CSI4142 Fundamentals of Data Science
Deliverable E, Part A: Preprocessing
Group 20

Charie Brady (300043672)
William Huynh (300018851)
Samet Yilmaz (7262963)

Submitted April 14, 2021

- First we queried 30 attributes from our database, duplicating the rows that represented more than one case. Checked the row counts = 76, 456, consistent with our database.
- Each column was checked for missing values and there were none.
- Checked for feature imbalance in our chosen label, outbreak_related, and found a 1:4 split for True:False. Decided not to undersample as this is not an extreme skew.
- Checked categorical attributes for their cardinality. Gender had four values. Removed the rows for 'unspecified' and 'gender diverse' as they represented 0.6% of cases (having a binary 1/0 is easier for our purposes than having to one-hot-encode four attributes).
- Converted binary variables like gender, city, holiday, weekend, outbreak_related to numeric 1/0.
- Age was converted to an ordinal, normalized numeric of equal spacing. [0.1, 0.9] for '<20' to '90+'.
- Used the get_dummies() Pandas function for one-hot-encoding of two categorical attributes: zone_measures, acquisition_group.
- All numerical columns were min-max normalized using MinMaxScaler from sklearn package (10 attributes).
- Labels were deleted and stored in a separate column, leaving a final feature set of 26 attributes with values [0,1] and 76, 062 rows.
- Feature selection was performed on the attribute values, first with the Low Variance method using VarianceThreshold() from sklearn package with threshold parameter = 0.8*(1-0.8). This method did not reduce any of our attributes. A second method of tried, tree-based, using ExtraTreesClassifier() from sklearn package with n_estimators=50. This reduced our features to only 5: { 'age', 'CC', 'MISSING INFORMATION', 'NO KNOWN EPI LINK', 'OB'}. Four out of the five are one-hot-encoded values from the acquisition_group variable, indicating that many of the outbreak related cases are linked by known contact transmission, community transmission, "OB" meaning outbreak, which is a trivial relationship in our dataset.
- We decided to remove acquisition_group as a feature because "OB" is a trivial predictor of whether a case is outbreak_related. We end up with 20 features. When feature selection is run a second time using the tree based method, it selects only 'age'.
- The data was split into training and test sets for the features and label (using a 20/80 split), stratified on the labels.
- Overall, there were not too many issues with the data, aside from the high number of redundant features identified through tree-based feature selection.