Capstone Project - 2

Bike Sharing Demand Prediction

Faraz Ahmad
(Individual)

What is bike sharing system ?

A bike sharing system, is a shared transport service in which bicycles are made available for shared use to individuals on a short-term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system. Docks are special bike racks that lock the bike, and only release it by computer control. The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place.

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 becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Data description

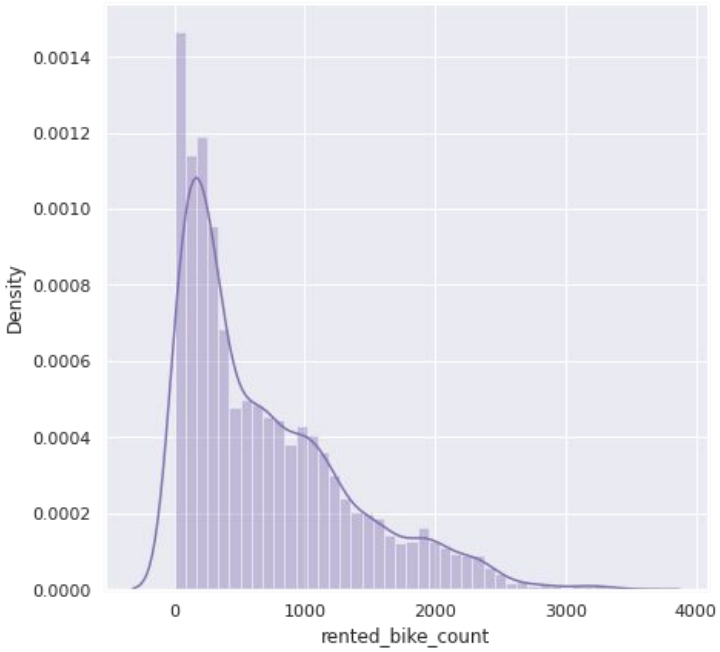
Dependent variable

- ❏ Rented Bike count - Count of bikes rented at each hour

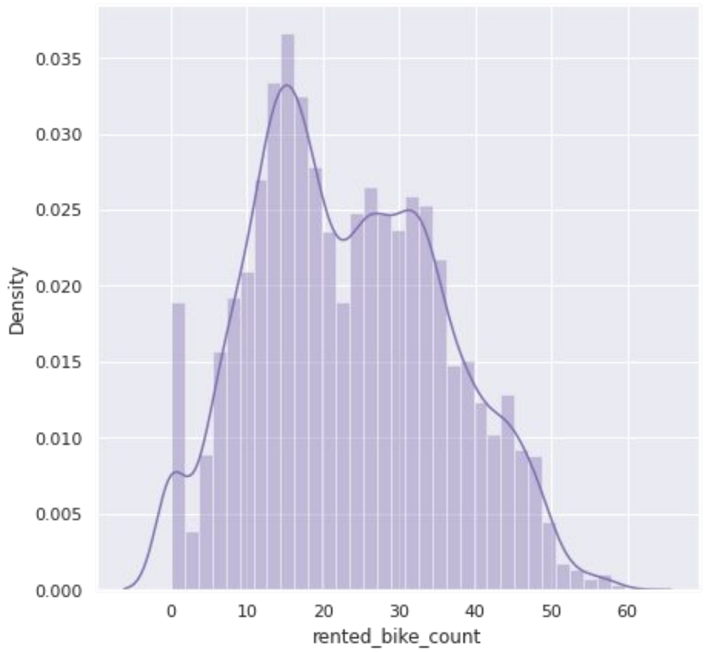
Independent variables

- ❏ Date : year-month-day
- ❏ Hour - Hour of he day
- ❏ Temperature-Temperature in Celsius
- ❏ Humidity - %
- ❏ Windspeed - m/s
- ❏ Visibility - m
- ❏ Dew point temperature - Celsius
- ❏ Solar radiation - MJ/m2
- ❏ Rainfall - mm
- ❏ Snowfall - cm
- ❏ Seasons - Winter, Spring, Summer, Autumn
- ❏ Holiday - Holiday/No holiday
- ❏ Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

Data distribution of target variable

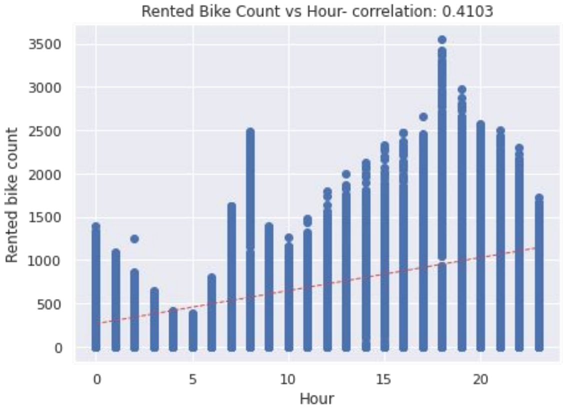


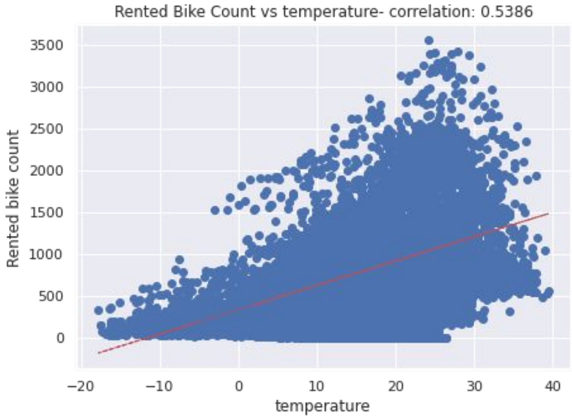
Distribution before square root transformation

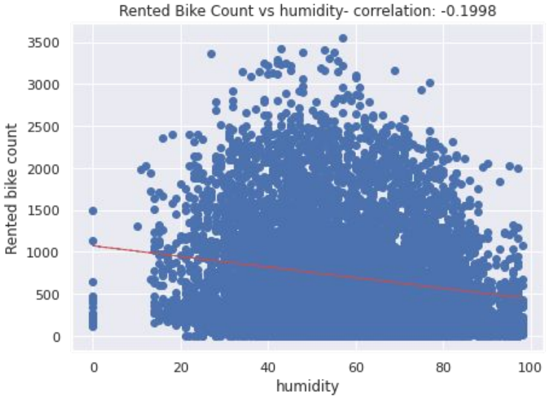


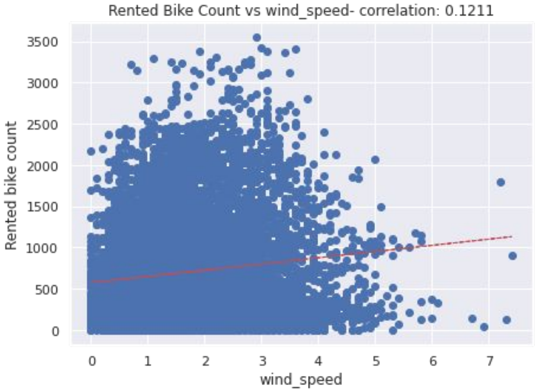
Distribution after square root transformation

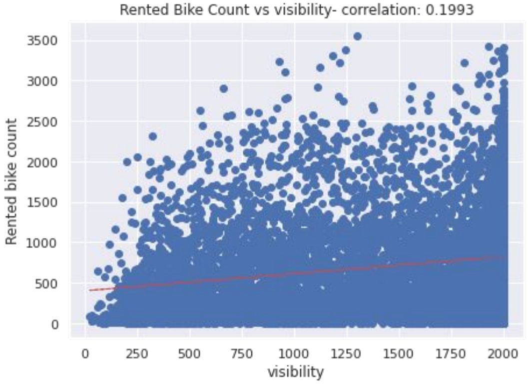
Relationship between dependent & independent variables

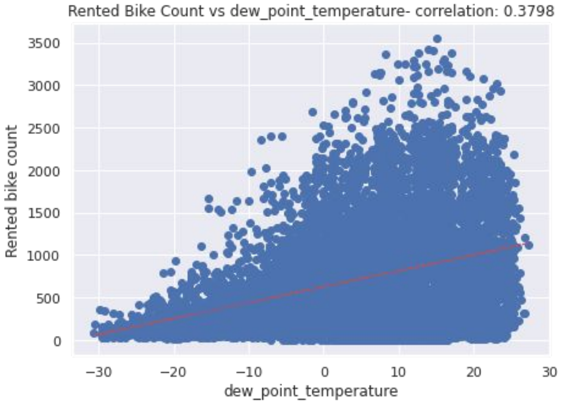




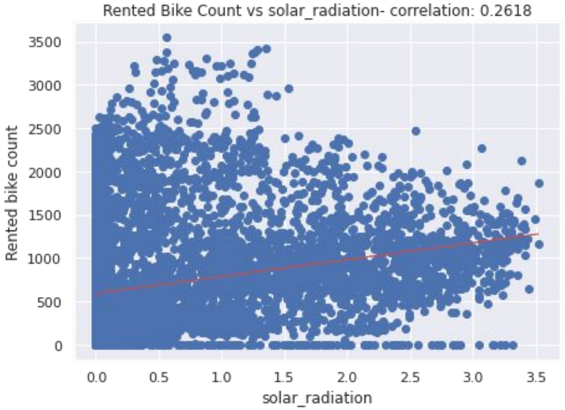


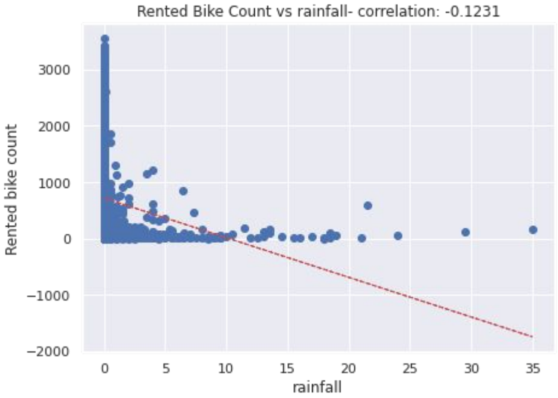


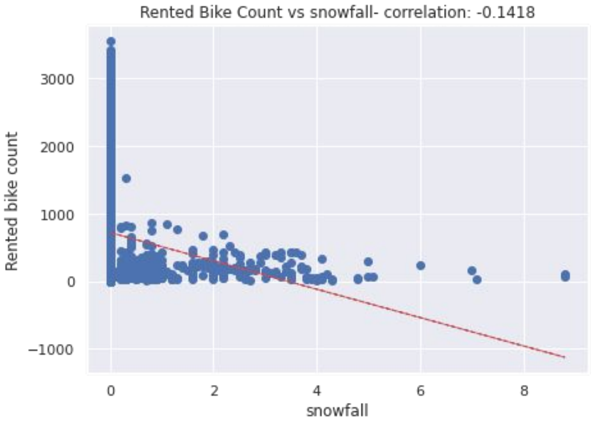


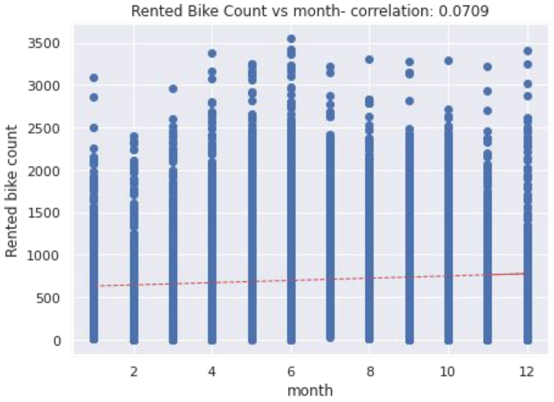


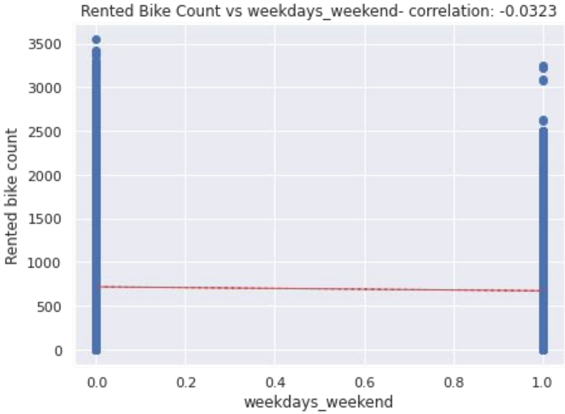
Contd...


Rented Bike Count vs solar_radiation- correlation: 0.2618
Rented bike count
solar_radiation
0 500 1000 1500 2000 2500 3000 3500
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5

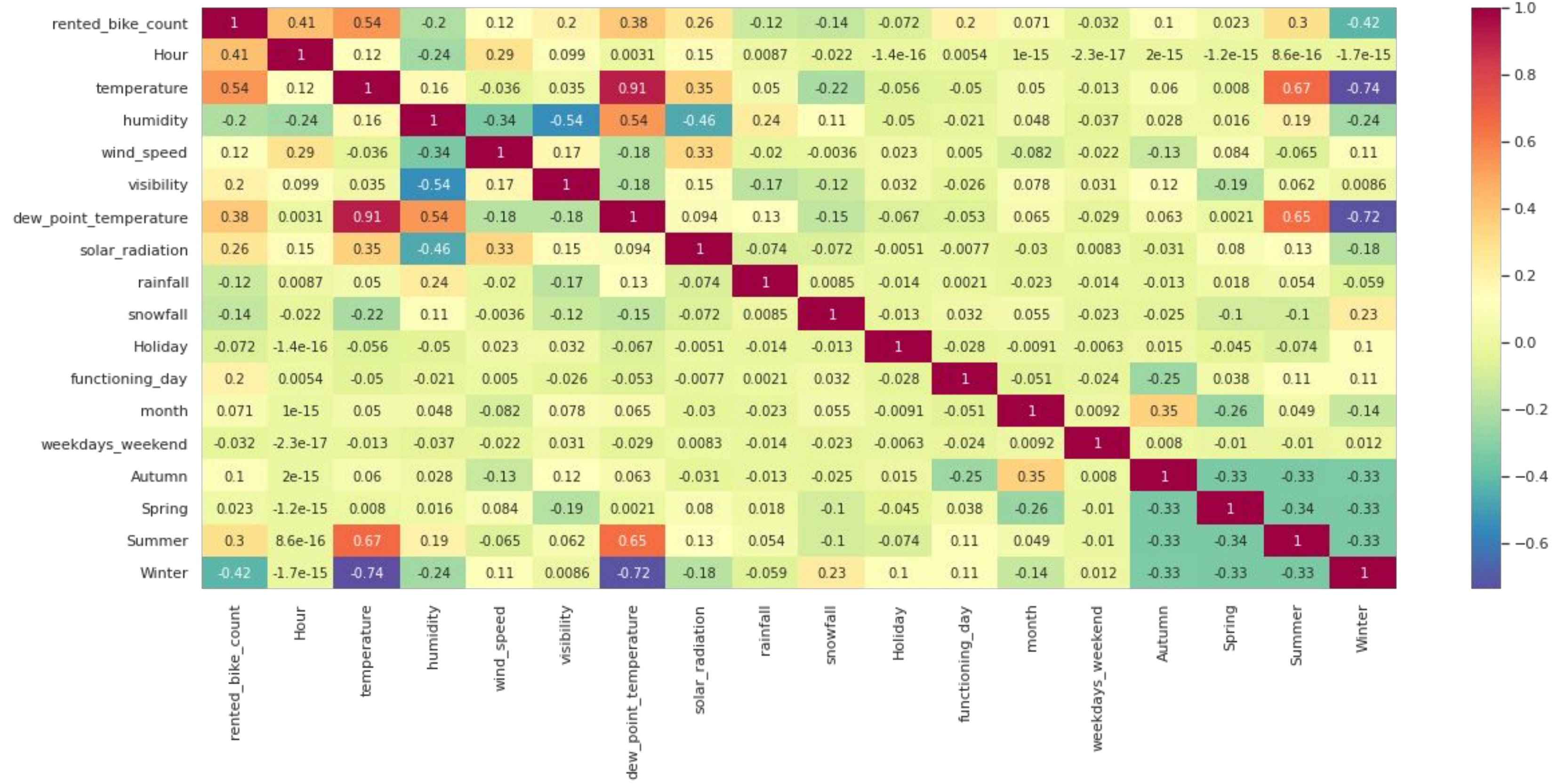

Rented Bike Count vs rainfall- correlation: -0.1231
Rented bike count
rainfall
−2000 −1000 0 1000 2000 3000
0 5 10 15 20 25 30 35


Rented Bike Count vs snowfall- correlation: -0.1418
Rented bike count
snowfall
−1000 0 1000 2000 3000
0 2 4 6 8

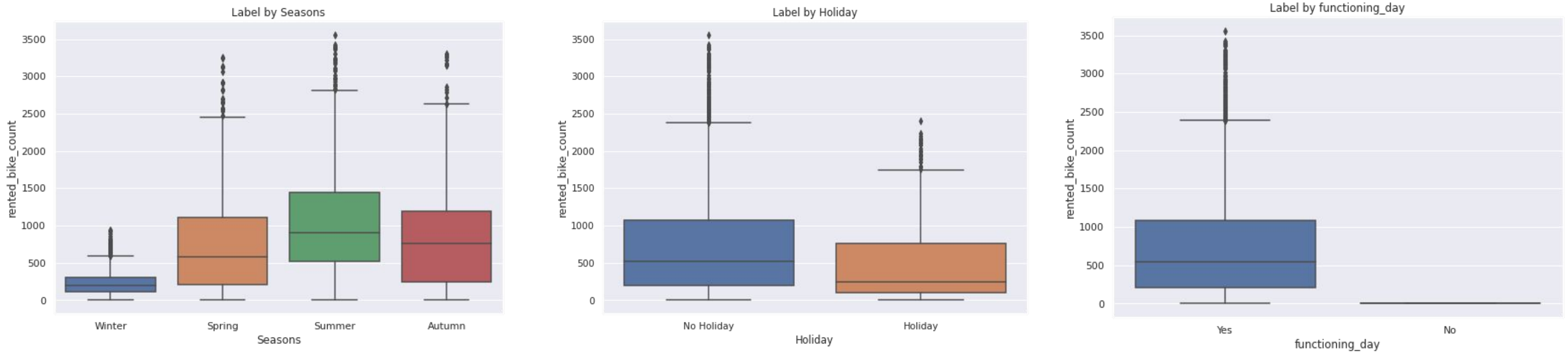

Rented Bike Count vs month- correlation: 0.0709
Rented bike count
month
0 500 1000 1500 2000 2500 3000 3500
2 4 6 8 10 12


Rented Bike Count vs weekdays_weekend- correlation: -0.0323
Rented bike count
weekdays_weekend
0 500 1000 1500 2000 2500 3000 3500
0.0 0.2 0.4 0.6 0.8 1.0

Checking Multicollinearity

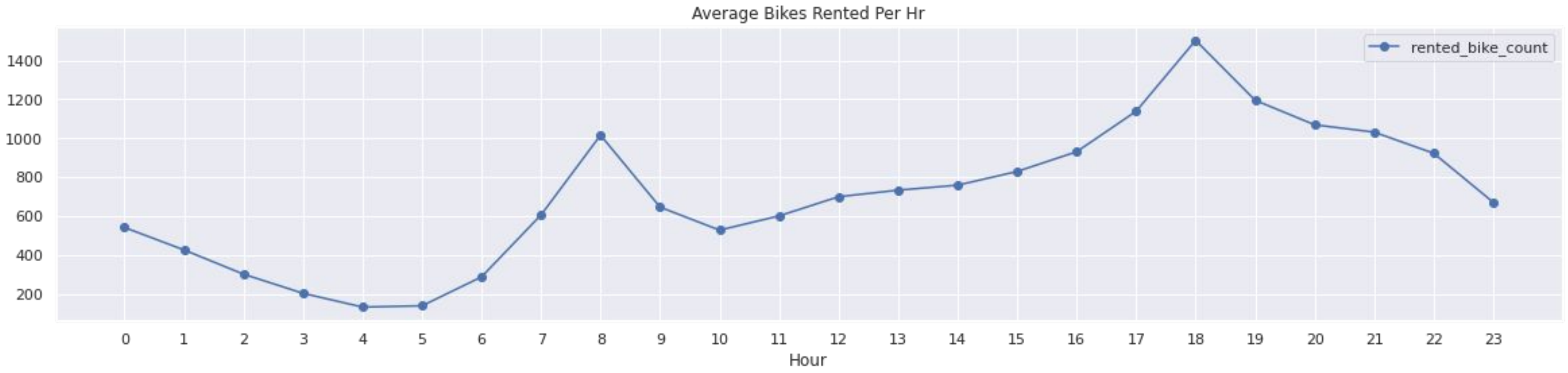


Seasons, Holiday & Functioning day effect on bike demand



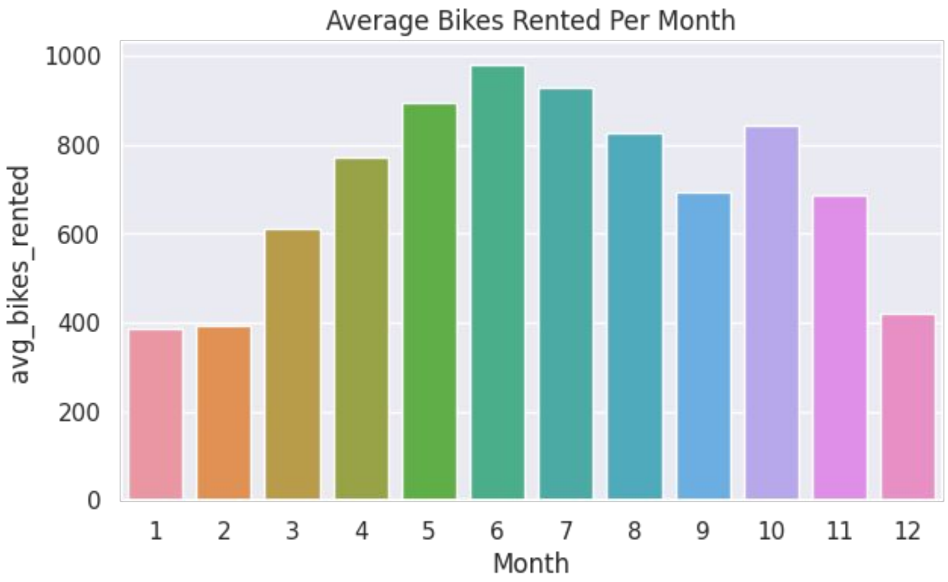
- The demand of bikes is lowest in winter while highest in summer.
- Slightly Higher demand during Non holidays as compared to Holidays.
- Almost no demand on non functioning day.

Average bikes rented per hour



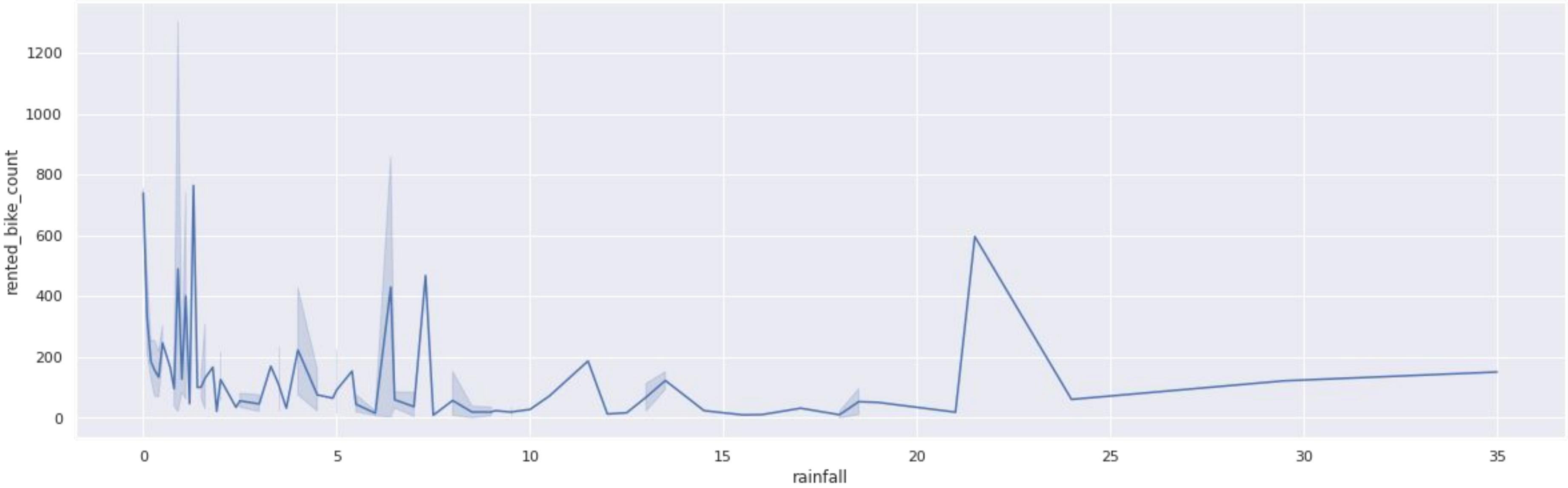
- During rush hour, rented bikes are in high demand from 8 a.m. to 9 p.m.
- It is easy to see that demand is highest at 8 a.m. and 6 p.m., so we can say that during office opening and closing time, there is high demand of bikes.

Average bikes rented per month



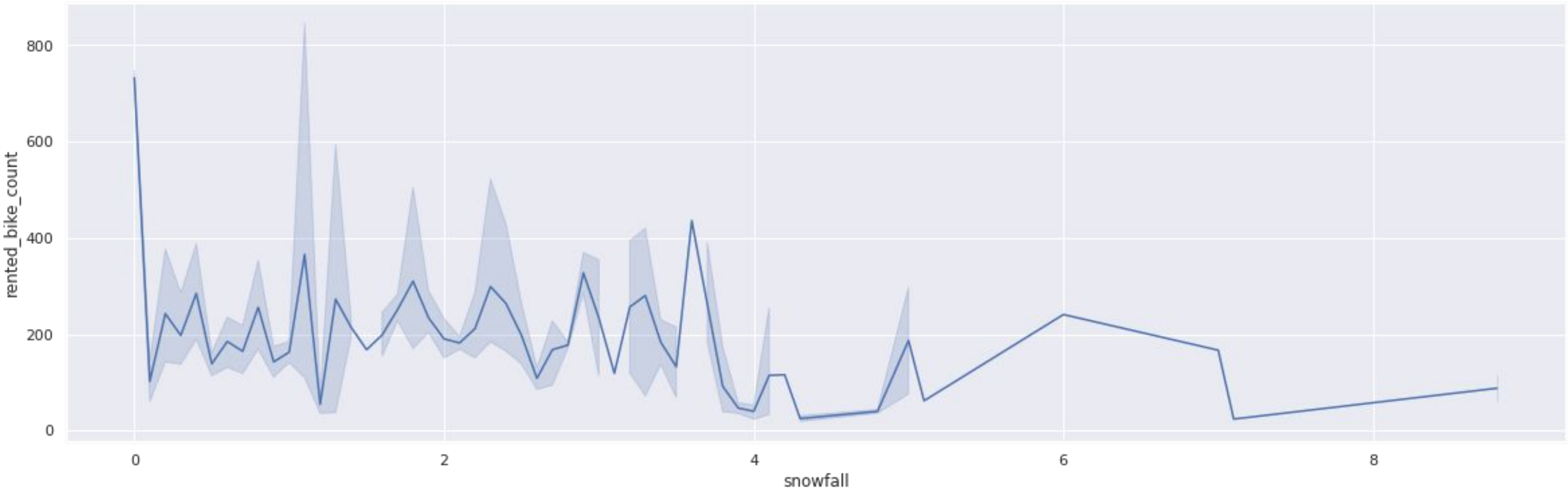
- Rented bikes are less in demand in December, January and February i.e. during the winter.
- In addition, the demand for bikes is highest during the summer months of May, June and July.

Demands of bikes during rainfall



- The demand of bikes decreases with the increase in rainfall.

Demands of bikes during snowfall



- The demand of bikes also decreases with the increase in snowfall.

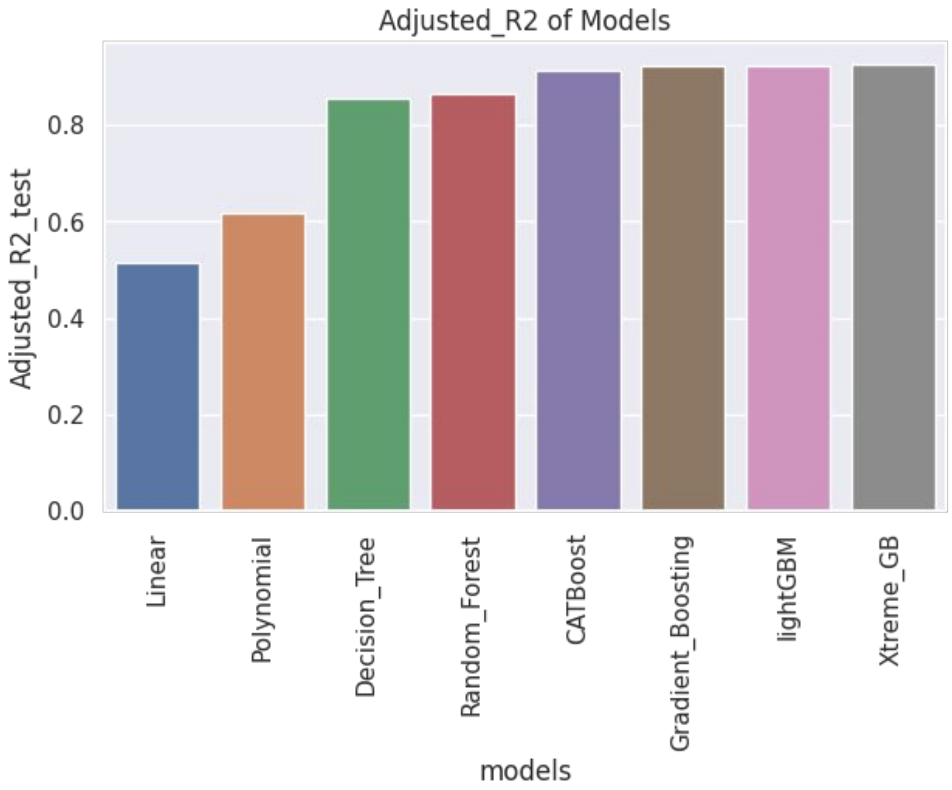
Algorithms used in modelling

- Linear regression
- Polynomial regression
- Decision tree
- Random forest
- Gradient Boosting
- eXtreme Gradient Boosting
- CatBoost
- lightGBM

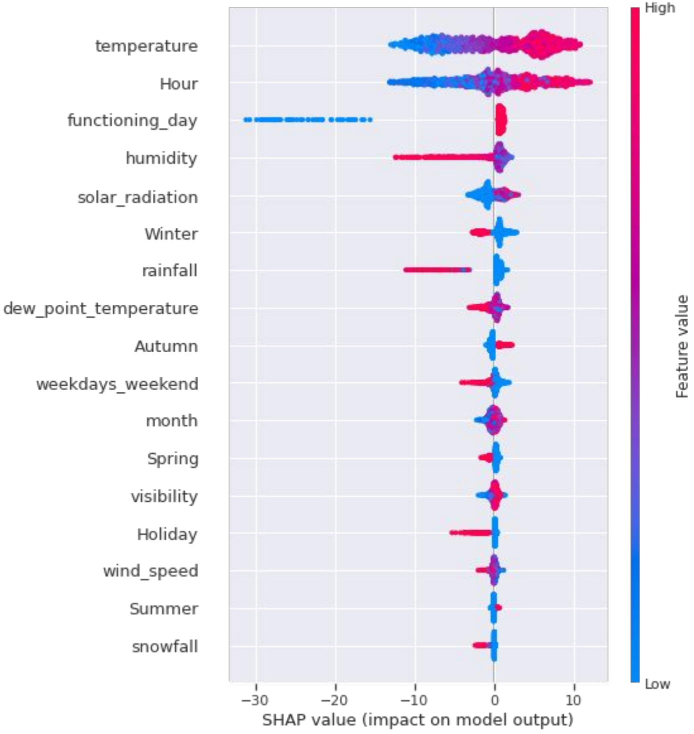
Evaluation metrics of models

	models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2_train	Adjusted_R2_test	Mean_absolute_error
0	Linear	207376.116147	455.385678	0.518573	0.539002	0.515251	303.872069
1	Polynomial	163740.992452	404.649221	0.619873	0.643577	0.617250	267.104322
2	Decision_Tree	58127.005836	241.095429	0.857359	0.905541	0.855960	150.684809
3	Random_Forest	54957.983294	234.431191	0.865135	0.884092	0.863813	151.911644
4	Gradient_Boosting	31402.176669	177.206593	0.922940	0.964966	0.922185	109.075398
5	Xtreme_GB	30448.593203	174.495253	0.925280	0.981728	0.924548	103.467846
6	CATBoost	35386.659612	188.113422	0.913163	0.931641	0.912311	117.644360
7	lightGBM	31284.773871	176.875023	0.923229	0.955437	0.922476	109.517147

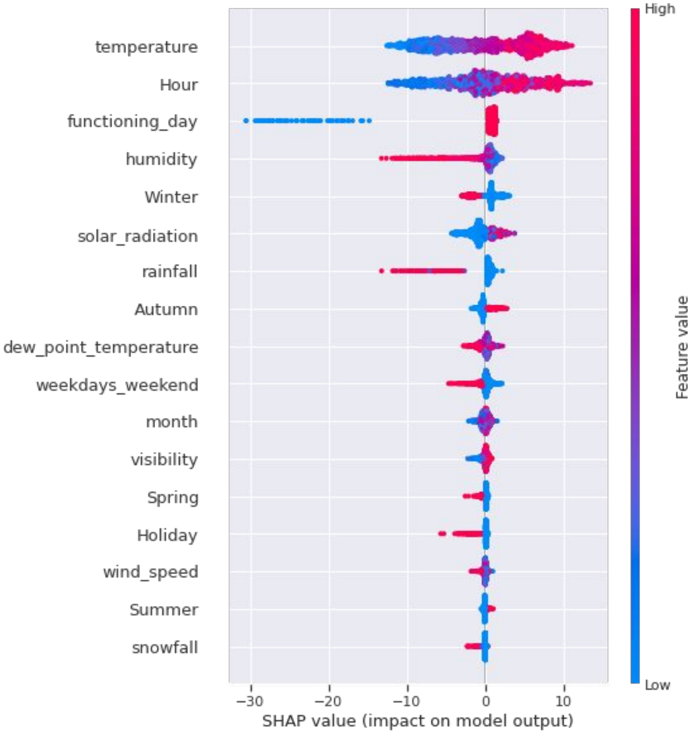
Adjusted R2 score (test) of different models



SHAP Summary Plot

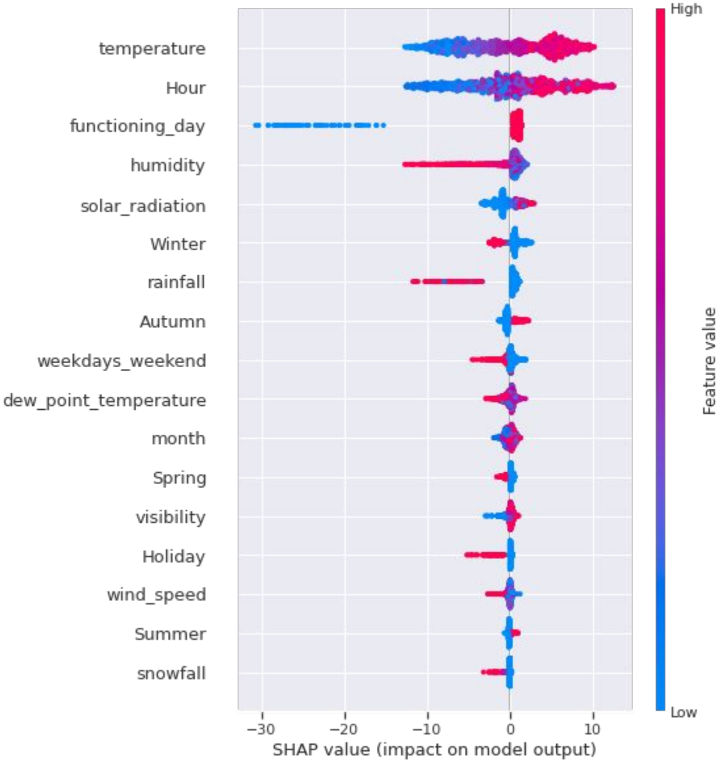


Gradient Boosting



XGBoost

SHAP Summary Plot



lightGBM

Conclusion

- Linear models are not performing well on this dataset as the relationship of target variable is not linear with the independent variables.
- Tree and ensemble based algorithms are performing well on this dataset.
- Temperature and hour are the most influencing features as interpreted by SHAP summary plot.
- It is found that Gradient boosting, XGboost and lightGBM are the best algorithms that can be used for Bike Sharing Demand Prediction since evaluation metrics show better performance on the models based on these algorithms.
- XGBoost has the least Root Mean Squared Error and highest adjusted R2 value. So, it can be considered as the best model for the given problem.