Bike Sharing Demand Prediction

Vikas Kumar Manjhi, Ankit Upadhyay, Raghvendra Singh
Team- Data Pirates
Data science trainees,
AlmaBetter

Abstract:

Bike Sharing Demand Prediction is a project which will be highly required in the coming time in countries like India. We started with loading the data, then we did Exploratory Data Analysis (EDA), null values treatment, feature selection, encoding of categorical columns, and then model building. In all of these models, our accuracy ranges from 62% to 90%, which can be said to be good for such a large dataset. This performance could be due to various reasons like the proper pattern of data, large data, or because of the relevant features

1.Problem Statement

Bike sharing systems are a means of renting Bikes where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these Bike Sharing systems, people rent a bike from one location and return it to a different or same place on need basis. People can rent a bike through membership (mostly regular users) or on demand basis (mostly casual users). This process is controlled by a network of automated kiosk across the city.

2. Introduction:

Bike sharing systems are innovative ways of renting bikes for use without the onus of ownership. A pay per use system, the bike sharing model is either works in two modes: users can get a membership for cheaper rates or they can pay for the bikes on an ad-hoc basis. The users of bike sharing systems can pick up bikes from a kiosk in one location and return them to a kiosk in possibly any location of the city. With more than 500 bike renting schemes across the globe, and popular bike renting programs functional in London (Boris bikes), Washington (Capital bikeshare) and New York (Citi bikes) which are used by millions of citizens every month; these schemes provide rich dataset for analysis. This prompted the authors of this paper to take up the interesting problem of inventory management in bike sharing system, which can be formulated as the ‘Bike sharing demand’ problem wherein given a supervised set of data, you have to create a model to predict the number of bikes that will be rented at any given hour in the future.

3. Literature Review:

Bike sharing systems:

It is important to distinguish between three generations of services. According with several authors there are three generations of services of bike-sharing: free bike system, coin-deposit systems.

The free bike-sharing system is characterized by a set of Bikes (with unusual colors and/or shapes) that are available without costs to the user. Typically the stations are located near public facilities that have their own staff which are responsible for the users' identification, reducing the needs of human resources of the system.

The use of the Bike is, in the most cases, free to the user. The first bike sharing system was emerged in Amsterdam, the Netherlands in 1965. A set of fifty free Bikes was seen as the solution for traffic problems. However the Witte Fietsen (white bikes) Plan failed after its launch due to the Bike damages and thefts. In the coin-deposit systems the Bikes are not freely available, once the users have to use a coin to unlock the Bike from the docking stations. At the same time, some concerns about the location of the stations are introduced to ensure the efficiency of the operation. Although some significant changes on the motorized transportation patterns in some cities the coin-deposit system did not solved the thefts problem. To overcome this problem, the third generation of bike-sharing emerged based on automatic services. This generation uses smart technology (mobile phones, mag-stripe cards, smartcards or codes) to unlock the Bikes from the stations allowing the automatic identification of the users (with a code for instance). The casual users pay a security deposit to ensure the return of the Bike, and the use of the Bikes is paid depending on the time interval of the usage.

4. Demand studies for bike-sharing

One of the biggest concerns of the urban transportation planners is to provide the most adequate response to traveller's needs, estimating transportation demand and its variation. Planners are also aware of the strong relation between transportation and land use, and as this relation should be incorporated in demand studies. It is complex and risky to predict the number of Bike trips, especially in cities where the Bike is not yet widely used.
bookings using time series or decision trees.

5. Methodology:

The methodology focuses in the relation between the target public of bike-sharing, trip characteristics and the physical characteristics of the city paths. As previously referred, the Bike usage is mainly affected by the distance of the trip, the slope inclination, the purpose of the trip and lack of Bike paths. However it is admissible that, in an urban environment, all streets are

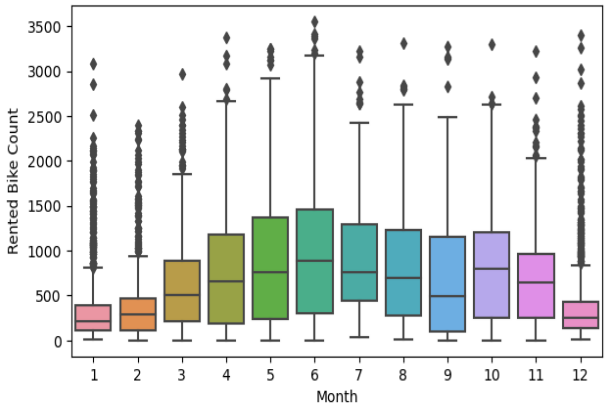
adaptable for Bike use, from minor to major improvements. Therefore, the main advantage of this methodology is not only the demand quantification (which usually is made by applying a Bike sharing users proportion to all the city trips – to all O-D pairs – only considering different purposes) but also modelling it according the studied area. The demand definition is studied considering two parts:

a) quantifying demand based on other case studies – obtaining the proportion of bike sharing users per trip purpose and

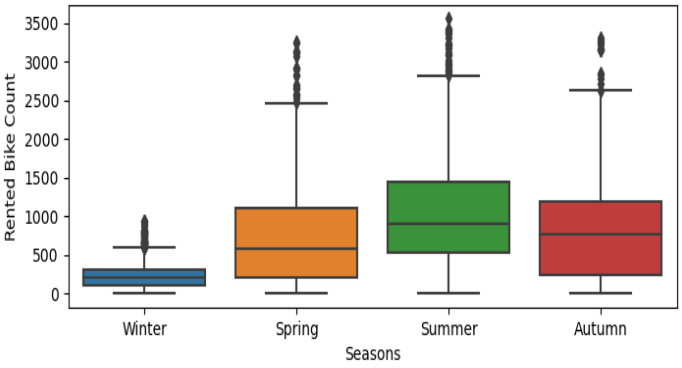
b) defining, sequentially, the effect on demand caused by the trip characteristics (travel time between traffic zones) and physical city characteristics (slopes).

6. Features of Dataset:

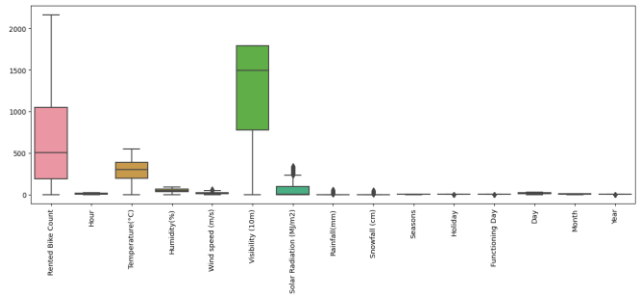
A. Month Wise Analysis



B. Season wise Analysis



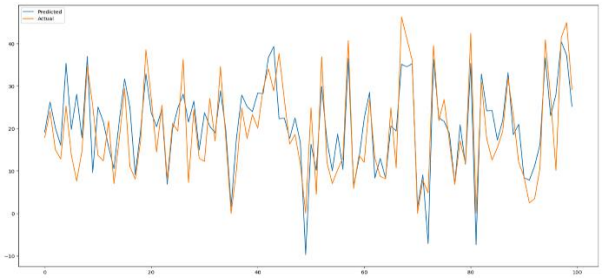
C. Checking Outliers



We see outliers in some columns like Solar radiation, Wind, Rainfall, and Snowfall but lets not treat them because they may not be outliers as snowfall, rainfall etc. themselves are rare event in some countries.

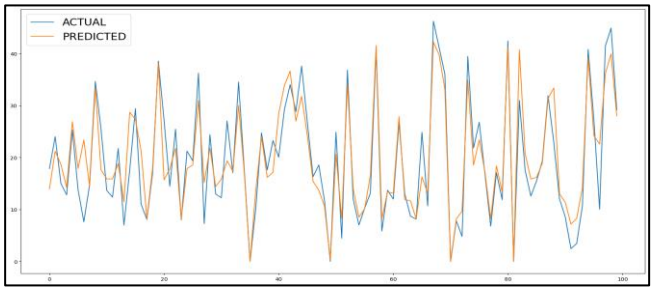
Linear regression:

Linear regression (LM) in the most simplest and nursing method, that is equated with the relationship between the Y attribute of the scalar output and one or even more X attributes of the input quantity. The case of an independent attribute is known as simple linear regression, and the method is called as multiple linear regressions when more than one independent attributes are considered. Data is designed using linear predictor functions in linear regression, and from data, the unknown model.



MSE : 122612.82171875217
RMSE : 350.16113679098106
R2 : 0.6275808915385301
Adjusted R2 : 0.6245792406931296

Random Forest Regressor:

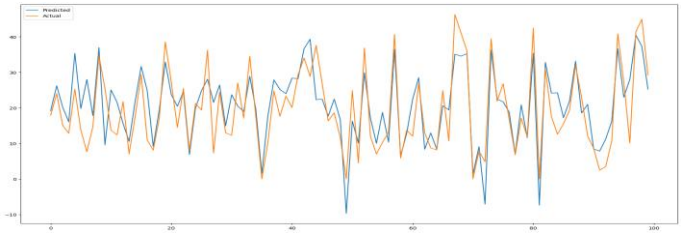



```
================Evalution Matrix==========================

MSE : 47630.46072929933
RMSE : 218.24403939008124
R2 : 0.8549059945631543
Adjusted R2 :  0.8539720561373539

================Evalution Matrix==========================
```

Lasso regression:

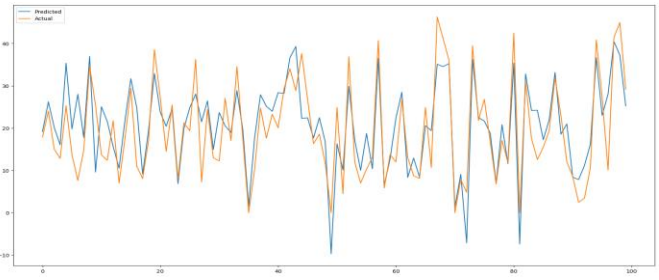


```
================Evalution Matrix==========================

MSE : 122612.82171875225
RMSE : 350.16113679098123
R2 : 0.6275808915385299
Adjusted R2 :  0.6245792406931294

================Evalution Matrix==========================
```

Ridge regression:



```
================Evalution Matrix==========================

MSE : 122707.52217188077
RMSE : 350.29633479652733
R2 : 0.6272932523028396
Adjusted R2 :  0.6242892831216305

================Evalution Matrix==========================
```

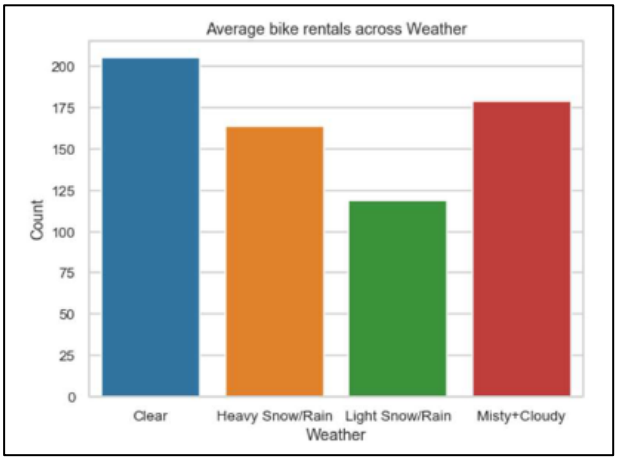
5. Benefits from Bike Sharing System.

The main reason municipalities around the world are so gung-ho about it is that bike-sharing offers a way to expand a city's transportation options and alleviate traffic at relatively low cost. It also enables people to bike to and from the bus and train, which helps cities get more value from existing public transit. • The environmental benefits of bike-sharing are equally attractive. By replacing a portion of trips that would otherwise be made by car with pollution-free bike trips, it cuts smog and global warming emissions. In addition, by lessening traffic, it reduces idling, which cuts emissions even more. Lastly, bike-sharing boosts public health by shifting people from passive to active transportation.

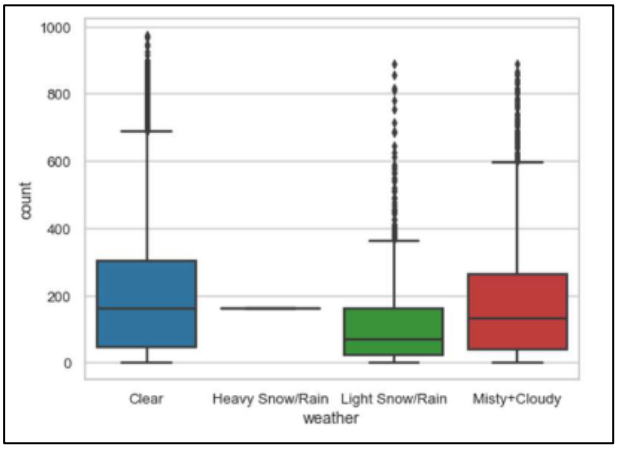
6. Steps Involved

- **Exploratory Data Analysis - Weather**

  Higher bike rental when weather is more clear and sunny.

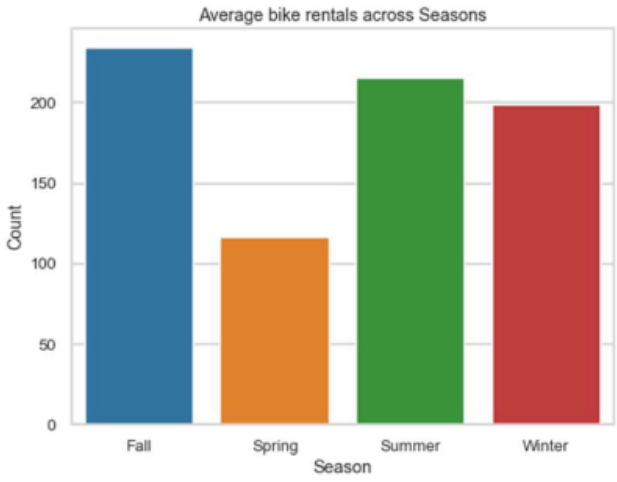


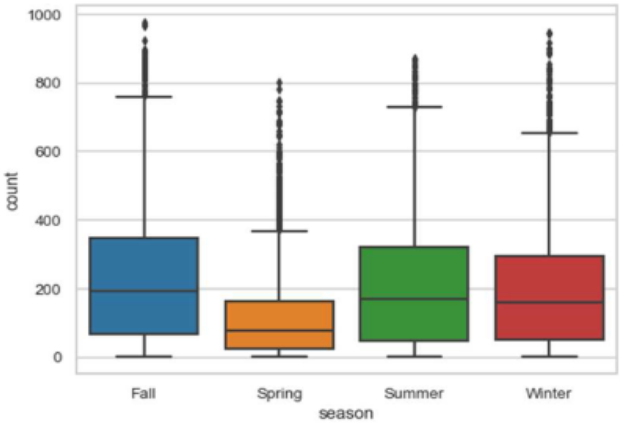
Single instance of a heavy snow /rain condition changed to light snow /rain condition.



- **Exploratory Data Analysis -Season**

  Highest bike reservations during summer (April to June) and fall (July to September) and lowest in spring (January to March)





7. Modeling.

- **Model building prerequisite**

  Standardization:

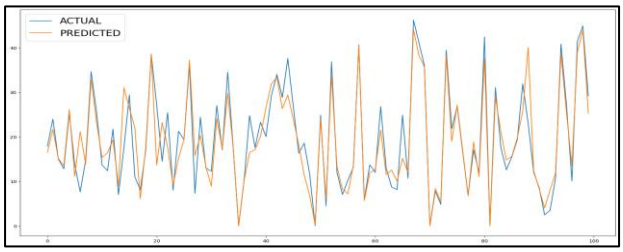
  It is a step of data pre processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

  Normalization:

  - Normalization scales our feature to a predefined range (normally the (0-1) or (-1 to 1) range), independently of the statistical distribution they follow. It does this using **minimum and maximum values** of each feature in the dataset.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Gradient Boosting

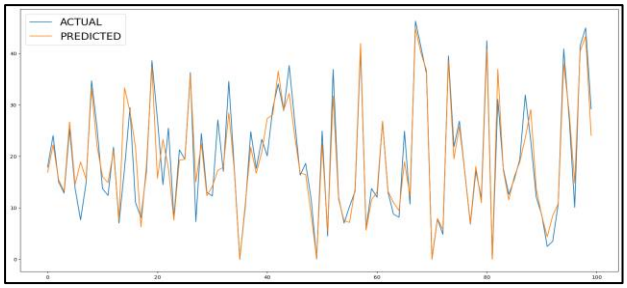



```
================Evalution Matrix==========================

MSE : 44134.998591562195
RMSE : 210.08331345340636
R2 : 0.8655540251438273
Adjusted R2 :  0.8646886257654427

================Evalution Matrix==========================
```

Extreme Gradient Boosting

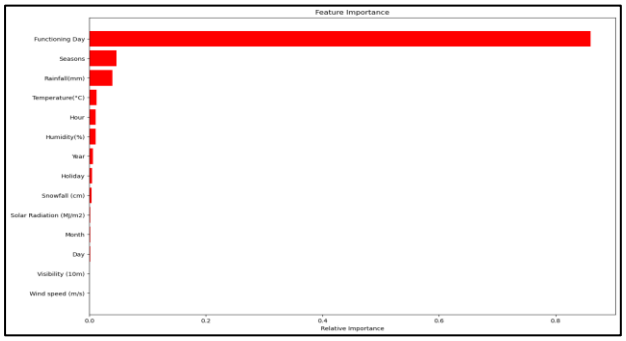


XG Boost regressor emerges as the best model according to the evolution matrix score.

```
================Evalution Matrix==========================

MSE : 32714.972084680037
RMSE : 180.8728063714389
R2 : 0.9003422124237208
Adjusted R2 :  0.8997007370094368

================Evalution Matrix==========================
```



8. Conclusion:

The independent variable in the data does not have a good linear relation with the target variable so the simple linear model was not performing good on this data. Tree based algorithm perform well in this case

Methodology

1.Majority of the analysis was EDA which was digging one level deeper and getting the data to answer the above questions

2.To answer few questions I sliced the data across various cuts

3.Another methodology included using different lenses to view data across segments

Insights

- There is a surge of high demand in the morning 8AM and in evening 6PM as the people might be going to their work at morning 8AM and returning from their work at the evening 6PM.
- After performing the various models the Gradient Boosting and Extreme Gradient Boosting found to be the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse,rmse) shows lower and (r2,adjusted_r2) shows a higher value for the Gradient Boosting and Extreme Gradient Boosting models

References

→Stack overflow
→ Kaggle…
→ Geeks for geeks