Seoul Bike Sharing Demand Prediction – Regression

Kajal Mahajan
Data Science Trainee,
AlmaBetter, Bangalore.

Abstract:

We will be leading in the data, do some exploratory analysis to get familiar with the features and target and their relationships, as well as look out for outlier or inaccurate data. After we explore and clean our dataset, we'll be trying out 6 regression models and see which one performs the best. We will also try a different method, where we use the split target columns ('casual' and 'registered' instead of 'cnt') and build two models that give us one final prediction. I named this the 'split method'. Performance will be measured using RMSE (root mean squared error) and using cross-validation, to get a distribution for each model. Once plotted, the boxplots will give us the winner - Random Forest Regressor, and will show that the split method does not perform noticeably better than the Random Forest Regressor on the single target.

Keywords : matplotlib, seaborn, barchart, supervised machine learning

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.

Introduction:

The bike sharing system is a way to rent a bike, where the process of membership, rental and return of the bike is automated throughout the city. These systems allow people to rent a bike in one place and return it to another if needed. The dataset is available on kaggle and github. Some pre-processing is required before using the data. Missing value completion / removal, normalization, etc. However, to get the right and useful functions, we apply correlations to all the functions and also use PCA (Principal Component Analysis) and shape libraries. A model is then developed using the Random Forest Regressor. This trains and tests the model to generate useful or needed insights.

Variables description:

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 - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day – No Fun(Non Functional Hours), Fun(Functional hours)

Steps Involved:

Data Collection:

Data collection is the process of collecting, measuring and analysing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyse them to make critical business decisions. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses. It refers to the process of finding and loading data into our system.
Pandas library is used to loading our data in our system in python. Using pandas we can manipulate data easily.

Data Cleaning:

Data cleaning refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. Such anomalies can disproportionately skew the data and hence adversely affect the results. Some steps that can be done to clean data are:

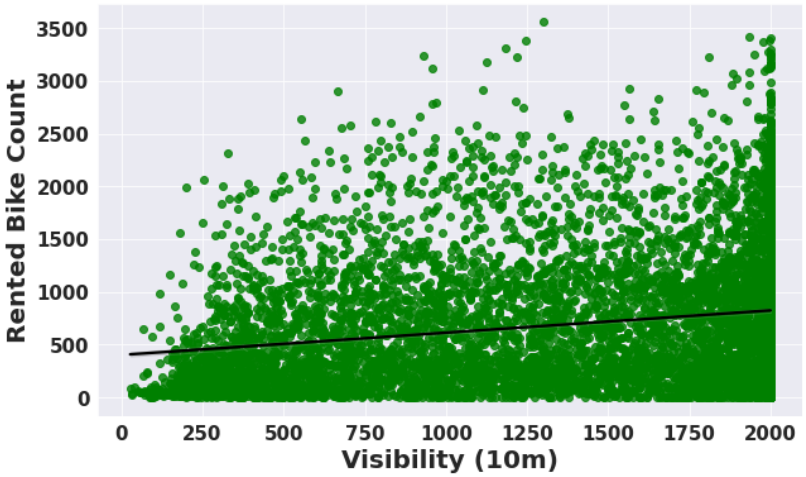
- Handling missing values: There are always some missing values in dataset. If we don't remove or handle those missing values then that can cause a trouble in our analysis. Removing or replacing those missing values with something meaningful is very important so that our data will have no missing values.
- Breaking date column
- Changing Data Type

Exploratory Data Analysis (EDA): Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows you to get a better feel of your data and find useful patterns in it.
It is crucial to understand it in depth before you perform data analysis and run your data through an algorithm. You need to know the patterns in your data and determine which variables are important and which do not play

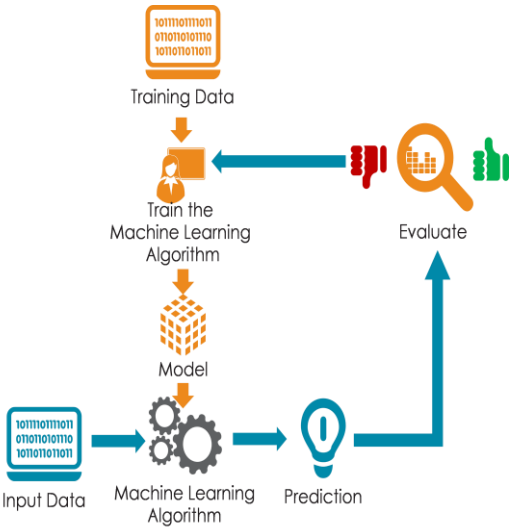
a significant role in the output. Further, some variables may have correlations with other variables. You also need to recognize errors in your data.
All of this can be done with Exploratory Data Analysis. It helps you gather insights and make better sense of the data, and removes irregularities and unnecessary values from data.

Regression Plots :
Regression lines are the best fit of a set of data. You can think of the lines as averages; a few data points will fit the line and others will miss. A residual plot has the Residual Values on the vertical axis; the horizontal axis displays the independent variable. Plots can aid in the validation of the assumptions of normality, linearity, and equality of variances. Plots are also useful for detecting outliers, unusual observations, and influential cases. Regression lines can be used as a way of visually depicting the relationship between the independent (x) and dependent (y) variables in the graph. A straight line depicts a linear trend in the data

For.Ex



Model Training: Model training is the process of fitting a data into machine learning model from which model learns the patterns in data to predict the dependent variable. Model do it so by assigning a weight to each variable. After our model is trained, we test our model on test data to check how our model is performing.



We used six different types of model to train and test performances.

- Linear Regression
- Lasso Regression (Lasso)
- Ridge Regression (Ridge)
- Elastic Net Regression
- Decision Tree
- Random Forest
- Gradient Boosting

Linear Regression :

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line. Linear regression uses a linear approach to model the relationship between independent and dependent variables. In simple words its a best fit line drawn over the values of independent variables and dependent variable. In case of single variable, the formula is same as straight line equation having an intercept and slope.

$$y\_pred=\beta0+\beta1x$$

where

$$\beta0 \text{ and } \beta1$$

are intercept and slope respectively.

In case of multiple features the formula translates into:

$$y\_pred=\beta0+\beta1x1+\beta2x2+\beta3x3+.....$$

where x_1,x_2,x_3 are the features values and

$$\beta0,\beta1,\beta2.....$$

are weights assigned to each of the features. These become the parameters which the algorithm tries to learn using Gradient descent. Gradient descent is the process by which the algorithm tries to update the parameters using a loss function . Loss function is nothing but the difference between the actual values and predicted values(aka error or residuals). There are different types of loss function but this is the simplest one. Loss function summed over all observation gives the cost functions. The role of gradient descent is to update the parameters till the cost function is minimized i.e., a global minima is reached. It uses a hyperparameter 'alpha' that gives a weightage to the cost function and decides on how big the steps to take. Alpha is called as the learning rate. It is always necessary to keep an optimal value of alpha as high and low values of alpha might

make the gradient descent overshoot or get stuck at a local minima. There are also some basic assumptions that must be fulfilled before implementing this algorithm

Lasso Regression:

This is a regularization technique used in feature selection using a Shrinkage method also referred to as the penalized regression method. Lasso is short for Least Absolute Shrinkage and Selection Operator, which is used both for regularization and model selection. If a model uses the L1 regularization technique, then it is called lasso regression.

In this shrinkage technique, the coefficients determined in the linear model from equation are shrunk towards the central point as the mean by introducing a penalization factor called the alpha α (or sometimes lamda) values.

$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m}|\hat{\beta}_j|.$$

Alpha (α) is the penalty term that denotes the amount of shrinkage (or constraint) that will be implemented in the equation. With alpha set to zero, you will find that this is the equivalent of the linear regression model and a larger value penalizes the optimization function. Therefore, lasso regression shrinks the coefficients and helps to reduce the model complexity and multi-collinearity.

Alpha (α) can be any real-valued number between zero and infinity; the larger the value, the more aggressive the penalization is.

Ridge Regression :

Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square.

$$L_{hridge}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m} w_j\hat{\beta}_j^2.$$

Elastic Net Regression :

Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training. Using the terminology from “The Elements of Statistical Learning,” a hyperparameter “alpha” is provided to assign how much weight is given to each of the L1 and L2 penalties. Summary. The elastic net method performs variable selection and regularization simultaneously. The elastic net technique is most appropriate where the dimensional data is greater than the number of samples used. Groupings and variables selection are the key roles of the elastic net technique

What is a Decision Tree?

The basis for the Random Forest is formed by many individual decision trees, the so-called Decision Trees. A tree consists of different decision levels and branches, which are used to classify data.

The Decision Tree algorithm tries to divide the training data into different classes so that the objects within a class are as similar as possible and the objects of different classes are as different as possible. This results in multiple decision levels and response paths,

Random Forest :

Random Forest is a supervised machine learning algorithm that is composed of individual decision trees. This type of model is called an ensemble model because an "ensemble" of independent models is used to compute a result.

Gradient Boosting :

The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term 'gradient' is related here.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc.

Intuitively, gradient boosting is a stage-wise additive model that generates learners during the learning process (i.e., trees are added one at a time, and existing trees in the model are not changed). The contribution of the weak learner to the ensemble is based on the gradient descent optimisation process. The calculated contribution of each tree is based on minimising the overall error of the strong learner.

Gradient boosting does not modify the sample distribution as weak learners train on the remaining residual errors of a strong learner (i.e, pseudo-residuals). By training on the residuals of the model, this is an alternative means to give more importance to misclassified observations. Intuitively, new weak learners are being added to concentrate on the areas where the existing learners are performing poorly. The contribution of each weak learner to the final prediction is based on a gradient optimisation process to minimise the overall error of the strong learner.

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 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.

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:

A bike-sharing system is a means of renting bikes that automates the process of membership, rental, and bike return through a network of kiosks around the city. These systems allow people to rent bikes at one location and return them to another if needed. There are currently over 500 bike-sharing programs worldwide. The data generated by these systems are attractive to researchers because they explicitly record travel time, departure location, arrival location, and elapsed time. The bike-sharing system therefore acts as a sensor network that can be used to study urban mobility. The competition asks participants to combine historical usage patterns with weather data to determine demand for bike rentals.

Reference:

https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression

https://www.mygreatlearning.com/blog/gradient-boosting/

https://towardsdatascience.com/introduction-to-random-forest-algorithm-fed4b8c8e848