Bike Sharing Demand Prediction

Vishal Raul
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 this project, we chose to analyze a dataset pertaining to Rental Bike Demand from the South Korean city of Seoul, consisting 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. 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 to rent a bike from one location and return it to a different place on an as-needed basis.

Prediction of bike sharing demand can help bike sharing companies to allocate bikes better and ensure a more sufficient circulation of bikes for customers. This presentation proposes a real-time method for predicting bike renting based on historical data, weather data, and time data. This demand prediction model can provide a significant theoretical basis for management strategies and vehicle scheduling in public bike rental systems. The design of the learning algorithm includes preprocesses of feature explanation and data selection, modeling and validation.

2. 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. Our task is to predict the demand of bike rent based on the historical usage over different factors such as seasons, weather, temperature, humidity etc.

3. Data

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.

Data fields:

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)

4. Steps involved

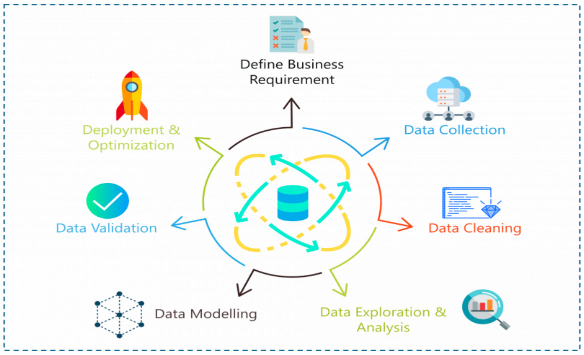

Define Business Requirement
Data Collection
Data Cleaning
Data Exploration & Analysis
Data Modelling
Data Validation
Deployment & Optimization

Image credit: edureka.com

- **Define Problem Statement**

  Before you even begin a Data Science project, you must define the problem you’re trying to solve. At this stage, you should be clear with the objectives of your project.

- **Data Collection**

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

- **Data Cleaning**

  Data cleaning is the process of removing redundant, missing, duplicate and unnecessary data. This stage is considered to be one of the most time-consuming stages in Data Science. However, in order to prevent wrongful predictions, it is important to get rid of any inconsistencies in the data.

- **Data Analysis and Exploration**

  Once you’re done cleaning the data, it is time to get the inner Sherlock Holmes out. At this stage in a Data Science life-cycle, you must detect patterns and trends in the data. This is where you retrieve useful insights and study the behavior of the data. At the end of this stage, you must start to form hypotheses about your data and the problem you are tackling.

- **Data Modelling**

  This stage is all about building a model that best solves your problem.

A model can be a Machine Learning Algorithm that is trained and tested using the data. This stage always begins with a process called Data Splicing, where you split your entire data set into two proportions. One for training the model (training data set) and the other for testing the efficiency of the model (testing data set).

- **Optimization and Deployment**

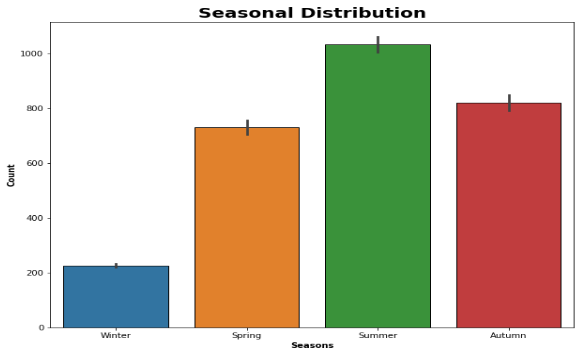
This is the last stage of the Data Science life-cycle. At this stage, you must try to improve the efficiency of the data model, so that it can make more accurate predictions. The end goal is to deploy the model into production or production-like environments for final user acceptance. The users must validate the performance of the models and if there are any issues with the model then they must be fixed in this stage.

5. Exploratory Data Analysis:

5.1. Seasons

Winter, spring, summer, autumn these seasons bike demand records are in the dataset. From Above Visualization, in the summer season bikes have more demand & less demand in the winter season.

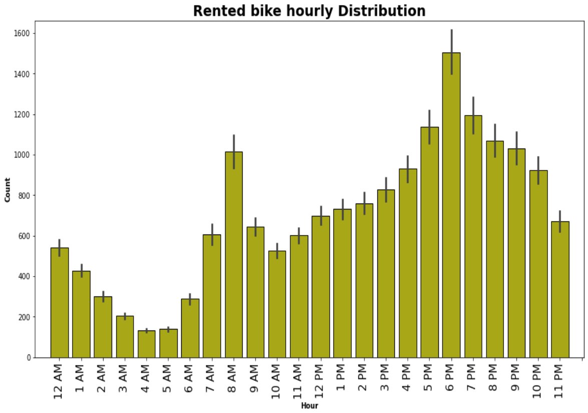
Graph 1. Distribution of demand for seasons



5.2. Hour

High rise of Rented Bikes from 8:00 a.m. to 9:00 p.m. means people prefer rented bikes during rush hour. We can clearly see that demand rises most at 8 a.m. and 6:00 p.m. so we can say that during office opening and closing time there is much high demand.

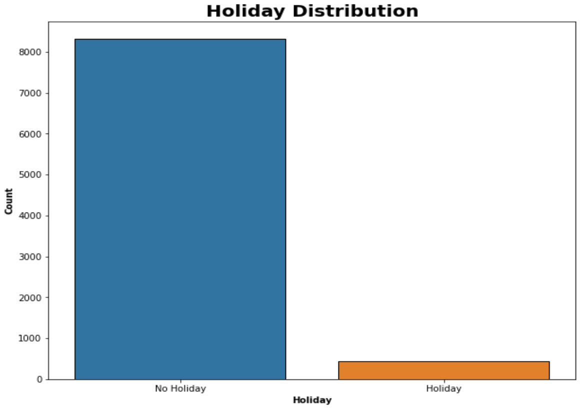
Graph 2. Hourly distribution of bike demand



5.3. Holiday

Bike demand is more on No-holiday as compared to holiday.

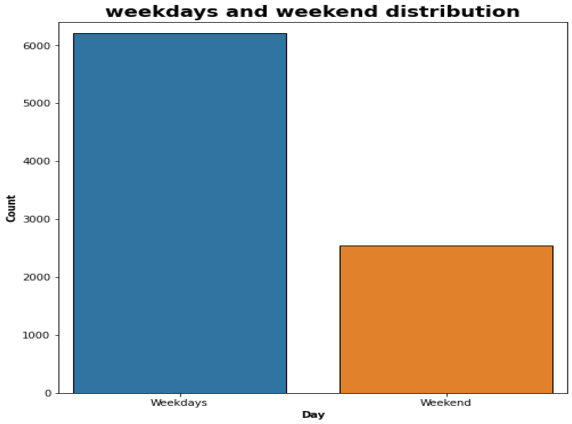
Graph 3. Holiday bike demand analyze



5.4. Weekdays_weekend

Users demand more bikes on non-holiday as compared to holidays.

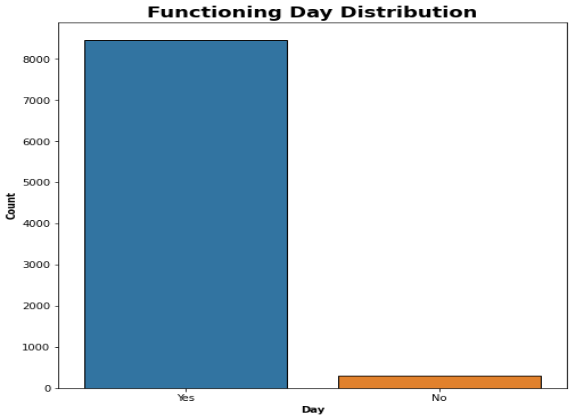
Graph 4. Weekdays_weekend bike demand



5.5. Functioning Day

On Functioning Day Bike demand is more as compared to other days

Graph 6. Functioning Day Distribution



6. 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.

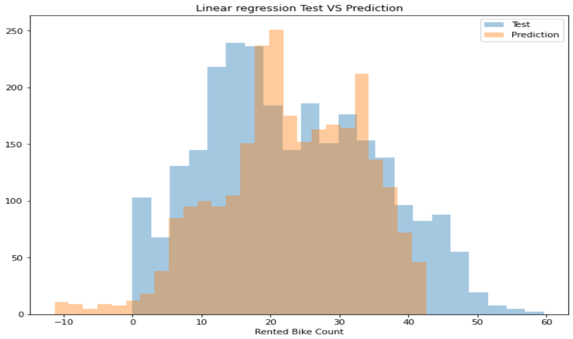
7. ML Model Building

A machine learning model is an expression of an algorithm that combs through mountains of data to find patterns or make predictions.

7.1. Linear Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.
From the above Viz. we can clearly identify that the Linear Regression isn't performing well. The Actual Data (in Grey) and Predicted values (in Yellow) are very different. We can conclude that Linear Regression doesn't seem like a right choice for Trip duration prediction

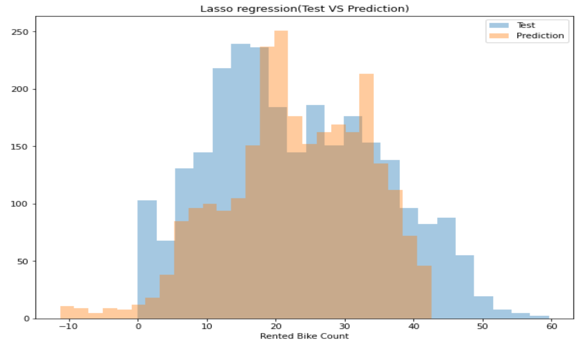
Graph: Test vs Prediction Linear Regression model



7.2. Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

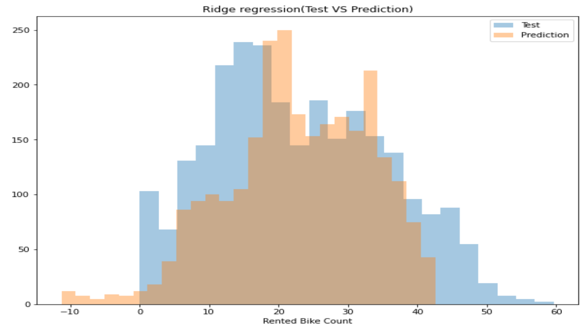
Graph: Test vs Prediction lasso Regression model



7.3. Ridge Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.

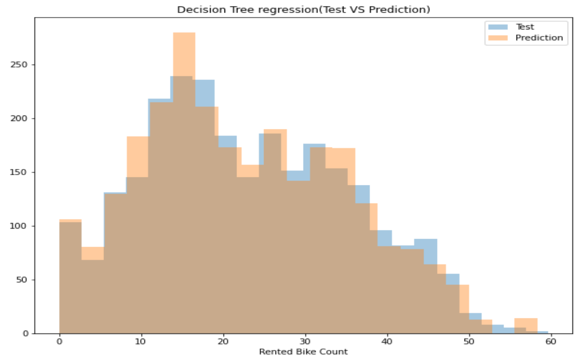
Graph: Test vs Prediction Ridge Regression model



7.4. Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

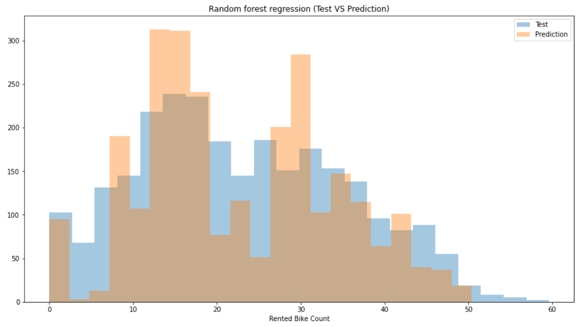
Graph: Test vs Prediction Decision tree Regression model



7.5. Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.
From the above Viz. We can clearly identify that the Random Forest Algorithm is also performing well. The Actual Data (in Grey) and Predicted values (in Green) are as close as possible. We can conclude that Random Forest could be a good choice for Trip duration prediction.

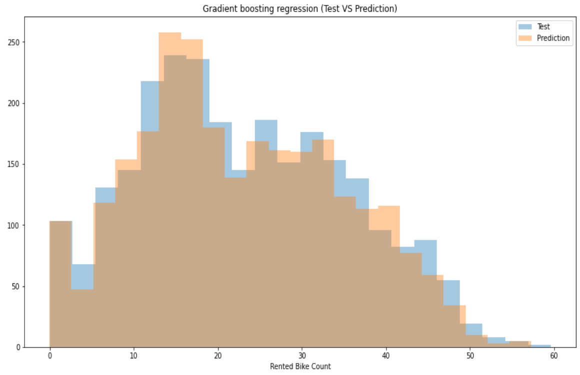
Graph: Test vs Prediction Random forest Regression model



7.6. Gradient boosting Regression

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

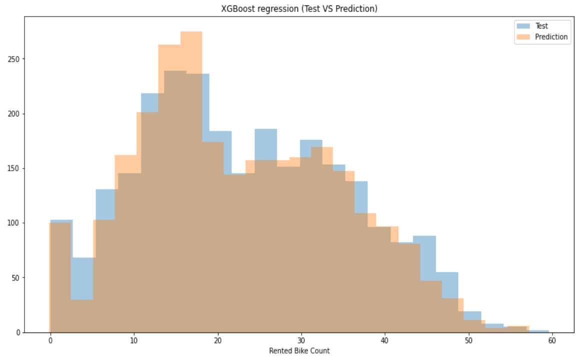
Graph: Test vs Prediction Gradient boosting model



7.7. XGBoost Regression

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners.

Graph: Test vs Prediction XGBoost Regression model



8. Model performance

The model can be evaluated by various metrics such as:

8.1. Mean square error

The MSE of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

8.2. Root mean square error

RMSE is just the root of MSE. It is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient. One can compare the RMSE to observed variation in measurements of a typical point

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}}$$

8.3. R square

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model. It has one limitation that its value increases as the number of Parameters increases even if that parameter does not improve the model.

$$SSE = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{m}(y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$\hat{y}$ is the predicted value, y is the actual value and $\bar{y}$ is the mean .

8.4. Adjusted R Square

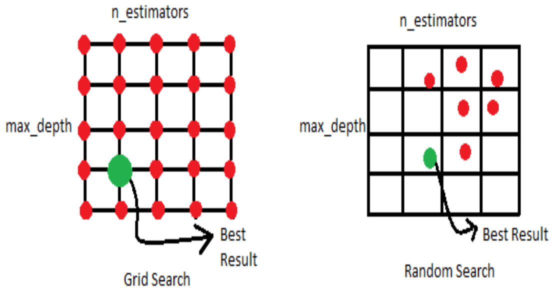
Adjusted R-squared is a modified version of R-squared that overcomes the problem of r2 and has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

9. Hyperparameter 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 the 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.

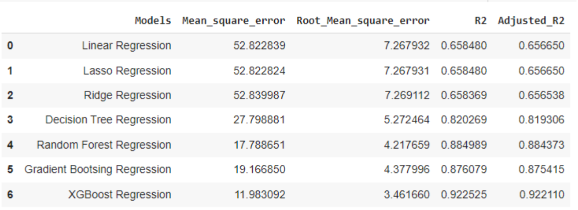
9.1. Grid Search CV-Grid:

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.

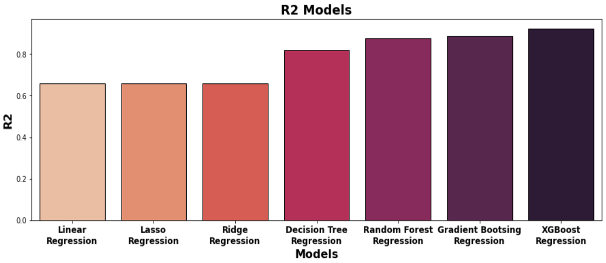


Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection, and then model building.
In all of these models, our accuracy revolves in the range of 56 to 91%.
And there is some amount of improvement in the adjusted R2 score after hyperparameter tuning
So the accuracy of our best model is 91% which can be said to be good for this large dataset. This performance could be due to various reasons like the proper pattern of data, large data, or because of the relevant features.

10. ML Model Evaluation Result



	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear Regression	52.822839	7.267932	0.658480	0.656650
1	Lasso Regression	52.822824	7.267931	0.658480	0.656650
2	Ridge Regression	52.839987	7.269112	0.658369	0.656538
3	Decision Tree Regression	27.798881	5.272464	0.820269	0.819306
4	Random Forest Regression	17.788651	4.217659	0.884989	0.884373
5	Gradient Bootsing Regression	19.166850	4.377996	0.876079	0.875415
6	XGBoost Regression	11.983092	3.461660	0.922525	0.922110



11. Conclusion:

As we can see the total amount of bike rentals increases with the temperature per month. Whereas it seems that the rentals are independent of the wind speed and the humidity, because they are almost constant over the months. This also confirms on the one hand the high correlation between rentals and temperature and on the other hand that nice weather could be a good predictor. So people mainly rent bikes on nice days and nice temperature. This could be important in planning new bike rental stations.

It is quite evident from the results that XGBoost is the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse, rmse) shows lower and (r2,adjusted_r2 = 92 %) show a higher value for the XGBoost model. So, finally this model is best for predicting the bike rental count on a daily basis.

12. References:

1. Aditya Singh Kashyap, Swastika Swastik 'Regression Model to Predict Bike Sharing Demand', (2021), International Journal of Innovative Science and Research Technology ISSN No:-2456-2165.
2. **https://www.kaggle.com/**
3. **https://www.analyticsvidhya.com/**
4. **https://www.geeksforgeeks.org/**
5. **https://learn.almabetter.com/**