Bike Sharing Prediction

Satyam Jyoti Sankar

Data Science trainee,

Alma Better, Bangalore

Abstract :

Bike Sharing rental is becoming a important now days as we can see from the market survey the traffic situation and the cost of the petrol is very high so the bike sharing could be an solution to that problem. Also many people don't have the personal vehicles and public transportation is also congested and time consuming that's why peoples are mostly moving toward the rental bikes.

The main goal of project is to maximize the availability of bikes to the customer and minimize the time of waiting to get a bike on rent. With the help of the Exploratory Data Analysis on the given data set we finding the correlation of the various feature with the target variable. The objective of the model building from the given data we perform the various regression technique to get the best result for that we perform the train test split with the important feature. This gives us the good predictive module to reach out our main goal of the supply of rental bike count on the hourly base through the year.

1. Problem statement :

It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. This project tackles the problem of predicting the number of bikes which will be rented at any given hour in a given city. The main objective is to build various regression models and analyze their performance with respect to each other so as to get the best performing model, which could help in predicting the number of bikes required at each hour for the stable supply of rental bikes.

2. Data set Description:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- 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)

To make an available the requirement of the bike as per the require time will be a good approach.

The crucial part is the prediction of bike count required at each hour for stable supply of rental

bikes. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. The goal of this project is to build a model from the historical bike usage patterns with the weather data to forecast bike rental demand using machine algorithm .

3. Factors Affecting :

Following are the factors affecting to the number of bike rentals:

1. **Seasons** : Bike rental is high during the summer and least during the Spring season.

2. **Temperature** : As the temperature increases the bike count is gradually increases .

3. **Hours** : We observed that there is a peak in the bike rentals counts at around 8am morning and at around 5pm evening.

4. **Weather** : When the weather is clear or sunny the bike rental is high where as in the heavy rain and snowfall the bike rental is very low.

5. **Working Day** : Bike rental counts on working and non-working days and we observed that the outliers are present in working day.

6. **Holiday** : Bike rental counts on holidays and non-holidays. Holidays correspond to non-working days. Also outliers are present in non holidays.

4. Steps involved :

The following steps are involved in the project

1. **Exploratory Data Analysis** :
   After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is bike rentals counts with other independent variables.
   This process helped us figuring out various aspects and relationships among the target and the independent variables and also we observed the distribution of variables.
   It gave us a better idea that how feature behaves with the target variable.

2. **Preprocessing data :**
   Dataset contains a no null values also no duplicate values are found to disturb the accuracy .
   Changing column names for easy handling.
   Dropping the unwanted columns from the dataset.

3. **Features selection :**
   With the help of exploratory data analysis we analyzed the categorical as well as numerical features in the dataset.

4. **One hot encoding :**
   In this dataset some categorical variables like seasons, holiday and function day, we change it with numerical database.

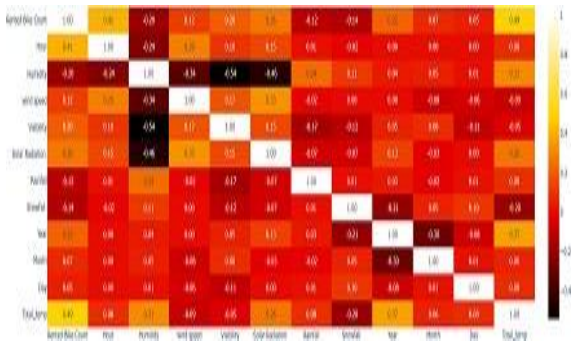
5. **Correlation Analysis :**
   We plot the heatmap to find the correlation between both dependent variable and independent variables.

6. **Train test Split :**
   In train test split we take ‘x’ as dependent variables and ‘y’ take as independent variable then train the model.

7. **Models :**
   We uses 7 modeling to train the data and for predicting the accuracy, RMS and R2.
   1. Linear regression
   2. Lasso regression
   3. Ridge regression
   4. Elastic net
   5. Decision Tree
   6. Random Forrest
   7. XGboost

5. Correlation Analysis :

We plot the heatmap to find the correlation between all the columns and observed that:

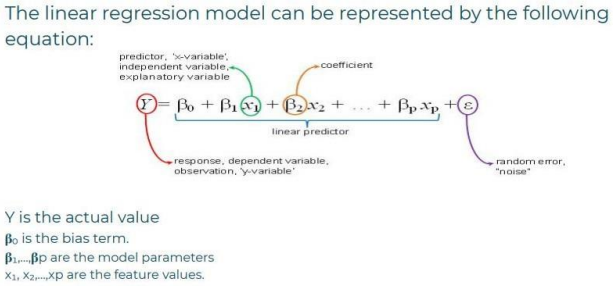


1. The highest correlation with respect to the bike count is Temperature which shows as the temperature increase the number of bike count also increases.
2. The lowest correlation is wind speed.
3. There is correlation with humidity also as the humidity increases the bike count decreases.
4. Also the increase in rainfall and snowfall decrease the count of bike.
5. More the humidity, the less people prefer to rental bikes.
6. There is very low correlation between the wind speed which indicates it does not effect the bike rental that much compare to other feature.

6. Modeling Results :

1. Linear Regression :

The hypothesis of linear regression model represented below:



We have model by linear regression and we get results as follows:

- MSE : 37.273629840755824
- RMSE : 6.1052133329438885
- R2 : 0.750827145595724
- Adjusted R2: 0.745240832185446

1. **Lasso regression:**
   By performing lasso regression we get the results are as follows :
   - MSE : 37.13614730823984
   - RMSE : 6.0939434940143515
   - R2 : 0.7629142641092735
   - Adjusted R2 :0.7575989370085006

2. **Ridge regression:**
   By performing ridge regression we get the results are as follows :

- MSE : 37.26764694572257
- RMSE : 6.104723330808905
- R2 : 0.7508671410305504
- Adjusted R2 :0.7452817242951307

3. **Elastic Net :**
   By performing Elastic net regression we get the results are as follows :
   - MSE : 36.469868314473736
   - RMSE : 6.039028755890614
   - R2 : 0.756200262048822
   - Adjusted R2 :0.7507344108476746

4. **Decision Tree :**
   By performing Decision tree we get the results are as follows :
   - MSE : 34.876743604873276
   - RMSE : 5.9056535290239704
   - R2 : 0.7668502425580741
   - Adjusted R2 :0.7616231578512958

5. **Random Forrest :**
   By performing random forrest we get the results are as follows :
   - MSE : 15.868355509166264
   - RMSE : 3.9835104504904044
   - R2 : 0.8939206228689478
   - Adjusted R2 :0.8915423836805824

6. **XGBoost:**
   By performing XGBoost we get the results are as follows :
   - MSE : 19.446866450146597
   - RMSE : 4.4098601395221815
   - R2 : 0.8699984079011416
   - Adjusted R2:0.8670838462847262

7. **Gradient Boosting :** By performing Gradient boosting we get the results are as follows :
   - MSE : 16.08789459082017
   - RMSE : 4.010971776367938
   - R2 : 0.8924530121247648
   - Adjusted R2: 0.8900418699397992

7. Conclusions :

1. We see that people generally prefer to bike at moderate to high temperatures. We observed highest rental counts between 32 to 36 degrees Celsius.

2. Bike rental count is mostly correlated with the time of the day as it is peak at 10 am morning and 8 pm at evening.

3. We observed that bike rental count is high during working days than non working day.

4. It is observed that highest number bike rentals counts in Autumn/fall Summer Seasons and the lowest in Spring season.
5. We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
6. We observed that with increasing humidity, the number of bike rental counts decreases.
7. Hour of the day holds most importance among all the features for prediction of dataset.
8. When we compare the root mean squared error and mean absolute error of all the models, Random_forest and Gradient Boosting gives the highest R2 score ending with the accuracy of 89% . So, finally this two model are best for predicting the bike rental count on daily basis