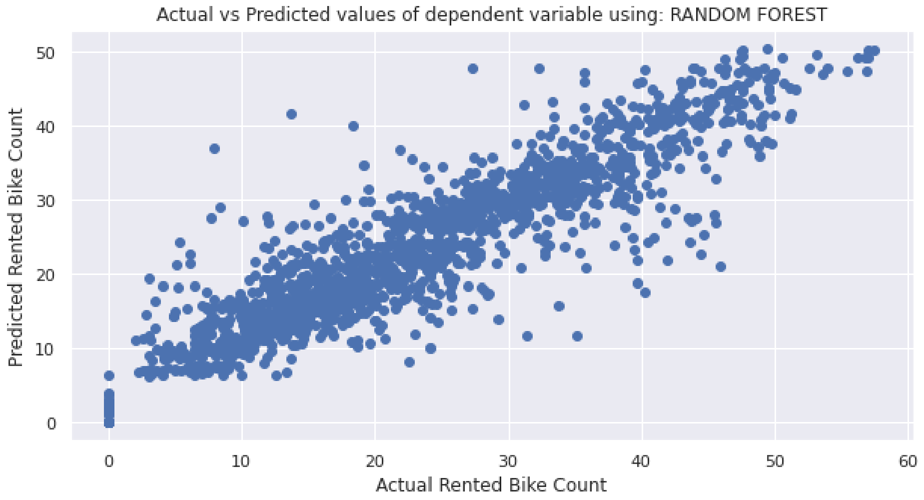
Parameters:

- N_estimators = 500
- Min_samples_leaf = 25

Evaluation metrics:

- Train RMSE = 255.13
- Test RMSE = 279.28
- Train R2 Score = 0.8432
- Test R2 Score = 0.8136





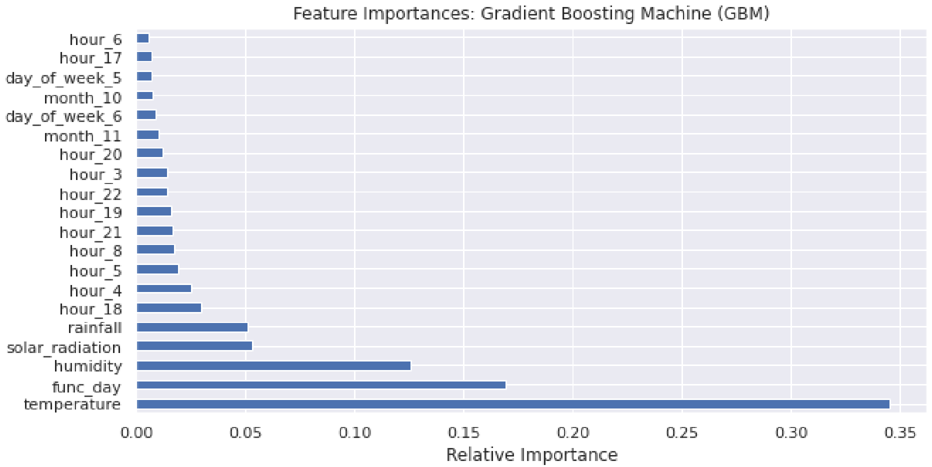
Gradient Boost

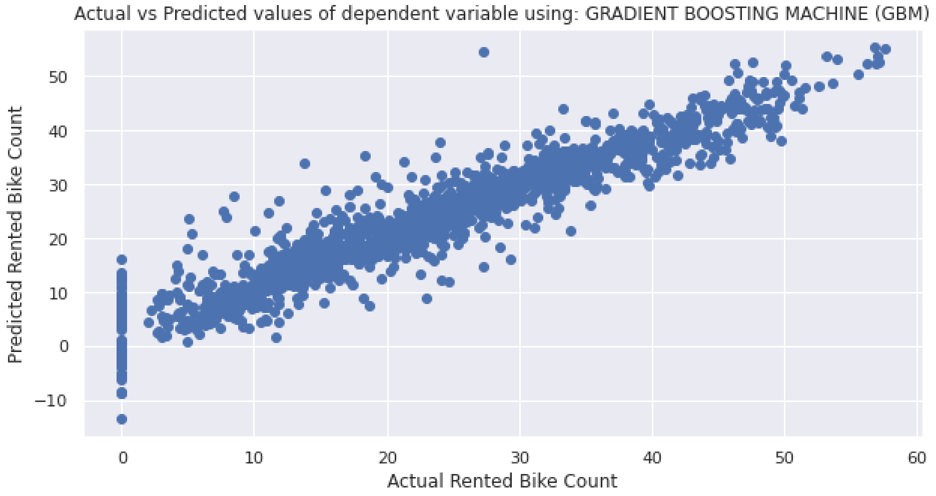
Parameters:

- N_estimators = 500
- Min_samples_leaf = 25

Evaluation metrics:

- Train RMSE = 171.52
- Test RMSE = 204.5
- Train R2 Score = 0.9291
- Test R2 Score = 0.9





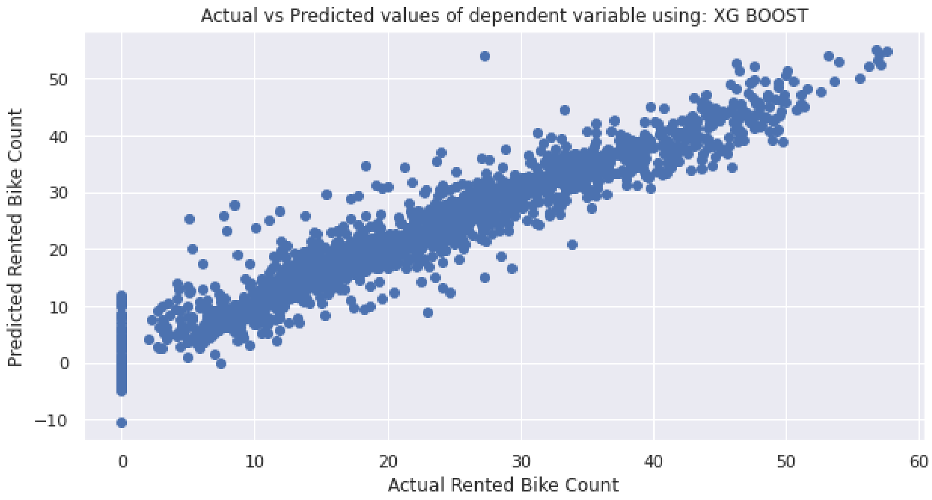
XG Boost

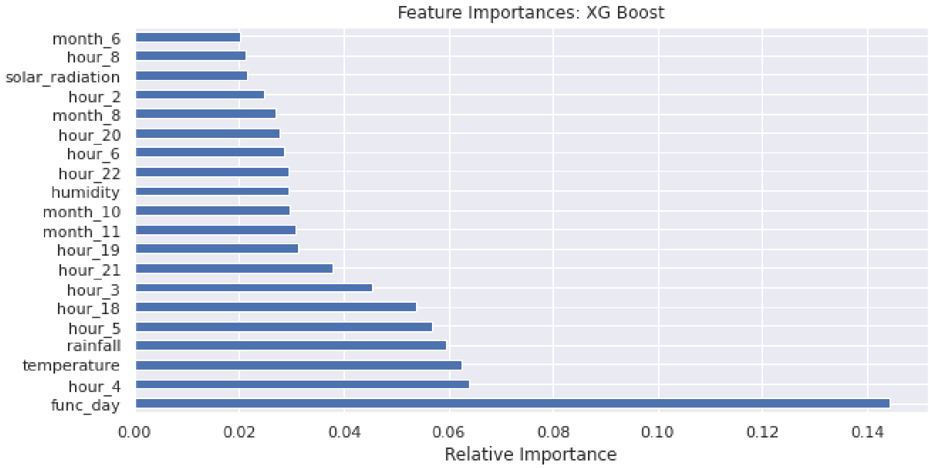
Parameters:

- N_estimators = 500
- Min_samples_leaf = 25

Evaluation metrics:

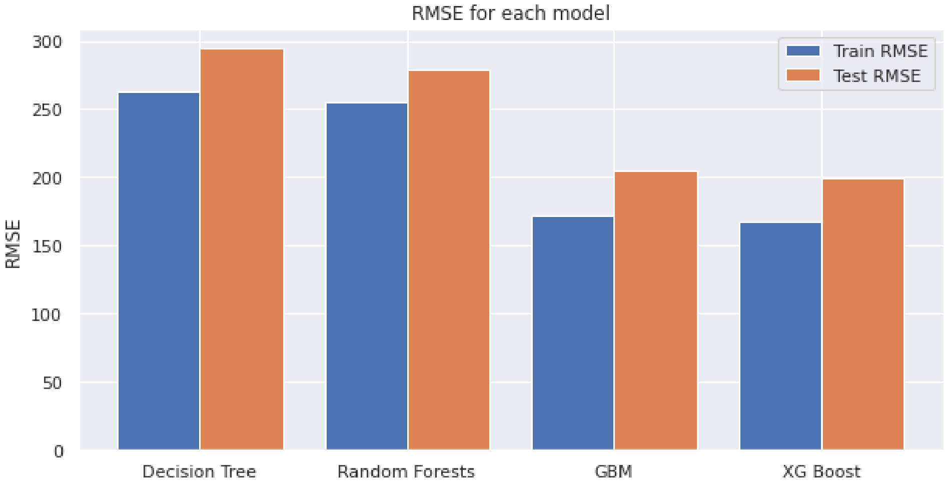
- Train RMSE = 167.93
- Test RMSE = 199.72
- Train R2 Score = 0.932
- Test R2 Score = 0.9046

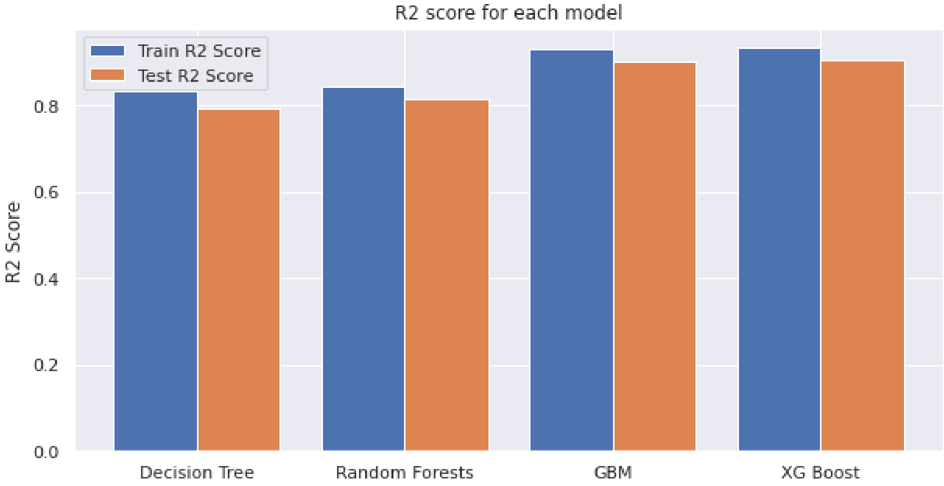




Model Comparison

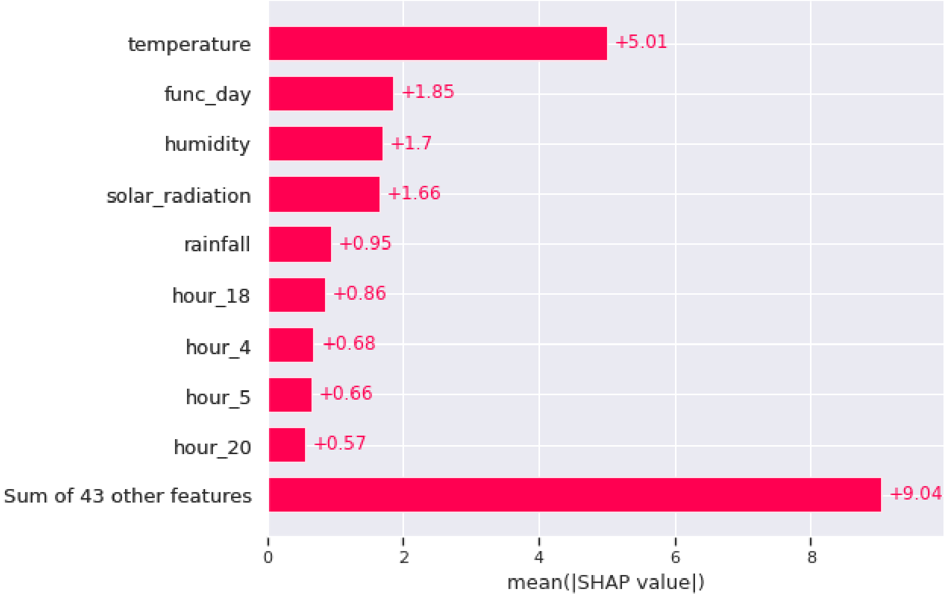
- The **XG Boost** model was able to give best predictions for the demand of rental bikes.

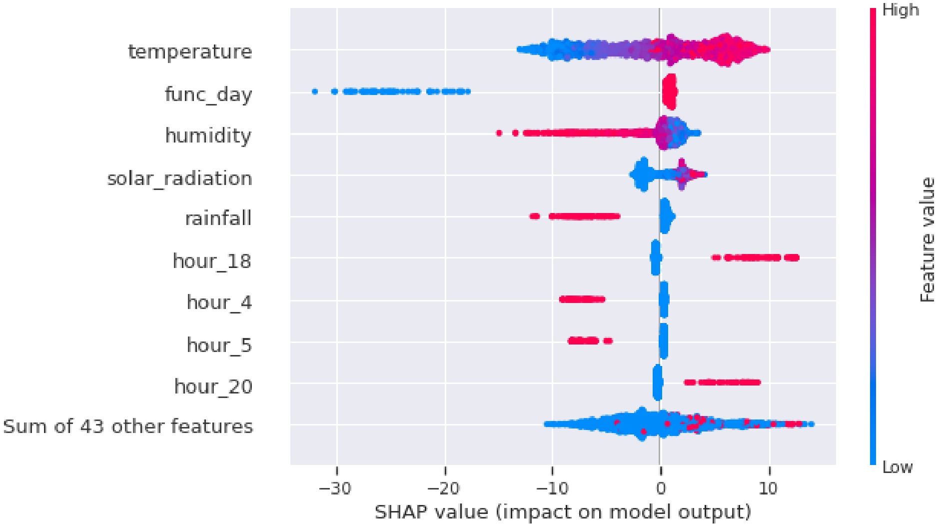




XG Boost Model Explanation Shapley Values

- **Temperature** is the most important feature in determining the value of the dependent variable followed by **functioning day**, **humidity**, **solar radiation**, and **rainfall**.





Challenges Faced

- Comprehending the problem statement, and understanding the business implications
- Feature engineering – deciding on which features to be dropped / kept / transformed
- Choosing the best visualization to show the trends among different features clearly in the EDA phase
- Deciding on how to handle outliers
- Choosing the ML models to make predictions
- Deciding the evaluation metric to evaluate the models
- Choosing the best hyperparameters, which prevents overfitting

Conclusion

- We have successfully built predictive models that can predict the demand for rental bikes based on different weather conditions and other factors and, they were evaluated using RMSE
- The XG Boost prediction model had the lowest RMSE
- We developed Shapely value plots to understand the predictions obtained from the XG Boost model
- The final choice of model for deployment depends on the business need; if high accuracy in results is necessary, we can deploy XG Boost model
- If the model interpretability is important to the stakeholders, we can choose deploy the decision tree model.

Thank You!