Jeanne Reppert
2/23/2020

***IAF 604, Machine Learning and Predictive Analytics, Project 1***

**Step 3:**

*Do the basic necessary things to understand the dataset, such as datatypes, check null values, and statistical information.*

The carwood dataset has 2048 rows and 68 columns. The first 67 columns are named f1 – f67. The last column is named label and contains the values 0 and 1 which are used to label the two datasets that have been combined to make this set of data. All datatypes in f1 to f67 are considered to be "double" types (according to apache documentation double refers to: "Represents 8-byte double-precision floating point numbers.") The last column 'label' is an integer type. There are no null values.

A review of the data visually shows that the distributions of the features are normal and there are no outliers observed in the boxplots. The overall distribution of the data set is also normal which is observed in the histogram that represents the overall representation although it does exhibit some multimodality. Line graph that shows the closeness of the mean and median also indicating normality of the distribution. The overall mean of the dataset is 123.47 and the overall median is 120.70. The range of the dataset is 166.09 with a standard deviation of 32.93. The minimum value is 45.67 and the maximum value is 211.76.

**Step 4:**

*In the previous assignment, most of you have no idea why we have to normalization and standardization the data (but these two comes under statistical information). Understand when we to perform these things. Does the dataset required to perform these two things?*

*Further discussion of the nature of the dataset for the purpose of considering normalization or standardization:*

All features have a similar kurtosis of between negative .68 and .91. This indicates that distributions of features are platykurtic having lighter tails and heavier peaks. Given that these values are between 0 and 1 and there is very little range between kurtosis values one can inference that there is a reasonable normal distribution among the features.

Likewise, skewness values are between .11 and .21 indicating very little skewness in each feature (acceptable levels being less than 1 or greater than -1) Again, one could

inference that based on these values and the relative small range of skewness values between features that there is a reasonably normal distribution among the features.

The boxplots of the features visualized below show a fairly consistent spread and median among the features. The histograms also show that the plots are fairly similar in terms of distribution. Some features do have a few differences in terms of the degree to which they are bimodal or multimodal. In the same way, the overall histogram of the data shows a relatively normal distribution.

**Should we use normalization or standardization?**

Normalization (Min-max) is used when the objective is to bring the values of a feature or features to a range of 0 to 1 through rescaling. Standardization (z-score or standard scaler), on the other hand, also rescales the data but instead brings the values of a feature or features to have a mean of zero and a standard deviation of 1, effectively resulting in the distribution of the data being between -1 and 1.

Typically, normalization or standardization are used to transform data when these types of changes to the data would correct issues that make the analysis of the data difficult to do. For example, if values are extremely different in terms of range, minimum and maximum values, then rescaling these values would make them easier to compare. Another reason to rescale data would be if there were significant outliers in the data or if variables that were to be compared in a model had different units of measure. Finally, sometimes certain machine learning algorithms (like principal component analysis) respond better when the data is normalized or standardized first.

Presently, since our data has a normal distribution and no outliers it does not appear that normalization or standardization is necessary at this point. Furthermore, all of our data features represent values that had the same unit of measure and the range of the numbers is fairly compact. At this time then these techniques would not need to be applied. However, should we utilize these sets in certain machine learning algorithms it could become necessary or advisable to standardize the data set before performing the machine learning algorithm.

**Step 5:**

*There are some duplicate columns in this dataset, think how we can find which are duplicates (repeated). What are you going to do with those columns?*

Three duplicate columns were located in the dataset. The duplicate columns are f18 and f61, f45 and f60, and f49 and f57. Because these three columns are duplicates (and I know that they have been added since we did the homework on the same sets) I have dropped one of each of the pairs (f18, f45 and f49) because otherwise they would negatively impact the accuracy of any results that would be achieved with later tests. Without the prior knowledge of what a dataset might look like,

however, in other scenarios, it would be more difficult to make the decision to drop these columns without further verifying why these three duplicate columns exist.

**Step 6:**

*Count label values. Discuss whether the data is Imbalance, Inaccurate, and Incomplete data. Provide you discussion.*

IMBALANCED: The previous cells identified that there were more labels with the value of 1 than the value of 0. There were 1027 cases of '1' in the label category and 1021 cases of '0'. In the case of this dataset, since we constructed it during the last assignment, the expectation would be that there would be an equal number of 0's and 1's (1024 each.) Hence, it is likely that one of two things has happened to the dataset: either entire rows have been deleted and then others duplicated or some 0's have been changed to 1.

INACCURATE: In this case, I would inference that there are three rows that have been altered since the cells used to check for duplicate rows did not indicate any duplicated rows. Unfortunately, I don't think there would be a way to detect which ones had been altered. Thus, in this case, I would find that the dataset is both inaccurate and imbalanced. Additionally, the presence of three duplicated columns also indicates that the data is inaccurate (since I know that there should be 64 columns plus a label and that three columns have been copied.)

COMPLETE: There are no null values in the dataset. In the case that the rows were completely deleted, and others added, then I would consider the possibility that the data was also incomplete since information would be missing. However, given the above discussion, and since there are no NaN, null or missing values observed the data is complete (in terms of missing items) but it is imbalanced and inaccurate.

**Step 9:**

*Discuss what type of machine learning approach would be best for this dataset.*

This dataset represents collections of pixel values of two photographs, one of a carpet and one of a hardwood floor. In this case the type of task that needs to be completed is to learn from the dataset how to accurately predict based on a row of values whether that row represents values from the hardwood set or values from the carpet set. Essentially, this would be an image recognition task or more specifically a pattern recognition task. Ultimately, this would involve classification of images. Deep learning with tensor flow could be one way of approaching the image recognition or classification job. Another method that might be applied to this dataset would be a neural network model.