Adil Khan, Sunil Panigrahi, Shubham Kumar,Vivek Singh , Sharaffin B
Data Science Trainees, Almabetter , Bangalore

Seoul Bike Sharing Demand Prediction

ABSTRACT:

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world. Bike-sharing systems have widely spread over many cities in the world as an environmentally friendly means to reduce air pollution and traffic congestion.

For survival amongst today's fierce competition, companies need to upgrade their prediction model to better predict customer demand in a more accurate manner. This study explores a new feature for bike share demand prediction models that resulted in an improved accuracy score. Our experiment can help understand what could be the reasons affecting the Rentals and prediction with machine learning algorithms taking into account previous trends..

.

INTRODUCTION

Users can check their travel details (distance, duration) and measure their physical activities (calories burned). Due to such smart technology and convenience along with the increased travel, the use of rental bikes is increasing day by day. Therefore, you need to be able to manage the demand for rental bikes and manage the continuous and convenient service to users effectively. This study proposes a machine learning-based approach that includes different metrics to predict the rental bikes demand across the city. The ML model is used to predict the number of rental bikes required per hour. Rental bikes demand is modeled using the available independent variables. The management can use this to understand exactly how the demand vary with different features. So, they can manipulate business strategies to meet the demand levels and customer expectations. In addition, this model is a great way for management to understand the demand dynamics of a new market.

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 SUMMARY

The provided data set has following different columns of variables necessary for bike sharing

- **Date** : year-month-day

- **Rented Bike count** - Count of bikes rented at each hour
- **Hour** - Hour of the day
- **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)

DATA PIPELINE

- .Data Processing

  In the first part, we have imported necessary libraries and data set. We then used these libraries to understand the data.
- Data Cleaning

  After understanding the data, we got to know that there are no null values or duplicate values in our data set.
- Data Preparation

  For the EDA, We refactored the datetime feature. We can't analyse non numerical values. We transformed it ("yyyy/mm/dd") into date, hours, day of year and year.
- Exploratory Data Analysis

  After preparing the data set, we did some exploratory data analysis using tables and graphs to derive the observations from the data and to better understand the problem statement, and make ways to the solution to the problem statement.

REFACTORING DATETIME

Datetime is a string. This is a problem because strings cannot be processed mathematically. We transformed the string into a date and then extracted the features hour, Day of the year, week day and year. Extraction of feature year is what improved performance the most.

The hour feature was a particularly challenging problem. It can be considered a categorical feature on its own. There are few rentals at 1, 2, 3 am and at 10 and 11 pm. The relationship between the hour and the number of rentals is not linear.

EXPLORATORY DATA ANALYSIS

While doing the Exploratory Data analysis we tried finding the factors affecting the rental bikes demand. The factors affecting the Rental Bikes Demand are:

- Temperature
- Humidity
- Wind Speed
- Visibility
- Rainfall
- Snowfall

From the below graphs, it is evident that the major factor affecting the Rental bikes demand are Temperature, Visibility, Rainfall and Snowfall.

DATE WISE ANALYSIS

While doing Date-wise analysis of the given rental bike demand dataset, we answered the following questions:

- At what time of the day the Rental bike demand is the highest?
- On which day the Rental bike demand is the highest?
- On which date the Rental bike demand is the highest?
- Which Year shows the most demand for Rental Bikes?

:

DISTRIBUTION OF FEATURES

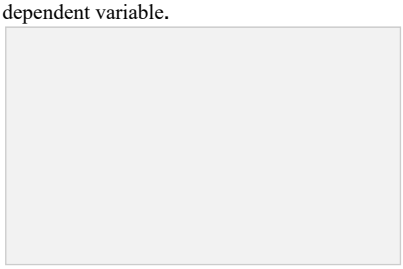
- ‘Temperature’ and ‘Humidity’ columns follows uniform distribution.
- ‘Dew Point Temperature’ and ‘Visibility’ are negatively skewed.
- ‘Wind Speed’ , ‘Solar Radiation’ , ‘Rainfall’ and ‘Snowfall’ are having positively skewed distribution.

CORRELATION

Below are the graphs of positive
Negative correlation between the variables

MULTICOLLINEARITY

From the below graph we can see that Temperature and Dew_point_temperature are highly correlated, keeping the factor of 0.91 . And, then we have hour in the graph which is having good correlation with our dependent variable.

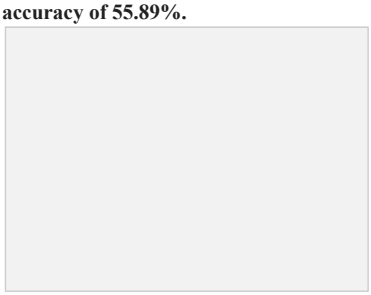


MACHINE LEARNING - SUPERVISED LEARNING – REGRESSION

We have used different ML Models to determine the prediction of Rental Bikes needed per hours.

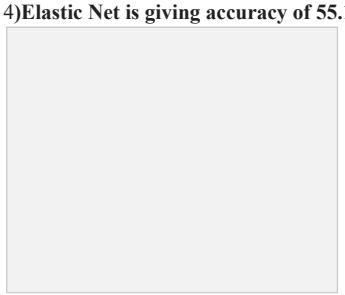
1) The Linear regression model is giving an accuracy of 55.18%.

2)The Lasso regression is giving the accuracy of 55.89%.



3)The Ridge Model is giving the accuracy of 55.73%.

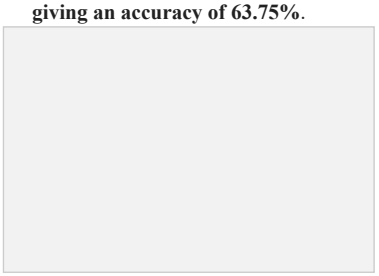
4)**Elastic Net is giving accuracy of 55.18%**



5)Decision Tree is giving accuracy of 83%.

6) **Random Forest Regression Model is giving an accuracy of 87.62%.**

7)The ADABoost Regression model is giving an accuracy of 63.75%.



8) the Gradient Boosting is giving an accuracy of 89.23%

9) **XGboost is giving an accuracy of 89%**.