BIKE SHARING DEMAND PREDICTION

RAHUL KUMAR GUPTA
Data science trainees,
AlmaBetter, Bangalore

Abstract

Rental Bike Sharing is the process by which bicycles are procured on several basis- hourly, weekly, membership-wise, etc. This phenomenon has seen its stock rise to considerable levels due to a global effort towards reducing the carbon footprint, leading to climate change, unprecedented natural disasters, ozone layer depletion, and other environmental anomalies.
In our project, we chose to analyse a dataset pertaining to Rental Bike Demand from South Korean city of Seoul, comprising of climatic variables like Temperature, Humidity, Rainfall, Snowfall, Dew Point Temperature, and others. For the available raw data, firstly, a through pre-processing was done after which a Here, hourly rental bike count is the regress and. To an extent, our linear model was able to explain the factors orchestrating the hourly demand of rental bikes.

Keywords: *Data Mining, Linear Regression, Correlation Analysis, Bike Sharing Demand Prediction, Carbon Footprint.*

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

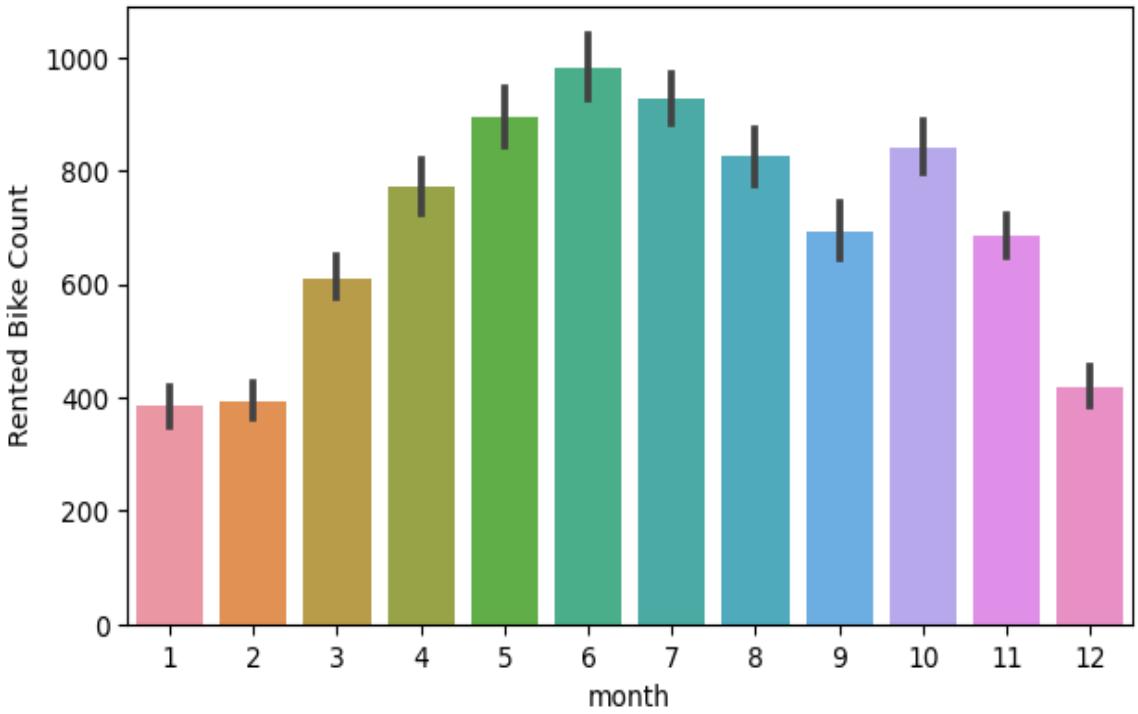
2. Introduction

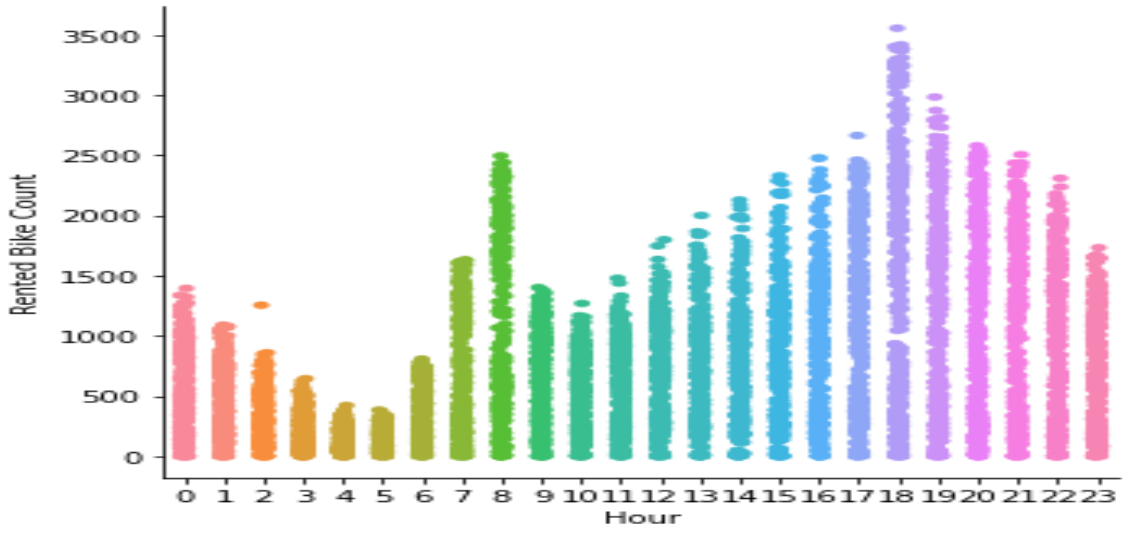
Bike Sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis.
The first bike-share programs began in 1960s Europe, but the concept did not take off worldwide until the mid- 2000s. In North America, they tend to be affiliated with municipal governments, though some programs, particularly in small college towns, centre on university campuses. The

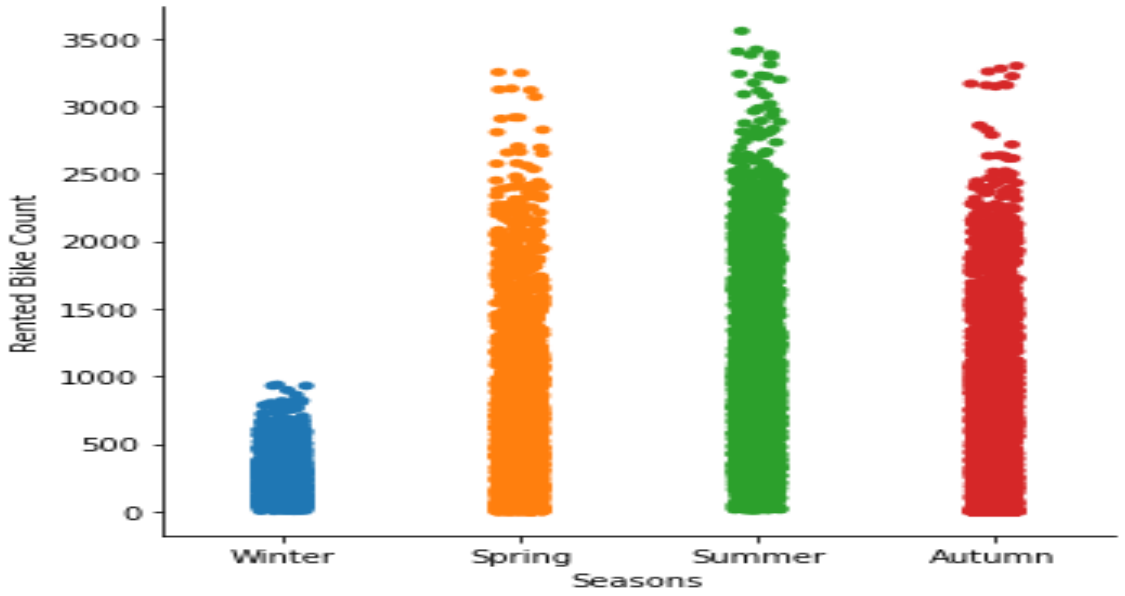
typical bike-share has several defining characteristics and features, including station-based bikes and payment systems, membership, and pass fees, and per- hour usage fees. Programs are generally intuitive enough for novice users to understand. And, despite some variation, the differences are usually small enough to prevent confusion when a regular user of one city's bike-share uses another city's program for the first time.
Bike sharing has been gaining importance over the last few decades. More and more people are turning to healthier and more live able cities where activities like bike sharing are easily available. there are many benefits from bike sharing, such as environmental benefits. It was a green way to travel.
This dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 in Capital bike share system with the corresponding weather and seasonal information. The dataset contains 8760 rows (every hour of each day for 2017 and 2018) and 14 columns (the features which are under consideration)

3. Steps involved:

- **Exploratory Data Analysis**
  After loading the dataset we performed this method by comparing our target variable that 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.







- **<u>Null values Treatment</u>**

  Our dataset contains if any number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project inorder to get a better result.

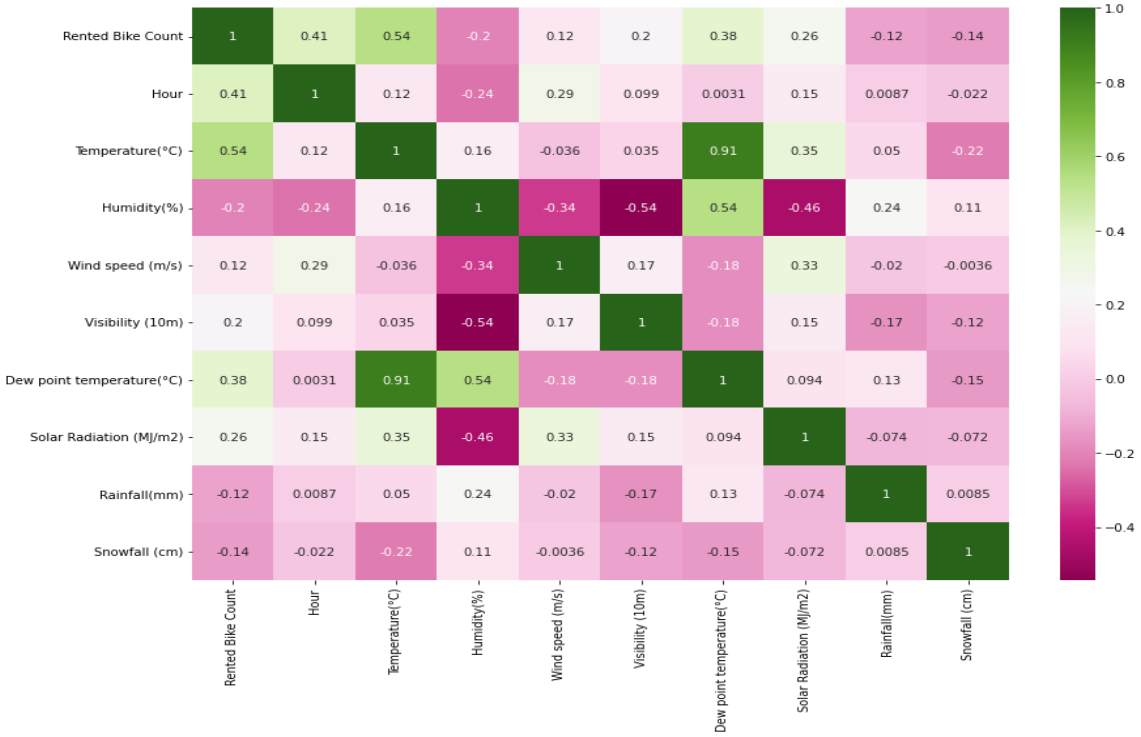
- **<u>Encoding of categorical columns</u>**

  We 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.

- **Feature Selection**

  In these steps we used algorithms like ExtraTree classifier to check the results of each feature i.e which feature is more important compared to our model and which is of less importance.

  Next we used VIF and Correlation Heat Map for numerical features to select the best feature which we will be using further in our model. From the correlation graph with Heat map we saw that dew point temp and temperature is highly correlated. Then we checked VIF and concluded that these two features are affecting VIF score also. so we decided to drop one of these feature and to do this we checked which feature is least correlated with Dependent variable and we identified it to be Dew point temperature and therefore we dropped the Dew point temperature



- **Standardization of features**

  Our main motive through this step was 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 was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

- **Fitting different models**
  For modelling we tried various classification algorithms like:

1. **Linear Regression**
2. **Lasso Regression**
3. **Ridge Regression**
4. **Elastic Net Regression**
5. **Polynomial Regression**
6. **KNN Regression**
7. **Decision Tree Regression**
8. **Random Forest**
9. **Gradient Boosting**
10. **Catboost**
11. **Light GBM**

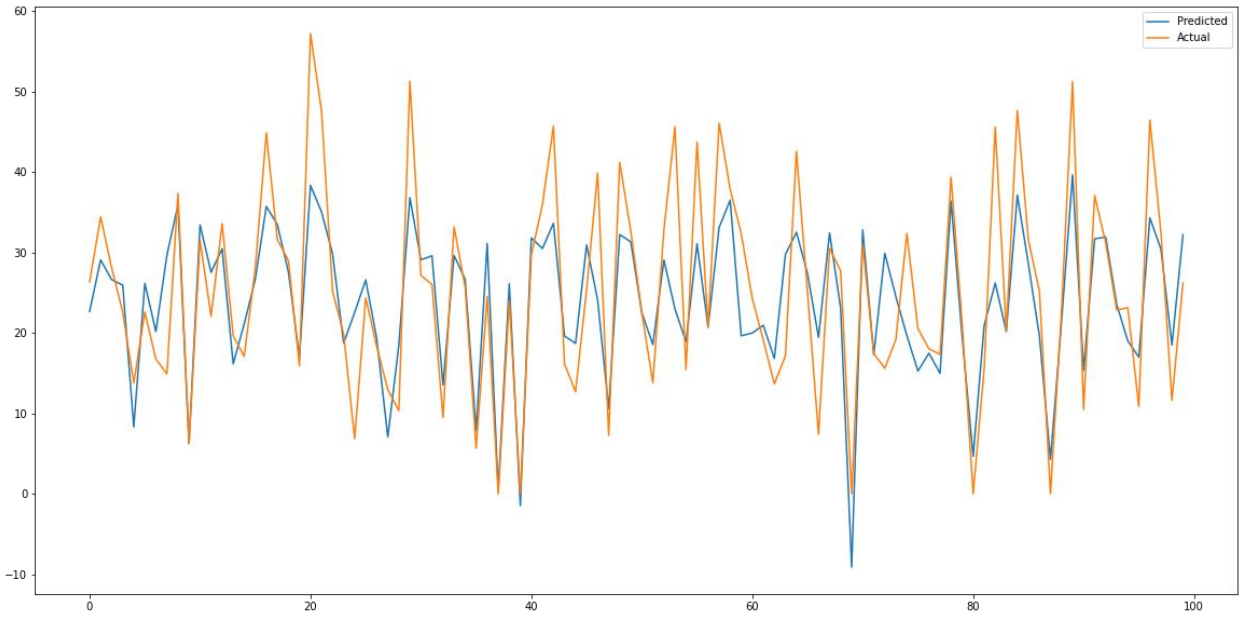
- **Tuning the hyperparameters for better accuracy**
  Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models
  like Random Forest Classifier and XGBoost classifier.

- **SHAP and Eli for features**
  We have applied SHAP and Eli value plots on the Random Forest model to determine the features that were most important while model building and the features that didn't put much weight on the performance of our model.

7.1. Algorithms:

1. Linear Regression:

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. We found r2: 0.5791



.

2. Lasso, Ridge, ElasticNet Regression:

Sometimes we need to choose between low variance and low bias. There is an approach that prefers some bias over high variance, this approach is called **Regularization**.

In Ridge regression, we add a penalty term which is equal to the square of the coefficient. The *L2* term is equal to the square of the magnitude of the coefficients. We also add a coefficient to control that penalty term. In this case if is zero then the equation is the basic OLS else if then it will add a constraint to the coefficient. As we increase the value of this constraint causes the value of the coefficient to tend towards zero. This leads to tradeoff of higher bias (dependencies on certain coefficients tend to be 0 and on certain coefficients tend to be very large, making the model less flexible) for lower variance.

Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds penalty term to the cost function. This term is the absolute sum of the coefficients. As the value of coefficients increases from *0* this term penalizes, cause model, to decrease the value of coefficients in order to reduce loss. The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

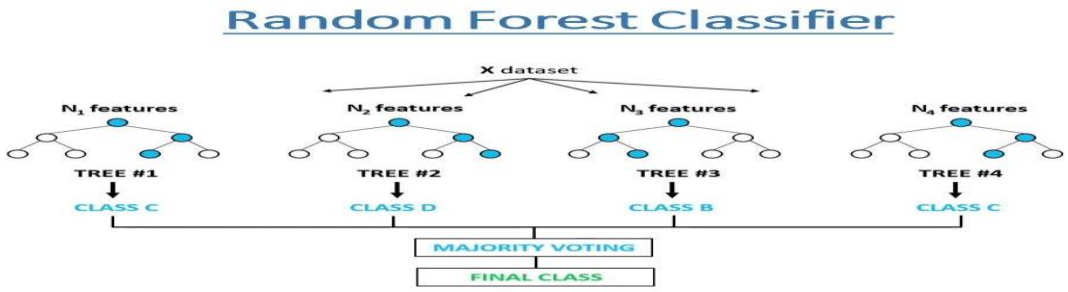
Sometimes, the lasso regression can cause a small bias in the model where the prediction is too dependent upon a particular variable. In these cases, elastic Net is proved to better it combines the regularization of both lasso and Ridge. The advantage of that it does not easily eliminate the high collinearity coefficient.

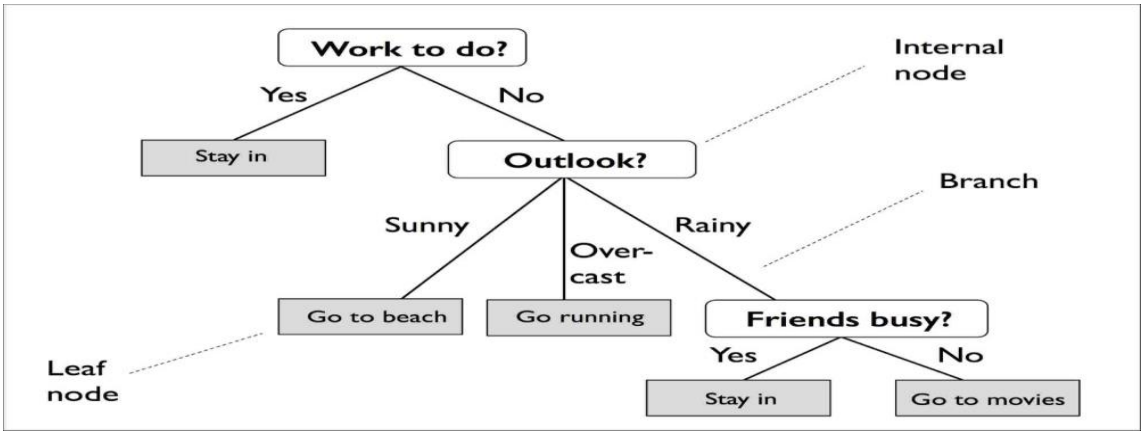
3. Decision Tree and Random Forest Classifier:

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility.

Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

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.



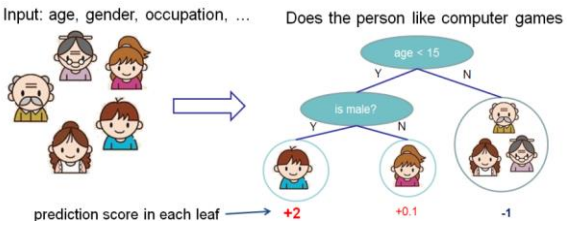


4. XgBoost, CatBoost, LightGbm-

To understand XGBoost we have to know gradient boosting beforehand.

- **Gradient Boosting-**

  Gradient boosted trees consider the special case where the simple model is a decision tree



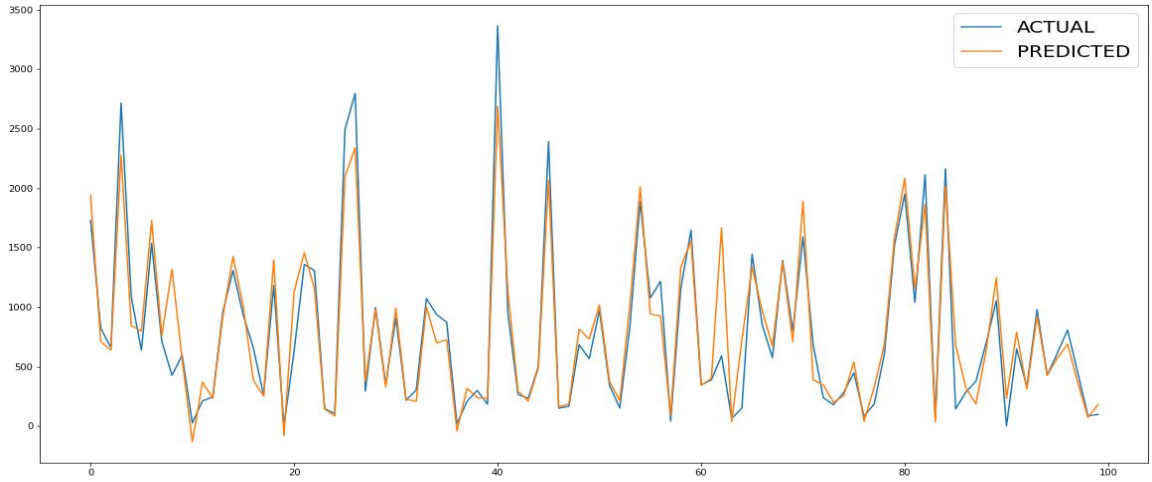
In this case, there are going to be 2 kinds of parameters P: the weights at each leaf, w, and the number of leaves T in each tree (so that in the above example, T=3 and w=[2, 0.1, -1]).

When building a decision tree, a challenge is to decide how to split a current leaf. For instance, in the above image, how could I add another layer to the (age > 15) leaf? A 'greedy' way to do this is to consider every possible split on the remaining features (so, gender and occupation), and calculate the new loss for each split; you could then pick the tree which most reduces your loss.

XGBoost is one of the fastest implementations of gradient boosting. trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XGBoost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

CatBoost or Categorical Boosting is an open-source boosting library developed by Yandex. In addition to regression and classification, CatBoost can be used in ranking, recommendation systems, forecasting and even personal assistants.

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.
It uses two novel techniques: **Gradient-based One Side Sampling** and **Exclusive Feature Bundling (EFB)** which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks



This Plot is for Catboost with higher accuracy

5. Model performance:

R Value is the coefficient between the Predicted and Observed values of the dependent variable.
R-Square Value is the goodness-of-fit and a statistical measure of how close the data are fitted to the regression line.

Adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. It calculates R-Square of only Independent Variables those are statistically significant.
A minute difference between R-Square and Adjusted R- Square suggests all our Independent Variables being significant, despite both values being on a relatively lower side.
R-square change, which is just the improvement in R- square when the second predictor is added. The R-square change is tested with an F-test, which is referred to as the F- change. A significant F-change means that the variables added in that step significantly improved the prediction.

Mean Square Error:

It is simply the average of the square of the difference between the original values and the predicted values.

	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear	175590.552873	419.035264	0.572911	0.569766
1	Lasso	175560.907118	418.999889	0.572983	0.569839
2	Ridge	175248.935066	418.627442	0.573742	0.570603
3	Elasticnet	175346.867499	418.744394	0.573504	0.570363
4	Polynomial	123952.860328	352.069397	0.698509	0.696289
5	K-Nearyest_Neighbor	83411.759209	288.810940	0.796159	0.794659
6	Decision_Tree	90012.614155	300.021023	0.780028	0.778409
7	Random_Forest	62747.974057	250.495457	0.846657	0.845528
8	Gradient_Boosting	54909.771579	234.328341	0.865812	0.864824
9	Xtreme_GB	40812.801816	202.021785	0.900262	0.899528
10	CATBoost	36706.535373	191.589497	0.910297	0.909637
11	lightGBM	35410.753754	188.177453	0.913464	0.912826

6. Hyper parameter tuning:

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. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

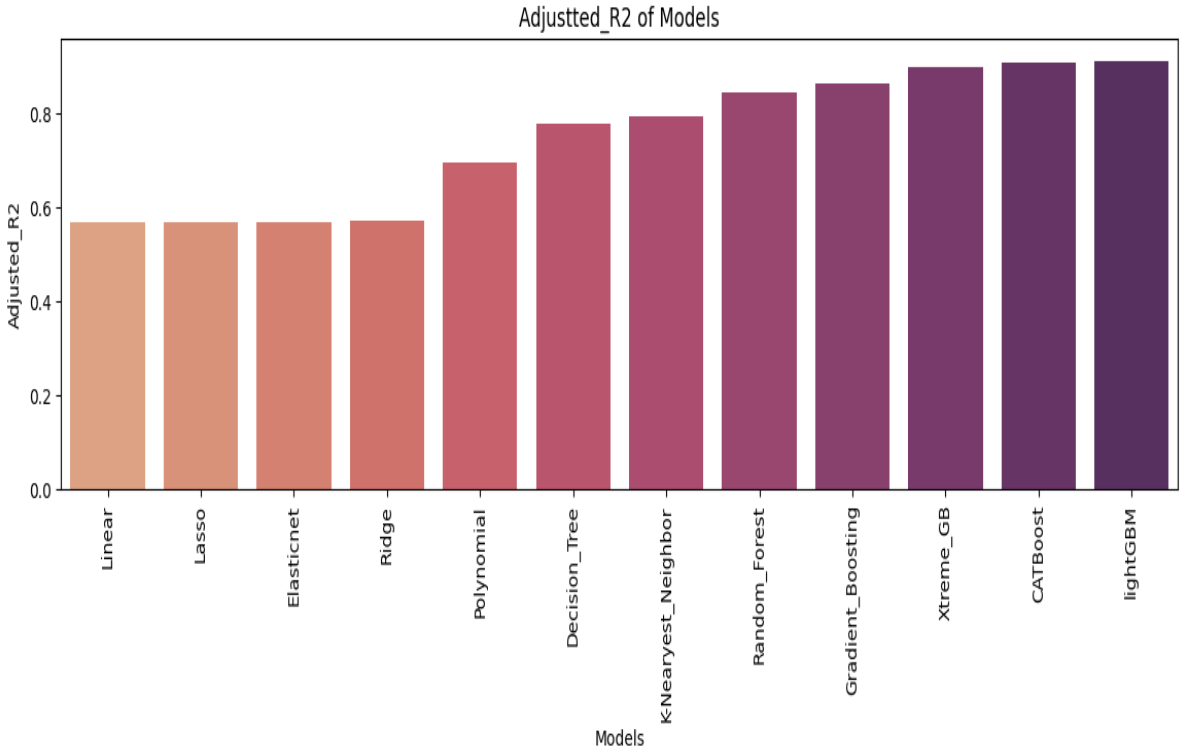
We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

1. **Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

2. **Randomized Search CV-** In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist's control

3. **Bayesian Optimization-** Bayesian Hyperparameter optimization is a very efficient and interesting way to find good hyperparameters. In this approach, in naive interpretation way is to use a support model to find the best hyperparameters.A hyperparameter optimization process based on a probabilistic model, often Gaussian Process, will be used to find data from data observed in the later distribution of the performance of the given models or set of tested hyperparameters. As it is a Bayesian process at each iteration, the distribution of the model's performance in relation to the hyperparameters used is evaluated and a new probability distribution is generated. With this distribution it is possible to make a more appropriate choice of the set of values that we will use so that our algorithm learns in the best possible way.

6. Conclusion:



. We observed that bike rental count is high during week days then weekend days.

• The rental bike counts is at its peek at 8 AM in the morning and 6pm in the evening, We can see an increasing trend from 5am to 8 am, the graph touches the peak at 8am and then there is dip in the graph. Later we can see a gradual increase in the demand until 6pm, the demand is highest at 6 pm, and reduces there after until midnight,

• We observed that people prefer to rent bikes at moderate to high temperature, and even when it is little windy,

• it is observed that highest bike rental count is in Autumn and summer seasons and the lowest is in winter season.

• We observed that the bike rentals is highest during the clear days and lowest on snowy and rainy days.

• when we compare the RMSE and Adjusted R2 of all the models, CATBoost, Light Gbm gives the highest Score where R2 score is 0.90 and Training score is 0.91 so this model is the best for predicting the bike rental count on daily basis.

7 References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya