Supervised ML (Regression) on Seoul Bike Dataset

Soma Pavan Kumar
Data science trainee,
AlmaBetter, Bangalore

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 rent a bike from a 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.

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.

Our Experiments will help us understand, what the data has to offer and what Machine Learning Algorithm model works the best to predict how many rental bikes should be available at a given particular time depending upon various variables such as, Time, Date, Weekday, Temperature, Rainfall of a given day and much more variables.

Keywords: Data Wrangling, Pandas, Matplotlib, Seaborn, Machine Learning, Decision trees, Random Forests, Multiple Linear Regression, Ridge, Lasso.

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.

2. Introduction

It is important to make the rental bike available and accessible to the public, as it provides many alternatives to commuters in metropolises. There are a lot of advantages to bike rents, it is convenient because it permits people not to keep the bike all day long, whether it is at work or at school.

Furthermore, it is the healthiest way to travel and it has environmental benefits. The studied dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the target is the number of bikes rented per hour and date information. The dataset presents the company's data between December the 1st of 2017 and finishes one year later. This study could have many aims for the company that could be seeing the results of the past year. It could also help them ameliorate themselves to become better and have full satisfaction from customers.

3. Data Preprocessing

Data preprocessing is a data mining technique which consists in transforming the data in order to make it understandable. It could be changing the type, the format, splitting the data, verifying that there are no missing values but also creating new columns thanks to columns we initially have. In machine learning, the data processing 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 and data description. We also changed the date type to DateTime which was initially a str object. Then, we created a column which takes the hour of the day and returns if it is the day of the week night or day moment of the day (because we remind you that data is collected by hour), in order to do data visualization with the target. From the date, we also created two columns with the day of the week and the month of the year corresponding. And finally also did an encoding on the day to convert 'WeekDays' in order to visualize it and make it a feature for ML models.

Pandas:

It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science. Pandas are also able to delete rows that are not relevant, or contain wrong values, like empty or NULL values. This is called cleaning the data.

Pandas library also provides us with functions such as describe () which provides us with statistics of the data including mean, mode, median and also Interquartile ranges.

Pandas library also gives us some functions like shape, which gives us the number of rows and columns in the whole data. Also, some functions like shape which gives us all the number of cells or elements in the whole of data.

Pandas also has some features like drop () which drops the whole column or a row as specified by the user.

4. Data Visualization

Data Visualization is an interdisciplinary field that deals with graphic representation of data. It is a particular efficient way of communication when the data is numerous as for example a time series

The first plot is maybe one of the most important because it shows all the correlations of the features. To complete this visualization, we created a ranking of the features which are the most correlated to the target. So, it gives an idea of which features we have to focus on.

Moreover, in the notebook, you can see that as expected the rise of rainfall and snowfall comes with a decrease of the rents which is totally logical. We also created a feature:" Night/Day" in order to see the distribution following the moment of the day. From 8pm to 5am, we decided to qualify this moment as 'night' and the rest is 'day'. Finally, we wondered what were the hours during which the rents were the highest, and we found that it was around 8am and around 6pm which confirms that people take bikes to go to school or work and go home at night. This analysis is very interesting because it shows that Koreans take global warming very seriously and respect the earth.

All of the Data Visualizations In this project are done using Matplotlib.pyplot, which is an inbuilt library in python that deals with plotting graphs including, bar graphs, histogram, all the way till pie charts and scatter plots.

Matplotlib:

Matplotlib is a low-level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter.

Matplotlib is open source and we can use it freely. Matplotlib is mostly written in python, a few segments are written in C, Objective-C and JavaScript for Platform compatibility.

Plotting in Matplotlib:

The plot () function is used to draw points (markers) in a diagram. By default, the plot () function draws a line from point to point.

The function takes parameters for specifying points in the diagram. Parameter 1 is an array containing the points on the **x-axis**. parameter 2 is an array containing the points on the **y-axis.**

In this project we've had 5 visualizations:

1. The First one is to find the Correlations between all the independent variables/Features.

2. The Second is to find the maximum number of rides happen at what time of the day.

3. The Third is to find the frequency of rides taking place according to the day of the week.

4. The fourth is to find out what is the sweet spot of temperature in which Koreans like riding their bikes.

5. Also, we found what part of the day the riders like riding their bike. Either day or night.

6. Also plotting if people like riding in too much rainfall and snowfall.

Groupby Function:

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

Any groupby operation involves one of the following operations on the original object. They are −

- Splitting the Object
- Applying a function
- Combining the results

5. Data modeling:

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. The indicator we decided to choose is the R2. This choice is because we wanted to be able to compare these algorithms between them and to choose which one is the most efficient. Let's apply regression techniques to our problem.

5.1 Multiple Linear Regression:

We ran a multiple linear regression; we assume that all the features have a linear relationship with the target. We also have to assume that these features have a Gaussian distribution and that features are not highly correlated between them, it is called multi-collinearity. The goal of this model is to get the R2 close to 1 or -1. We first fit the model on the training data and training target. Then we first predicted the training data then the test data in order to get indicators. For the training set, we got a R2 score = 0.55 , which is not that huge knowing that we have 6394 values predicted. For the testing set, we got the same R2 score = 0.55 which is not a very good score. The test and train scores are very close (almost the same). The RMSE value is around 420.

5.2 Decision Tree Regressor:

Then we applied a decision tree regressor on our data. We first scaled our X training and X testing sets. Then we applied a grid search on the model by tuning the feature 'max depth'. We fit our grid search on the X and y training sets. Then we kept the best model (with the best estimators) and got the score with the X and y testing sets. The best max_depth value is 12. The result of the testing sets is 0.84 or 84% and is much higher. The training dataset is 0.77 or 77%. The RMSE value obtained is around 300.

5.3 Random Forest Regressor:

Then we applied a random forest regressor on our data. We also take the scaled X training and X testing sets. We fit our model on the X and y training sets. After fitting our models on the training datasets we then use the model and get the score with the X and y testing sets. The R2 score on the testing dataset is 0.98 or 98%. The R2 score of the testing sets is 0.89 or 89% which is much higher and is a pretty good score as compared to other models. While the RMSE value is around 200. So, the R2 scores and RMSE values are much better than all the previous models.

5.4 XGBoost Regressor:

Finally, we applied an XGBoost regressor on our data. For this model, we also take the scaled X training and X testing sets to fit the model. Then we made a grid search on the model by tuning the features like ‘learning rate’, ‘no. of estimators’(which is the number of trees). We fit our grid search on the X and y training sets. Then we kept the best model (with the best estimators) and got the score with the X and y testing sets. The score is pretty similar to the random forest regressor. The best result of the training sets that we got is around 0.90 or 90%which is a pretty good score with a testing score of 0.885 or 88.5% which is pretty similar to the testing score. The RMSE value is around 200. As compared to the Random Forest Regressor the training dataset R2 score is lower than Random Forest Regressor.

5. Conclusion:

The Dataset I choose to Explore is the data on Seoul Bike Sharing Dataset. Which includes the data which influences people renting a bicycle may it be Temperature, Humidity and so on.

The Parameter that does not affect the number bikes ridden is the week day. Regardless of which week day it is, the number of rides taken remains more or less the same.

The metric we have chosen is the R2 score and RMSE value in order to compare all the models we did. All the scores we compare are best estimators grid search scores. The algorithm that we have to keep is the algorithm which presents the best test score. Hence, we would like to finalize Random Forest Regressor as our final algorithm for further predictions.

The final predictions were made on the Scaled X_test (X testing dataset). The results obtained (the rented bike count) was plotted against the ‘hour of the day’ for all the days to find out which hour of the day the demand for bikes is high. This study is necessary in order to make sure that there is availability of bikes to meet the demand in order to reduce the waiting time.

References:

- W3 schools
- Geek for geek
- Stack overflow