Bike Sharing Demand Prediction

Shubham Srivastava
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Predicting bike sharing demand can help bike sharing companies to allocate bikes better and ensure a more sufficient circulation of bikes for customers. In this project a real-time method for predicting bike renting and returning in different areas of a city during a future period based on historical data, weather data, and time data.

Our experiment can help understand what could be the reason for the classification of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the correct classification.

Keywords: machine learning,bike sharing,seoul,demand prediction

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.

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 - %
- Windspeed - 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

Over the past two decades sharing economy has not only revolutionized the organization of economic activity but also unleashed the consumption and production potentials of a variety of tourism and hospitality businesses. These businesses include but are not limited to sharing accommodation exemplified by Airbnb, sharing transportation pioneered by Uber and Lyft, as well as various online booking platforms such as Booking.com and OpenTable. There are even more localized sharing businesses, such as bike sharing provided by private enterprises or governments as an alternative to the so-called “last-mile” public transportation. Bike sharing has been popular in many countries, due to the fact that environmental proception organizations proposed environmental sustainability transportation methods such as electric vehicles and bicycles. Bike sharing provides benefits in various aspects and is achieving world-wide popularity. For instance, the number of renters in US was larger than 28 million in 2006. All these businesses share one commonality, for which consumer demand is upon request. Namely, suppliers need to immediately, if not instantaneously, deploy goods and services as soon as demand is generated.

While studies using machine learning techniques to predict consumer demand are proliferating in tourism and hospitality, there are very few devoted to predicting demand for bike sharing. A wealth of studies that indeed addressed bike sharing are primarily from the field of computer sciences . In fact, modeling tourism demand is disproportionately devoted to predicting tourist arrivals using either machine learning or a combination of machine learning and search query data . However, sharing economy has not only changed the way we model tourism demand but also extended what is modeled to reflect the nature of sharing economy in various areas. In this regard, we aim to use machine learning techniques to predict consumer demand for bike sharing. We also aim to advance previous research on bike sharing by incorporating a wide range of features other than weather to increase prediction accuracy.

3. Steps involved:

- **Exploratory Data Analysis**
  After loading the dataset we performed this method by comparing our target variable that is Rented Bike Count with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

- **Null values Treatment**
  Our dataset may contain a large number of null values which might tend to disturb our accuracy hence we drop them at the beginning of our project in order to get a better result.

- **Standardization of features**
  Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.
  The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

- **Fitting different models**
  For modelling we tried various classification algorithms like:

1. **Linear Regression**
2. **Lasso Regression**
3. **Ridge Regression**
4. **Polynomial Regression**
5. **Decision Tree Regression**
6. **Random Forest Regression**
7. **Gradient Boost**
8. **XGBoost Regressor**
9. **lightGBM Regressor**

- **Tuning the hyperparameters for better accuracy**
  Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models
  like Random Forest Regressor, Gradient Boosting Regressor, lightGBM Regressor and XGBoost Regressor.

4.1. Model performance:

Model can be evaluated by various metrics such as:

1. **Mean Absolute Error**-
   Mean Absolute Error of your model refers to the mean of the absolute values of each prediction error on all instances of the test data-set. Prediction error is the difference between the actual value and the predicted value for that instance.
   Statistically, Mean Absolute Error (MAE) refers to a the results of measuring the difference between two continuous variables.
2. **Mean Squared Error**-
   The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss.
   Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.
3. **R2 Score**-
   R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.
   R2 Score is a comparison of residual sum of squares with total sum of square. Total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line.

4.2. Hyper parameter tuning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds.

Grid Search CV-Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

5. Conclusion:

That's it! We reached the end of our exercise.
Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns and then model building.
We selected our best Machine Learning Model based on highest R2 score of 0.89 which we get from lightGBM Regressor.
We also drew some conclusions from the dataset we were provided:

1. In holiday or non-working days there is high demands for bike.
2. People preferred more rented bikes in the morning than the evening.
3. When the rainfall was less, people have booked more bikes.
4. The Temperature, Hour & Humidity are the most important features that positively drive the total rented bikes count.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya
4. medium