CSE 805: Machine Learning and Data Mining

Class Project: Phase 2, Data Preprocessing

Dr. Balu Gokaraju

Robert Akinie

Fall 2021

Objective

The purpose of the group project is to give students an opportunity to apply course learning objectives in formulating their own Machine Learning projects, using selected datasets. This report details the data preprocessing stage of the overall project.

Analysis

Data preprocessing is essential in any Machine Learning (ML) problem, in making data and datasets possess functionality and useability for building of ML models for deployment. Data preprocessing is mainly cleaning the dataset, by looking for missing/zero data, outliers and taking the relevant actions, as well as data visualization, to determine key information associated with the data.

Problem Description

The data for this problem is populated from TripAdvisor.com. It contains reviews on destinations in 10 categories. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0) and average rating is used against each category per user. Each category represents user feedback on a certain destination.

Preprocessing Steps

For this project, the remaining sections show the output of the various steps taken in preprocessing the data, as well as some questions that arise from these steps. The appendix contains the code for each step, respectively.

Peek of Data:

Graphical user interface

Description automatically generated with medium confidence

Figure 1. Data attributes and values

Descriptive Statistics:

A picture containing application

Description automatically generated

Figure 2. Descriptive Statistics.

Data Size:

Text

Description automatically generated

Missing Data check:

Text

Description automatically generated

Some of the attributes in the dataset contains zero data. This is part of the data, as shown by the comments in the data description. The attribute “User ID” is just and index representing each user’s ratings. As this is not relevant, it will be removed in the feature engineering phase.

Outliers:

Outliers in this dataset will be defined as values greater/less than 3 standard deviations from the mean. Box plots will be displayed to reveal outliers. If the outliers are small relative to dataset size, these outliers will be removed/ Following that, the following data visualization methods will be shown, based on an outlier free data: Correlation Matrix, Histogram, Density plot, Skew, and Scatter Matrix plot respectively.

Box plot:

A picture containing calendar

Description automatically generated

Figure 3. Box plot of data.

Data shows a few outliers. The following figure will display the number of outliers per attribute. Text

Description automatically generated

Since the outliers are small relative to the dataset size, they will be removed. The next figure shows the descriptive statistics of the dataset after removing the outliers.

A picture containing application

Description automatically generated

Figure 4. Descriptive Statistics of modified data.

A second box plot on this new data reveals outliers in the data. This could be because the descriptive statistics are applied to this new data, independent of previous data manipulations. Subsequent box plots prove the case.

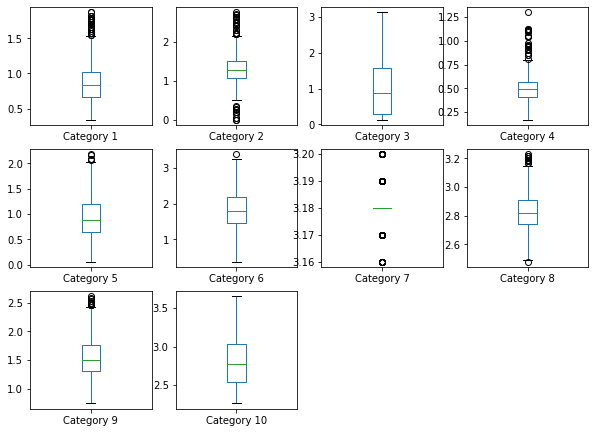


Figure 5. Box plot of modified data.

Category 7 has an interesting data, as shown by its standard deviation value. Most of the values are centered very tightly around the mean.

Correlation Matrix:

Table

Description automatically generated

Figure 6.Correlation matrix of data.

Categories 5 and 6 seem to be in slight correlation (0.58), whilst Category 7 has above average correlations with categories 3 and 10 (0.75, -0.71). Category 10 also has some correlation with Category 4. The remaining attributes have very low correlation with each other.

Correlation Matrix plot:

Chart

Description automatically generated with low confidence

Figure 7. Visual plot of correlation matrix.

This shows a visual representation of the correlations of the attributes with each other.

Histogram:

A picture containing crossword puzzle, scoreboard

Description automatically generated

Figure 8. Histogram of data.

Most of the attributes seem to possess a skewed Gaussian distribution, with tails on one end. Category 3 has a more exponential curve, and Category 7 has its observations very close to the mean. Normalizing these attributes would make them more Gaussian.

Density plot:

Graphical user interface

Description automatically generated

The density plot is similar to the histogram, except it does not show number of observations per bin.

Skew:

Text

Description automatically generated

The skew measures the asymmetry of the attribute values around the mean. Most of the ML models assume a Gaussian distribution, with no skew. Methods such as Box-Cox transformation or log transformations might improve the skew of each attribute. Category 4 possesses the most skew.

Scatter Matrix:

Diagram, timeline

Description automatically generated

The scatter plot matrix shows the visual relationship between all the features in a dataset, as dots. Category 7 has an interesting data pattern. A lot of users gave ratings between 3.16 and 3.20.

Conclusion

This phase reveals some interesting trends of some attributes, like Category 7. It also shows some of the steps to consider in the feature engineering phase, such as feature transformations and data normalization. The next section would decide whether such an attribute is to be kept. Table III contains information regarding the skew of these attributes.