**Capstone Project**

**Supervised Ml-Regression**

**Seoul Bike sharing Demand Prediction**

## **Project Overview**

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 over this the rented bike is a good option for transportation which is quick & cheap. The aim of this project is to check the availability of the bike for a particular time in the day

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 he day

Temperature-Temperature in Celsius

Humidity - Humidity in the air in %

Wind speed - 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

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. We did hyper parameter tuning and evaluated the performance of the model.

I analyze the data and build the model by considering the below **problem description**

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.

**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 the summer season the use of rented bikes is higher and lower in the winter season 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 related 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 for rented bikes is uniformly distributed despite wind speed but when the speed of wind was 7 m/s then the demand of bikes also increased. This clearly means people love to ride bikes when it’s little windy and when it is pretty hot, around 25°C on average.

I implemented 6 machine learning algorithms: Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest and Gradient Boost. i 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 Regressor** and **Gradient Boosting using 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 columns on a daily basis.

**Github Link**

https://github.com/YashwantRaul/Supervised-ML-Regression

**Drive Link**

https://drive.google.com/drive/folders/1IoZODRB8wYh79cu4MeDt5btOixaMM-dI?usp=sharing