BIKE SHARING DEMAND PREDICTION

Chouyin Zhang

Hunan University
China

2022-11-11

Overview

Problem Definition

- Problem Introduction

Exploratory Data Analysis

Data Preprocessing

- Data Preprocessing–Logarithmic transformation
- Data Preprocessing–Drop variables
- Data Preprocessing– Replace and Split

Train and Apply Models

Conclusion

Problem Definition

Problem Introduction

Defn

Problem Background

- Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city.
- Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis.

Defn

Problem introduction

- In the data generated, the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded.
- Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.
- In this competition, participants are asked to combine historical usage patterns with weather data in order to *forecast bike rental demand*.

Exploratory Data Analysis

Exploratory Data Analysis

- Distribution of variables
  - Distribution of dependent variables *count*, which represents bike usage.
  - Distribution of *count* with *season*, *holiday*, *weather*,*month*, and *working day*.

Figure 1: Description of Count

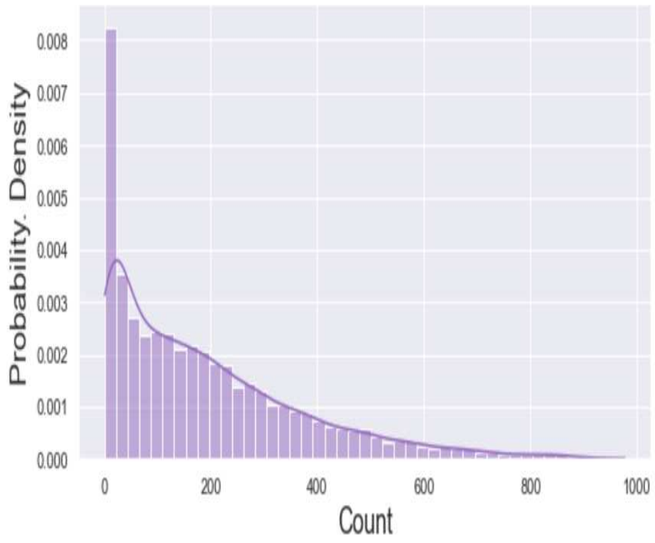


Figure 2: Description of Count and Month

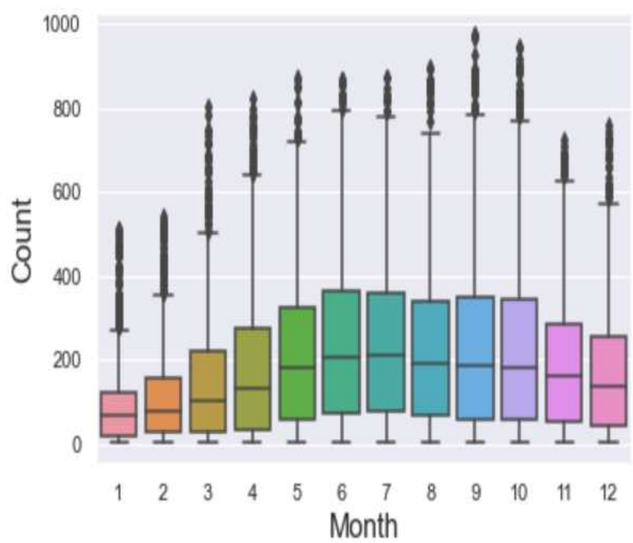
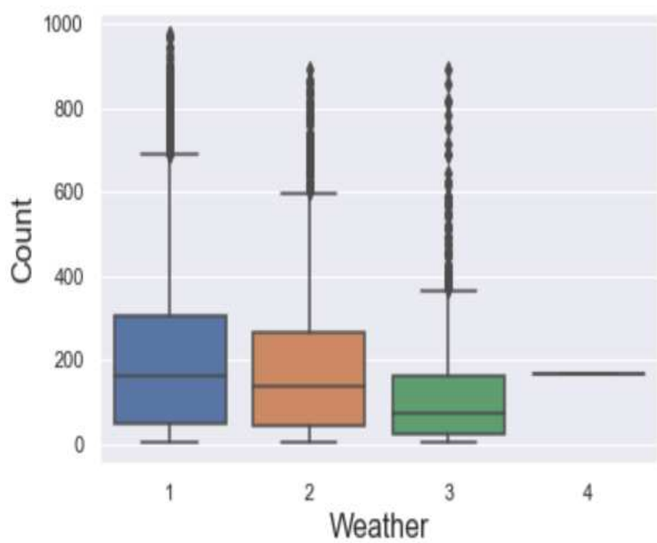


Figure 3: Description of Count and Weather



Exploratory Data Analysis

- Distribution of variables
  - Distribution of dependent variables *count*, which represents bike usage.
  - Distribution of *count* with *season*,*holiday*, *weather*,*month*, and *working day*.

Figure 4

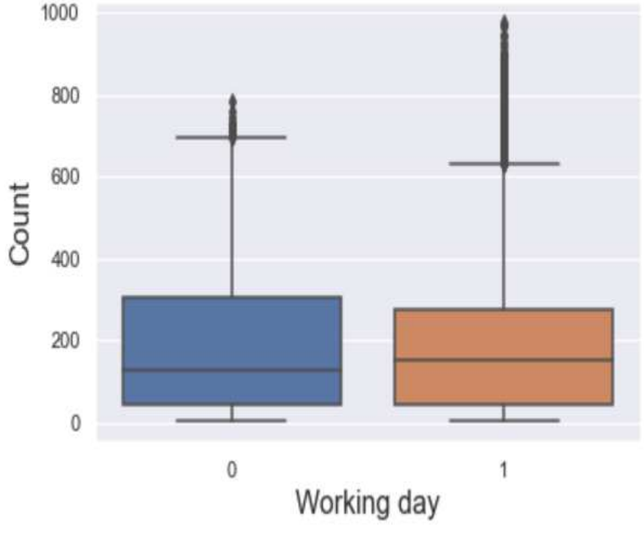


Figure 5

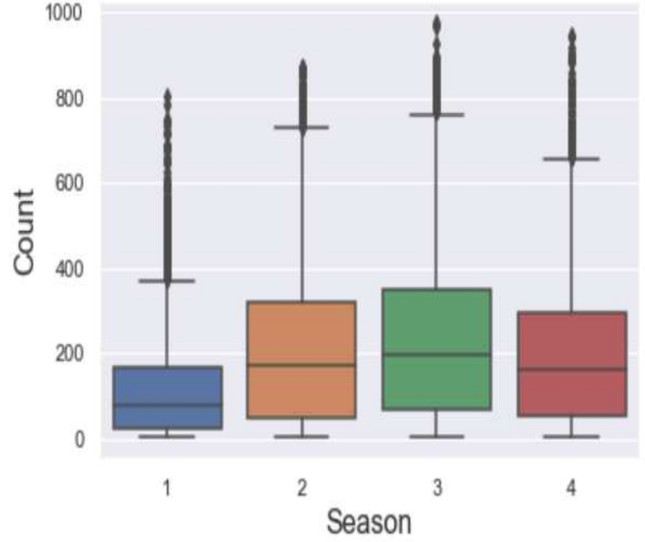
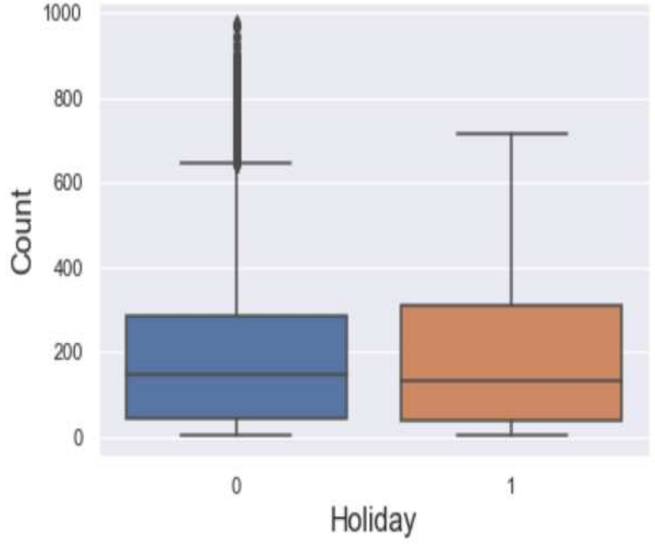


Figure 6



Exploratory Data Analysis

- Distribution of variables
  - *Four jointplots* showing how daily usage varies with the continuous variables.

Figure 7

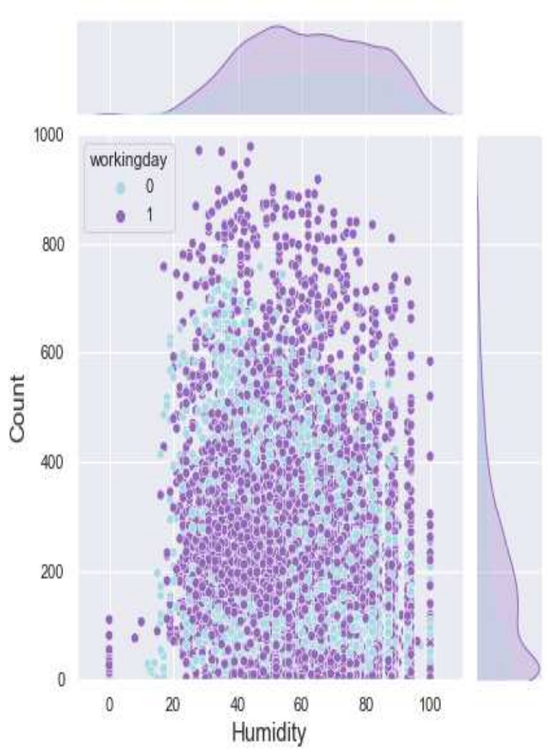


Figure 8

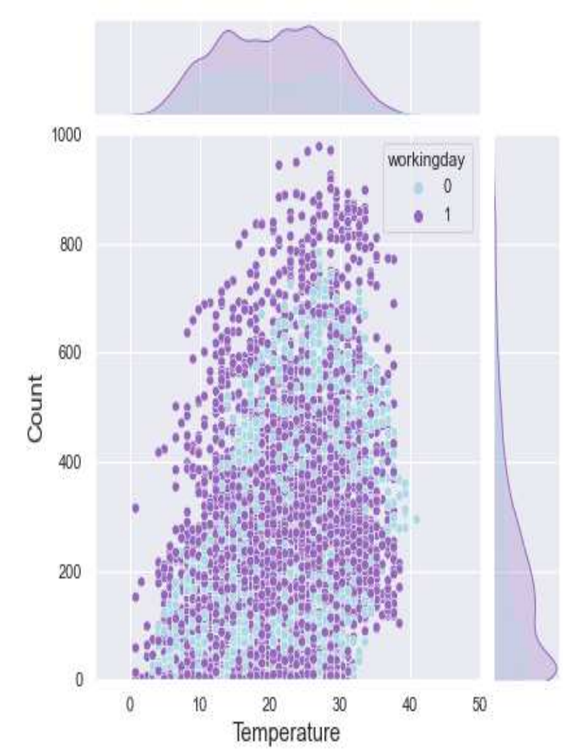
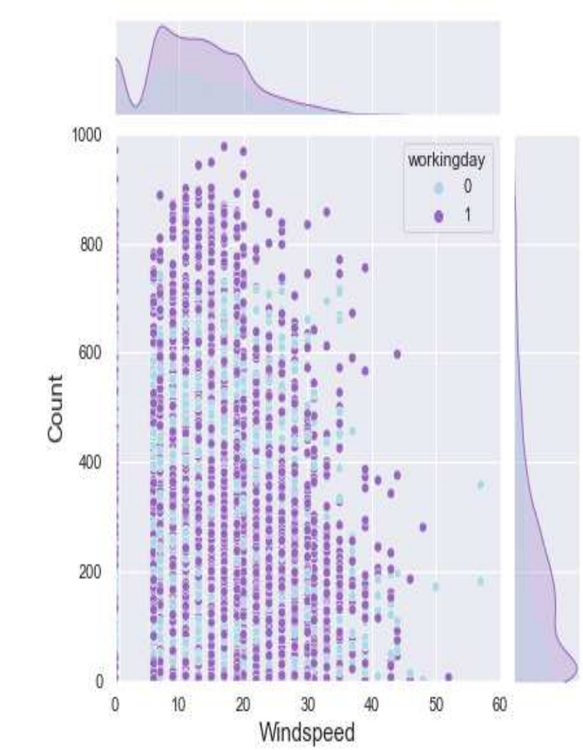


Figure 9



Exploratory Data Analysis

- *A correlation heatmap* showing the *2D correlation matrix* for the features in our dataset.

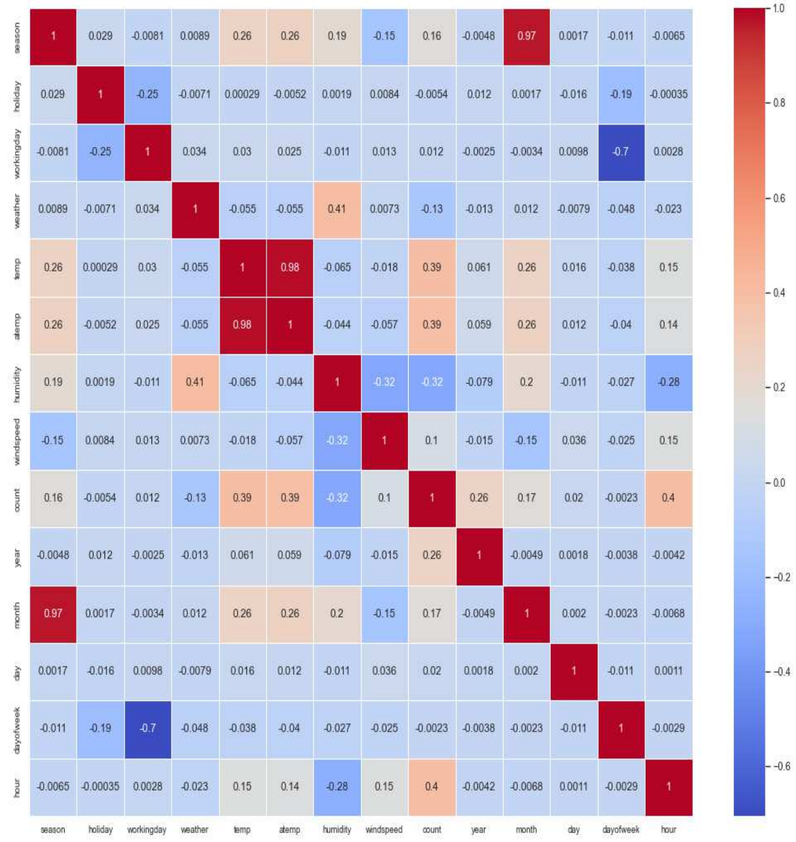


Figure 10: Correlation heatmap

Data Preprocessing

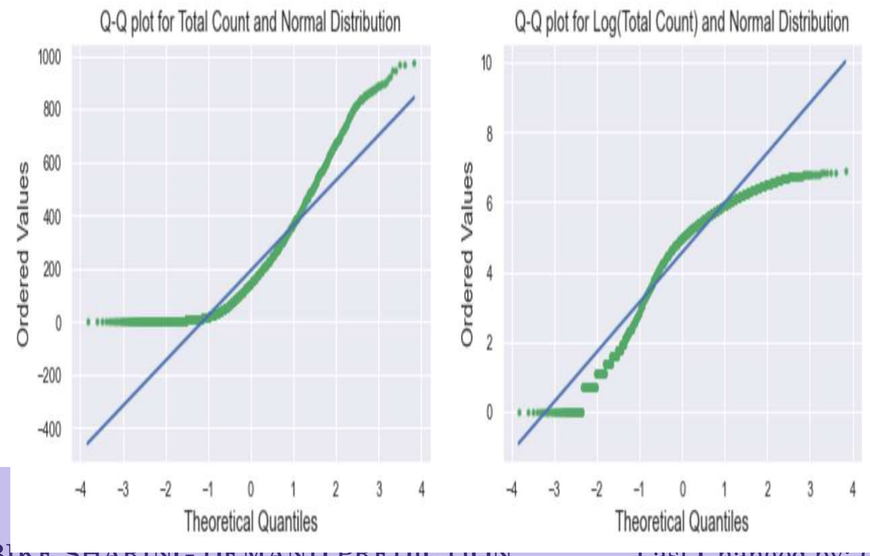
Data Preprocessing–Logarithmic transformation

Defn

Logarithmic transformation

- First we can look at the distribution of count. The distribution is heavily skewed right, which is not well approximated by a normal distribution.
- Taking the natural log of the count data gives a distribution slightly more normally distributed, as shown above, and also in the quantile-quantile (Q-Q) plots below.

Figure 11



Data Preprocessing–Drop variables

Defn

Drop some variables from our dataframe

- 1.Drop casual and registered as the test set does not contain data for the casual count or the registered count.
  2. Drop atemp since it has almost perfect correlation with temp.
  3.Drop all day of the month data from the training set before training our models.

Defn

Replace the target variable count with the natural logarithm of count

- As the natural logarithm of count is more normally distributed, Let us replace the target variable count with y=ln(count+1).

Data Preprocessing– Replace and Split

Defn Replace the categorical data with binary dummy variables

- The categorical data is expressed as an arbitrary numerical value. Therefore we can use dummy coding, replacing the categorical data with binary dummy variables using the pandas function.

Defn Split the labelled training set into two sets

- One for training our models, and a validation set for determining the best set of hyperparameters.

Defn Define a prediction function

- Finally, let's define a function predict that will report prediction scores for a given model.

Train and Apply Models

Train and Apply Models

- Build and apply some chosen model

Table 1: $\alpha = 4$

Models	RMSLE value
$\{LinearRegression\}$	0.136
$\{Ridge\}$	0.137
$\{Lasso\}$	0.148
$\{RandomForestRegressor\}$	0.084
$\{GradientBoostingRegressor\}$	0.080
$\{XGBRegressor\}$	0.085
$\{LGBMRegressor\}$	0.078

It can be observed that the LGBMRegressor model can get the best score. So, we apply this best model to test data.

Train and Apply Models

- The output of the prediction is as below

Figure 12

	Unnamed: 0	datetime	count
0	0	2011-01-20 00:00:00	10.0
1	1	2011-01-20 01:00:00	4.0
2	2	2011-01-20 02:00:00	3.0
3	3	2011-01-20 03:00:00	3.0
4	4	2011-01-20 04:00:00	2.0

Conclusion

Conclusion

- In this data mining task, I use the dataset from Kaggle. According the problem background given, I apply EDA, data preprocessing to the dataset, and finally apply the best model to test data, successfully finish this task.

Contact Information

Chouyin Zhang
Hunan University, China