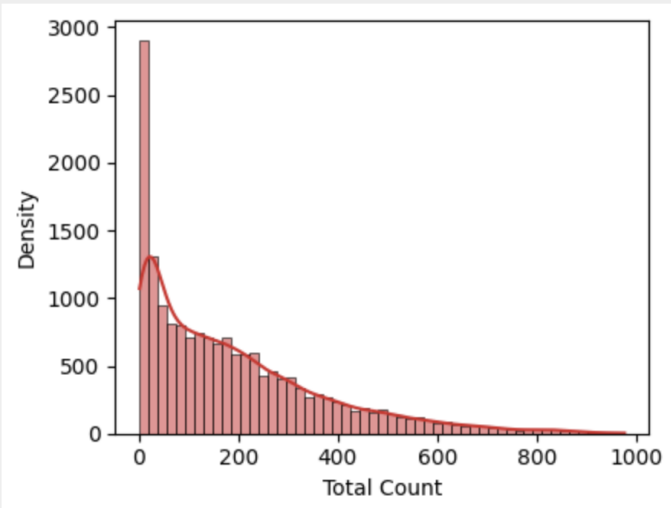
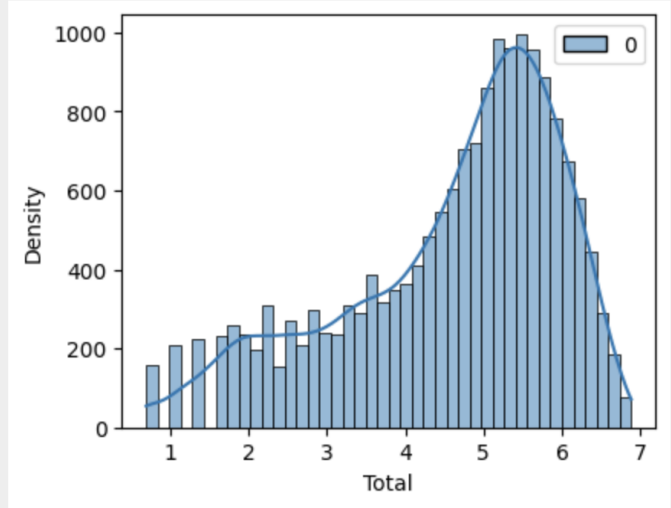
Bike Rent Per Hour

MEHDI VALINEJAD, *Student*



Density per hour.

Fig. 1. Density per hour



Density per hour transformed to normal.

Fig. 2. Density per hour transformed

***Abstract*—The current report concentrates on the data analysis on the bike rent dataset to deduct the relation of the weather, season, time of year and the other dataset information to analyze and predict the number of bikes per hour. https://github.com/TxCorpi0x/data_mining under asgn2 directory.**

***Index Terms*—rent bike, prediction, k-fold, box plot, histogram plot, scatter plot, violin plot.**

I. Introduction

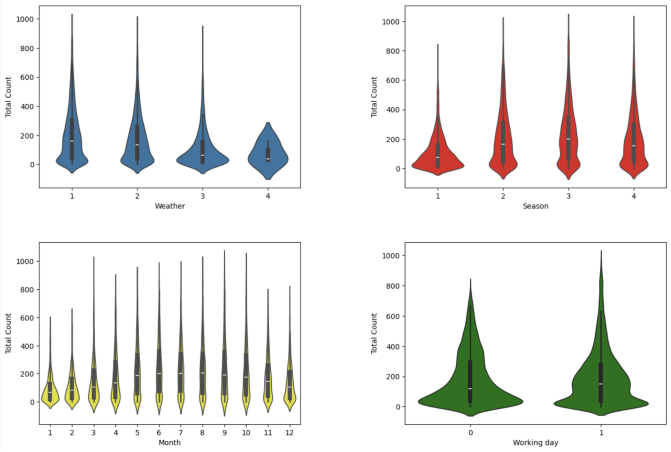
THE assumptions we have are not necessarily the correct match for a dataset, in the current report, the data analysis and predictions, do not necessarily have the same outcome, the prediction may result in a over-fitted prediction with low error value, we need to reconsider the prediction and feature engineering according to the data analysis.

II. Explanatory Data Analysis

The following histogram plot, shows the pure data of the Count per hour of the raw data. the data is skew to the right with a big skewness Fig(1). To overcome the skewness of the data, we can use data transformers based on the logarithmic calculations to convert the data to a normal distribution Fig(2).

Distribution

There is an obvious correlation between total counts of rent bikes and **weather**, where the participation increases, the amount of rent bikes decreases. it make sense for the people to prefer other types of transportation when the weather has participation or very hot. This can also inspired by the **Season** and **Month**, in the time of the year which the weather is mild, the number of rent bikes increases.



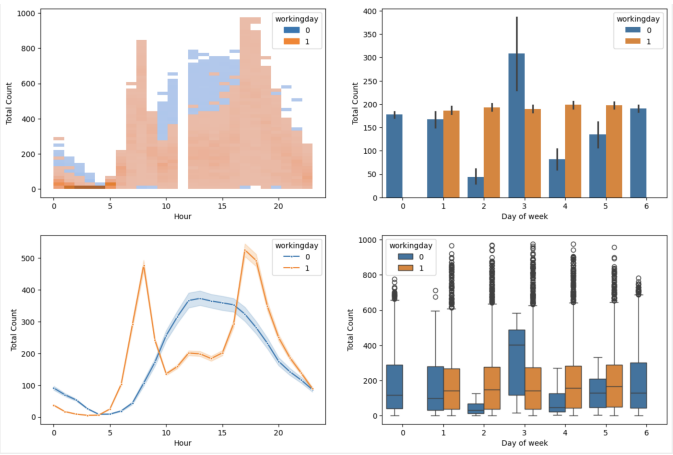
Distribution.

Fig. 3. Distribution violin plot

The amount of bike rental is also a function of **working days**, in the weekends or vacations, the number of rent biked decreases which means people which use it for commuting are not using it in the vacations. but the correlation between these features are much less than the weather condition Fig(3).

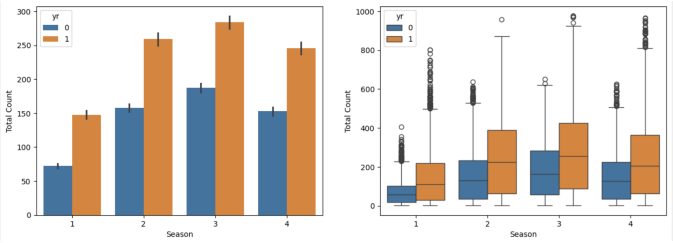
Bikes Per Hour

- **Weekday:** In the working days, rent bike count is higher thant he weekends, and specifically,in the commuting time range it increases meaningfully. A minor correlation



Bikes per Weekday.

Fig. 4. Weekday



Bikes per season.

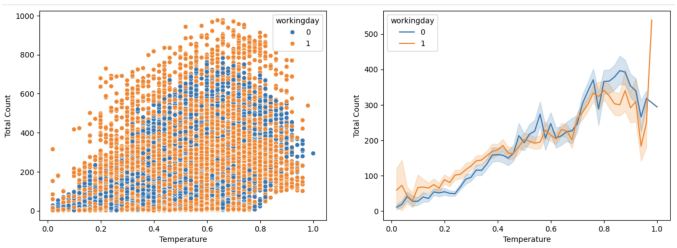
Fig. 5. Season

is visible between the bike rend count and the day of the week. this means that event there is a correlation between working days and non-working days, it is a very small correlation. In fact, users, choose a different time of the day to rent a bike during the working days Fig(4).

- **Season:** There is a positive correlation between the season and the number of bikes rent. in the following plot, we can obviously see that the second year has more bikes rent ber hour, this means the overall increment in the popularity of bike renting program Fig(5).
- **Temperature:** There is a positive correlation between temperature and bikes rent, people prefer to use bikes in a warm-mild weather instead of Hot or cold. Fig(6).
- **Humidity:** There is a negative correlation between number of rent bikes and humidity, as it increase, the number of rent bikes decreases. Fig(7).
- **Wind Speed:** There is a negative relation between rent bikes and wind speed, it also has different behavior in different seasons, where the wind speed between 0.4 and 0.6 has a different bike usage rate in th different seasons, in the warmer seasons, the usage is much more than cold seasons. Fig(8).

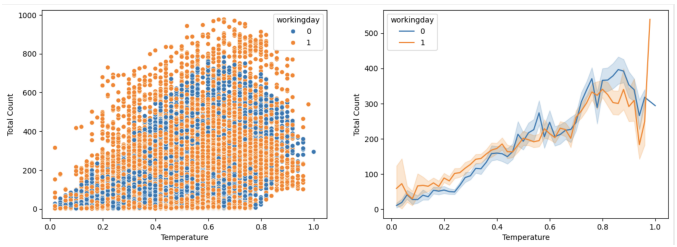
III. Prediction

The prediction of the total count of bikes rent per hour can be done by feature engineering and leverage the models such as Knn, Random Forest, Decision Tree and Gradient



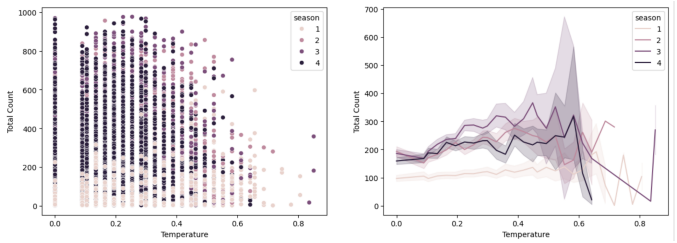
Bikes per temperature.

Fig. 6. Temperature



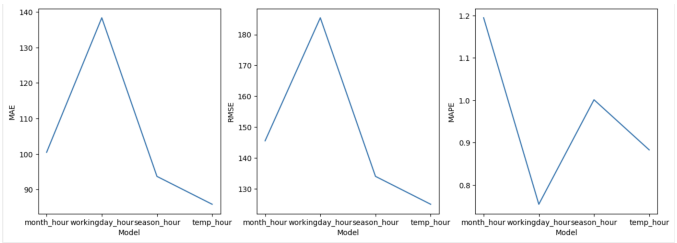
Bikes per humidity.

Fig. 7. Humidity



Bikes per wind speed.

Fig. 8. Wind Speed



Feature Engineering for Knn.

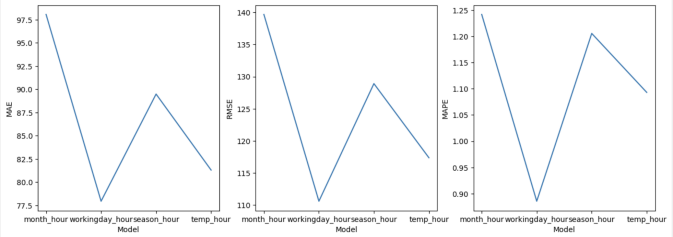
Fig. 9. Knn

Boosting. one of the most important steps to be considered in the prediction, is to choose which feature to include in the prediction.

The used 10-Fold used for the prediction and the result have been added to the dictionary of the results and finally the comparison have been done between the all results.

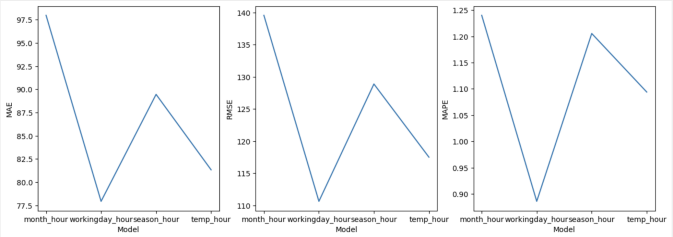
Conclusion

Using various prediction models, the lowest error occurs when considering the **working day** and hour as features. However, this deduction is not generally correct. In the data



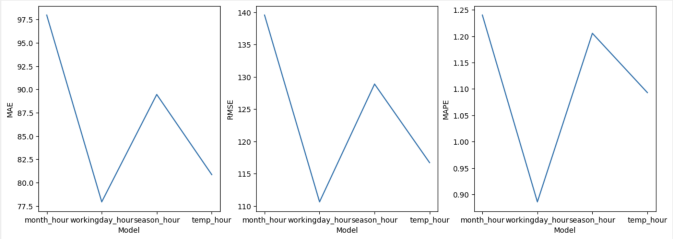
Feature Engineering for random forest.

Fig. 10. Random Forest



Feature Engineering for decision tree.

Fig. 11. Decision Tree



Feature Engineering for Gradient Boosting.

Fig. 12. Gradient Boosting

analysis section, the found result is that there is not a strong correlation between working days and the number of rented bikes. Therefore, using this feature would lead to over-fitting, as it lacks relevance to the bike count. Instead, a better choice would be to focus on the **temperature**. During the data analysis phase, it was clear that there is a significant positive or negative correlation between weather, temperature, and seasonal variations in bike rentals.

Mehdi Valinejad Received the B.S. degree in industrial engineering from the Azad University (South Tehran Branch) in 2012, and is currently working Master's. degree at the University of Bahcesehir at Istanbul.