**BIKE SHARING DEMAND PREDICTION**

**Sampanna Mishra**

**Data science trainee**

**AlmaBetter, Bangalore**

**Abstract:**

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated throughout a city.Predicting bike sharing demand can help bike sharing companies to allocate bikes better and ensure a more sufficient circulation of bikes for customers. This project forecasts bike rental demand which involves predicting bike renting and returning in different areas of a city during a future period based on historical data, weather data, and time data.

Statistical regression models were trained with different ML algorithms, their best hyperparameters using repeated cross-validation and the performance is evaluated using a testing set.In this regression analysis feature selection, data analysis and prediction with machine learning algorithms carried out taking into account previous trends to determine the correct rented bike demand.

***Keywords:machine learning,bike demand,regression model,cross-validation***

**1.Problem Statement**

In the problem the dataset provided is the Seoul Bike Sharing Demand.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 kiosks across the city.

The main objective is to build a predictive model, which could help them in predicting bike count required at each hour for the stable supply of rental bikes.The features present in the dataset are as follows:

### Date: days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY

* Rented Bike count - Count of bikes rented at each hour
* Hour - The hour of the day, starting from 0-23 it's in a digital time format, type : int, we need to convert it into category data type.

### Temperature-Temperature in Celsius

### Humidity - Humidity in the air

### Wind Speed - Speed of the wind in m/s

### Visibility - Visibility in m

### Dew point temperature - Temperature at the beginning of the day

### Solar radiation - Sun contribution (MJ/m2)

### Rainfall - Amount of raining in mm

### Snowfall - Amount of snowing in cm

### Seasons - Season of the year

* Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

### Holiday -If the day is a holiday period or not.

**2. Introduction**

Hyper urbanization, coupled with decentralization, has caused a host of transportation problems: gridlock, increasing travel demand,tailpipe emissions, and decreasing accessibility.Hyper-motorization and expanding urban form have contributed to problems of congestion and degrading levels of transit. While in recent years, dozens of cities have implemented bike share systems in efforts to mitigate some of these problems.

### The goal here is to build a predictive model, which could help bike count required at each hour for the stable supply of rental bike

**3. Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset, I 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,trends 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.Data used include weather information Functional Day,

Temperature,Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall,Holiday the number of bikes rented per hour and date information

* **Null values Treatment**

This dataset does not contain null values.Having this might tend to disturb our accuracy.Methods like isnull() and notnull() functions from the pandas library used to determine null values.

* **Encoding of categorical columns**

Used One Hot Encoding to produce 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.Features like

Holiday,Functioning Day,Season were encoded.

* **Feature Selection**

Correlation matrix has been used and heatmap plotted using seaborn library.Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features.It also depicts results of each feature i.e which feature is more important compared to the model and which is of less importance.

* **Outlier Treatment**

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one important task is to detect and filter these outliers. The target variable Rented Bike Count had outliers which were removed using square root transformation.

* **Standardization of features**

The main motive through this step is to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal is to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

* **Fitting different models**

For modeling I tried various regression algorithms like:

1. **Linear Regression**
2. **Lasso Regression**
3. **Ridge Regression**
4. **ElasticNet Regression**
5. **Decision Tree Regressor**
6. **Random Forest Regressor**
7. **Gradient Boosting Regressor**

* **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 like Gradient Boosting using GridSearchCV

* **Regularization Technique**

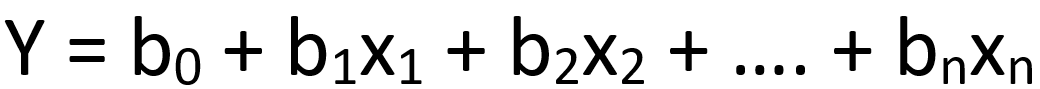
To enhance the performance of Linear Regression models specifically three regularization techniques have been used. These Lasso(L1), Ridge(L2),ElasticNet regularization techniques have been used.

**4. Algorithms:**

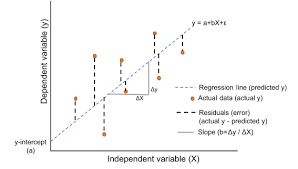
1. **Linear Regression:**

As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable of interest).

The function used in Linear Regression is:



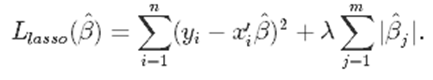
Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable.



.

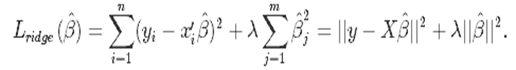
1. **Lasso Regression:**

It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso:



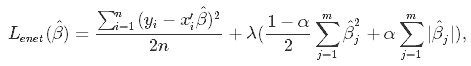
1. **Ridge Regression:**

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issueof multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.



1. **ElastiNet:**

The benefit is that an elastic net allows a balance of both penalties, which can result in better performance than a model with either one or the other penalty on some problems. The elastic net adds a quadratic part to the penalty.The solution is to combine the penalties of ridge regression and lasso to get the best of both worlds. Elastic Net aims at minimizing the following loss function:

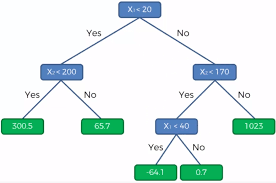


where *α* is the mixing parameter between ridge (*α* = 0) and lasso (*α* = 1)

1. **Decision Tree:**

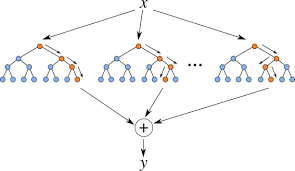
A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.



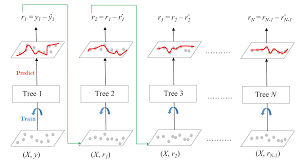
1. **Random Forest:**

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.

****

1. **Gradient Boosting:**

Gradient boosting is a machinelearning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.Its a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error..

****

**5. Hyperparameter Tuning:**

**GridSearchCV with Gradient Boosting:**

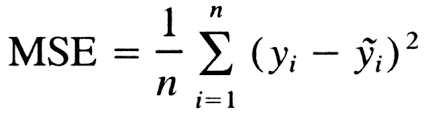
Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters.In this project I have used GridSearchCV with Gradient Boosting Algorithm.Its advantage is that it is a simple technique that will go through all the programmed combinations

**6. Model performance:**

Model can be evaluated by various metrics such as:

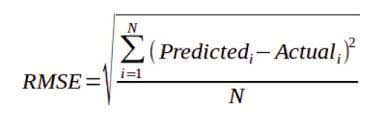
1. **Mean Square Error**-

The simplest and most common loss function is the difference between your model’s predictions and the ground truth, square it, and average it out across the whole dataset.The MSE will never be negative.

.

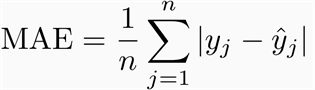
1. **Root Mean Square Error-**

Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.



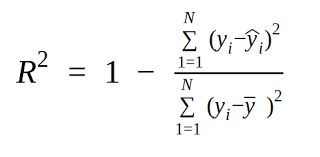
1. **Mean Absolute Error-**

To calculate the MAE, take the difference between the model’s predictions and the ground truth, apply the absolute value to that difference, and then average it out across the whole dataset.The MAE, will never be negative since in this case we are always taking the absolute value of the errors.



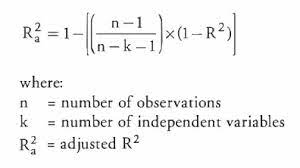
1. **R2 Score**-

is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted. It works by measuring the amount of variance in the predictions explained by the dataset.



1. **Adjusted R2 score-**

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit**.**

****

**8. Conclusion:**

Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building.The accuracy and performance has been compared between the models using Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 and Adjusted R2

Considering the R2 score following results can be concluded.Out of all the models Random forest gave the highest r2 score of 89 %.Also, there is much improvement in r2 score after hyperparameter tuning on gradient boosting.

So the r2 score of our best model is 90% which can be said to be good for this large dataset.

**References-**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya
4. Towards Data Science
5. ai-ml-analytics