CSC 177 – Data Analytics and Mining
California State University, Sacramento.

Project - Data Preprocessing:

Detailed Rubric Explanation:

Testing the data preprocessing Jupyter notebook tutorial4 on CANVAS:
Export tested program with result data into a PDF file. 40 pts. (these are just like bonus points)

Remaining 60% (steps 3(a) or 3(b)):

Detailed breakup of remaining 60 pts:

1. Choosing domain area, finding dataset, preparing data: 15 pts.
   a) Choosing the domain area of value.
   b) Finding the dataset in any domain.
   c) If the dataset is tool clean, use any technique to modify the CSV file manually or programmatically for pre-processing.

2. Documenting summary in PPT and a short report: 45 pts.
   a) Documenting or describing a comprehensive set of techniques for preprocessing.
   b) Showing all steps in proper order for preprocessing and why you think the order you used is important.
   c) Developing and documenting human insights with human interpretation on preprocessed data and possible effect on predictions.
   d) Splitting dataset into two sets. Document your rations for splitting and results. Compare results for training and test sets and developing comparative intuition on the meaning of your results of splitting, with respect to statistical parameters.

**** 5% extra points for understanding and responding to any imbalanced dataset issues with clear documentation ****

Imbalanced dataset is a dataset for example, where there are disproportionate records of each kind say, 80 percent records on women and 20 percent records on men.

Finally, Submit the jupyter file with results as a PDF or if you are using any other language, submit code and export your code into PDF file and submit all the files in a zip folder.
Submit all documentation with it.

**** Also mention each team member's contribution. *****

CAVEAT:

It is possible that your data set is not rich enough to perform all preprocessing steps on a single data set. In this case, it is okay for you to use different data sets so you may apply partial set of

pre-processing steps on each data set and thus complete all kinds of preprocessing operations across all the data sets.

In any case, you may describe how you would order the pre-processing steps on a single dataset, so that the operations are consistent with each other, such as dependencies, because of the ordering.

Some FAQs:

1. We all have gone through the tutorial and were able to get the right results for the code. We have looked at all the different techniques that were mentioned in the tutorial. So, everything from how to deal with missing values all the way to PCA. We were talking about trying to implement all of the techniques in the tutorial in our project. Will this be sufficient or is it possible that we only need to apply a few of the techniques mentioned in the tutorial instead of all of them? Do we need to do the TensorFlow stuff? Should we do all the stuff that has the red code above it?

Answer: All the techniques should be used for the assignment. Try to apply as many of the techniques in the tutorial to your data set. Some might not be applicable depending on your data set. If data is not ready for machine-learning then a machine learning algorithm cannot be applied. Hence this is why we do data preprocessing.

2. What is meant in part 1(a) of the instructions where it says "Choosing the domain area of value"?

Answer: You do not specifically choose the domain area. The domain area is chosen based on the datasets you are using. Apply some judgment to what datasets you want to use and justify why you chose the data set you did.

3. What is meant in part 2(b) of the instructions where it says "Showing all steps in proper order for preprocessing and why you think the order you used is important"? Is there a specific order in which we should apply the techniques? If we apply the techniques in a different order than what is shown in the tutorial would that be wrong?

Answer: Don't do the techniques in a haphazard way. You can't do a z-score before filling in the missing values and so on.

4. Is it safe to say that the proper order to do the preprocessing techniques is the order displayed in the lab tutorial?

Answer: After doing the missing values do a describe function. Might not need to do concatenation. In terms of the order to apply techniques to your dataset depends on the data set itself, use your judgment. What to do with missing values will depend on your data set as well. You can get rid of the missing values or use the mode or mean to replace the missing values.

One hot encoding is when independent variables in your table are categorical (that is not numerical, such as color and its values red, blue, yellow).

In a table the column for which values are predicted is the dependent variable and the columns used to predict are called the independent variables.

Here you would ask python to create new columns color_red, color_blue, color_yellow and

for example, if the color is red the values are:

color_red	color_blue	color_yellow
1	0	0

This is one-hot encoding and python has a function to do this for you from the dataset.

Labeling or label encoding is applied only to the dependent variable also called the output or class variable.

In this case, the labels are unique distinct values. Say the dependent variable is color just for discussing label encoding. (In the earlier example we considered color as an independent variable). In this case, your labels may be turned into 1 for red, 2 for blue and 3 for yellow. Then when you predict a 2 it means you predicted blue color. This kind of encoding is label encoding.

Aggregation is very important as in statistics, doing statistical analysis on aggregated values is more reliable. For example, if you have sales data for each day, aggregating by week or by month and then performing analysis gives more stable and reliable results.

Outliers are in simple terms data that is abnormal or not normal for the domain of analysis. For example, a height of 10 feet for a person is abnormal. So, this value is an outlier. Box plots give you an idea of the median and the percentiles and also visually depict the outliers. Outliers are values outside the IQR range: IQR = Q3-Q1 where Q3 = the 75th percentile and Q1 is the 25th percentile values.

An outlier is statistically below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers.

Some statisticians may use their own outlier ranges.

For data with normal distributions (you can check the shape by plotting the column) Z-scores are used to detect outliers.

z-score = (x - mean)/standard deviation

with Z-scores, the outliers are considered to range below -3 z-score and above 3 z-score.

5. How would we take a clean data set and make it unclean?

Answer: Create missing values to make the data set unclean randomly. Pick a random number in the row and column. Remove the data in that row and column if it is not already missing. Do not remove more than ten values!

6. For 2(d) it originally said to calculate the standard deviation and mean of both the training and test data sets. Now it says to calculate the 25th percentile, the 50th percentile, and so on. Are we required to calculate these extra things now?

Answer: Yes, calculate the extra things added to the instructions. Use the describe function in python on the data set to get all of these calculations automatically. Then provide comments of your analysis of the results from the describe function.

7. Part 2 of the instructions says “documenting summary in PPT and a short report”. Should we make a PowerPoint describing our findings or can we just write the report in a Word document? Will we be needing to do a presentation on our project?

Answer: At the end of the course, you will create a PowerPoint on all of the previous projects. It's only recommended to create a PowerPoint for each specific project to help you remember certain information about the project. Reference your code comments and other report material to help you with the final PowerPoint presentation. Everything can be documented inside the Python code. A standard report can be written but it is not necessary if everything is documented in the Python code. (Modify Instructions of the Assignment). You are encouraged to provide a separate document for the report as this builds the skill in you to prepare you in your career.

8. What exactly should we turn in for the assignment? Should we submit everything in a zip file?

Answer: Turn everything in via a zip file. Python code, reports, PowerPoint presentations, whatever you have for the project. PowerPoint Presentation is optional though it is valuable to make them in preparation for your semester-end-of-course submission.