**Capstone Project**

**Supervised ML Regression**

**Bike Sharing Demand Prediction**

* **Abstract**

The contents of the data came from a city called Seoul. It is the capital city of South Korea and has a population of around 9.7 million people. It was the 4th largest metropolitan economy in 2014. It has a humid continental climate influenced by monsoons.

Quick transportation is a big need in most cities. Ola & Uber are providing good transportation but nowadays the rates are getting high. The rented bike is a good option for transportation which is quick & cheap.

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.

* **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.

* **Data Description**

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 - Humidity in the air in %

Windspeed - Speed of the wind in m/s

Visibility - Visibility in m, 10m

Dew point temperature - Dew point temperature in Celsius

Solar radiation - Energy radiated by Sun in MJ/m2

Rainfall - Amount of raining in mm

Snowfall - Amount of snowing in cm

Seasons - Winter, spring, Summer, Autumn

Holiday - Holiday/No holiday

Functional Day - NoFunc (Non Functional Hours), Fun (Functional hours

* **Project Description**

  First I explore the data, cleaned and preprocessed the data and then I performed the exploratory data analysis to extract information, in which we identified certain trends, relationships, correlation and found out the features that had some impact on our dependent variable and plotted the graph to visualize the impact on dependent variable. I also encoded the categorical variables.

I build the various machine learning algorithms on our split and standardized data. I tried different algorithms namely; Linear regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest and Gradient boosting algorithm. I did hyper parameter tuning and evaluated the performance of the model.

* **Steps Involved**

**Data Wrangling**

After loading the dataset, we performed this method by cleaning, organizing, and transforming raw data into the desired format to understand the data clearly. This process helped to tackle the unwanted data, to produce accurate results, to make better decision.

**Exploratory Data Analysis**

After Data wrangling, we performed EDA. Comparing the 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 I observed the distribution of variables. It gave us a better idea of how the feature behaves with the target variable.

**Continuous and categorical Features Analysis**

With the help of exploratory data analysis we analyzed the categorical as well as numerical features in the dataset.

**Analysis of Dependent Variable**

We analyze our dependent variable, a dependent variable is a variable whose value will change depending on the value of another variable.

**Regression Plot**

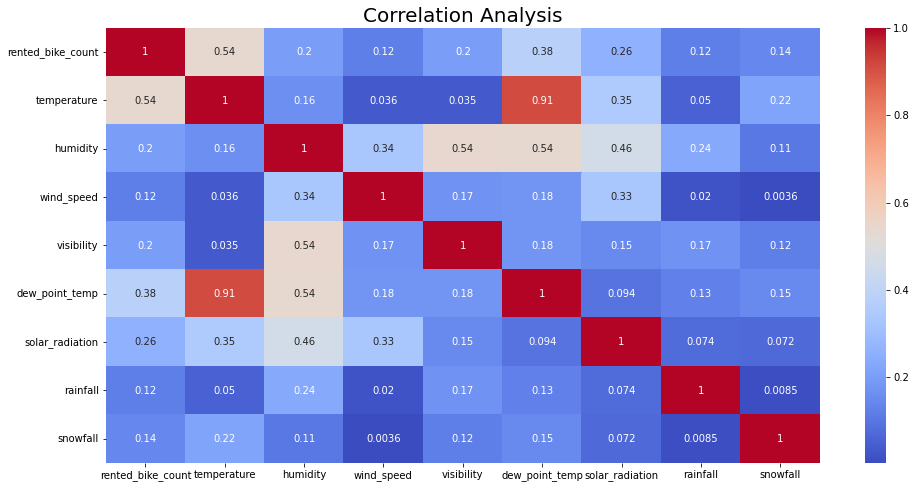
The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses. Regression plots as the name suggests creates a regression line between 2 parameters and helps to visualize their linear relationships.

**Create Dummy Variable and Data encoding-One hot encoding**

In this dataset some categorical variables like seasons, holiday and function day, we change it with a numerical database.

**Correlation Analysis**

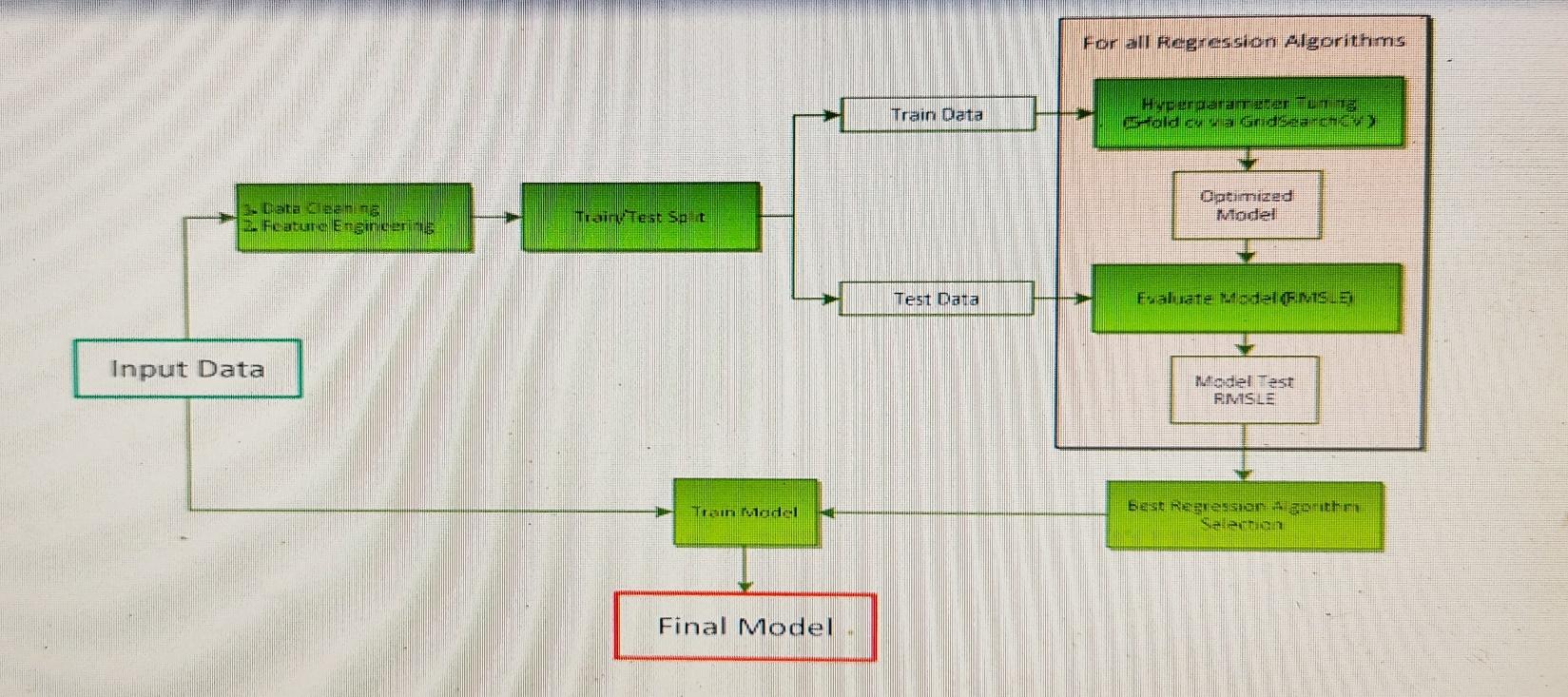
We plot the heat map to find the correlation between both dependent variables and independent variables.



* **Model Training**

**Train test Split**

Before fitting any model it is a rule of thumb to split the dataset into a training and test set. The proportions but mostly used is 80:20 for training and testing respectively.



* **Model Building**

. **Models**

I uses 6 model to train the data and for predicting the accuracy

1. **Linear regression**
2. **Lasso regression**
3. **Ridge regression**
4. **Decision Tree**
5. **Random Forest**
6. **Grading Boosting**

**Hyper parameter Tuning**:

To improve the performance of model

GridSearchCV helps to loop through predefined hyper parameters and fit the model on the training set. So, in the end, we can select the best parameters from the listed hyper parameters.

**Performance Indicator**

**The mean squared error (MSE)** tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. It’s called the mean squared error as you’re finding the average of a set of errors. The lower the MSE, the better the forecast.

**MSE formula = (1/n) \* Σ (actual – forecast)2** Where:

n = number of items,

Σ = summation notation,

Actual = original or observed y-value,

Forecast = y-value from regression.

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors).

**Mean Absolute Error (MAE**) are metrics used to evaluate a Regression Model. ... Here, errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable.

**R-squared (R2)** is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

**Formula for R-Squared**

**R2=1− (Unexplained Variation / Total Variation)**

**​**

**Adjusted R-squared** is a modified version of R-squared that has been adjusted for the number of predictors in the model.​

* **Conclusion**

**Normally distributed attributes:** temperature, humidity.

**Positively skewed attributes:** wind, solar\_radiation, snowfall, rainfall.

**Negatively skewed attributes:** visibility.

Generally people use rented bikes during their working hours from 7am to 9am and 5pm to 8pm.. The demand for rented bikes is low especially in the morning hour but when the evening starts from 4 pm to 8 pm the demand slightly increases. In summer season the use of rented bikes is higher and lower in winter because of snowfall. The rented bike count is higher on working days than on non-working days. In holiday people uses the rented bike from 2pm-8pm

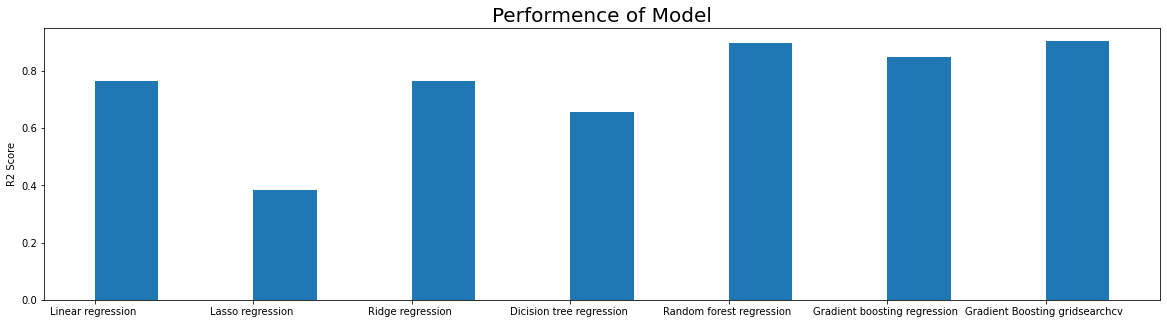
From the above regression plot we see that the columns 'Temperature', 'Wind\_speed','Visibility', 'Dew\_point\_temperature', 'Solar\_Radiation' are positively relation to the target variable, which means the rented bike count increases with increase of these features. Rainfall’,’Snowfall’,’Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

The demand of rented bike is uniformly distribute despite of wind speed but when the speed of wind was 7 m/s then the demand of bike also increase that clearly means peoples love to ride bikes when its little windy and when it is pretty hot around 25°C in average

We implemented 6 machine learning algorithms: Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest and Gradient Boost. We did hyper parameter tuning to improve our model performance.

The **results** of our evaluation are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **R2 Score** | **Adjusted R2** |
| Linear Regression | 4.66 | 37.13 | 6.09 | 0.76 | 0.76 |
| Lasso Regression | 7.443 | 97.08 | 9.8 | 0.38 | 0.37 |
| Ridge Regression | 4.66 | 37.13 | 6.09 | 0.76 | 0.79 |
| Decision Tree | 5.40 | 54.28 | 7.36 | 0.65 | 0.65 |
| Random Forest | 2.605 | 16.382 | 4.04 | 0.89 | 0.89 |
| Gradient Boosting | 3.62 | 23.65 | 4.86 | 0.85 | 0.85 |
| Gradient Boosting using GridsearchCV | 2.65 | 15.05 | 3.87 | 0.90 | 0.90 |



**Random forest Regres**sor and **Gradient Boosting gridsearchcv** gives the highest **R2 score** of **89%** and **90%** respectively.

Feature Importance value for Random Forest and Gradient Boost are different.

We can use **Random forest Regres**sor and **Gradient Boosting gridsearchcv** for predicting bike rented column on a daily basis.