### Supervised ML Regression Capstone Project

Title: Bike Sharing Demand Prediction

by: Sneha Raikar

  Data Science Trainee, AlmaBetter

**Abstract:**The objective of this work is to predict the trip duration of rental bikes in the Seoul Bike sharing system. The data used include weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Feature engineering is done to extract additional features from the data. Four statistical models are used to predict the trip duration.

1. Linear regression
2. Lasso regression
3. Decision Tree
4. Extreme Gradient Boosting (XGBoost)
5. Random Forest (RF).

**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 main objective is to make a predictive model, which could help them in predicting the bike demands proactively. This will help them in stable supply of bike wherever needed.

The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

**Attribute Information:**

• Date: year-month-day

• Rented Bike count - Count of bikes rented at each hour

• Hour - Hour of the day

• Temperature-Temperature in Celsius

• Humidity - %

• Wind speed - 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. Introduction**

Today, bike-sharing systems are blooming across more than 1000 cities around the world, particularly in big or large cities like New York City, Paris, Washington DC, London, Beijing, and Barcelona. To complete a short trip renting a bike is a faster way when compared to walking. Moreover, it is eco-friendly and comfortable compared to driving.

Due to global warming, continuous pollution, and depletion of sources of energy. Many countries have been focused on using renewable energy which does not harm environment and can be reused as well. South Korea is one the country which has adapted to it and their most used service is rented bikes in Seoul. But in order to avoid

any difficulties such as waiting time it is necessary to have an estimate of future demand. Our goal here is to build model that can predict bike sharing demand considering all the factors which have their effects.

**3. Major Factors Affecting Bike Demand**

**i) Rainfall:** People tend to use rented bikes quite frequently due to the fact that They can be easily rented from any place and can be dropped off any other place, cheap enough to rent daily, but conditions like Rainfall affects its rental count a lot.

People don’t rent bikes during the rainy season. So, we can say rainfall is negatively correlated with rented bike count.

**ii) Snowfall:** Similarly, as rainfall, snowfall negatively affects rented bike count

as it's hard to drive on snowy roads.

**iii) Visibility:** At times when one can’t see properly, it’s natural for them to avoid

driving, and this is what affects the rented bike count. Although in Seoul the cases of these are quite low.

**iv) Temperature:** It is seen that people avoid renting bikes at low temperatures.

Seoul is a place with an average temperature of 27 to 32 degree Celsius.

So, when temperature become warm people tend to enjoy it which has an effect in

renting bikes as well.

**v) Working Day or Not:** Compared to an Off day, people rent bikes more on a working day. Reason behind this is being the same I.e., they can be easily rented from any place and can be dropped off any other place, cheap enough to rent daily

**vi) Traffic:** Even though this isn't mentioned in data, traffic also supports renting bike count indirectly. If traffic is high or large people visiting nearby walk or rent a bike for purpose.

**4. Steps Involved**

**I. Exploratory Data Analysis:**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations. It gives us better idea of which feature behaves in which manner compared to target variable.

**II. Data Cleaning:**

Our Data contains some null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project to get better results.

**III. Encoding of Categorical Columns:**

We used one Hot Coding 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 machine and needs to be encoded.

**IV. Feature Scaling:**

Feature scaling is essential for machine learning algorithms that calculate distances between data. If not scale, the feature with a higher value range starts dominating when calculating distances.

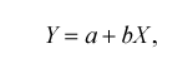
**V. Fitting Models:**

At first, we tried with basic linear regression and also with Lasso regularization technique but soon realised we will need a much more complex model and so we then used a Decision tree Regressor, XGB Model and Random Forest Regressor compared the results.

**5. Algorithms**

**I. Linear Regression:**

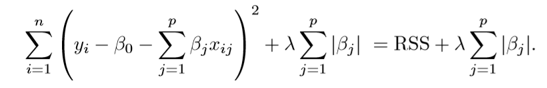
Linear regression (LR) is the simplest statistical regression method for identifying the linear link between the independent and the dependent variables. It is done by fitting a linear equation of line to the observed data. For fitting the model, it is utmost important to check, whether there is a connection between the variables or features of interest, which is supposed to use the numerical variable, that is the correlation coefficient. The following equation defines an LR line:



where X is the independent variable whereas Y is a dependent variable. b is the slope of the line and a is the intercept (the value of y when x = 0).

**Regularization:**

**II.** **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. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity. Lasso Regression uses L1 regularization technique.



This variation differs from ridge regression only in penalizing the high coefficients. It uses |βj|(modulus)instead of squares of β, as its penalty.

**III. 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 branches/edges represent the truth/falsity of the statement and makes a decision based on that in the example below which shows a decision tree.

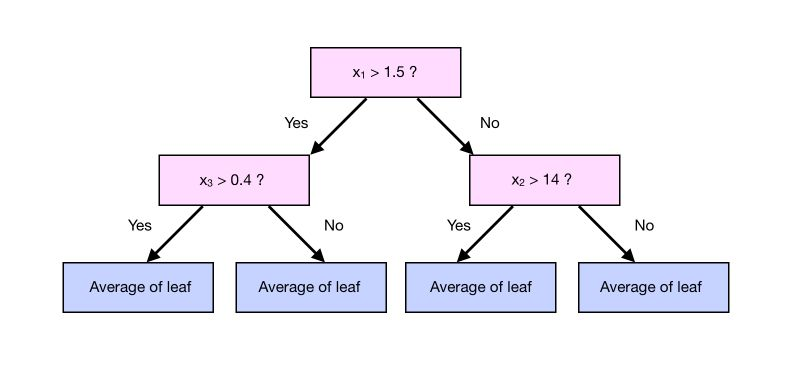


Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

**IV. XGBoost:**

To understand XGBoost we have to know gradient boosting beforehand.

* Gradient Boosting:



Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

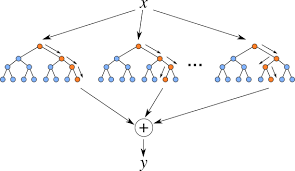
How are the targets calculated? The target outcome for each case in the data depends on how much changing that case's prediction impacts the overall prediction error.

**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.

**V. Random Forest:**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.



**6. Evaluation Metrics to check Model Performance**

**1)Mean Squared Error (MSE):** MSE or Mean Squared Error is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model.

MSE = mean ((observed - predicted) ^2)

**2)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.

RMSE = sqrt(MSE). The lower the RMSE, the better the model.

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

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

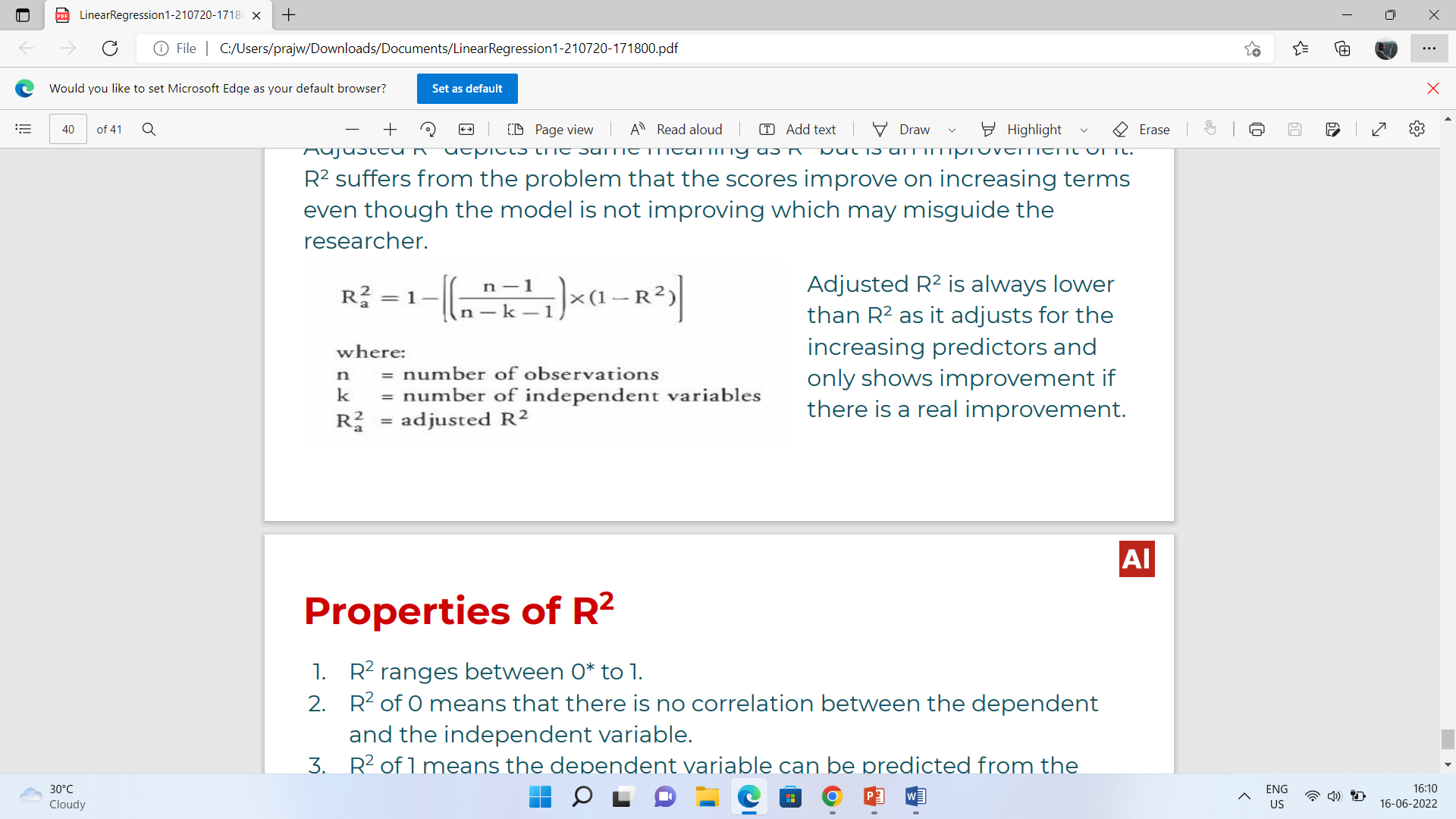
and predicted outcomes.

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

**4)R-squared (R²):** which is the proportion of variation in the outcome that is explained by the predictor variables. In multiple regression models, R2 corresponds to the squared correlation between the observed outcome values and the predicted values by the model.

R² = 1- MSE(model)/MSE(Baseline). The Higher the R-squared, the better the model.

**5)Adjusted R²:** Adjusted R² depicts the same meaning as R² but is an improvement of it. R² suffers from the problem that the scores improve on increasing terms even though the model is not improving which may misguide the researcher.



Adjusted R² is always lower than R² as it adjusts for the increasing predictors and only shows improvement if there is a real improvement

**7. Hyper Parameter Tuning**

Hyper-parameters are those sets of information that are used to control our parameters in order to get good results. We used Grid Search CV for hyper parameter 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 hyper parameters. Note that there is no way to know in advance the best values for hyper parameters 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.

**8. Conclusion:**

The analysis is done with Seoul Bike data. Four regression techniques Linear Regression, Decision Tree, XG Boosting and Random Forest are used to predict the trip duration. This statistical data analysis shows interesting outcomes in prediction methods and also in an exploratory analysis.

**The experimental results show that:**

* Most numbers of Bikes were rented in **Summer**, followed by **Autumn**, **Spring**, and **Winter**. **May-July** is the peak Bike renting Season, and **Dec-Feb** is the least preferred month for bike renting.
* **Majority** of the **client** in the bike rental sector belongs to the **Working class**. This is evident from EDA analysis where bike demand is more on weekdays, working days in Seoul.
* **Temperature** of **20-30 Degrees**, **evening time 4 pm- 8 pm, Humidity** between **40%-60%** are the most favorable parameters where the Bike demand is at its peak.
* **Temperature, Hour** of the day, **Solar radiation**, and **Humidity** are major driving factors for the Bike rent demand.
* Feature and Labels had a weak linear relationship, hence the prediction from the linear model was very low. Best predictions are obtained with a **Random Forest** model with an accuracy of **0.875**