Capstone 2 Report
Justin Holladay, 2021

Anomaly Detection in ABA Behavior Data

Applied Behavior Analysis is a data driven method of behavior change that began in the 1970s.

Problems:

1. In ABA data is collected and examined. In many companies entering data from paper into spreadsheets has been common for many years, but this leads to different methods and templates being used. Collating this data to be inputted into modern collection and examination tools can jumpstart companies' transition from paper to digital data.

Collating this data can:

- Allow examination of a clients historical data by modern software

2. There are many factors that can correlate to a clients' progress, but it is very time consuming to constantly examine all the possible relationships to a clients' behavioral progress. There are relationships that are not obvious as well as relationships that change over time.

Finding these relationships can help discover things like:

- If there are any relationships between bx and a client
- If more bx is observed with a particular staff member
- If a client reacts differently to one therapist than others
- If a therapist observes more of a type of bx between clients
- Inter rater reliability

Data Wrangling

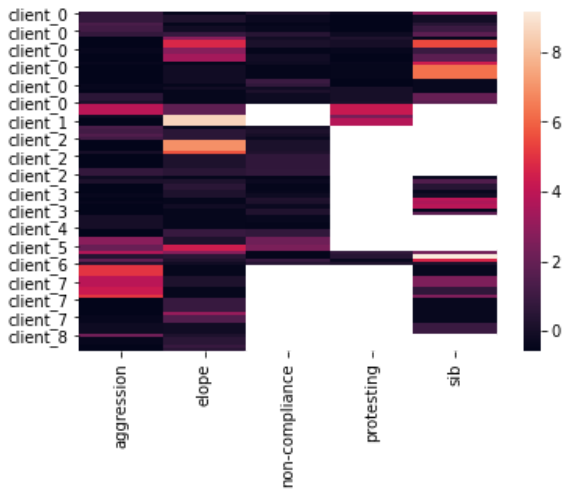
ABA data excel files were collected from many different templates and styles. Observations were generated between a staff and client pair containing the target behaviors. Behavior frequency data was generated for clients based on previously collected data.

The FuzzyWuzzy library in python was used to pull columns with approximations of target behavior names from many spreadsheets. Each threshold for the fuzz matching was generated after manual examination of different possible inputs. Matching observations were combined into a single standardized data frame containing all observations from all sheets.

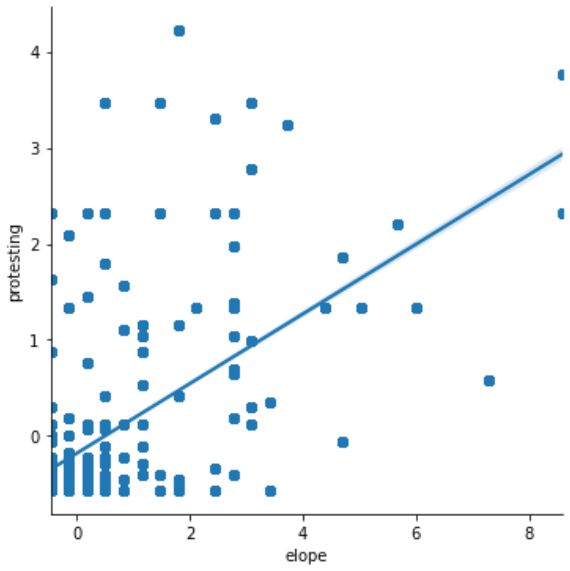
Input: Differently formatted excel files containing observations of behavior data
Output: Standardized template containing all observations

EDA

Looking initially through the dataset we can see there are some relationships between our variables. This heatmap shows some at a glance.



Through exploring our newly created dataset we can see an interesting relationship between eloping and protesting have the strongest correlation with overall behaviors. Another strong correlation is between aggression and non-compliance.



A number of different data frames were generated including lists of extreme behavior observations per client and reported by staff.

Input: Data frame containing client/staff behavior observations
Output: Graphs of correlations between behaviors, lists of top behaviors for clients, data frame of observations where at least one bx was labeled extreme.

Preprocessing

Dummy variables were created for categorical variables and merged into the dataset. Dummy variables for categorical data were later dropped as it became out of the scope of the project. Missing values in observations were replaced with the behavior column mean for each client. The bx data was standardized using StandardScaler(). Using this standardized data, a manual anomaly detection method was created to define outliers within a session.

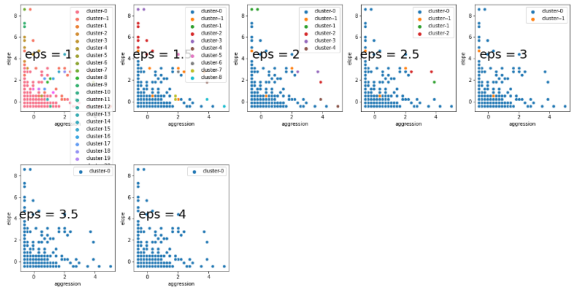
First, a threshold was generated using the overall average of each behaviors' mean. 2.5 standard deviations from the mean was used to define outliers. Session observations were extracted where at least one extreme behavior was observed into the extreme data frame. After manual review, these observations will serve as true values.

	Problem	Date	Day	Staff Member	Duration	aggression	elope	non-compliance	protesting	sib	client
client_3	elope	2021-06-17	3	staff_17	1.50	-0.585974	0.508188	-0.495307	NaN	0.405570	client_3
client_2	aggression	2021-04-05	0	staff_14	4.33	1.096018	0.508188	0.172368	NaN	NaN	client_2
client_6	na	2021-01-22	4	staff_7	7.00	0.348466	0.508188	-0.495307	0.411591	9.176212	client_6
client_3	na	2021-03-08	0	staff_21	4.00	-0.585974	-0.137430	-0.295005	NaN	3.860672	client_3
client_3	elope	2021-06-17	3	staff_19	4.00	-0.585974	0.508188	-0.495307	NaN	0.405570	client_3

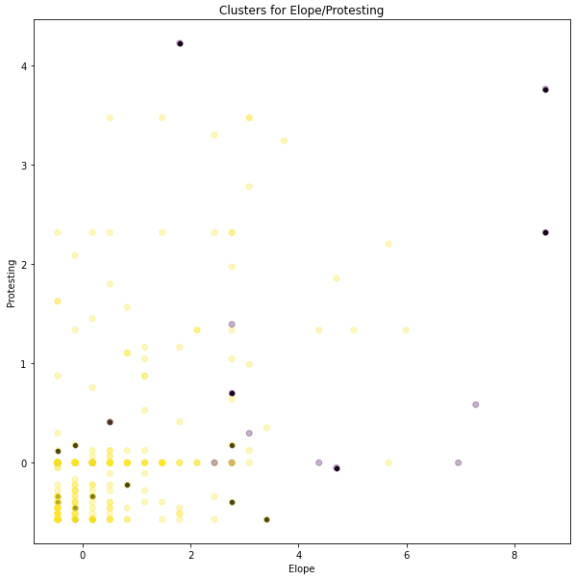
Input: data frame of extreme behavior data observations
Output: Behavior data standardized inside extreme data frame to be used for modeling

Modeling

Next DBSCAN and KMeans models were created to cluster the original dataframe. DBSCAN hyperparameters were tuned by generating many possible values for epsilon and 2.5 eps with minimum 20 samples. An elbow plot was generated for KMeans, though 2 clusters were chosen, not-extreme and extreme.



The result from DBSCAN and KMeans methods was similar with a silhouette score of 0.599
and 0.0.494 respectively. Final Project report as PDF. The report should clearly explain the problem, your approach, and your findings. Include ideas for further research, as well as up to 3 concrete recommendations on how your client can use your findings. Neither of these were better than the manual method of filtering.



Input: Original Dataframe containing client/staff session observations
Output: DBSCAN and KMeans models clustering the sessions observations into extreme and not extreme.

Conclusion:

Anomaly detection in ABA can quickly bring light to anything unusual for a client, and discover unseen relationships in a number of ways.

Exploring the data lead to trying 3 different methods (Manual, DBSCAN, KMeans) for outlier detection on a dataset of behaviors. The manual model is chosen over the others due to its clear inner workings and ease of making changes. KMeans did not perform as well as DBSCAN. The DBSCAN model performed at nearly 60% and could be investigated further when more features of this project are added.

Future work:

- Extend detection to include skill program targets in addition to behaviors.
- Collect information on gender, observation location and time of session.
- Define more relationships between independent variables.