Seoul Bike Sharing Demand Prediction

Ankit Sharma
Pankaj Kumar

Data science trainee

Almabetter , Banglore

Introduction

Rental bikes is a new concept used in many cities now, it helps in reducing traffic as well as help in reducing cost of travel of individual. With this project we will try to identify demand of rental bike requiring for every hour so that there stays a near equilibrium in demand and supply of rental bikes. We have used data mining technique to ease our prediction for rental bike demand. We have used various variables like:-
(i)weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall),
(ii)The number of bikes rented per hour
(iii) Information and data of dates of bookings

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 understanding

- Date : year-month-day
- 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)

Data visualization

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

- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius

- Humidity - %
- Windspeed - m/s

Conclusion of EDA

✓ Bike is rented out for 18 hours mostly

✓ people are rented bike for hours ranging 10 to 24 the most might be because people are renting bikes for office work

✓ As the temperature and humidity increases users are increasing

- ✓ people are preferring to ride bike when its little windy but not too much windy
- ✓ Visibility increases user are increasing
- ✓ rain and snow increasing user are decreasing
- ✓ functioning days are very much as compared to nonfunctioning days no people can rent a bike on nonfunctioning days
- ✓ weekdays and weekends in both people are using bike almost equally

DATA Preprocessing

Null values Checking-Our dataset does not contain any null values this makes work easy.

Duplicate values- Our dataset does not contain any duplicate values**.**

One Hot Encoding- For categorical variables where no such ordinal relationship exists, the integer encoding is not enough.

In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

We used one hot encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features cannot be understood by the machine and needs to be converted to numerical format. One Hot Encoding is applied on Holiday, Functioning day, Season.

Splitting Data- The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm.

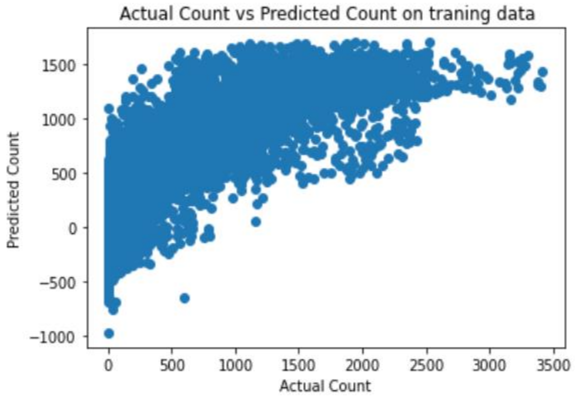
- **Train Dataset**: Used to fit the machine learning model.
- **Test Dataset**: Used to evaluate the fit machine learning model.

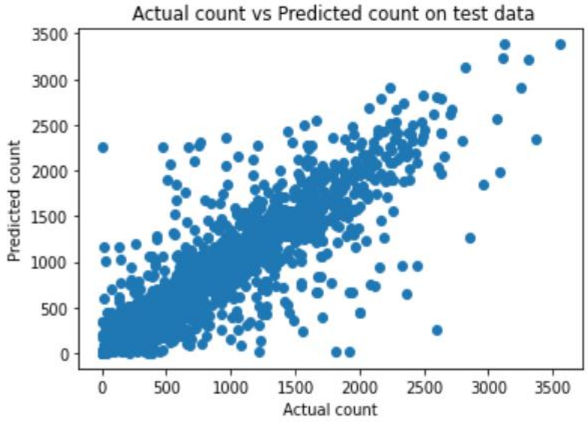
Model training and prediction

- ➢ Linear Regression
- ➢ Lasso Regression
- ➢ Random Forest Regressor
- ➢ Decision Tree Regressor

Linear regression:

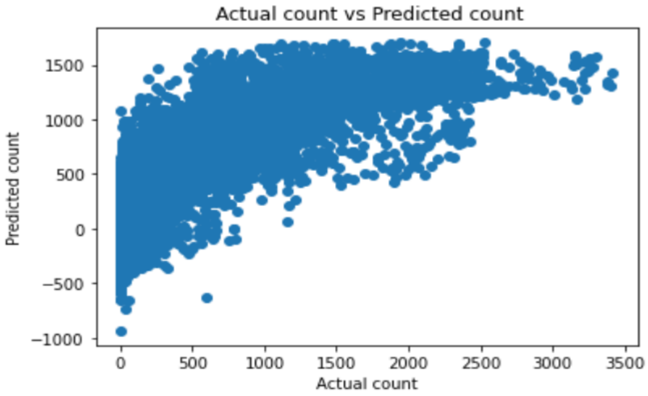
- ➢ Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.
- ➢ Performed Linear Regression on both the test data and training data
- ➢ R squared Error on training data id **0.55** and on test data is **0.53**

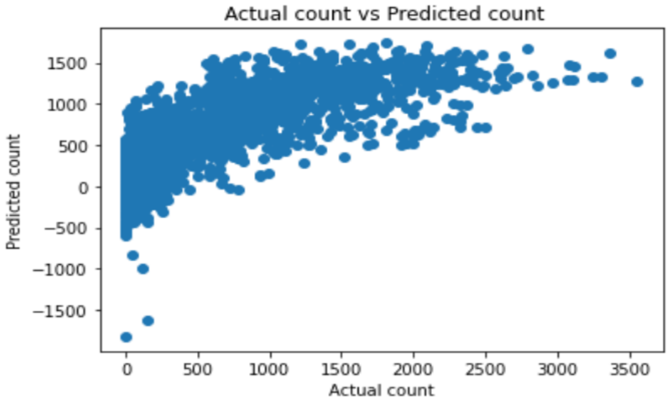




Lasso Regression: -

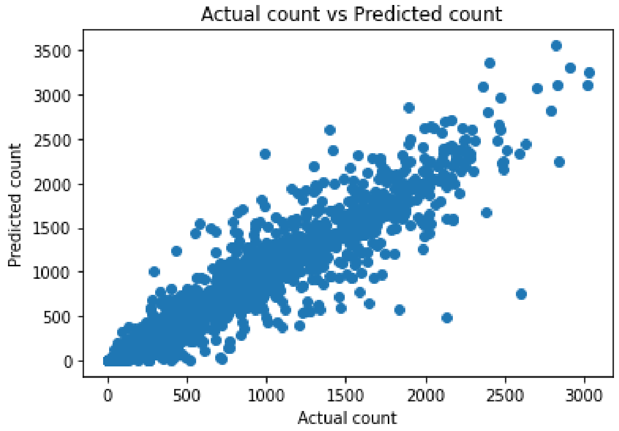
- **LASSO** stands for **Least Absolute Shrinkage and Selection Operator**. Lasso regression performs L1 regularization that is it adds the penalty equivalent to the absolute value of the magnitude of the coefficients.
- Applied lasso regression on training as well as test data
- R squared Error on training data id **0.55** and on test data is 0.53 same as linear regression





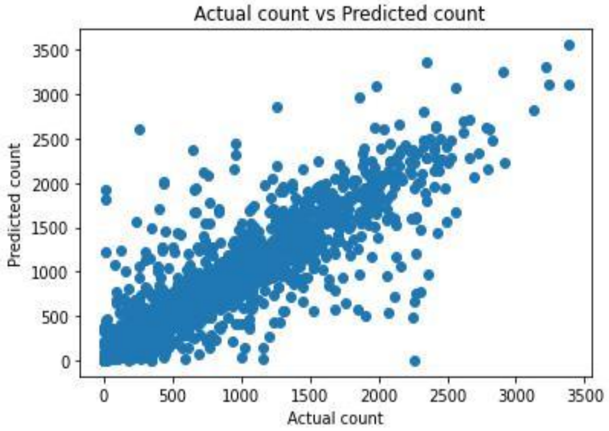
Random Forest Regressor: -

- Random forest is **a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems**. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.
- **R squared Error of random forest regressor is- 0.89**



Decision Trees (DTs)

- They are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.
- When we applied Decision tree
- R squared Error on test data is- 0.79



Basic parameters to evaluate the model performance

- To build and deploy a generalized model we require to Evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result.
- **Mean Absolute Error (MAE)-** MAE is a very simple metric which calculates the absolute difference between actual and predicted values.
- **Mean Squared Error (MSE)-** Mean squared error states that finding the squared difference between actual and predicted value.
- **Root Mean Squared Error (RMSE)-** it is a simple square root of mean squared error
- **Root Mean Squared Log Error (RMSLE**)- Taking the log of the RMSE metric slows down the scale of error. To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.
- **R Squared (R2)-** R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform. I have find out R squared value of all the models

Conclusion of Model

- R squared Error of linear regression is-0.77
- R squared Error of lasso regression is-0.37
- **R squared Error of random forest regressor is- 0.99**
- R squared Error of Decision tree is-0.71
- R squared Error of Ridge regression is 0.77

- R squared Error of Elastic net regression is 0.62

- **Random forest is performing very good as compared to another model**

- R2 score is between zero and one like 0.8 which means your model is capable to explain 80 per cent of the variance of data.

R squared Error of random forest regressor is- 0.99 which explains 99 percent of the variance of data so model is performing very good.