Seoul Bike Sharing Demand Prediction

Owes Khan
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Data analytics in Bike Sharing Demand Prediction is a growing strategy implemented to support business decisions designed to generate revenue or save costs. This study contains the real-world data record of bike sharing demand prediction of Seoul containing details Temperature, Humidity, Windspeed, Visibility, etc. Main aim of the project is to understand and visualize dataset from customer point of view and make predictions of rented bike counts and also provide suggestions to focus on the important factors which can improve the business model.

Keywords: Python, Data cleaning, Pandas, Matplotlib, Data visualization, Sklearn etc.

Problem Statement

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 discussion of data analytics would not be complete without an understanding of what is the data that is used in analytics. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore data, and possibly formulate hypotheses that might cause new data collection and experiments. EDA focuses more narrowly on checking assumptions required for model fitting and hypothesis testing. It also checks while handling missing values and making transformations of variables as needed filling the counts with.

EDA builds a robust understanding of the data, issues associated with either the info or process. It's a scientific approach to get the story of the data.

An interpreted, object- oriented programming language with dynamic semantics. Its high level, built-in data structures, combined with dynamic binding, make it very attractive for rapid application development, also as to be used as a scripting or glue language to attach existing components together. Python and EDA are often used together to spot missing values in the data set, which is vital so you'll decide the way to handle missing values for EDA.

Types of exploratory data analysis:

1. Univariate Non-graphical
2. Multivariate Non-graphical
3. Univariate graphical
4. Multivariate graphical

Dataset used for EDA

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 he 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 - NoFunc(Non Functional Hours), Fun(Functional hours)

Methodology:

Business understanding

Analytics in the vehicle renting world today is important, and nowadays this business cannot be run with some sensible and smart use of data.

Here I demonstrate how to use data to analyse business important concepts in the fields of revenue management and marketing.

The analysis tries to answer below questions

1. Which are the most important factors deriving the predictions?
2. Can a model explain the predictions easily?

Installations

Below is a list of the modules used in the analysis is shown

pandas, numpy, matplotlib, seaborn, sklearn, etc.

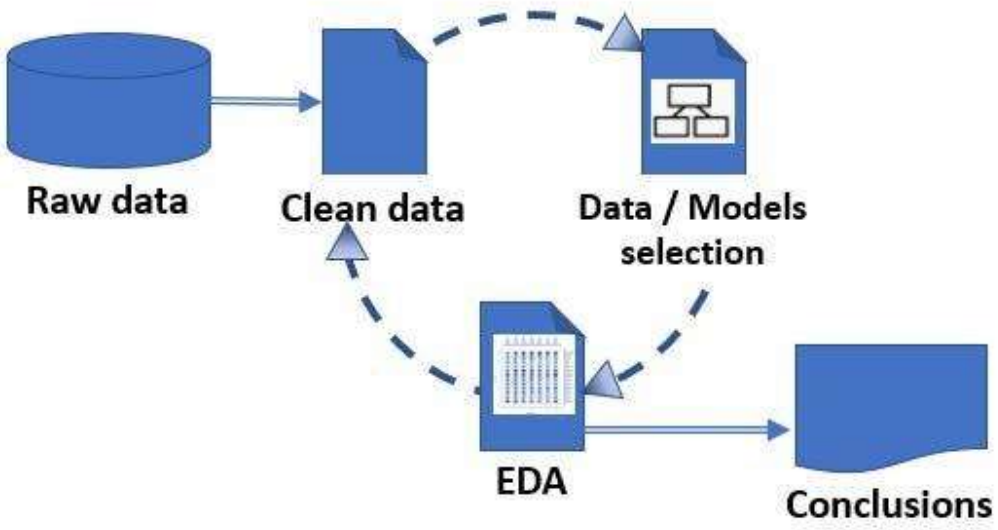
Handling mismatched data types and outliers.

(1) Converting columns to appropriate data types

(2) Checking for outliers

- The mean and median difference is quite large for most of the features.

Exploratory Data Analysis



Data visualization

Data Visualization is the process of analysing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data. There are various types of visualizations –

- **Univariate analysis:** This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
- **Bi-Variate analysis:** This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
- **Multi-Variate analysis:** When the data involves three or more variables, it is categorized under multivariate.

Mainly performed using Matplotlib and Seaborn library and the following graph and plots had been used:

- Bar Plot..
- Box Plot
- Count plot
- Heat map

Bar Plot:

A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

Subplots:

Subplots mean **groups of axes that can exist in a single matplotlib figure**. subplots() function in the matplotlib library, helps in creating multiple layouts of subplots. It provides control over all the individual plots that are created.

Countplot:

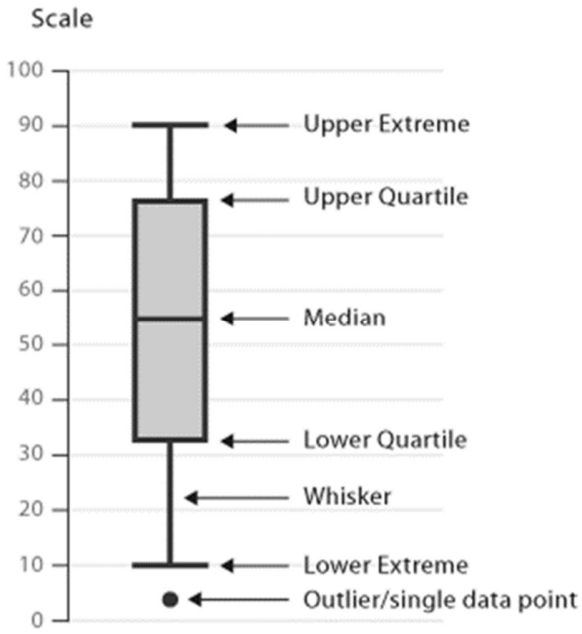
A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot(), so you can compare counts across nested variables.

Heat map:

A heatmap contains values representing various shades of the same colour for each value to be plotted. Usually, the darker shades of the chart represent higher values than the lighter shade. For a very different value a completely different colour can also be used.

Box and Whisker Plot (or Box Plot):

It is a convenient way of visually displaying the data distribution through their quartiles. The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally.



Metrics

To evaluate the performance or quality of the model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics.

- **R Squared Score**: R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
- **Adjusted R Squared Score :** Adjusted R^2 is a modified version of R^2 that has been adjusted for the number of predictors in the model. The adjusted R^2 increases when the new term improves the model more than would be expected by chance.

Models

Linear Models :

- Linear Regression
- Polynomial Regression

Non Linear Models :

- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor

Summary

- The job is to predict the number of rented bike counts each hour.
- The dataset contains 14 features in which "Rented Bike Count" is a response variable and others are predictor variables.
- The dataset contains no missing values.
- The dataset contains a "Date" feature which has an object dtype, we changed the dtype of this feature to correct one.
- The "Date" feature is splitted to "Year", "Month", and "Day" features for better understanding.
- The features "Dew point temperature" and "Temperature" are highly correlated. So we used the combination of both to remove correlated features from the dataset.
- Each numerical feature is less correlated with the dependent variable and follows a non-linear relationship.
- The response variable is positively skewed, so to remove the skewness different transformations are used. The most effective is square root transformation.
- The numerical features are scaled and categorical features are one hot encoded before passing to the model.
- The key metrics to be noticed are R2 and adjusted R2.
- Both Linear and Non Linear Models are used in the task.
- The Linear Regression, gives R2 score as 0.65 and adjusted R2 score as 0.65.
- The Decision Tree Regressor gives R2 score as 0.95 for train and .79 R2 score for test data prone to over fitting.
- The Random Forest, GradientBoosting Regressors give around R2 as score 0.98,0.975 for train data and a R2 score as 0.89,0.92 for test data respectively.
- The highest scores can be seen for ensemble models.

Conclusion

- Random forest regressor performs really good when compared to linear regression with high model performance. But it's performance is low when compared to gradient boosting regressor. However, time taken for hyperparameter tuning and training the model is much low for random forest regressor than gradient boosting regressor. Thus, there's a tradeoff of accuracy and time in between random forest and gradient boosting regressor. It's up to us and business domain to which algorithm to use.
- Out of all above models Gradient Boosting Regressor gives the highest R2 score of 97.5% for Train Set and 92.0% for Test set and no overfitting is seen.

Take Home Messages

1. Always start by exploring a dataset with an open mind for discovery.
2. EDA allows us to better apprehend the features and possible issues of a dataset.
3. EDA is a key step in generating research hypotheses.

Challenges

- The dataset contains skewed predictor variables.
- The response variable was also skewed.

Reference

- [1] Kaggle.com
- [2]Analytics Vidhya
- [3] towardsdatascience
- [4]Greeksforgreeks