Michael Albers

Project 5: Data Gathering

Wrangle Report

Summary

The Twitter data was comprised of three files, enhanced archive data, image predictions, and twitter Api data. The data quality issues were found in the enhanced archive data. The image predictions file data was normalized meaning its prediction column data were converted into rows to make future analysis and visualization manageable. The new image predictions table design allows for the additional predictions to be added to neural network model without making structural changes to the image predictions table by adding additional columns.

Quality Issues

The text data in over 400 records from the twitter enhanced archive file had a reference to profanity, a four-letter word in the form of “h*cking”. While this word is not explicitly offensive, we see this text pattern of “*ck” everywhere on social media and even on clothing. It’s meaning is the 4-letter F-word and this is what’s triggered in the reader’s mind when he/she sees it. There is no place for this on a business report and so the resolution was to remove all 400+ records from the twitter enhanced archive file with this text pattern. A regular expression was used to find all the records containing this text pattern. It would have been good if I could have replaced this part of the text with spaces or # signs but there were too many records to check manually in the testing phase.

There were 11 rating numerators that were over 100, a couple of them were over 1,000. The resolution was to remove these outliers from the twitter archive enhanced file. I thought about dividing these records by 100 but that would have meant altering some of the ratings while not touching other ratings which would have brought bias into the analysis. It would have lessened the credibility of future findings if the consumer of this study knew that some of the ratings were altered.

There were about 23 denominators not equal to 10. These outliers were removed from the twitter enhanced archive dataset. It would have been easy to replace these denominators with 10. But it would have altered the perception of the final analysis knowing that some of the ratings altered.

The final instance in which records were removed was for the entries that were not dogs. The Twitter “We rate dogs” is only for dogs, so records containing “Please only send in dogs” in the text field were removed.

There were other minor changes made to the twitter enhanced archive for the timestamp data type, missing data in columns, and replacing ‘None’ with empty text in the name column.

Tidiness

The image predictions file had its first, second, and third prediction data values stored in separate columns, i.e. p1, p2, p3. This made it very difficult to accommodate adding future predictions to the neural model. An additional prediction would have meant altering the image predictions table with three additional columns. Having separate columns for each prediction meant that each prediction

exists on a different access for visualizations. I wanted all predictions to be on one axis (either x or y axis) so that I could summarize breeds of dogs by each prediction (p1, p2, p3) and then generate bar charts showing the count of breeds in each prediction. This new layout makes histograms easier to generate for an average of confidence percentages in p1, p2, or p3