Capstone Project

Seoul Bike Sharing Demand Prediction

Technical Documentation

By

Team

Nidhi Pandey

Date: June 10, 2023.

Table of Contents

Abstract:

This technical document represents a rule-based regression predictive model for bike-sharing demand prediction. In recent days, Pubic rental bike-sharing is becoming popular because of increased comfort and environmental sustainability. The dataset which was provided to us is 'Seoul Bike Data'. This dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, DewpointTemperature, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information. While doing this project various models have been created. These various models are being analyzed and we tried to study various models to intuitively get the best performing model for our project.

An analysis with variable importance was carried to analyze the most significant variables for all the models developed with the given data sets considered. We are getting the best results from LightGBM and CatBoost.

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 dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, 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 - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - water in mm
- Snowfall - Thickness cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

Steps involved:

1. Exploratory Data Analysis

After loading the dataset we compared our target variable that is the Rented Bike count Type with other independent variables. This process helped us figure out various aspects and relationships among the dependent and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the dependent variable.

2. Null values Treatment

Our dataset didn't have any null values to be treated.

3. Encoding of categorical columns

We used One Hot Encoding(converting to dummy variables) to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to the numerical format.

4. Feature Selection

In these steps, we used correlation, VIF analysis to check the results of each feature i.e which feature is more important compared to our model and which is of less importance.

5. 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 performingfitting 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.

6. Fitting different models

For modeling, we tried various classification algorithms like:

- Linear regression with regularization (lasso, ridge)
- KNN
- Decision Tree
- Random Forest regression
- Gradient boosted
- CatBoost regression
- LightGBM

7. Tuning the hyperparameters for better accuracy

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in the case of tree-based models like Random Forest Classifier and XGBoost classifier.

8. Features Explainability

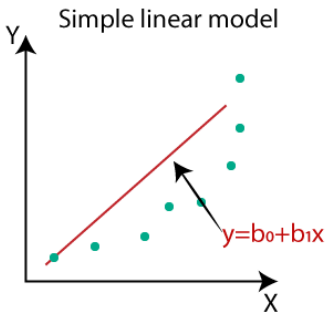
We have applied SHAP on the XGBoost and CatBoost model to determine the features that were most important while predicting an instance and also build a feature importance graph to find out which features were important and which were redundant in a model

Algorithms:

1. Linear Regression:

Linear regression is a linear approach for modeling the relationship between a scalar response and one ormore explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

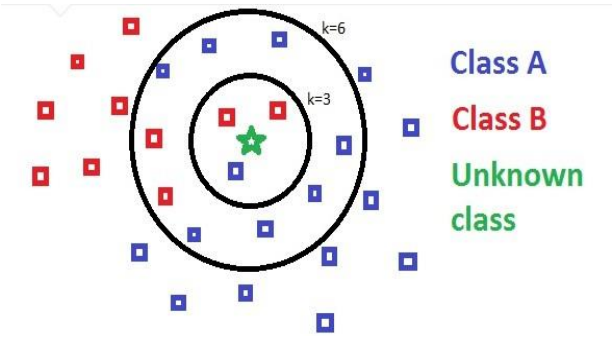
The optimization algorithm used is Gradient Descent. We also performed different types of regularization techniques to prevent overfitting in the model.



2. KNN:

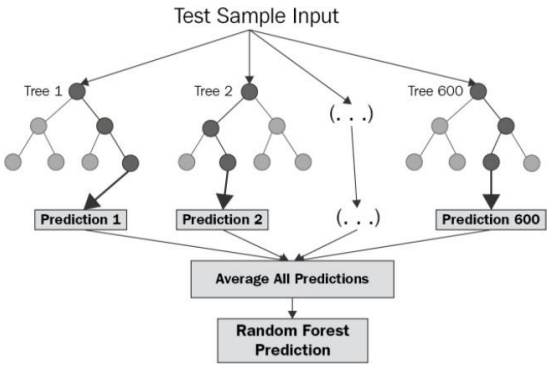
KNN regression is a lazy learning andnon-parametric method that uses ‘feature similarity’ to predict the values of new data points. It calculates the distance between test data and each row of training data with the help of any of the methods namely: Euclidean, Manhattan, or Hamming distance. Assign the points which are nearest to itand approximate the result with mean or mode valuesof its neighbors for regression or calcification respectively. The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation.

the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.



3. Random Forest Regression:

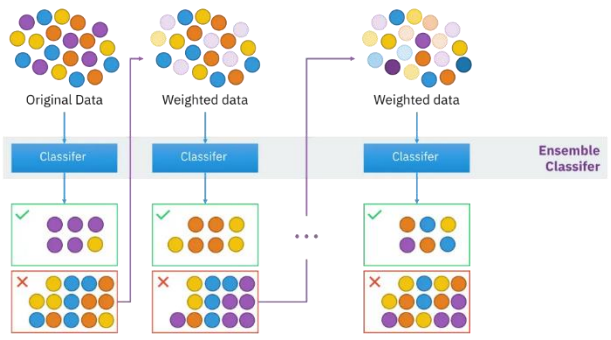
Random Forest is a bagging type of Decision Tree Algorithm that creates several decision trees from a randomly selected subset of the training set and n features, collects the values from these subsets, and then averages the final prediction out of all n number of decision trees



4. Gradient Boosting:

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

This approach supports both regression and classification predictive modeling problems.



7. CatBoost:

CatBoost is a recently open-source machine learning algorithm from Yandex. It can easily integrate with deep learning frameworks like Google's TensorFlow and Apple's Core ML. It can work with diverse data types to help solve a wide range of problems that businesses face today. To top it up, it provides best-in-class accuracy.

It is especially powerful in two ways:

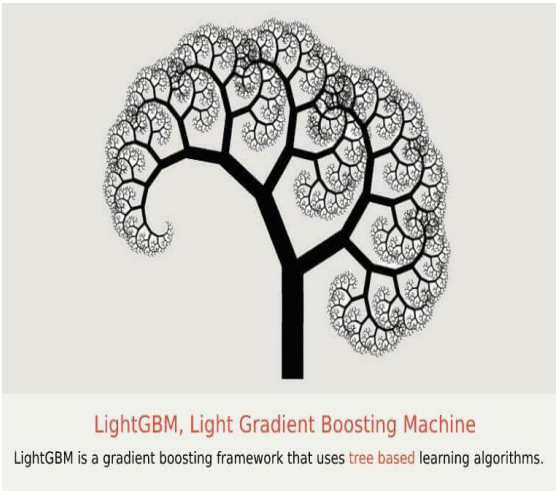
- It yields state-of-the-art results without extensive data training typically required by other machine learning methods, and
- Provides powerful out-of-the-box support for the more descriptive data formats that accompany many business problems.

"CatBoost" name comes from two words "Category" and "Boosting".

8. LightGBM:

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel, distributed, and GPU learning.
- Capable of handling large-scale data.



Model performance:

The model can be evaluated by various metrics such as:

1. Mean square error

The MSE of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

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

2. Root mean square error

RMSE is just the root of MSE. It is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient. One can compare the RMSE to observed variation in measurements of a typical point

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

3. R square:

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model. It has one limitation that its value increases as the number of Parameters increase even if that parameter does not improve model

$$SSE = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{m}(y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$\hat{y}$ is the predicted value, y is the actual value and $\bar{y}$ is the mean .

4. Adjusted R Square:

Adjusted R-squared is a modified version of R-squared that overcomes the problem of r2 and has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

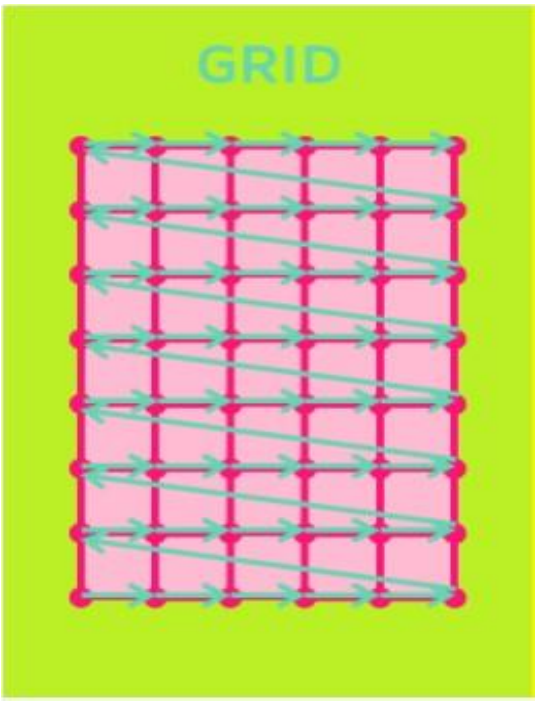
Hyperparameter tuning:

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 the 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, Randomized Search CV, and Bayesian Optimization for hyperparameter tuning. This also results in cross-validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

1. 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.



Conclusion:

Finally, we landed at the end of our exercise.

- Upon Exploratory Data Analysis, we found that the bike rentals follow an hourly trend where it hits the first peak in the morning and the highest peak later in the evening.
- We also found that these trends are prominent only during weekdays and working days, leading us to make a safe assumption that office-goers make a notable contribution to bike sharing demand.
- In addition, seasons were observed to have a notable effect on bike rentals with high traffic during summer and a significantly lower demand in winter.
- It is quite evident from the results that lightGBM and Catboost is the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse, rmse) shows lower and (r2, adjusted_r2) show a higher value for the lightGBM and Catboost models !
- So, we can use either lightGBM or catboost model for the above problem
- Also, it can be concluded that the lightGBM and CatBoost models are the best performing models for our project.