**SEOUL BIKE SHARING DEMAND PREDICTION**

**Shashank Kumar Yadav, Amitha K, Mohd Khalid, Apoorva Vishwakarma, Abhishek P Sadawarte**

**Data science trainees, Alma Better, Bangalore**

**Abstract:**

A Bicycle-sharing, Bike-sharing programs, public bicycle schemes or public bike share PBS scheme is a shared transport service where bicycles are available foe shared use by individuals for a short-term at low or zero cost.

The programs themselves include both docking and dock less systems, where docking systems allow users to borrow a bike from a dock, i.e., a technology-enabled bicycle rack and return at another node or dock within the system and dock less systems, which offer a node-free system relying on smart technology. In either format, systems may incorporate smartphone web mapping to locate available bikes and docks.

**1. 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.
* The main objective behind this project is to explore and analyze data and to predict the demands in a bike sharing system using machine learning methods. And secondary objective is to analyze the collective inter-station and time varying effect on predicting the demands for a bike station.

**2. Data Summary**

The dataset contains different columns of variables important of bike sharing demand prediction. Dependent variable is ‘Rented bike count”

* 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)

**3. Introduction**

This project presents a rule-based regression predictive model for bike sharing demand prediction. In recent days, public rental bikes sharing is becoming popular because of increased comfortableness and environmental sustainability. Data used include Seoul Bike and have data associated with it for each hour. It is important to compare the results of conventional machine learning algorithms to signify the importance pf positives and negatives of each method considerations in the study. There are multiple algorithms used, which must be optionally applied for every scenario. Therefore, five prediction algorithms were considered in this project to compare their performance with each other.

**4. Steps Involved**

* **Exploratory Data Analysis**

After loading the dataset, we performed this method by comparing our target variable that is Rented Bike Count with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

Our dataset contains a no null values so we do not have to concerned about missing values.

* **Outlier Treatment**

Our dataset does not contain any outliers so we do not have to worry about outliers.

* **Explore our Numerical columns**

1. **Skewness** - We have seen some of the features are skewed positively or negatively. we should treat skewness as it will mislead the results while applying algorithms.

2. **Correlation** - we are able to see this temperature(°C) and dev point temperature(°C) column are highly correlated i.e. 0.91.

We need to drop this column then it will not affect the outcome of our analysis and also having same variation.

3. **Multicollinearity** – we have seen there is high multicollinearity between columns and also having high VIF and affects the results.

* **Encoding of categorical columns**

We used One Hot Encoding to produce dummy variable which uses binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

Our categorical values are Hour, Seasons, Holidays, Functional days etc.

* **Normalization of features**

Our dependent variable ‘Rented bike count” is right skewed so it needs to be normalized and the methods used for normalization is log10, square, square root.

**5. Model Training**

* **Test Train Split**

We need to split the data into train and test data for estimating the performance of machine learning algorithms. We split in 75-25 % of the dataset.

* **Fitting different models**

For modelling we tried various regression algorithms like:

1. Logistic Regression
2. Lasso Regressor
3. Ridge Regressor
4. Decision Tree Regressor
5. Random Forest Classifier

* **Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree-based models.

1. Gradient Boosting
2. Ada Boosting

3. XG Boosting

* **Evaluation metrices**

1. Mean Absolute error
2. Mean Squared error
3. Root Mean Squared error
4. R Squared
5. Adjusted R Squared

**6. Conclusions**

* In holidays or non-working days there is demand in rented bikes.
* There is a surge of high demand in the morning 8AM and in evening 6PM as the people might be going to their office at morning 8AM and returning from their office at the 6PM.
* People preferred more rented bikes in the morning compared with evening.
* When the rainfall was less, people have booked more bikes except some few cases.
* The temperature, Hour are the most important features that positively drive the total rented bikes count.
* It is observed that highest number of rental bikes counts in Autumn and summer seasons and the lowest in winter season.
* We observed that the highest number of rental bikes counts on a clear day or little windy day and the lowest on a snowy and rainy day.
* In the given dataset there was no strong relationship present between dependent variable "Rented bike count" and independent variables.
* Out of all models we apply Decision tree and Random Forest model are most accurate and the reason is there are no specific relation between features.
* Random Forest worked best in predicting the count of rented bikes as its R2 score is maximum from the tried model.
* We are getting best results using Gradient Boost Regressor.