**BIKE SHARING DEMAND PREDICTION**

**Navneet Keshri**

**Data Science Trainee**

**AlmaBetter**

**Abstract:**

A rule-based regression predictive model for predicting demand for bike sharing is presented in this technical documentation. A bike-sharing system has positive effects on the user and the environment and gives people a sustainable mode of transportation. In recent days, Pubic rental bicycle sharing is becoming well known on account of its expanded ease and natural maintainability. The Seoul Bike data were utilized in this study. Each hour's worth of data is linked to weather information. We train and optimize hyperparameters for the dataset using cross-validation regression models, and the testing set is used for evaluation. The prediction accuracy of the regression models is evaluated using a variety of evaluation indices, including R2 and the Root Mean Square error. The time interval used to transform the data influences the model's performance.

**Introduction:**

The expanded use of private vehicles in metropolitan regions has brought about a critical ascent in fuel utilization that affect the environment. People in today's society have come to accept issues like traffic jams as normal. As a result, to deal with the problem, the government and other organizations began implementing strategies to encourage sustainable development.

There are bike-sharing programs in a lot of countries, like the bike-sharing program in South Korea, which started to solve all of these problems and make Seoul a healthier place to live. To address the issue of public transportation, the Bike Share initiative was launched in that setting. It offered the populace a low-cost alternative to a sustainable mode of transportation for short distances. and allowed individuals to use the service on their own. A user can borrow a bike from any bike station and return it to a bike station near their destination in a bike-share system, which has positive health effects because it involves pedalling the bike. Additionally, more bike-friendly areas were made accessible thanks to the city-wide installation of bike stations. Computerized stands that are used to pick up and drop off rental bikes are known as docking stations. At any docking station, public bike users can rent and return rental bikes. At My Page, users can check the details of their trip (distance, duration) and the measure of their physical activities (calories burned).

Rental bikes are becoming more and more popular daily because of their smart technology and ease of use. As a result, it is necessary to control the demand for bike rentals and ensure that users have access to a consistent and convenient service. A method based on data mining that incorporates weather data is proposed in this study for predicting public bike demand across the city. Predicting the number of rental bikes needed each hour is done with a rule-based model.

**Problem Statement:**

At present Rental bikes are presented in numerous metropolitan urban communities for the upgrade of portability solace. Because it shortens the number of time people has to wait, the public must have access to the rental bike at the right time. In the end, giving the city a steady supply of rental bikes turns into a main pressing issue. The most important part is to predict the number of bikes needed each hour to maintain a steady supply of rental bikes.

The primary objective is the creation of a predictive model that may assist them in anticipating bike demand in advance. This will help them in a stable supply of bikes in any place required.

The primary objective is the creation of a predictive model that could assist them in anticipating surge pricing types. They would then be able to match the right cabs with the right customers more quickly and effectively thanks to this.

**Data Description:**

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information.

## **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:** Non-Functional Days, Functional Days

**Breakdown of Datasets:**

We need to complete the following actions before moving on to the data visualization:

1. Import necessary packages for upcoming analysis.
2. Reading data files from Google Drive and mounting the drive.
3. Removing future seaborn warning plans.
4. Displaying each DataFrame column by column.
5. View all information in the data.
6. Drop any duplicates that are found.
7. Examining the percentage of null values, data types, and unique values.
8. Data filtering.
9. Categorical and numerical data are separated.

**Examining Null/Missing Values:**

The actual data frequently contains numerous missing values. Missing values could be brought on by corrupted or missing data. During the pre-processing of the dataset, the treatment of missing information is vital because many AI calculations don't uphold missing qualities. As a result, we begin by looking for missing values.

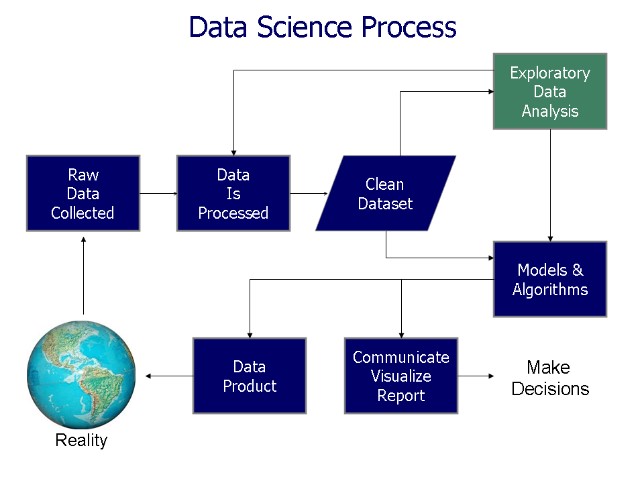
In deep learning and machine learning, the issue of null values is a major one. Assuming you are utilizing sklearn, TensorFlow, or some other AI or profound learning bundles, it is expected to tidy up invalid qualities before you pass your information to the AI or profound learning system. If you don't, it will display a lengthy and unattractive error message. As a result, we search for empty or missing values. The provided dataset contains no null or missing values.

**Data Cleaning:**

The first step in any data science project is to clean the data. Although not all of the data is clean, the majority is useful. Identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data is what is meant by the term "data cleaning." Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. To begin cleaning the data, we first look for duplicate values in the dataset and find that none exist. After that, we convert data types and perform exploratory data analysis to determine the dataset's best-fit model.

## **Exploratory Data Analysis and Visualization:**

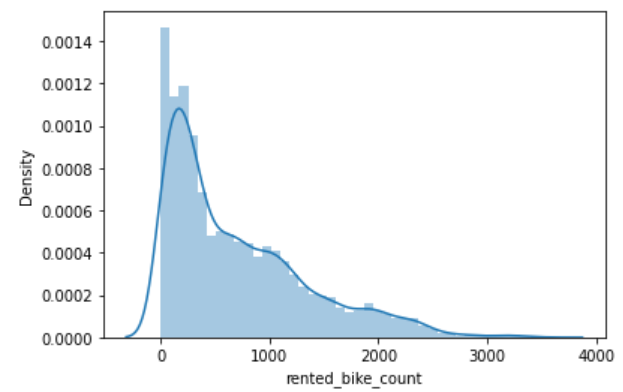
## Exploratory data analysis (EDA) is a method in statistics that uses statistical graphics and other data visualization techniques to summarize the main characteristics of data sets. A measurable model can be utilized or not, yet basically, EDA is for seeing everything that the information can say to us past the proper demonstration and in this manner contrasts conventional speculation testing. We were able to determine various aspects and relationships between the target and the independent variables with the assistance of EDA.



* **Observation 1: Positively Skewed Dependent Variable**

When one tail is longer than the other, there is a skewed distribution. A distribution's asymmetry is referred to as its skewness. These distributions are asymmetric, in contrast to the familiar normal distribution with its bell-shaped curve. Because the data are not evenly distributed on both sides of the distribution's peak, the two halves are not mirrored images.

Long tails on the right side of a right-skewed distribution are called right-skewed distributions. They are also referred to as positively skewed by analysts. Because probabilities decrease more slowly with increasing values, this condition arises. As a result, extreme values far from the peak are more likely to be found on the high end than on the low end. Our dependent variable rented\_bike\_count follows a positively skewed distribution.

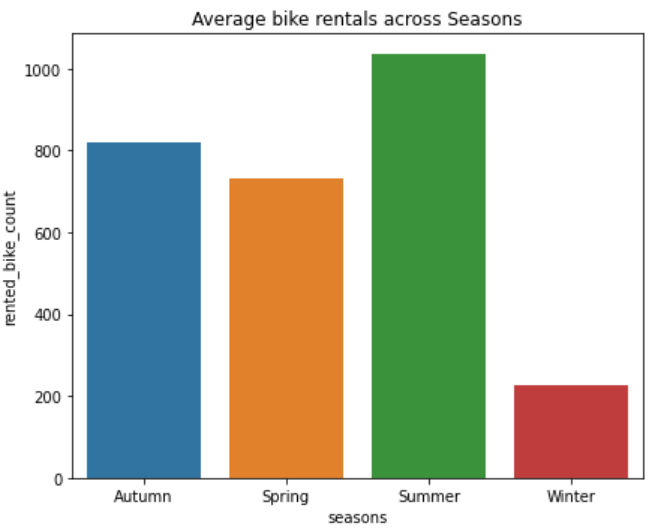


* **Observation 2: People prefer warm seasons more than cold ones**

Summer has the highest rented\_bike\_count preference, while winter has the lowest, indicating that people prefer to rent bikes in warm weather.

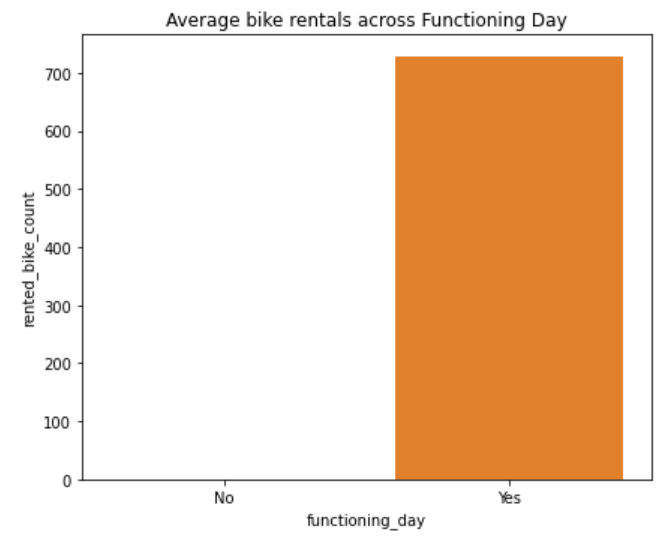
When compared to the summer and autumn seasons, bike reservations are lower in the spring.

There are numerous outlier points for specific seasons. This is most likely because of the day-to-day variation in distribution.



* **Observation 3: There is no demand when the day is not functioning**

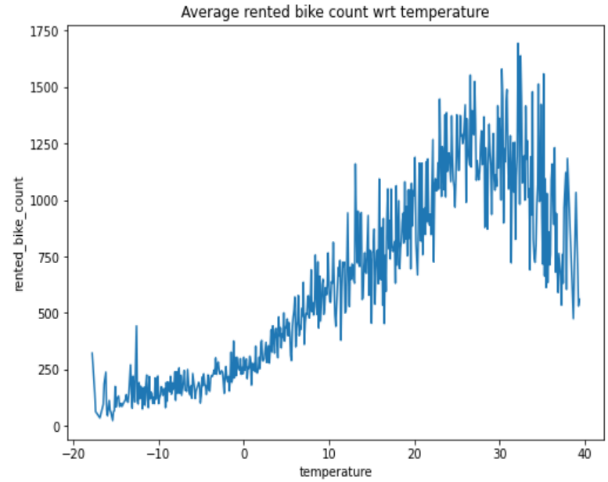
I used a bar chart to show the number of rented\_bike\_count by functioning\_day. Bike-sharing services are in high demand during functioning days, which could be because many customers use these bikes to get to work. Renting bicycles is not in demand during non-functioning days.



* **Observation 4: Warm temperature is mostly preferred**

I chose a line chart to display the temperature column because the temperature is a continuous variable and we must determine the ideal temperature for rented\_bike\_count.

The line plot above demonstrates that as the temperature rises, the average number of bikes rented decreases slightly at the highest temperature.

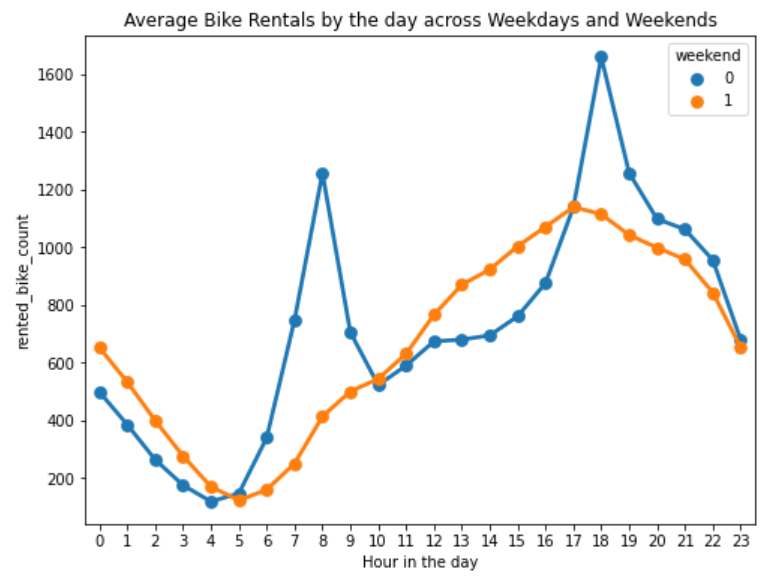


* **Observation 5: Hourly distribution follows 2 types of rent distribution**

Point charts are useful for clearly displaying quantitative data. Plotting data along an ordinal axis with point charts requires multiple points. A point chart is equivalent to a line chart without lines.

**Working Hours**: The first pattern is one in which rentals reach their highest point around 8 a.m. and their lowest point around 5 p.m. These are local bikers who are employed and typically go to work Monday through Friday.

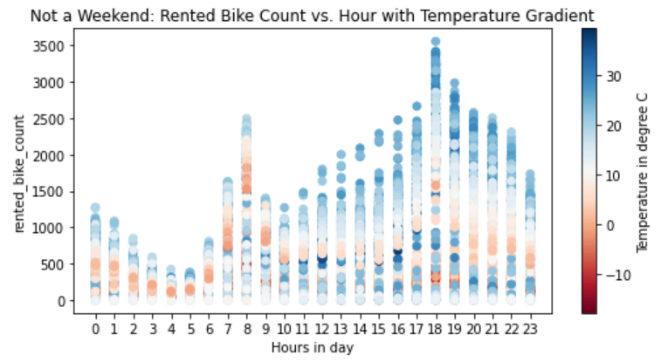
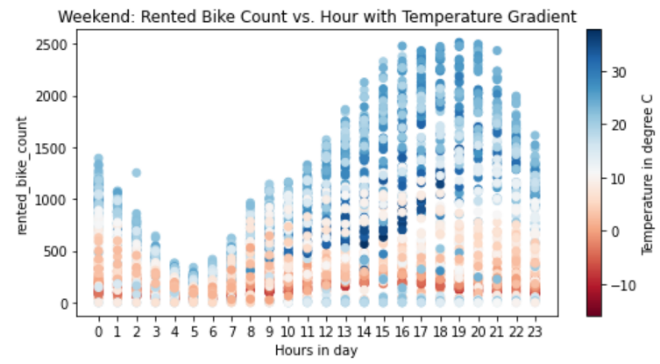
**Non-Working Day:** Second pattern where there are more or less uniform rentals across the day with a peak at around noon time. These correspond to probably tourists who typically are casual users who rent/drop off bikes uniformly during the day and tour the city on nonworking days which typically are Saturday and Sunday



* **Observation 6: Temperature, hour, and weekday are the most important features**

For multivariate analysis, we understand the relationship between various variables with the help of a scatter plot.

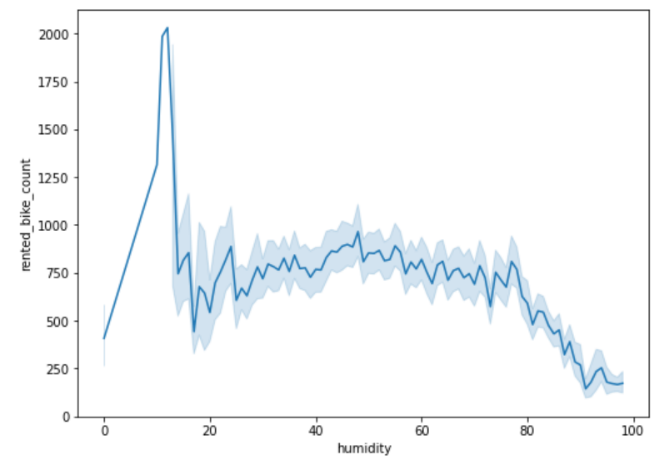
As can be seen from the previous, biking is preferred by a greater number of people in temperatures between moderate and high; however, if the temperature is too high, the count will slightly decrease (the darkest of the blue dots).

****

* **Observation 7: Humidity can be a deciding factor in high demand**

I chose a line chart to monitor the humidity column because the humidity is a continuous and numerical variable.

The line plot above demonstrates that the average number of bikes rented during humid conditions fluctuates significantly. The most desired humidity range is between 0 and 20 for the number of rented bicycles in demand. When the humidity is above 20 and 80 percent, there is a significant drop in demand for the rented bike.



**Advantages of Visualization:**

* The processing of visual data is quicker and simpler.
* Better experiences of the information are drawn which might be missed in traditional reports.
* Assists us with envisioning patterns that improve performance.
* Sales and productivity both rise when data are visualized.

**Outlier Analysis:**

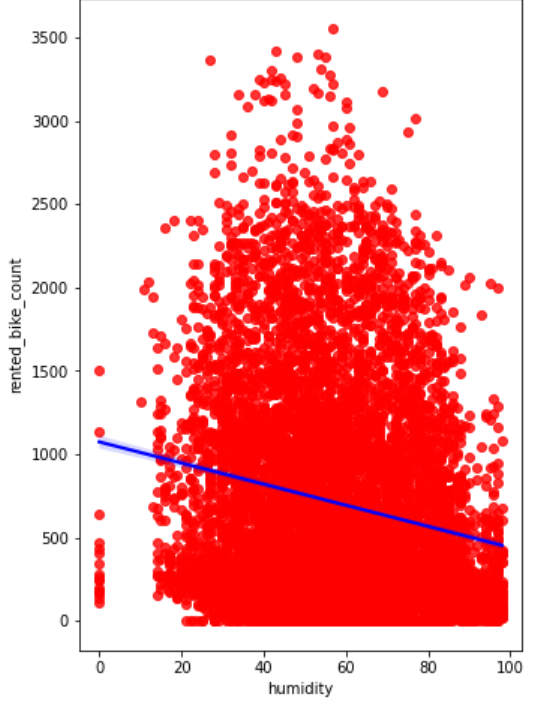
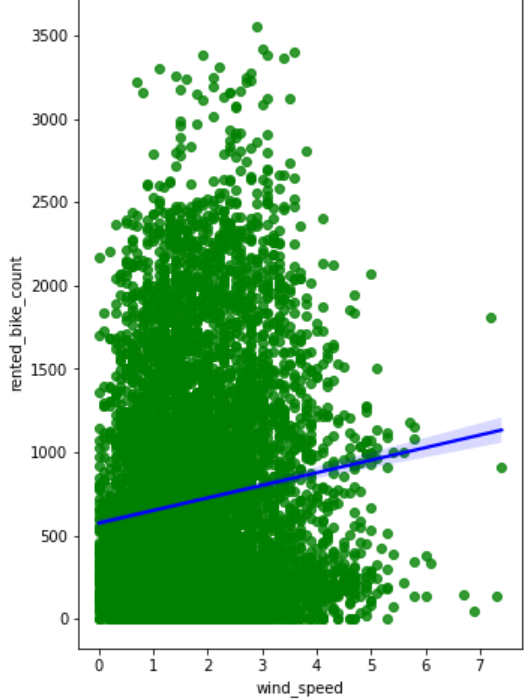
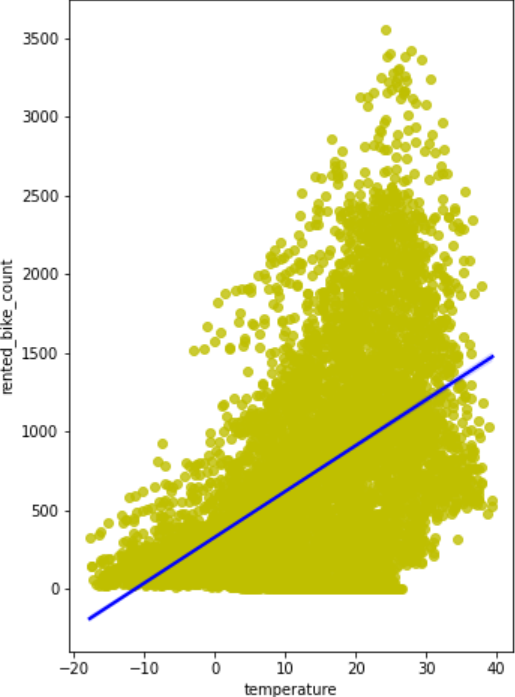
* **Z score > 4 Pruning**

The number of standard deviations from the mean is represented by the Z-score scaling variation. To ensure that your feature distributions have mean = 0 and std = 1, we would use the z-score. When there are a few outliers that aren't so extreme that clipping is necessary, this method is useful.

The following is the formula for determining a point's z-score:

* **Correlation Analysis**

Seaborn regression plots serve primarily as a visual guide for highlighting patterns in a dataset during exploratory data analyses. Regression plots help to illustrate the linear relationships between two parameters by drawing a regression line between them, as the name suggests.

As a result, I used regression plots to determine the relationship between wind speed, humidity, and temperature.

* **Heatmap**

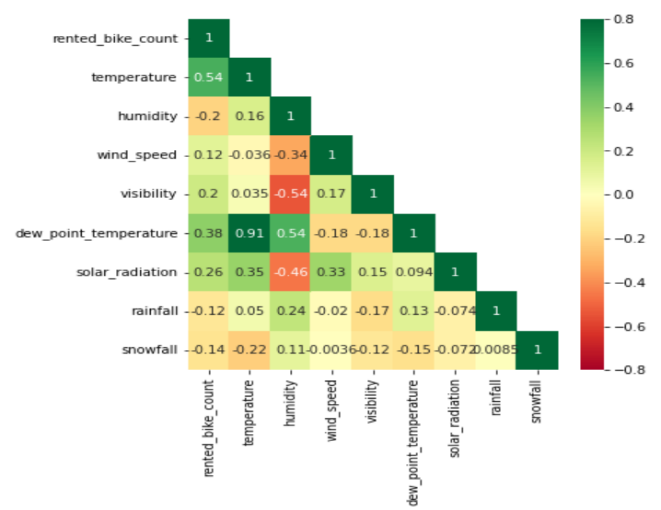
Density can be displayed using a heat map. When conducting analysis, it also makes it simpler to visualize the relationship between the variables and the dependent variable. The linear relationship between all of the variables and the dependent variable was found with the help of a heat map.

**Very Highly Correlated (0.7 - 0.9):** temperature and dew\_point\_temperature are very highly correlated as expected.

**Moderately Correlated (0.5 - 0.7):** We see a moderate correlation between humidity and dew\_point\_temperature and temperature and rented\_bike\_count. This is probably only true for the range of temperatures provided.

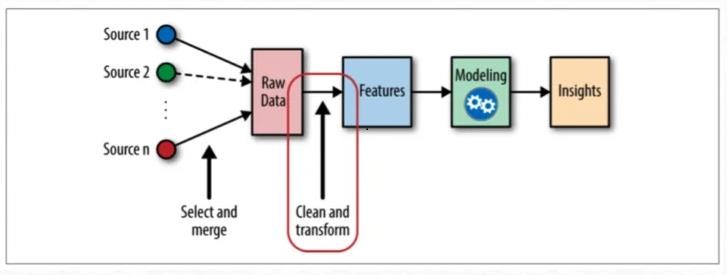
**Negative Correlation (less than 0):** We see a negative correlation between visibility and humidity and solar radiation and humidity. The more the humidity, the fewer people prefer to bike.

**Low Correlation (near zero):** rented\_bike\_count has a weak dependence on windspeed, snowfall, and rainfall.



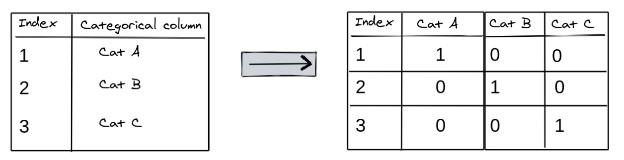
**Feature Engineering & Data Pre-processing:**

Using statistical or machine learning methods, the process of transforming raw observation into desired features is known as feature engineering. Our dataset's manipulation—including addition, deletion, combination, and mutation—to enhance the training of machine learning models is referred to as "feature engineering." A solid understanding of the business problem and the available data sources underpins effective feature engineering.



* **One Hot Encoding:**

In machine learning, one-hot encoding is used to quantify categorical data.



* **Label Encoding:**

The process of changing the labels into a numerical form so that they can be read by machines is known as label encoding. The operation of those labels can then be better decided by machine learning algorithms. In supervised learning, it is a crucial preprocessing step for the structured dataset.

Autumn

spring

Summer

Winter

0

1

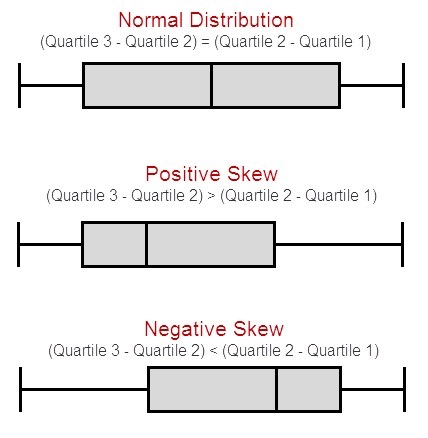
2

3

* **Target Variable Normalization:**

The feature of the dataset that you want to learn more about is the dataset's target variable.

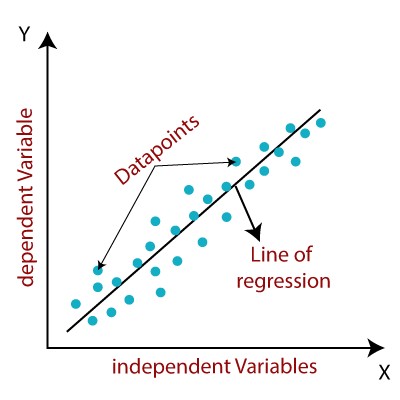
In our information, the segment 'rental\_bike\_count' contains the worth we want to foresee for example the objective variable is 'rental\_bike\_count'. The Rental Bike Count distribution analysis plot is positively skewed, and this is the target variable.



When your data are moderately skewed, the square root method is typically used. Now, a transformation that has a moderate effect on distribution shape is using the square root, such as sqrt(x). Typically, it is used to reduce data with the right skewness. Last but not least, counted data are the most common application for the square root, which can be used on zero values.

**Fitting different ML Models:**

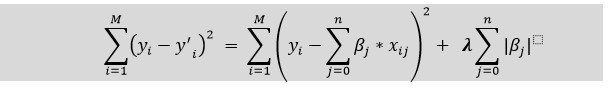
1. **Linear Regression –** One of the most common and simple Machine Learning algorithms is linear regression. The predictive analysis makes use of this statistical technique. LR makes predictions for both numerical and continuous variables.



The relationship between a dependent variable and one or more independent variables is depicted by linear regression. An LR line is defined by the equation: 𝑌=𝑎+𝑏𝑋

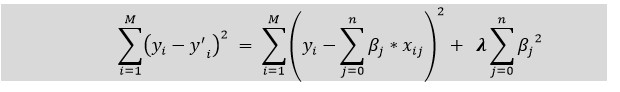
It is finished by fitting a direct condition of line to the noticed information. Checking whether there is a connection between the variables or features of interest, which is supposed to use numerical variables, i.e., the correlation coefficient, is more important for fitting the model.

1. **Lasso Regression –** Least Absolute Shrinkage and Selection Operator is LASSO. Lasso regression seeks to identify the subset of predictors for a quantitative response variable with the lowest prediction error. To accomplish this, the lasso places a constraint on the model parameters, causing the regression coefficients for some variables to decrease until they are close to zero. Additionally, it serves as L1 regularization. The following is the Lasso regression cost function equation:

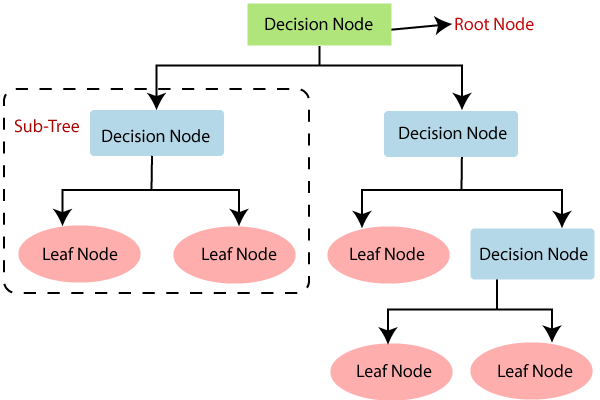


1. **Ridge Regression –** Ridge regression is a model strategy utilized to examine any information that experiences multicollinearity and diminishes the model's intricacy. When multicollinearity arises, least-squares are impartial, and variances are substantial, the predicted values diverge significantly from the actual values. It is likewise utilized as L2 Regularization.

In ridge regression, the cost function will be solved as follows:

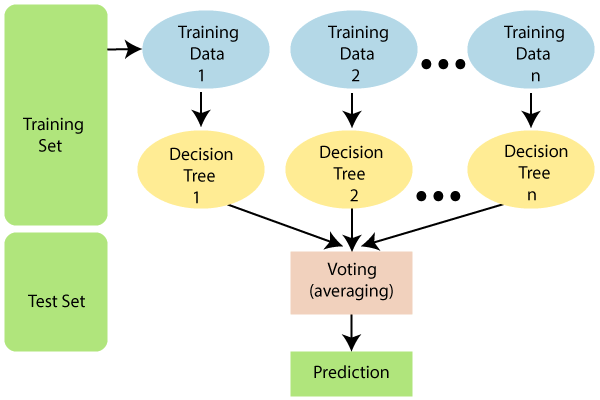


1. **Decision Tree –** Data mining employs a Decision Tree, a supervised learning technique, for classification and regression methods. A tree helps us with dynamic purposes. The decision tree is steadily developed while a data set is divided into smaller subsets. The decision nodes and leaf nodes make up the final tree.

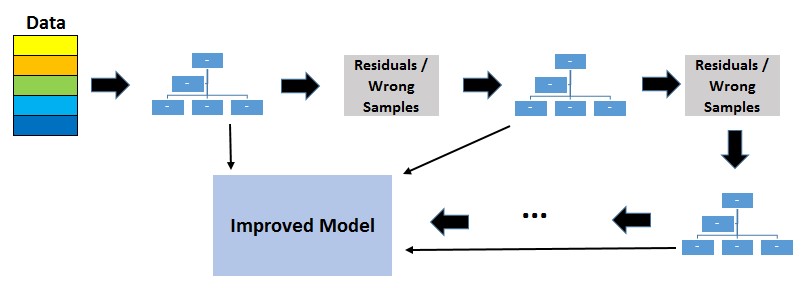


1. **Random Forest Regression –** The Random Forest Decision Tree Algorithm is a bagging version of the Decision Tree Algorithm. It creates several decision trees from a randomly selected subset of the training set, collects the labels from these decision trees, and then averages the final prediction based on the label that has been predicted the most times out of all of them.

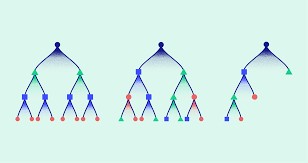
Accuracy increases and the issue of overfitting is avoided when there are more trees in the forest.



1. **Gradient Boosting Regression–** One of the most widely used machine learning algorithms for tabular datasets is gradient boosting. It has great usability and can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment. It is powerful enough to find any nonlinear relationship between your model target and features. Even though popular libraries like XGBoost and LightGBM allow you to build barebone gradient boosting trees without knowing much about the algorithm, you still want to know how it works when you start tuning hyper-parameters, customizing loss functions, etc. to get a model of better quality.



1. **XGBoost Regressor –** An open-source library called XGBoost stands for Extreme Gradient Boosting offers an effective and efficient gradient-boosting algorithm implementation. In machine learning competitions, XGBoost quickly established itself as the preferred method and frequently served as the key component in the winning solutions to a variety of issues. Predicting a numerical value, such as a height or a dollar amount, is the goal of regression predictive modeling problems. Regression predictive modeling can be performed directly with XGBoost.

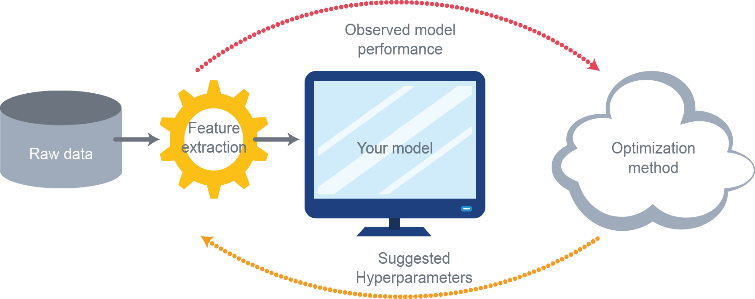


XGBoost is one of the fastest implementations of gradient boosting. This is accomplished by addressing one of the major shortcomings of gradient-boosted trees: taking into account the loss that could result from any possible splits when creating a new branch—especially if there are thousands of features and, consequently, thousands of possible splits. This inefficiency is fixed by XGBoost, which uses information about a leaf's feature distribution to narrow the search space for potential feature splits.

**Hyper Parameter Tuning:**

A mathematical model with several parameters that must be learned from the data is known as a machine learning model. We can fit the model parameters by training them with existing data.

Hyperparameters, on the other hand, are different kinds of parameters that cannot be learned directly from the regular training process. Usually, they are fixed before the actual training begins. The model's complexity and expected learning speed are both expressed by these parameters.

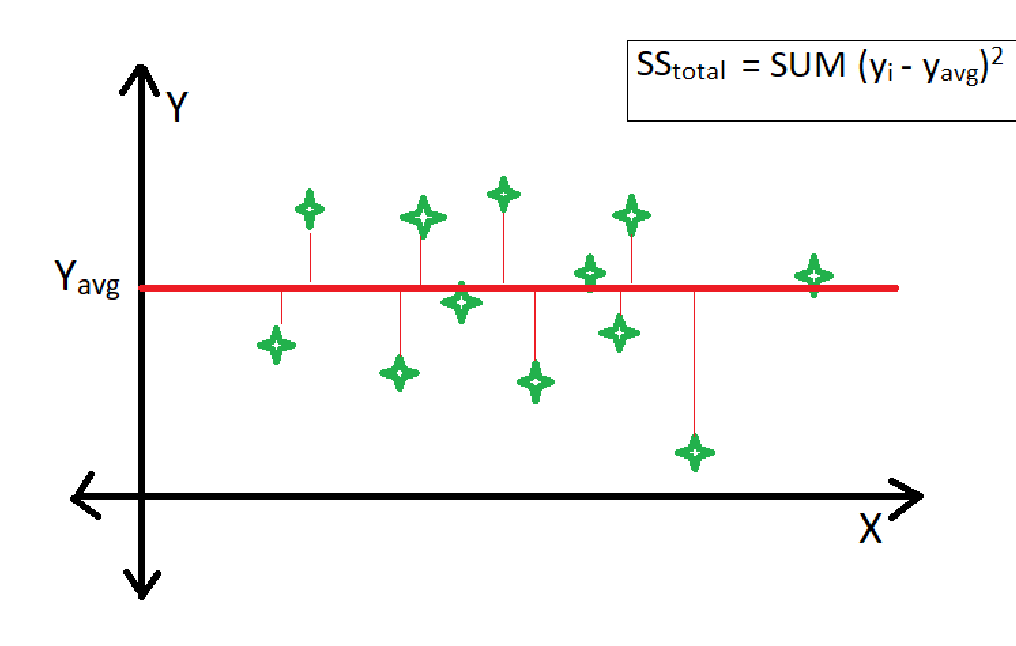


* **Grid Search CV –** The hyperparameter combination that gives the lowest error score is chosen by the Grid Search Method. This technique is explicitly valuable when there are just hyperparameters to improve. However, as the Machine Learning model becomes more complex, it loses out to other weighted-random search strategies.

**Model Evaluation:**

* **R-squared (R𝟐)** - A statistical measure of a regression model's goodness of fit is R-squared. The ideal R-square value is 1. The model fitted is better the closer the value of the R-square is to 1.

R-square is an examination of the leftover number of squares with the complete number of squares. The sum of squares of the perpendicular distance between data points and the average line is used to calculate the total sum of squares.



* **Root Mean Square Error (RMSE) –** The transformation that occurs between the values that a model predicted and the actual values is called RSME (Root mean square error). In other words, it is one of these mistakes in measuring the precision and error rate of any regression problem's machine learning algorithm.

The mean square error (RMSE) is the value divided by the square root. It makes it easier to plot a difference between the model parameter's estimated and actual value.

Utilizing RSME, we can undoubtedly quantify the effectiveness of the model.

* **Mean Square Error (MSE) –** The MSE risk method makes it easier to indicate the average squared difference between a feature or variable's predicted and actual values.

The general loss equation from earlier is used to calculate the Mean Squared Error. We'll also take into account the bias value because it's another parameter that needs to be changed during training.

**Conclusion:**

We began our analysis by performing EDA on all of our datasets. First, we looked at and changed our dependent variable, "Rental Bike Count." After that, we looked at categorical variables and eliminated those that represented the majority of one class. We also looked at numerical variables and discovered their correlation, distribution, and connection to the dependent variable. Additionally, we hot-encoded the categorical variables and removed some numerical features that primarily had 0 values.

Following that, we examine several well-known individual models, ranging from straightforward ensemble models like Random Forest and Gradient Boost to more complex ones like the Linear Regressor and Regularization Models (Ridge and Lasso). A single, unified model for working and nonworking days was also one of the few model formulation options tested.

Linear Regression, Lasso, Ridge, Decision Tree, Random Forest, and XGBoost were the next eight machine learning algorithms we used. To enhance the performance of our model, we performed hyperparameter tuning.

**Here are some suggestions to manage Bike Sharing Demand -**

* Create a portfolio of regular customers.
* The majority of rentals are for daily commutes to workplaces and colleges. Therefore, open additional stations near these landmarks to reach their primary customers.
* While planning for extra bikes to stations the peak rental hours must be considered, i.e. 7–9 am and 5–6 pm.
* Start a new renting program for premium customers to increase business.
* Utilize the ML model to cater to demand efficiently.
* Be ready for 2 kinds of patterns in demand which are for a working day and a non-working day.
* Maintenance activities for bikes should be done at night due to the low usage of bikes during the night time. Removing some bikes from the streets at night time will not cause trouble for the customers.
* Try to get the bookings as early as possible to manage the demand.
* May start giving discounts to bookings if they book a bike in advance.
* Be proactive with communication. Ask for feedback often.
* Periodically throw Offers to retain customers.
* Look at the customers facing problems with the service.
* Lean into the best customers.
* Solving Poor Network Connectivity Issues.
* Define a roadmap for new customers.
* Stay competitive.

We see 2 rental patterns across the day in bike rentals count - first for a Working Day where the rental count is high at peak office hours (8 am and 5 pm) and the second for a Non-working day where the rental count is more or less uniform across the day with a peak at around noon.

Hour of the day: Bike rental count is mostly correlated with the time of the day. As indicated above, the count reaches a high point during peak hours on working days

and is mostly uniform during the day on non-working days.

Temperature: People generally prefer to bike at moderate to high temperatures. We see the highest rental counts between 32 to 36 degrees Celcius

Season: We see the highest number of bike rentals in the Spring (July to September) and Summer (April to June) Seasons and the lowest in the Winter (January to March) season.

Weather: As one would expect, we see the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day

Humidity: With increasing humidity, we see a decrease in the bike rental count.

I have chosen the Random Forest model which is above all else I want better expectations for the rented\_bike\_count and time isn't compelling here. As a result, various linear models, decision trees, Random Forests, and Gradient Boost techniques were used to improve accuracy. I compared R2 metrics to choose a model.

No overfitting is seen in our model.

Due to less no. of data in the dataset, the training R2 score is around 99% and the test R2 score is 92%. Once we get more data, we can retrain our algorithm for better performance.

**References-**

1. StackOverflow
2. GeeksforGeeks
3. Analytics Vidhya