Bike Sharing Demand Prediction

By – Ranajay Biswas

Data Science Trainee,

Almabetter, Bangalore

Abstract:

Bike renting is now-a-days very common and desirable method of transportation for many. Whether you're visiting for a short trip or you are doing your daily chores, bike sharing has quickly become one of the most popular forms of travelling.

The benefits of bike sharing includes transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals. But the most special quality of public bicycles is the idea of sharing.

Bike share programs increase the visibility of cyclists, making riding safer for everyone. Studies also show that more people riding bikes in urban areas leads to improved bicycling and walking infrastructure.

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 most crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Data Description:

The dataset (Seoul Bike Data) 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 - %
- 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)

Approach:

In order to predict the bike demand, we can use the Rented Bike Count column. Since, the number of bikes rented can be any number or in other words, continuous type, we will be using regression models for prediction.

To understand important features, we need to perform exploratory data analysis.

Steps Involved:

- **Loading & Understanding the Data**

  The csv file containing the Seoul Bike Data was loaded in our work environment as a data frame using pandas. We checked top 5 rows and bottom 5 rows to get an initial idea about the data.

- **EDA & Data Pre-processing**

  This process is one of the most important of all. In this step, we did the following-

  1. We checked the different column names and then replaced their names with the appropriate and suitable for using names.
  2. Checked for null values and duplicate entries. But there were no null or duplicate entries in the data.
  3. Performed Descriptive Statistics and checked the data distributions.
  4. We performed univariate & bivariate analysis on the data and used various visualizations to understand the dependency of different features.
  5. Created a few new variables which we considered would be helpful in our project. We obtained these variables from the columns that we already had. Dropped the columns that we didn't.

- **Encoding**

  In this part, we find the categorical columns and perform some kind of encoding based on requirement & feature type. For columns like ‘Holiday’, ‘Functioning Day’ and ‘Weekend Days’, we used one hot encoding. For other categorical columns such as ‘Hour’, ‘Seasons’ and ‘Month’, we performed dummification.

- **Feature Selection**

  To pass the data to any of our ML algorithms, we select the most relevant features. After dropping the unimportant columns, we also looked for multicollinear features. ‘Dew point temperature’ and ‘Temperature’ columns were highly correlated with each other. So, we dropped the Dew point temperature column since it was having less correlation with the dependant column. We also dropped the ‘Functioning Day’ column and the observations that were having 0 values for the dependant column, since we found that 0 values only occur in case of a non-functioning day and keeping those would only hamper our model performance.

- **Data Scaling**

  We performed MinMaxScaler on the data to normalize it and make sure all the values are on same scale. It is very important to have the data in the same scale because it will help the gradient descent to converge much faster and distance based algorithms work better with scaled data.

- **Fitting Different Models**

  For solving the regression problem, we used models such as – **Linear Regression, Ridge, Lasso** and **Elastic Net Regression, KNN, Decision Tree, Random Forest, Gradient Boosting** and **XGboost 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 Decision Trees & Random Forest.
  We performed **GridSearch** with 3 fold Cross Validations for a set of hyperparameters and trained the models to obtain generalized models. 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.

- **Performance Metrics & Model Scores**

  1. **R2 Score:** It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset.
  2. **MSE Value:** Mean squared error (MSE) measures the amount of error in models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero.
  3. **RMSE Value:** Root mean square error or root mean square deviation is shows how far predictions fall from measured true values using Euclidean distance.
  4. **Adjusted R2 Score:** The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable.

Scores For Regression Models (For Base Models)**:**

We tried out 9 different models without any hyper-parameter tuning. The model scores are provided below. We saw that ensemble learning algorithms outperformed all the

basic algorithms and the linear models. Linear models were particularly not good since they heavily rely on the linear assumption.

The absolute best model was **XGBoost** model. The R-squared value for it was 0.9 which is the maximum that we find. The Adjusted R2 score was also very close 0.88 and the Mean Squared Error value was also very low 0.024

The predictions made by the model were really close to the original values. Since, we do not need explainability all that much for this kind of problem statement rather we want better and more accurate predictions, we can go with this model.

Model	R2 Score	Adjusted R2	MSE Value	RMSE Value
XGBoost Regressor	0.9	0.88	0.024	0.15
Gradient Boosting	0.88	0.88	0.026	0.16
Random Forest Regressor	0.86	0.86	0.032	0.18
Decision Tree Regression	0.79	0.79	0.048	0.22
Linear Regression	0.74	0.62	0.062	0.25
Ridge Regression	0.74	0.62	0.062	0.25
Elastic Net Regression	0.73	0.6	0.064	0.25
Knn Regressor	0.62	0.65	0.089	0.3
Lasso Regression	0.007	-0.26	0.237	0.48

After Cross Validation & Hyper-parameters Tuning :

After trying out many different hyper-parameters for multiple models, we find out that the best scores we get from the Random Forest model.

The R-Squared value for this model was 0.78 and MSE value was 0.052.

Though the MSE value looks good and we can be sure that the model is not an overfit model, we still didn't find the results as satisfactory as our XGboost model. Since, Xgboost performs cross-validation internally and it was giving a better performance, we decided to make **XGBoost** our final model for this problem statement.

Conclusion:

With that, we have come to an end of our project. To summarize the entire process -

- We performed data cleaning and EDA. Used visualizations to better understand the relations between the various features that were present in our data.
- We come to an understanding of how different features impact the number of bikes that are being rented, the effect of weather conditions, temperature and holidays bike rents. We also figured out which were the busiest hours of the day for business.
- In the pre-processing step, we removed multi-collinearity and unimportant features from the data. Encoding and feature scaling were done. So that the data can be passed through our ML models for training and validation.
- Finally models were trained to make predictions with the least margin of error possible.

The best model we found for this dataset was the XGboost model with an R2 score 0.9 and MSE value 0.024. This model is being able to predict the number of bike counts confidently. This model can now be deployed to make predictions and assist the business to have a better understanding.

References-

1. AlmaBetter
2. Kaggle
3. MachineLearningMastery
4. GeeksforGeeks
5. Analytics Vidhya