**Seoul Bike Sharing Demand Prediction**

**Almabetter, Bangalore**

**Kajal Dhun**

**Navinkumar Sambari**

**Tanu Rajput**

****

**Abstract**

Bike Sharing is the process by which bicycles are procured on several basis- hourly, weekly, membership-wise, etc. In our 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 other variables like Rental Bike Count, etc. For the available raw data, firstly, a thorough pre-processing was done after which a Here, rental bike count variable is the target variable. The dataset presents the company’s data between 1st of December’ 2017 to 30th of November’ 2018 which finishes one year later. While doing this project various models have been created. These various models are being analyzed and we tried to study various models to intuitively get the best performing model for our project to predict the demand of bikes per hour.

***Keywords: Temperature, Hour, Rental Bike Count, dataset, model, predict, demand.***

**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 goal of the company Seoul Bike is providing the city with a stable supply of rental bikes. It becomes a major concern to keep users satisfied. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

**Introduction**

Many countries have bike sharing systems, such as Ddareungi is a bike sharing system in South Korea, which started in the year 2015, known as Seoul bike in English. It was started to overcome issues like greater oil prices, congestion in traffic and pollution in the environment and to develop a healthy environment for citizens of Seoul to live. There must be a system that can estimate enough bicycle supplies at certain docking stations, which will save customers’ time and allow them to pick up the bicycle at the nearest bicycle station. With an efficient and accurate model that can solve the bike supply problem, machine learning is very helpful in solving such a problem. .By using machine learning methods, these problems can be resolved to some extent. Forecasting the rental bike demand is a very crucial part not only for individuals, but as an example for other mobility share companies that are yet to come. In this study, multiple regression algorithms will be used to predict bike count. An evaluation metric will be used to measure the performances of the regressors, and the best regression will be selected.

**Project steps involved**

**Preprocessing**

Data preprocessing is a data mining technique which consists in transforming the data in order to make it understandable. In machine learning, the data preprocessing step is critical because it involves cleaning, integration, transformation, scaling, standardizing data and many other tasks, in order to have a good preparation for the application of models.To begin we first did some data exploration by checking types, missing values, duplicate values and data description. In this dataset there are neither null values nor duplicate values. We also changed the date type to DateTime which was initially a str object. From the date, we also created three columns with the day of the week and the month and the year corresponding. We also changed the datatype of the hour feature from int to object.

**Exploratory Data Analysis**

We performed this method by comparing our target variable that is Rented Bike Count with other independent variables visualizing in the form of boxplot and pointplots. Then we created a pie chart of all categorical variables to see the percentage distribution of value counts of each categorical variable. Then we saw the relation between variables through pairplots, regression plots, and correlation matrix. Also we saw the distribution of the target column which is rightly skewed. So, it needs to be transformed.

**Multicollinearity Detection**

We used the best technique called VIF(Variance Inflation Factor) for the detection of the collinearity among variables. And we found two variables which are multicollinear. So, we decided to remove one of the variables among those two.

**Encoding**

We used Pandas get dummies 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. And also we used square root technique on rental bike count during splitting the dataset.

**Standardizing the feature**

We used Yeo Johnson Power Transformation technique to standardize the 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.

**Fitting the models using regression algorithms**

We have a regression problem because our target is the number of rented bikes per hour. So the goal of this part is to apply many algorithms in order to find the algorithm with the best indicator.

For modeling we tried various regression algorithms like:

1. **Linear Regression**
2. **Regularization = Ridge**
3. **Regularization = Lasso**
4. **Polynomial with Linear Regression**
5. **Decision Tree**
6. **Random Forest**
7. **Gradient Boost**
8. **ExtraTreesRegressor**

**Tuning the hyper parameters for batter accuracy**

Tuning the hyper parameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in the case of tree-based models like Random Forest Regression and ExtraTreesRegressor.

**Regression Algorithms used**

1. **Linear Regression**

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

Yi =β0+β1xi1+β2xi2+...+βpxip+ϵ

1. **Lasso**

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.Ridge Regression. The only difference in ridge and lasso loss functions is in the penalty terms. Under lasso, the loss is defined as:



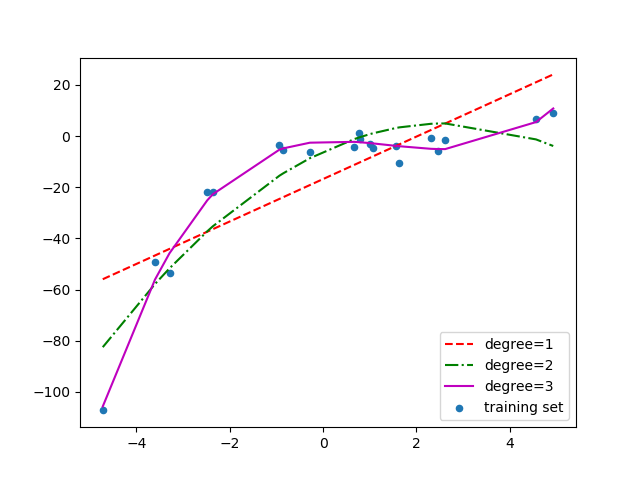
1. **Ridge**

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. 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. In Ridge Regression, the OLS loss function is augmented in such a way that we not only minimize the sum of squared residuals but also penalize the size of parameter estimates, in order to shrink them towards zero:



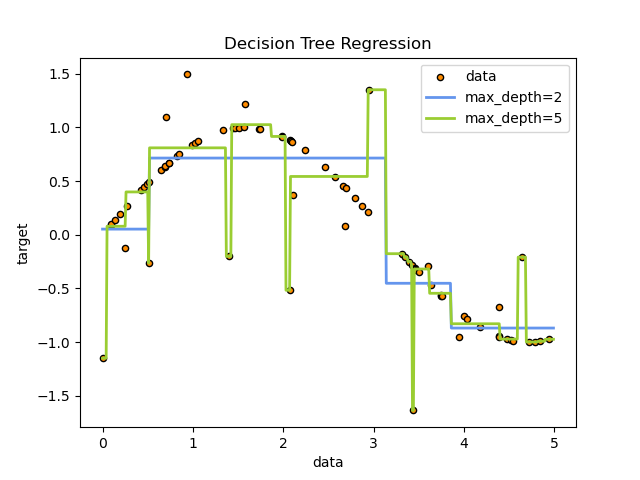
1. **Polynomial Regression**

Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial. Polynomial Regression is sensitive to outliers so the presence of one or two outliers can also badly affect the performance and here we use degree=2



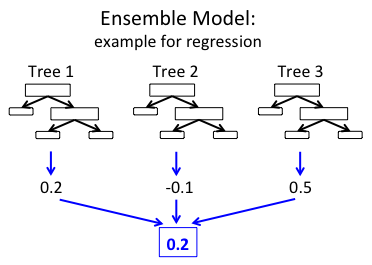
1. **Decision Tree**

A decision tree is a graphical representation of possible solutions to a decision based on certain conditions. It's called a decision tree because it starts with a single box (or root), which then branches off into a number of solutions, just like a tree



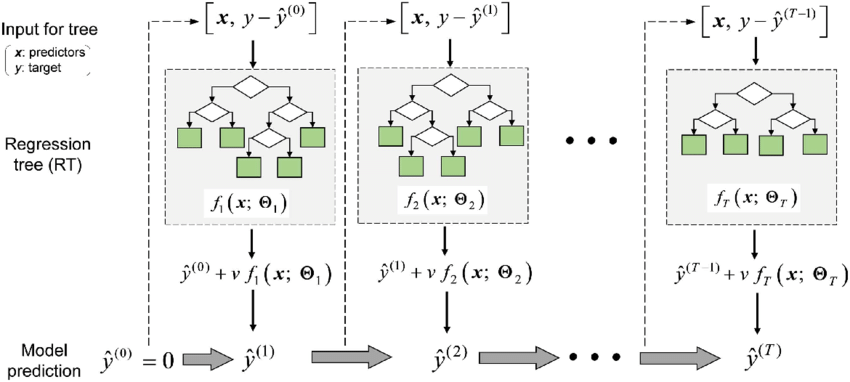
1. **Random Forest Regressor**

Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. ... A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees



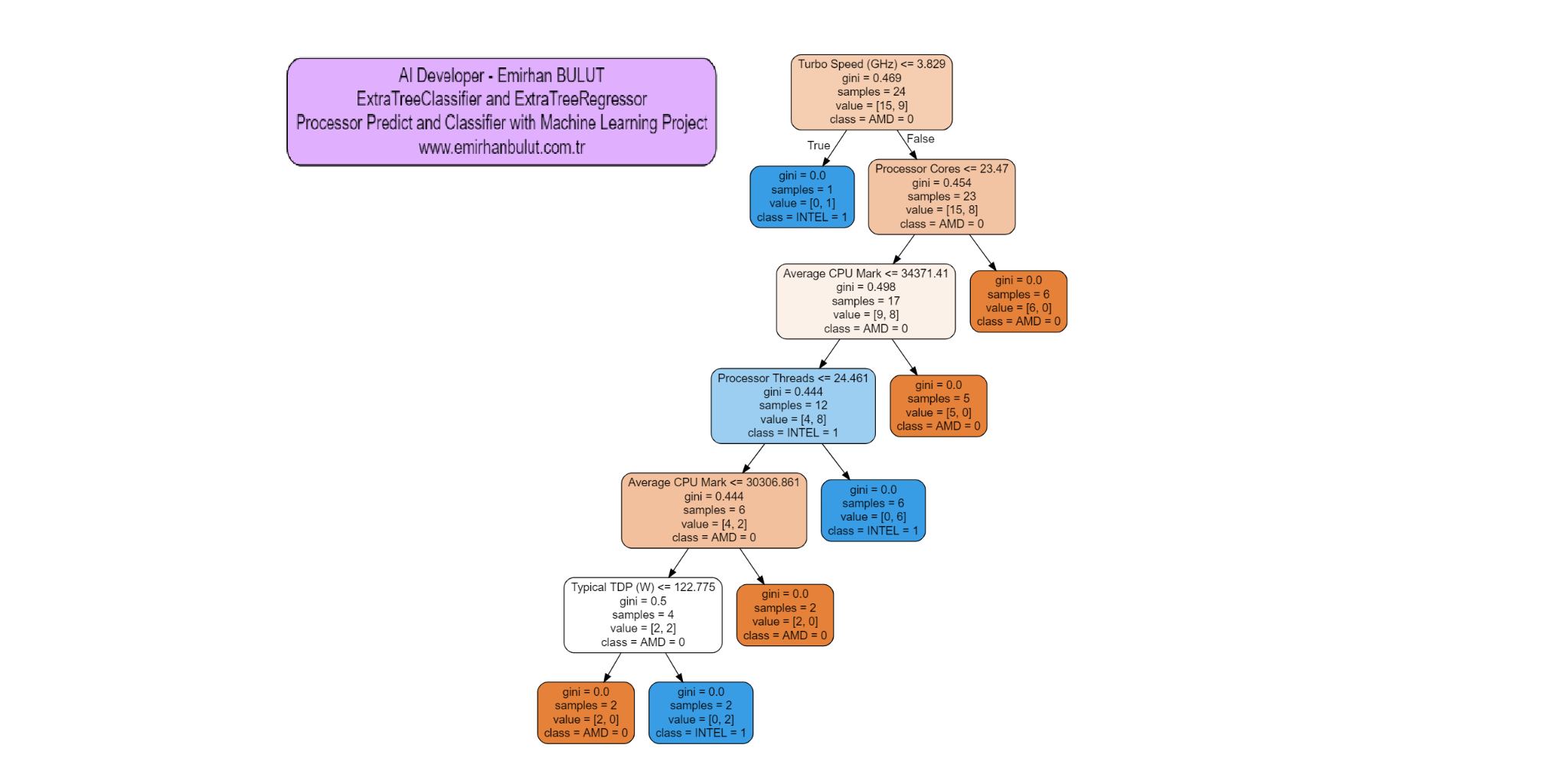
1. **Gradient Boosted Regressor**

Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.



1. **ExtraTreesRegressor.**

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

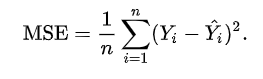


**Model Evaluation Metrics**

Model can be evaluated by various metrics such as:

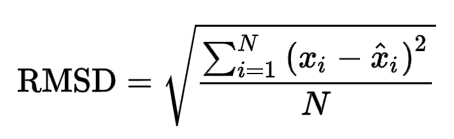
1. **MEAN SQUARE ERROR**

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.

****

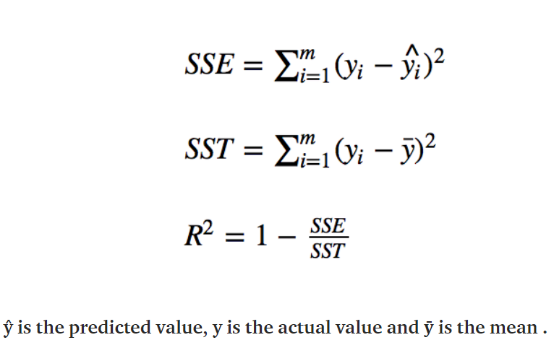
1. **Root mean square error**

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

****

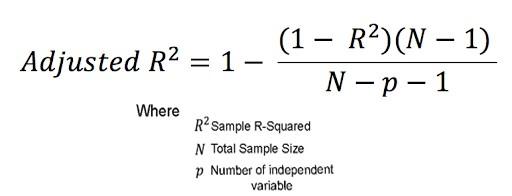
1. **R square:**

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 increase even if that parameter does not improve model



1. **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.



**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, for hyperparameter tuning. This also results in cross-validation.

* **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.

**Conclusion**

### We observed that bike rental count is higher during weekdays than weekend days.The rental bike counts are **at its peak at 8 AM** in the morning and **6pm in the evening.** We observed that people prefer to rent bikes during moderate to high temperatures. Highest rental bike count is during Autumn and summer seasons and the lowest in winter season. Comparing the Adjusted R2 among all the models, **ExtraTreesRegressor** gives the highest Score where the Adjusted **R2 score is 0.908699** and **Training score is 0.987167.** Therefore, this model is the best for predicting the bike rental count on hour basis