Bike Sharing Demand Prediction

Shubham Chougule : sc9395288@gmail.com

Data Science Trainee, Almabetter

Abstract

Currently rental bikes are increasing throughout the world and especially in fast growing cities, it helps individuals to Move easily without waiting for any public transport and it helps reduce dependency on public transport and reduces pollution.

In the world of rising new technology and innovation, the Transport industry is advancing with the role of Data Science and Analytics. Data analysis can help them to understand their business in a quite different manner and helps to improve the quality of the service by identifying the weak areas of the business. This study demonstrates the prediction of the rental bikes available during the given time period.

Our experiment can help understand and predict to get insights from this data based on which business decisions will be taken.

Problem Statement

The objective of the project is to perform exploratory data analysis, data pre-processing, data cleaning & imputation, and in the end, apply different Data Visualization techniques to get meaningful insights and apply different models to predict from the given data. 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

Transport industries are having an important reflection on the economy over the past few decades. Currently rental bikes are increasing throughout the world andespecially in fast growing cities. It helps individuals to Move easily without waiting for any public transport and it helps reduce dependency on public transport and reduces pollution.

Prediction and model building

Here the main goal is to predict the total number of rental bikes that should be made available at a certain period of time even during the peak periods and the low periods .

Understanding the Dataset can refer to a number of things including but not limited to extracting important "variables", identifying "outliers", "missing values", or "human error". Ultimately, maximizing our insights of a dataset and minimizing potential "errors" that may occur later in the process.

Dataset

The dataset contains weather information such as (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

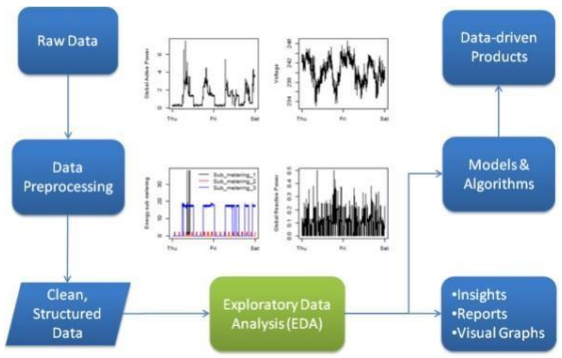
Data Exploration

This dataset has around 14 columns and it is a mix of categorical and numeric values.

The features description includes

Date	Specific date
Rented bike count	Total rented bikes
Hour	No of hours
temperature	Degree of hotness
humidity	Degree of water vapors level
visibility	Degree of visibleness
snowfall	Degree of snowfall
windspeed	Speed of wind
rainfall	Amount of rainfall
Dew point temperature	Hotness around dew
Solar radiation	Solar radiation levels
seasons	Summer, winter, spring autumn
holiday	holiday
Functioning day	Functioning day

Architecture



Data Pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for our analysis purpose, where we have to do a lot of Data Cleaning and handle the missing values by using appropriate imputation techniques and based on that variable nature i.e., either of categorical & numerical variable.

Substitution/imputation of missing values using either mean, median, mode or zero according to the nature of those variables. Here, in this project, we have imputed with zero. Moreover, we also removed the columns which do not participate in our analysis.

Data Preparation

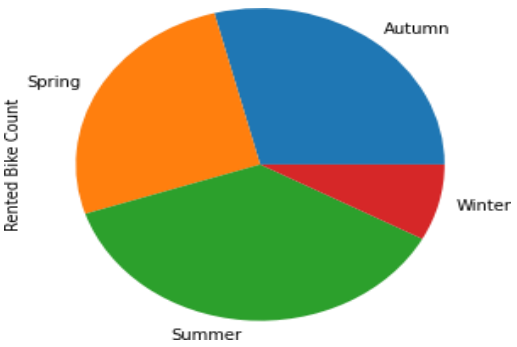
Data Preparation includes loading the dataset into a data frame, exploring the number of rows & columns, ranges of values, data types, descriptive summary of numerical features, correlation and distribution of features etc.

Data Cleaning

It includes identifying the missing values in the dataset and handling the same.

Analysis of crucial understandings explored include

1. **Total bike counts in respective season**

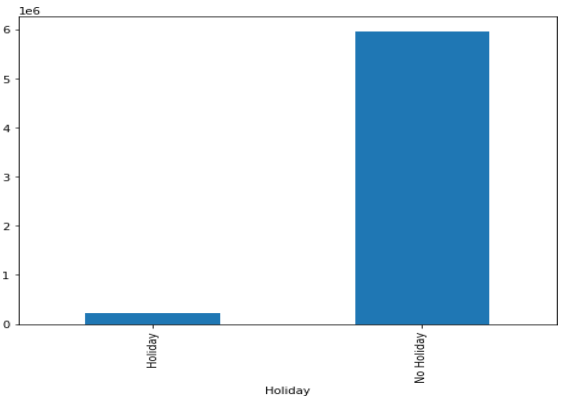


It can be observed that during summer the bikes rented were higher and it was almost around 2283234, during autumn it was around 1790002 and during spring it was around 1611909 and during winter the bikes rented were very low and it was around 487169. So, we can conclude that the bikes a rented more during summer.

2. Total bikes rented in each year

It can be clearly observed that the bike renting trend started to increase from the end of 2017 and it can be observed that in 2018 it was so much higher.

3. Total bike counts in respective season

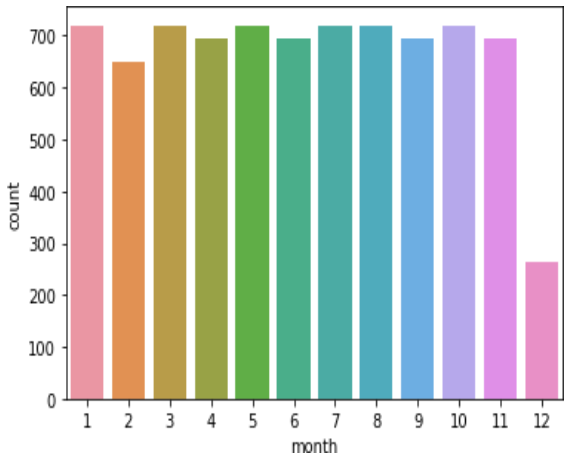


It can be clearly seen that during work days or non-holiday's the bike rented is higher compared to the days which were holidays.

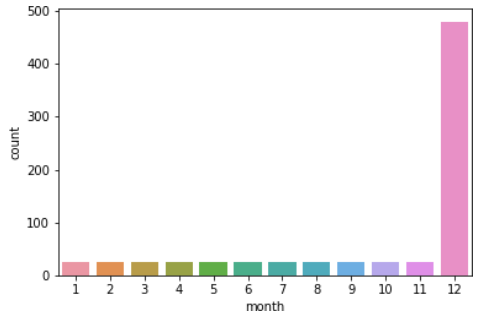
4. Total bike counts in respective season

This shows the density of bikes rented with respect to each year and from the figure we can clearly see that during holidays in 2017 and 2018 the density of bike rented is poor, but in 2018 during non-holidays the rented bikes are higher compared to rented bikes in 2018.

5. Total bike counts in respective years

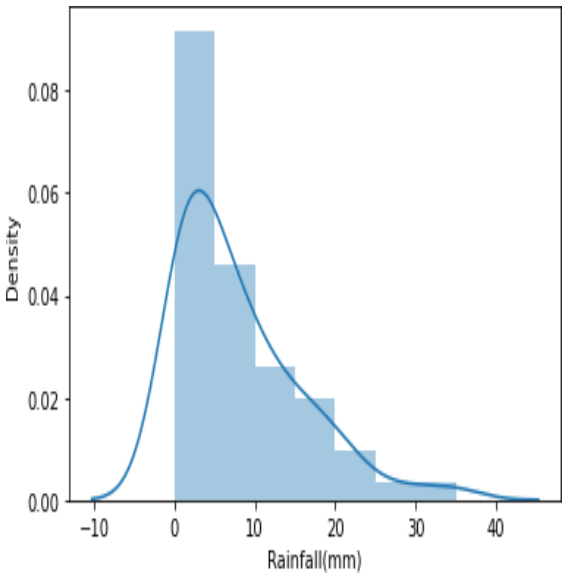


In 2018 we can observe that during most of the months the rented bike count is higher except December and we can clearly see that during summer months the rented bike count is higher.



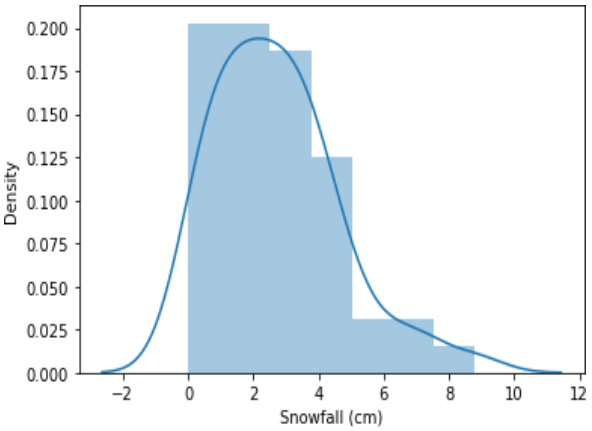
But in 2017 we can observe that except December, all the months have recorded very poor rental count because the trend of increase in rental bikes started from December 2017.

6. Distribution of bike rentals according to rain intensity



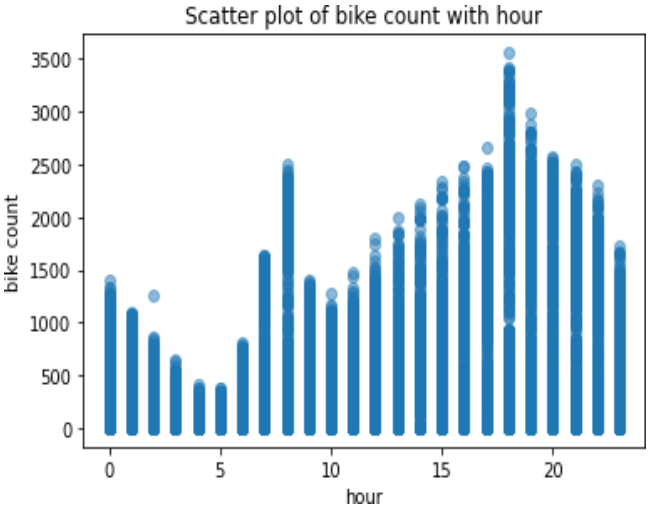
From this distribution we can clearly see that when the rainfall increases from around 5mm there is very less number of bike rentals so we can concur that during low or no rainfall there will be higher rentals compared to rentals during rain.

7. Distribution of bike rentals during snowfall



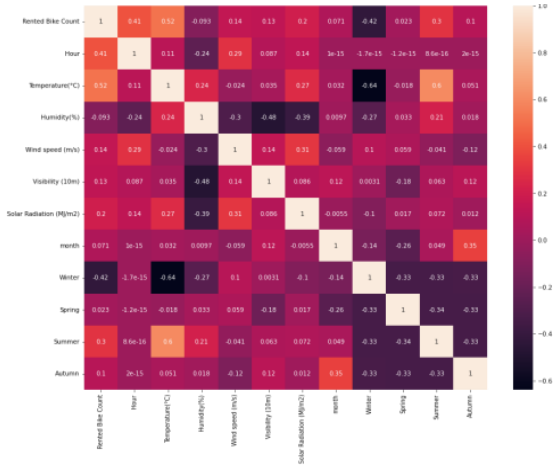
Here we can clearly observe that when the snowfall increases the demand for rental bikes decreases and where there is less snowfall the demand gradually increases.

8. Bike count with respect to hour



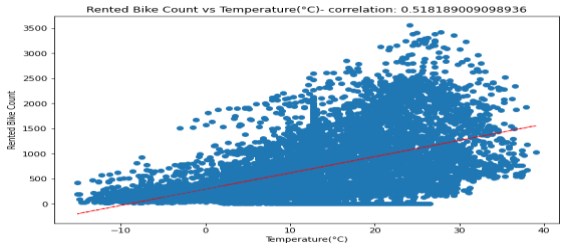
From the above scatter plot, we can observe that from morning 8am to 11am and from 4pm to 8 pm there was peak demand for the bikes; this might be attributed to the beginning and ending of work hours.

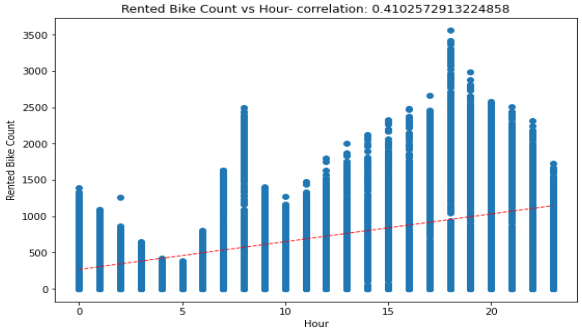
9. Correlation analysis



The correlation analysis has been done, variation inflation factor has been calculated and multicollinearity has been detected, the highly correlated features are removed, here summer feature has been removed due to higher multicollinearity compared to all other features.

10. Correlation plots between dependent and independent variables



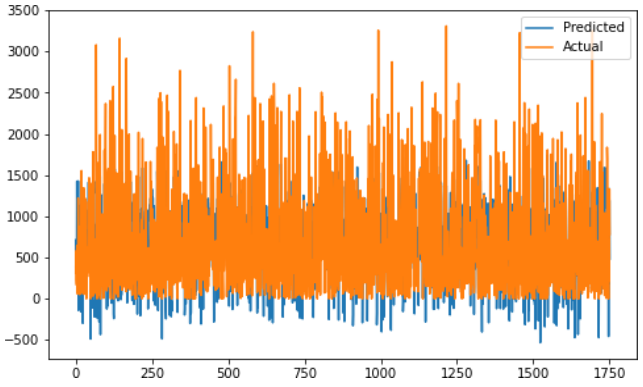


★ Model building and predictions

Linear regression model

After selecting the rented bikes feature as the independent feature and the rest other columns as dependent feature, we split the data into train set and test set , later we transformed the data using minmax scalar and we fitted the LINEAR REGRESSION MODEL to the dataset.

The following results were obtained,
MSE : 202937.66869256634
RMSE : 450.4860360683407
R2 : 0.5151094008043042
Adjusted R2 : 0.5117634047201476



Later to increase our model predicting capacity we introduced decision tree as it predicts much better than the linear regression model and we founded the following inferences in decision tree

Decision tree model

After applying the decision tree model we obtained the following results,
The best Decision Tree R2 score is :
0.7753673960792731 with max depth 10
The best R2 test score is :
0.7948642204947705 with max depth = 10

As we can clearly see that the decision tree test score and train score are higher compared to the linear regression model and we can clearly say that decision tree outperforms linear regression model

Random forest regression model

The best Random Forest R2 train score is : 0.8278994903973604 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

The best Random Forest R2 test score is : 0.8516768166059918 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

As we can clearly observe from the test score and the train score the random forest regressor model outperforms the linear regression model and the decision tree model with better results , so we can conclude that he random forest regressor has the best model of prediction compared to other two models.

Conclusion

The exploratory data analysis and modelling for Bike sharing demand prediction dataset has been successfully done and the following inferences have been made from the obtained visualizations and also from the dataset,

For the linear regression model the obtained results are

- ★ MSE : 202937.66869256634
- ★ RMSE : 450.4860360683407
- ★ R2 : 0.5151094008043042
- ★ Adjusted R2 : 0.5117634047201476

The exploratory data analysis and modelling for Bike sharing demand prediction dataset has been successfully done and the following inferences have been made from the obtained visualizations and also from the dataset,

For the decision tree model the obtained results are

The best Decision Tree R2 score is 0.7753673960792731 with max depth 10
The best R2 test score is : 0.7948642204947705 with max depth = 10

For the random forest model the obtained results are

The best Random Forest R2 train score is : 0.8278994903973604 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

The best Random Forest R2 test score is : 0.8516768166059918 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

The results are obtained and compared thoroughly and the best predictions are made.