Prediction of Successful Discharge

Children and Youth in Behavioral Health Treatment: Applying Machine Learning to Treatment Outcomes

Judith Calvo, PhD

June 25, 2020

Successful Discharge

- Hypothesized Predictors: age, diagnosis at admission, length of stay, initial scores on standard instrument, ethnicity, language and level of care
- Outcome of interest: child successfully completes treatment program

Why Project Was Chosen

- Effective treatment depends on ability to create interventions that are likely to lead to successful discharge
- Importance of ascertaining what leads to success in naturalistic settings where randomized clinical trials are not feasible. Machine learning allows for use of data collected in such naturalistic settings.
- Machine Learning Models are versatile and highly recommended at many levels of health care delivery: diagnosis, prediction and treatment planning
- Important to identify high risk clients at onset and intervene early and at the appropriate level of care to affect outcomes

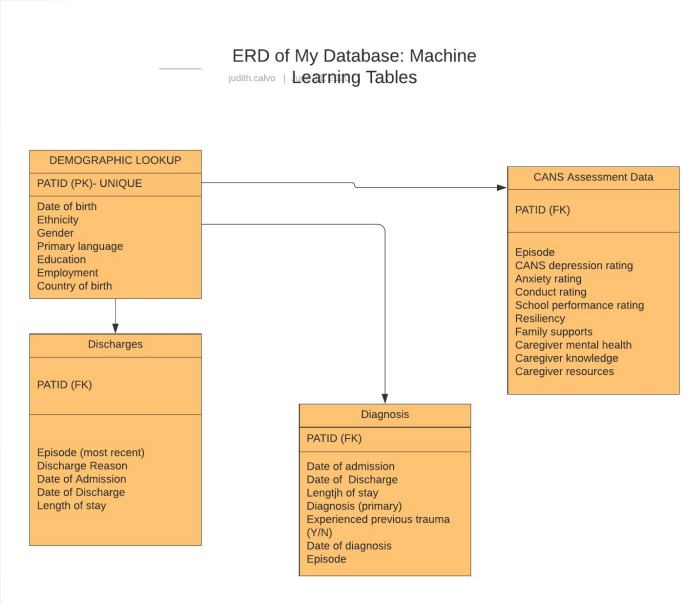
Data Source

- Agency EMR
  - Legacy system with limits on how data can be extracted
  - Use of Crystal Reports and SQL to extract data
  - Children and youth between 6-25 years old
  - EMR contains de-identified client data:
    - Demographics
    - Episodic data (admissions and discharges)
    - Diagnosis at admission (primary diagnosis only)
    - Program as a proxy for level of care
    - Discharge reason-outcome variable dichotomized to reflect successful vs, not successful reason
    - Scores on a measure of needs and strengths at intake (Child and Adolescent Needs and Strengths--CANS)

Questions We Hope to Answer

- What factors lead to successful completion of treatment?
- What client characteristics are indicative of high risk and failure to complete treatment?
- Can we create a streamlined, parsimonious model that can be replicated to predict treatment outcomes across other datasets, times and groups of clients?

First Step: ERD Data Relationships in Postgres



- Final data did not include Country of Birth or Employment as these did not vary much. Most youth were born in the U.S. and were not employed at the time of admission
- Caregiver scores on the CANS were also removed in the final model as they did not show significant relationship to the outcome. They did not predict nor play a role in classification based on testing several models

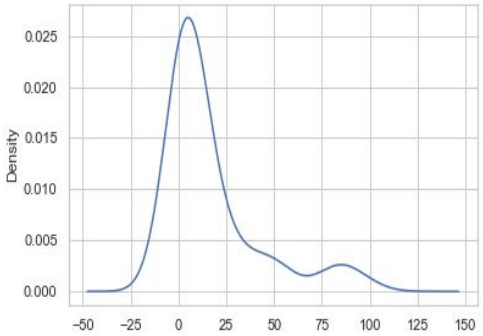
Tools Used

- Data preprocessing in Pandas
  - Encoding
  - Dummy variable creation
  - Matplotlib and Seaborn for visualizations
  - SMOTE to correct imbalance as needed
- Statsmodel to create full logistic model with betas and statistical significance
- Calculation of odds and probability for key features using Numpy exponentiation formulas
- Logistic model fitting using Scikit Logistic Regression Library
- ROC curve using Scikit AUC library
- Random Forest to compare results
- PCA and hvplot for interactive graphing
- Power BI for interactive dashboard using Python Scripting

Data Exploration

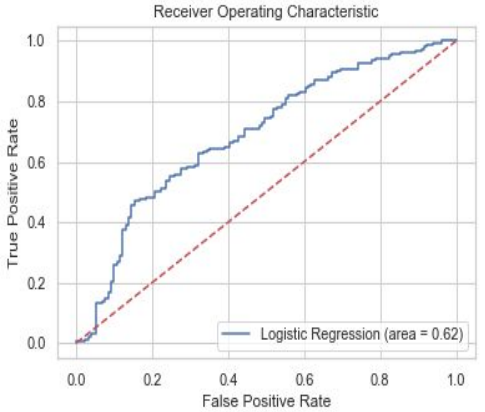
- Describe with summary statistics
- List categorical variables to narrow down
- Encoding of categorical variables
  - Scikit Encoder
  - Density plotting to ascertain binning categories
  - Custom encoding of level of care based on youth's program
  - Created dummy variables using Pandas Get Dummies Function
- Graphs
  - Visuals of data characteristics
- SMOTE to correct imbalance

Sample Density Plot



Data Analysis

- Logistic regression with feature output of beta weights
- Calculation of odds and probability for key features
- Logistic model fitting
- Metrics to evaluate model
- ROC curve after model fit (example)
- Random Forest to compare results
- PCA in unsupervised learning to create youth groupings/profiles based on entering characteristics
- Drop and add new features with next run to improve model fit and prediction, repeat analysis

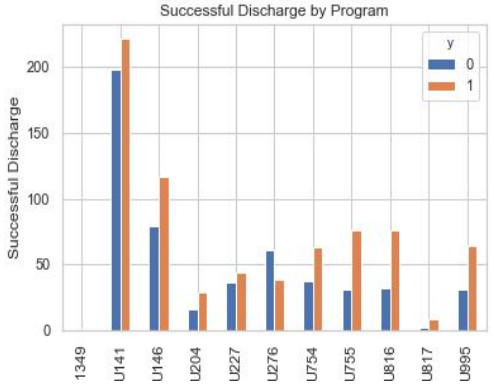


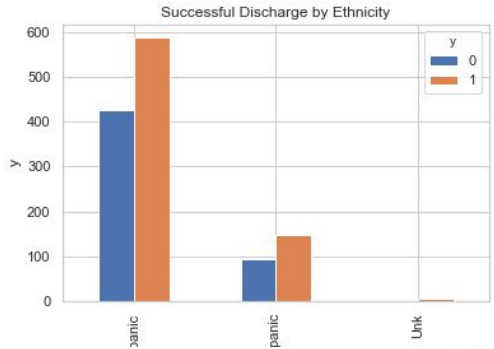
Dashboard Tools and Interactive Elements

- Power BI for interactive dashboard using embedded Python Scripting
- Interactive Elements
  - Slicers for exploratory data by discharge status, Level of Care and levels on initial CANS scores for key features (i.e., levels of oppositional behavior at intake, child resilience, family strength) that were posited to be predictive of outcome
  - Python script embedded in Power BI using matplotlib For Graphing
  - Importing of datasets using a python script with Pandas to allow for interactivity of visualizations.

Storyboard--Data exploration phase

- Data Stored in Postgres and .csv
- Cleaned and encoded
- Successful discharge by original program (1= successful) prior to custom encoding to create levels of care
- Successful discharge by ethnicity (no strong difference in proportions based on ethnicity--still higher proportions of success)





Storyboard--Analysis

- Logistic Regression Results
- Logistic model fit
  - Classification Reporting
- Random Forest
  - Classification and Feature Importance
- PCA Classification (possibly exclude)

Proposed Dashboard Layout