**Bike Sharing Demand Prediction**

**Naved mansuri**

**Data science trainees,**

**AlmaBetter**

**Abstract**

The aim of this work is to get a better idea of the time it will take to go for a ride in the bike sharing system. The number of bikes rented per hour and the date information are all part of the data used.

Predicting consumer demand for bike sharing in South Korea was done using machine learning models. The two models that were found to be the best are the randomforest and the Gradient boosting. The Temperature and Hour feature are the most crucial in model prediction. All features are not as important as they should be, since they help enhance the performance of the models.

From those Bar Charts, we observed the variance of the first 25 most frequently used word associated with each rating ranging from 1 through 5. To further investigate that, we created a dictionary of positive and negative words manually, using information from the dataset

**Introduction**

In big cities like New York City, Paris, Washington DC, London, Beijing and Barcelona, bike sharing systems are on the rise. Renting a bike is a better way to complete a short trip than walking. It is comfortable and eco-friendly compared to driving. Due to global warming and pollution, many countries have been focusing on using renewable energy that doesn't harm the environment and can be used again. Renting bikes in Seoul is their most used service in South Korea, which has adapted to it. It's important to have an estimate of future demand in order to avoid difficulties such as waiting time. We want the model to predict bike sharing demand considering all of the factors which affect it.

**Data Description**

| Attribute | Description |
| --- | --- |
| **Date** | ear-month-day |
| **Rented Bike count** | Count of bikes rented at each hour |
| Hour | Hour of the day |
| Temperature | Temperature in Celsius |
| Humidity | Humidity in percentage |
| Windspeed | Windspeed in m/s |
| Visibility | Visibility |
| Dew point temperature | Dew point temperature |
| Solar radiation | Solar radiation |
| Rainfall | Rainfall in mm |
| Snowfall | Snowfall in cm |
| Seasons | Winter, Spring, Summer, Autumn |

**Exploratory Data Analysis**

Data analysis is done in a way that's visual and statistical. Data can be better understood with the graphs plotted in exploratory data analysis.

**Null values Treatment**

After the data is loaded , The missing data is checked using is.na() or isnul() function . The output depicted that there was no missing values in our dataset.

So our dataset does not contain any missing values .

**Encoding**

The machine can't read categorical features, so we use One Hot Encoding to convert them to numerical format. The season, functioning day, and holiday variables were converted to numerical depictions to fit our model to predict bike rented count.

Functioning Day, Holiday, and Seasons were convertedcoded to fit our model to predict Bike rented count.

**Feature Engineering**

The season variable is divided into three parts, the spring, summer and winter.

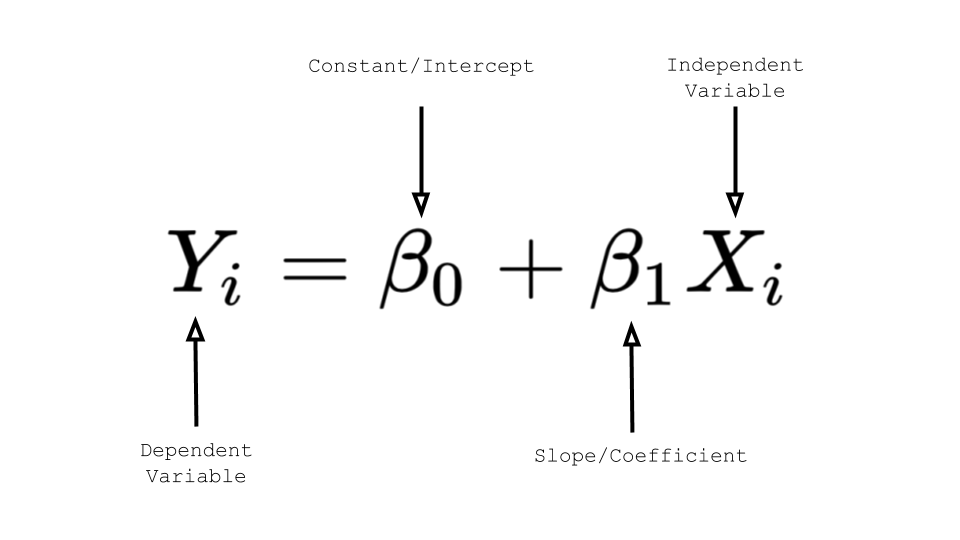
In order to make it easy to perform operations on the data, we need to separate day,month, year and hour into separate columns.

**Algorithms**

**Linear Regression**

Linear regression is the most efficient method for identifying the linear correlation between the independent and dependent variables. It's done by making a linear equation of line

For fitting the model, it is utmost important to check, whether there is a connection between the variables or features of interest.



**Regularised Linear Regression**

Machine Learning model building, the Regularization Techniques is an unavoidable and important step to improve the model prediction and reduce errors

Regularization techniques and followed by the implementation

**Ridge Regression (L2 Regularization)**

we are going to minimize the sum of squared errors and sum of the squared coefficients β. In the background,  
the coefficients β with a large magnitude will generate the graph peak and  
deep slope, to suppress this we using the lambda λ use to be called a  
Penalty Factor and help us to get a smooth surface instead of an irregular-graph. Ridge Regression is used to push the coefficients β value nearing zeroin terms of magnitude. This is L2 regularization, since its adding a penalty-equivalent to the Square-of-the Magnitude of coefficients.

Ridge Regression = Loss function + Regularized term

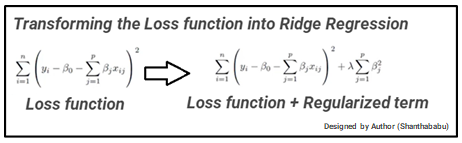
#### ***Lasso Regression (L1 Regularization)***

This is very similar to Ridge Regression, with little difference in Penalty Factor that coefficient is magnitude instead of squared. In which there are possibilities of many coefficients becoming zero, so that corresponding attribute/features become zero and dropped from the list, this ultimately reduces the dimensions and supports for dimensionality reduction. So which deciding that those attributes/features are not suitable as predators for predicting target value. This is L1 regularization, because of adding the Absolute-Value as penalty-equivalentto the magnitude of coefficients.

#### **Elastic-Net Regression Regularization**

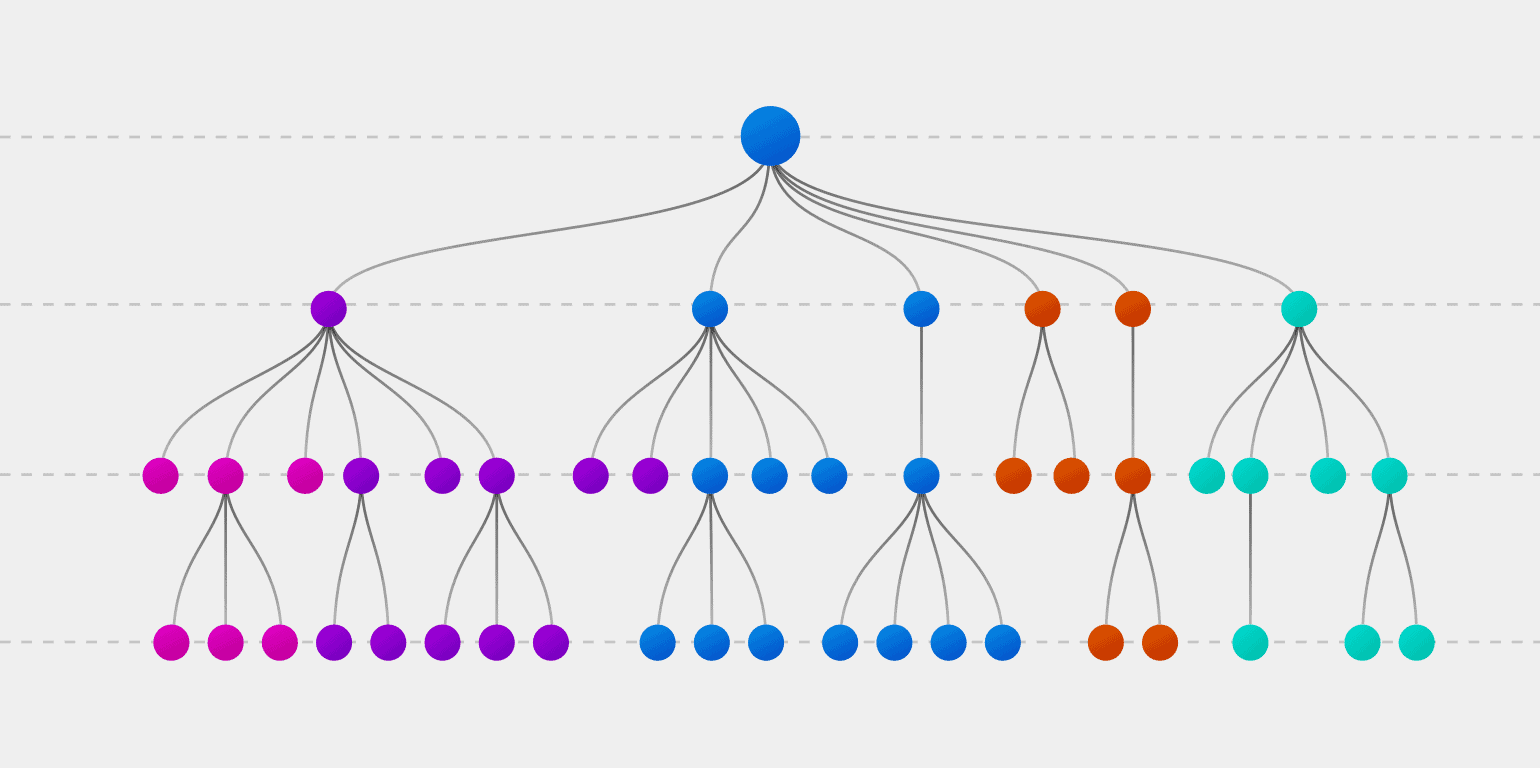
#### Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the [lasso](https://corporatefinanceinstitute.com/resources/knowledge/other/lasso/) and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

**Decision Tree**

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 fall under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]

2. Result [End Nodes]

The decision is made by the branches/edges based on the truth of the statement, in the example below which shows a decision tree.

**Random Forest**

Random forest is a almighty tool which ensembles decision trees and bagging .The base learner of random forests is a binary tree constructed by recursive partitioning (RPART) and then developed using classification and regression trees . Binary splits of the parent node of a random forest splits data into two children’s nodes and increases homogeneity in children nodes compared to parent nodes. Note that a random forest does not split tree nodes based on all variables; instead, it chooses random variable subsets as candidates to find the optimal split at every node of every tree . Then the information from the n trees is aggregated for classification and prediction . Random forests also provide the importance of each feature by accumulated Gini gains of all splits in all trees representing.

**Gradient Boosting**

Gradient Boosting like Random Forests is an ensemble learning method. Similar to latter, it uses multiple weak learners which are combined to form a strong learner. But unlike its Random Forests, Gradient Boosting as the name suggests uses boosting. Boosting methods work iteratively to create a new learner at every stage; these new learners are then trained on the error residuals at a current iteration to produce new learners which are stronger than the previous stage. Applied to decision trees, every decision tree is works on the error residuals of the previous iteration to produce a better decision tree. The collection of these decision trees is then used as the overall model for predicting values.

**Model Performance**

Model can be evaluated by various metrics such as:

**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 RSquare suggests all our Independent Variables being

significant, despite both values being on a relatively lower

side.

**Mean Absolute Error (MAE)** like the RMSE, the MAE measures the prediction

error. Mathematically, it is the average absolute difference between observed

and predicted outcomes,

MAE = mean(abs(observed - predicted)). MAE is less sensitive to outliers compared to RMSE.

**Root Mean Squared Error (RMSE**) which measures the average error performed by the model in predicting the outcome for an observation. Mathematically, the RMSE is the square root of the mean squared error (MSE), which is the average squared difference between the observed actual outcome values and the values predicted by the model. So, MSE = mean((observerd - predicteds)^2) and RMSE = sqrt(MSE). The lower the RMSE, the better the model.

**HyperParameter Tuning**

Hyper-parameters are information that we use to control our parameters in order to get good results. The Grid Search CV was used for tuning.

**Grid Search CV**

It is the process of performing hyperparameter tuning in order

to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.GridSearchCV is a function that comes in Scikit-learn’s(or SK-learn) model\_selection

package.So an important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

**Conclusion**

The best model is the random forest model and gradient boosting model in our study. This study shows that the rents of bikes are influenced by a lot of features. In this study, we understood that many koreans usually and mainly rent bikes during the week days, so we supposed that the main use is to go to school or work. There are also many conditions which contribute to the variation of number of rents like the the day of the week, the moment of the day and weather conditions. Weather conditions are also very important because there are more rents during spring and summer. And as we expected more people are set to rent bikes when the weather is favorable.