**BIKE SHARING DEMAND PREDICTION**

**SARANYA N**

**Data science trainee**

**AlmaBetter, Bangalore**

**ABSTRACT**

As an emerging mobility service, bike-sharing has become increasingly popular around the world. A critical question in planning and designing bike-sharing services is to know how different factors, such as land-use and built environment, affect bike-sharing demand. Most research investigated this problem from a holistic view using regression models, where assume the factor coefficients are spatially homogeneous. However, ignoring the local spatial effects of different factors is not tally with facts. Therefore, we develop a regression model with spatially varying coefficients to investigate how land use, social-demographic, and transportation infrastructure affect the bike-sharing demand at different stations to address this problem. Unlike existing geographically weighted models, we define station-specific regression and use a graph structure to encourage nearby stations to have similar coefficients.

Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system. 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 : Linear regression, Bike rental counts, correlation, Null values, regression model.***

**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 mid2000s. 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 perhour 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.

The goal of this project is to combine the historical bike usage patterns with the weather data to forecast bike rental demand. The data set consists of hourly rental data spanning two years.

1. **Problem statement :**

The main objective is to build a predictive model, which could help to train a model to predict the number of bike rentals of the year given the weather conditions. This would in turn help to predicting quickly and efficiently. The bike sharing demand prediction dataset from rented bike provider company from Seoul contains 14 features and 8760 observations of a complete year I.e. from 1.12.2017 to 31.11.2018

The dataset contain following columns :

* 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)

**2. Factors affecting :**

While doing this analysis , we recognize that some of the factors are affecting to the number of bike rentals;

* Weather : when the number of bike rental is high , at that time the weather is clear and sunny .
* Seasons: bike rentals are higher in summer seasons and least during spring season .
* Working day(weekdays and weekend) : during this analysis outliners are present in working days
* Temperature : We observed that there is increase in the bikes rented counts with temperature with a small decrease at the highest temperature.
* Hours : people are using rented bike between 6AM to 9AM And 5pm to 7PM ,that means they use bikes for reaching their office.

**3.Steps involved:**

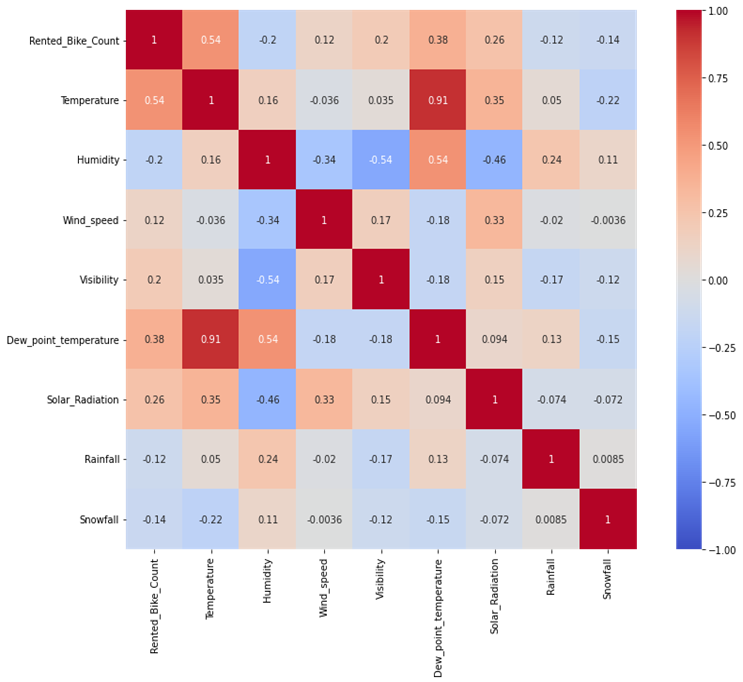
* **Exploratory Data Analysis** : After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is bike rentals counts with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables and also we observed the distribution of variables. It gave us a better idea that how feature behaves with the target variable.
* Exploring head and tail of the data to get insights on the given data.
* Data preprocessing : a dataset may contain missing values and inconsistent data so we need to check this .so if there is any missing values, the easiest way is remove them from the dataset . for removing we can use isnull() , notnull() functions from pandas library to determine null values . so this will not affect the performance of the model . here in our data set is not contain any null values to disturb the accuracy.
* **Handling outliners**: outliners are present which can disturb the accuracy .
* **Numerical and categorical features**: anaylysing the numerical and categorical features in the dataset .
* **One hot encoding** :In this dataset some categorical variables like seasons, holiday and function day, we change it with numerical database
* **Correlation Analysis** :We plot the heatmap to find the correlation between both dependent variable and independent variables.
* **Train test Split** :In train test split we take x as dependent variables and y take as independent variable then train the model.Splitting the model into train and test sets
* **Models** : using models to train the data and predicting the accuracy ,

**4.Correlation analysis:**

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour

We plot the heatmap to find the correlation between all the columns and observed that:

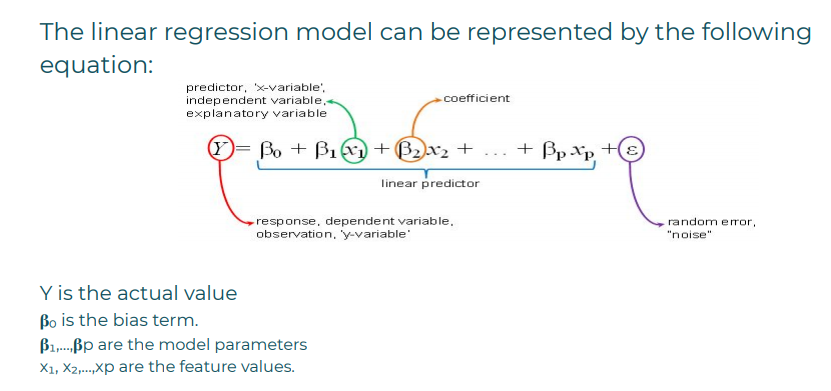
* Temperatures are highly correlated.
* We see that there is a positive correlation between columns 'Temperature' and 'Dew point temperature' i.e 0.91 so even if we drop this column then it does not affect the outcome of our analysis. And they have the same variations.. so we can drop the column 'Dew point temperature.

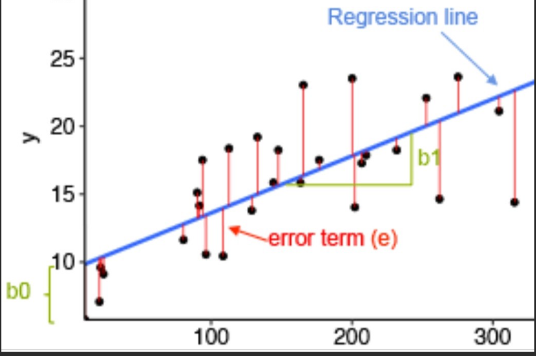


**5.Models**

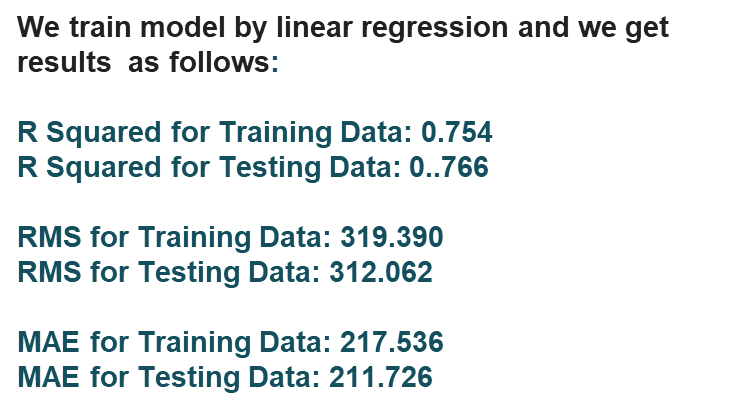
1. **Linear regression**

Linear regression is one of the most basic types of regression in supervised machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other. We try to find the relationship between independent variable(input) and a corresponding dependent variable (output). This can be expressed in the form of a straight line y = β0 + β1x . Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable

The hypothesis of linear regression ****

****

We train model by linear regression and we get results as follows:



**2**. **Lasso regression:**

Lasso regression is a type of linear regression that uses shrinkage is where data points are shrunk towards a central point , like the mean . Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge 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, which is never the case in ridge regression.

**We train model by lasso regression and we get results as follows:**

* **R Squared for Training Data: 0.754**
* **R Squared for Testing Data: 0..766**
* **RMS for Training Data: 319.390**
* **RMS for Testing Data: 312.062**
* **MAE for Training Data: 217.536**
* **MAE for Testing Data: 211.726**

**3. Ridge regression**

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

The cost function for ridge regression : Min(| |Y-X(theta)| |^2 – λ| |theta| |^2)

Lambda is the penalty term . λ is denoted as an alpha parameter in the ridge function . so by changimg values of alpha , we are controlling the penalty term . higher the value of alpha bigger is the penalty and therefore the magnitude of co eff is reduced

* It shrinks the parameter . so it is used to prevent multicollinearity
* It reduces the model complexity by coefficient shrinkage .

**By performing ridge regression we get the results are**

## R Squared for Training Data:0.7539

## R Squared for Testing Data :0.7904

## RMS for Training Data: 319.735

## RMS for Testing Data: 312.434

## MAE for Training Data: 217.696

## MAE for Testing Data: 211.917

**4. Decision tree regression model:**

Decision tree is the most powerful and popular tool for classification and prediction. 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 tree can be *“learned”* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called*recursive partitioning*. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

**We train model by Decision tree and we get results as follows:**

* **R Squared for Training Data: 0.689**
* **R Squared for Testing Data: 0.651**
* **RMS for Training Data: 359.273**
* **RMS for Testing Data: 381.502**
* **MAE for Training Data: 240.805**
* **MAE for Testing Data: 253.517**

**5. 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. Advantages of random forest are given below,

* Good for Large Datasets to handle
* The learning is fast and provides a high accuracy.
* Can handle plenty number of variables at once.
* Over-fitting is not a problem in this algorithm

Also it has some disadvantages too,

* Complexity is a major issue. Since the algorithm creates a number of trees and combines its output to produce the best output takes more computational time and resources.
* The time period usually for training a random forest model is greater since it generates a large number of trees.

We train model by Random forest and we get results as follows:

* R Squared for Training Data: 0.989
* R Squared for Testing Data: 0.906
* RMS for Training Data: 67.364
* RMS for Testing Data: 197.244
* MAE for Training Data: 40.276
* MAE for Testing Data: 112.071

**6. GRADIENT BOOSTING:**

**Gradient Boosting** is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc

**We train model by Gradient boosting and we get results as follows:**

**R Squared for Training Data: 0.845**

**R Squared for Testing Data: 0.835**

**RMS for Training Data: 253.748**

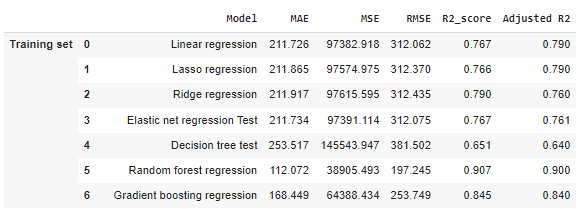
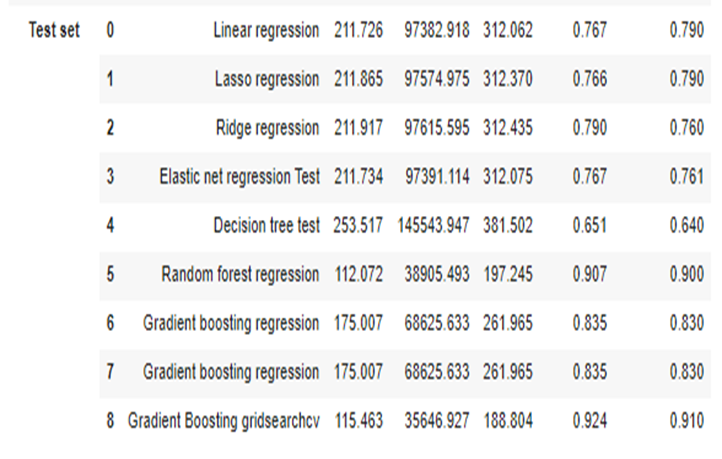
**RMS for Testing Data: 261.964**

**MAE for Training Data: 168.44**

**MAE for Testing Data: 175.007**

**6. CONCLUSIONS:**

* **We train the dataset to predict the number of rented bike is used in the given weather conditions .**
* **initially we did EDA on all the features of our dataset ,analyze dependent variable and categorical variables.**
* **Implemented 7 machine learning algorithms Regression,lasso,ridge,elasticnet,decission tree, Random Forest and XGBoost. All algorithms performed really well on both training dataset and testing dataset so we can say that variance is less and no issues of overfittings are present.**
* **"Random forest regression(90%)" and "Gradient Boosting regression(gridsearch cv) has highest R2 score(84%).**
* **DECISION TREE algorithm has comparatively less R2 score(65%)**
* **When we compare the root mean squared error and mean absolute error of all the models, Random forest Regression and Gradient Boosting gridsearchcv gives the highest R2 score of 99% and 95% respectively for Train Set and 92% for Test set. So, ﬁnally this model is best for predicting the bike rental count on daily basis.**

****

However ,this results are not the ultimate . as this data is time dependent , the values for variables like temperature, solar\_radiation, wind\_speed etc., Will not always be consistent.Therefore, there will be scenarios where the model might not perform well. As machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time. Therefore, having a quality knowledge and keeping pace with the ever evolving ML field would surely help one to stay a step ahead in future