Bike Sharing Demand Prediction

Sanchit Misra, Mohit Jain,
Tushar Hande
Data science trainees, Almabetter.

Abstract:

Bike-sharing programs have received increasing attention in recent years due to their positive impact on the environment in mitigating air pollution and traffic congestion. In addition to environmental advantages, cycling aids in improving public health. Users can pick up and drop off bicycles from the station. To improve system efficiency, system operators should establish appropriate repositioning strategies based on accurate predictions of demand for bicycles. This study focuses on the short-term forecasting of the shared bikes demand in Seoul using a Machine learning approach. It practically contributes to increasing operational efficiency and reducing operating costs by improving demand predictability in a bike-sharing system.

Keywords: Exploratory data analysis, Machine learning, Prediction, Accuracy.

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 the bike count required at each hour for the stable supply of rental bikes.

Data description:

The dataset contains weather information, day information, date, and bike count.
Dataset has 8760 attributes and 14 features. Bike count is the target variable and the other 13 are independent variables.

Know more about the data:

- **Date:** Date is given in the format Year-Month-Day.
- **Rented bike count:** This is an integer type target variable that shows the count of the bikes rented at each hour.

- **Hour:** This is an integer type independent variable that shows the hour of the day.
- **Temperature:** This is a float-type independent variable that shows temperature in Celsius.
- **Humidity:** Humidity is an integer type variable in the % form.
- **Windspeed:** Windspeed is float type variable that shows speed m/s.
- **Visibility:** Integer type variable that shows the distance in meters.
- **Dew point temperature:** It shows the dew point temperature in Celsius.
- **Solar radiation:** Float type variable that represents solar radiation in MJ/m2 (megajoules per square meter).
- **Rainfall:** Float type independent variable that shows rainfall in mm.
- **Snowfall:** Float type independent variable that shows rainfall cm.
- **Seasons:** Four types of seasons are given Winter, Spring, Summer, and Autumn. This is an object type.
- **Holiday:** Two categories are given No holiday and Holiday. The type of this variable is an object type.
- **Functioning Day:** Two types of object type categories are given. Yes (Functioning day), No (Non-functioning day).

Introduction:

Bike rental businesses give customers—who are often, but not necessarily, tourists—bicycles for a short period. Bikes are generally rented for a few hours to recreationally explore the locality. But the customer base might also consist of college students on campus or others who rent for practical reasons.

City bike rentals are particularly popular among tourists who like to explore their destination by bicycle. Usually, the customers of these businesses are most interested in an efficient, comfortable, and safe way of commuting from one place to another. Depending on the destination, weather conditions or the business can be seasonal. However, due to very seasonal industry, it can be negatively affected by environmental forecasts and various other variables. Bike sharing is increasingly attracting more riders in cities around the world for its benefits regarding the urban environment and public health. One critical issue that Seoul is currently facing is the serious air pollution levels. The city's PM10 and PM2.5 levels maintained considerably high levels in the past few years.

Our goal is to build a productive model, which could help to predict the appropriate bike count.

Steps involved:

- **Exploratory data analysis (EDA):**
  Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, and patterns, or to check assumptions with the help of statistical summary and graphical representations.

- **Handling Missing values:** Many machine learning algorithms fail if the dataset contains missing values. We may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly. Missing data can lead to a lack of precision in the statistical analysis.
  Two primary ways to handle missing values.
  1. Deleting the missing values
  2. Imputing the missing values

  We don't have any missing values.

- **Feature distribution:** The feature distribution helps in understanding what kind of feature you are dealing with, and what values you can expect this feature to have. In Machine Learning, data satisfying Normal Distribution is beneficial for model building. It makes math easier.
  We applied square root transformation to the target variable.

- **Encoding categorical variables:** The machine learning algorithms can't understand categorical data. It requires all independent and dependent variables i.e input and output features to be numeric.
  Some of the ways to transform the categorical data.
  1) Label encoding: The label encoder assigned a number to each category.
  2) One Hot encoding: Dummy or one hot encoder convert each category to a binary column that takes the value 0 or 1.

  We applied both ways to our data.

- **Handling Outliers:** An outlier is an observation that lies at an abnormal distance from other values in a random sample from a population. Data outliers can spoil and mislead

the training process resulting in longer training times, less accurate models, and ultimately poorer results. We will detect them using the boxplot and IQR methods.
Ways to handle outliers:

1) Trimming or removing the outliers
2) Mean, median imputation
3) Quantile-based flowering and capping.
4) Imputation with upper bound (Q3+1.5*IQR) and lower bound (Q1-1.5*IQR)

We identified the outliers for numerical features and imputed them with the upper bound and lower bound.

- **Standardization of the features:** Standardizing the features around the center and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias.
  Standardization improves the numerical stability of the model. It may speed up the training process. We used standardscalar to standardize the independent features.

- **Feature selection:** Feature selection improves the machine learning process and increases the predictive power of machine learning algorithms.
  We used `feature_importances_` attribute to identify the important feature to each model.

- **Multicollinearity:** Multicollinearity occurs in the regression model when there is a high correlation between two or more independent variables. Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model. Multicollinearity can be identified using heatmap and Variance Inflation Factor (VIF). The following formula calculates VIF.

$$VIF = \frac{1}{1 - R_i^2}$$

We identified multicollinearity using both methods and we dropped some features. We considered threshold of 5 for VIF.

- **Using different models**:

  We used different regression models to predict the target.

  1. **Linear Regression**
  2. **Ridge Regression**
  3. **Lasso Regression**
  4. **Elastic Net Regression**
  5. **Decision tree Regression**
  6. **Random forest Regressor**
  7. **Extreme Gradient Boosting (XGBoost)**
  8. **Categorical Boosting (CatBoost)**

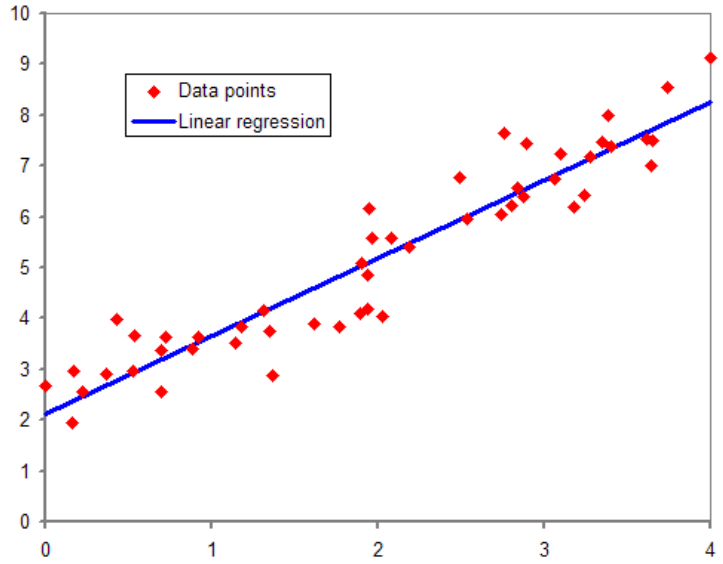
ML Algorithms:

1. Linear Regression:

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

The hypothesis function for liner regression is:

$$y = \theta_1 + \theta_2 x$$



The following cost function is used to calculate error or the cost in linear regression.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

2. Ridge Regression:

Ridge regression is almost identical to linear regression except we introduce a small amount of bias. In return, we get a significant drop in variance.

The bias added to the model is known as the Ridge Regression penalty.

The equation for ridge equation penalty is:

$$\text{the sum of the squared residuals} + \lambda \times \text{Slope}^2$$

3. Lasso Regression:

Lasso Regression is almost identical to Ridge Regression, the only difference being that we take the absolute value as opposed to squaring the weights when compounding the ridge regression penalty.

$$\text{the sum of the squared residuals} + \lambda \times |\text{the slope}|$$

λ is the tunning parameter.

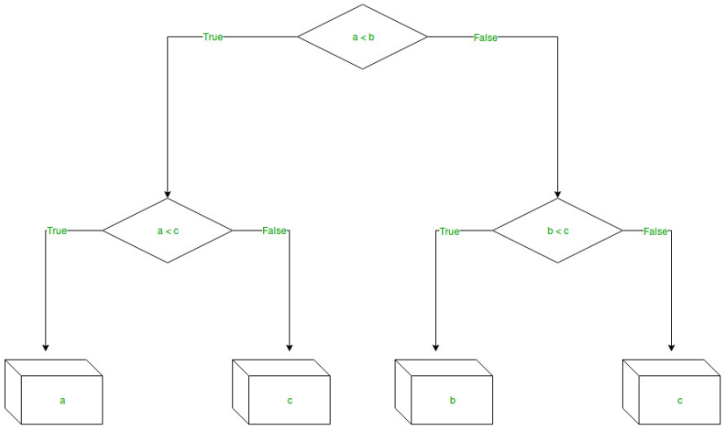
4. Elastic Net Regression:

Elastic-Net Regression is the combination of the Ridge and Lasso regression. Elastic-Net Regression groups and shrinks the parameters associated with correlated variables and leaves them in the equation or removes them all at once.

$$\text{the sum of the squared residuals} + \lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X| + \lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$$

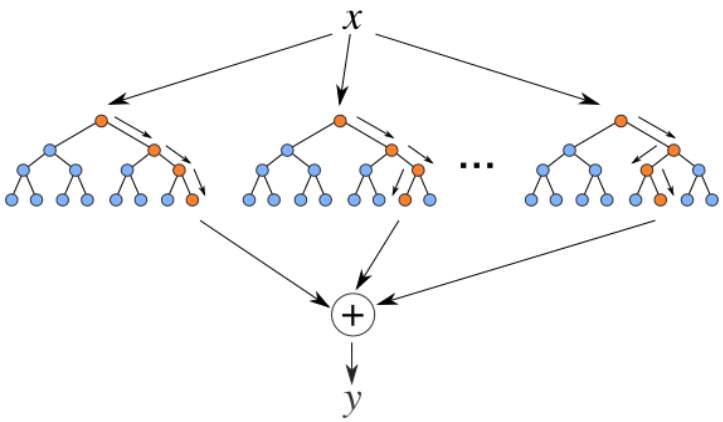
5. Decision Tree Regression:

A decision tree is a decision-making tool that uses a flowchart-like tree structure. Decision tree regression observes the features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.



6. Random Forest Regressor:

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

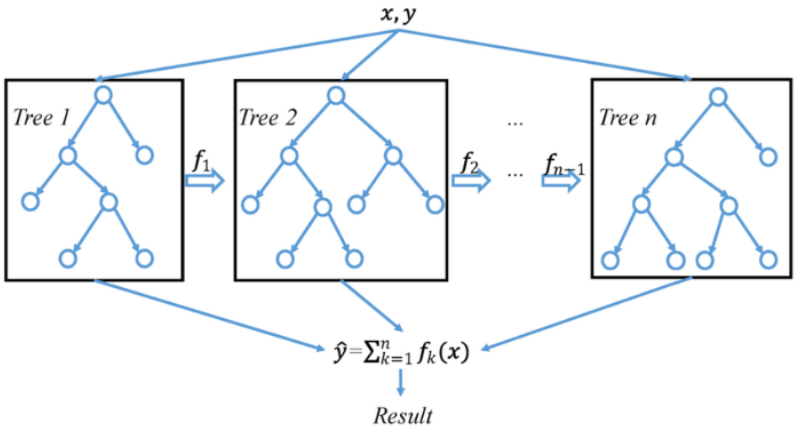


x is the input variable and y is the aggregate output variable.

7. Extreme Gradient Boosting (XGBoost):

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners.

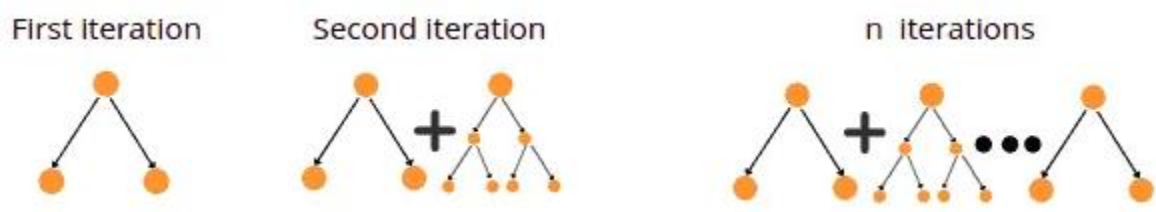
The objective function contains a loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e. how far the model results are from the real values.



8. Categorical Boosting (CatBoost):

CatBoost algorithm is designed to work with categorical features and it works similarly to Gradient and XGboost algorithms but it has some advanced features which make it more reliable, fast, and accurate. It is easy to use and works well with heterogeneous (more one than one type of data) data and even relatively small data. It essentially creates a strong learner from an ensemble of many weak learners.

CatBoost algorithm creates several binary decision trees each time trying to reduce the error: Once all iterations are complete, the algorithm stops creating new decision trees and finds the best-fitted model using the already familiar approach.



Hyperparameter tuning

Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

Strategies for hyperparametric tuning are:

1) GridSearchCV
2) RandomizedSearchCV

We used GridSearchCV.

1. **GridSearchCV:**

In GridSearchCV the machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for the best set of hyperparameters from a grid of hyperparameters values. It tries every possible combination of each set of hyper-parameters.

Model Performance:

The accuracy of the model can be evaluated using various metrics.

1. **Mean Squared Error (MSE):**

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.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

2. **Root Mean Squared Error (RMSE):**

Root Mean Squared Error (RMSE) is the square root of Mean Squared Error (MSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

3. **Mean Absolute Percentage Error (MAPE):**

Mean absolute percentage error is calculated as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{Actual\ Value_i - Predicted\ Value_i}{Actual\ Value_i}\right|$$

4. R^2 Score :

R^2 Score also called coefficient of determination. It is the amount of the variation in the dependent variable which is predictable from the independent variable(s). It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS: Sum of square of residuals, TSS: Total sum of squares.

5. Adjusted R^2:

R-square always increases as new variables are added to the model, no matter that they have a positive impact on the model or not.
Adjusted r-square is a modified form of r-square whose value increases if new predictors tend to improve model's performance and decreases if new predictors do not improve performance as expected.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square
p = Number of predictors
N = Total sample size.

Conclusion:

- The results clearly suggest that Cat Boost is the best model for predicting bike sharing demand, as the performance measure (MAPE, RMSE) is lower and (R2, adjusted_R2) value is greater for XG Boost and Cat Boost.
- We may say that the XG Boost model helps us to anticipate the number of rental bikes more accurately.
- Temperature is regarded as the most important variable for the linear regression model.
- The Cat Boosting algorithm's hour feature is considered to be its most significant factor.
- We discovered through EDA that people ride bikes more frequently in the summer season and also when the winds are strong.

- The months of May, June, and July can be considered to have a larger demand for rented bikes.
- There is a considerable demand for rented bikes during peak office hours. Therefore, base stations can be setup near office buildings.
- In order to reduce public waiting times, the number of bikes should be raised during the summer. As a result, they can easily rent bikes whenever they need to or at any time.

References :

- GreeksforGreek
- Medium
- Analytics Vidhya