PREDICTION OF SUCCESSFUL DISCHARGE

CHILDREN AND YOUTH IN BEHAVIORAL HEALTH TREATMENT: APPLYING MACHINE LEARNING TO PREDICTION OF TREATMENT OUTCOMES

JUDITH CALVO, PhD

DIRECTOR OF QUALITY IMPROVEMENT,
COMMUNITY SOLUTIONS

JUNE 25, 2020

Successful Discharge

- Hypothesized Predictors: age, diagnosis at admission, length of stay, initial scores on Key Items in 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 young adults between 6-27 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)
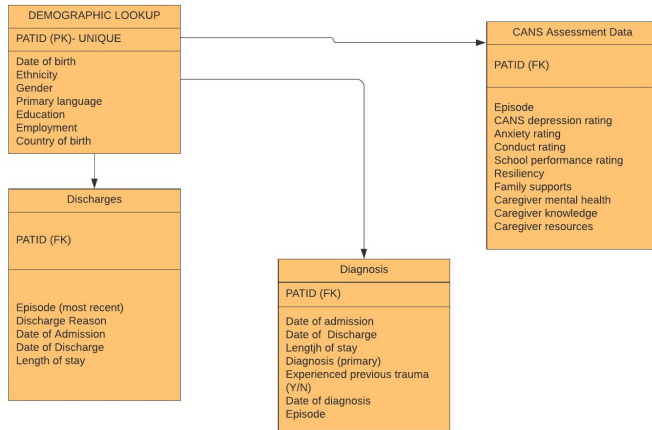
Questions I Hope to Answer

- 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

ERD of My Database: Machine Learning Tables
judith.calvo | Learning Tables

DEMOGRAPHIC LOOKUP

PATID (PK)- UNIQUE

Date of birth
Ethnicity
Gender
Primary language
Education
Employment
Country of birth

CANS Assessment Data

PATID (FK)

Episode
CANS depression rating
Anxiety rating
Conduct rating
School performance rating
Resiliency
Family supports
Caregiver mental health
Caregiver knowledge
Caregiver resources

Discharges

PATID (FK)

Episode (most recent)
Discharge Reason
Date of Admission
Date of Discharge
Length of stay

Diagnosis

PATID (FK)

Date of admission
Date of Discharge
Lengtjh of stay
Diagnosis (primary)
Experienced previous trauma (Y/N)
Date of diagnosis
Episode

- 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
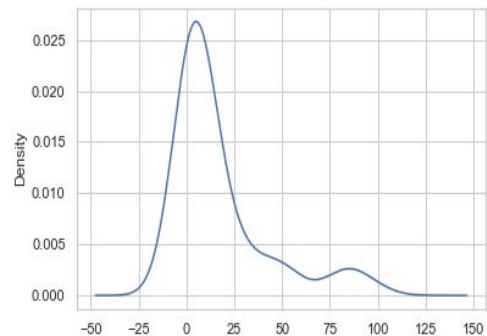
DATA EXPLORATION PHASE

- 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)

Successful Discharge by Program

Successful Discharge by Ethnicity
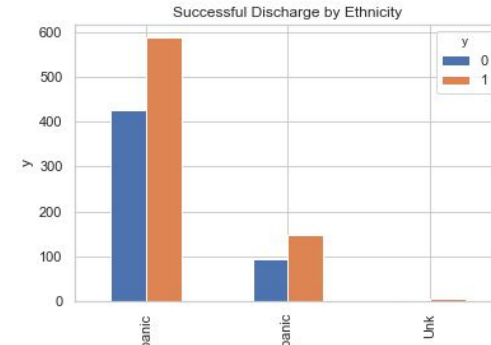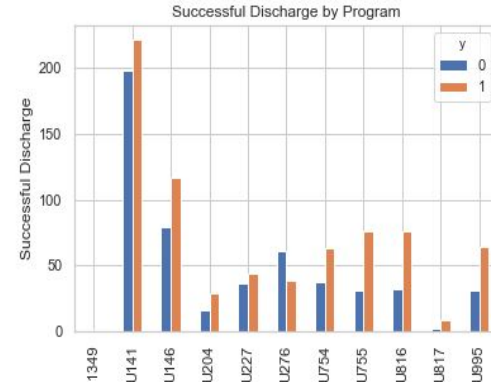
LEVEL OF CARE

LEVEL OF CARE IS AN IMPORTANT CONSIDERATION FROM THE BEGINNING AS A PREDICTIVE FEATURE. THIS WAS DEFINED DURING THE PRE-PROCESSING PHASE

- LEVEL 1--EARLY INTERVENTION/PREVENTION
- LEVEL 2--OUTPATIENT CARE (THERAPY/CASE MANAGEMENT)
- LEVEL 3--INTENSIVE CARE HIGHER SERVICE DOSAGE OF THERAPY AND CASE MANAGEMENT
- LEVEL 4--CRISIS CARE
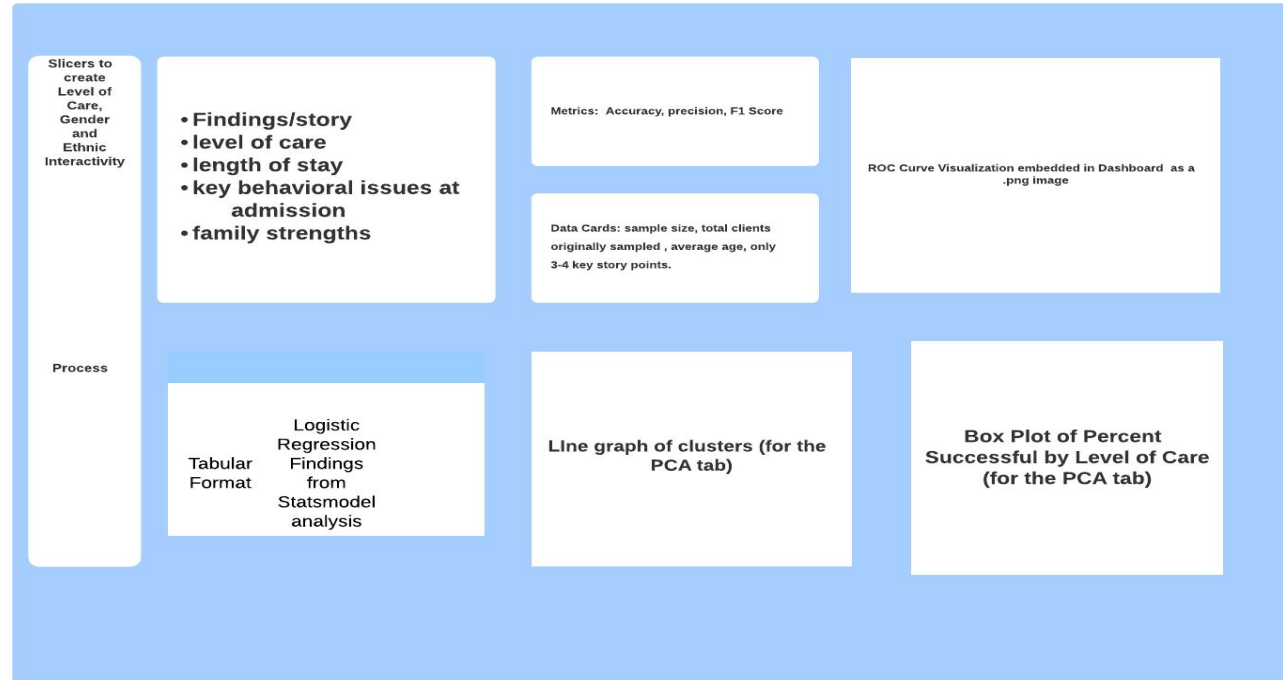
Dashboard Tools and Interactive Elements

- Power BI for interactive dashboard using embedded Python Scripting
- Interactive Elements
  - Importing of datasets using a python script with Pandas to allow for interactivity of visualizations and to have correct data.
  - 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 with interactive slicer to look at varying Levels of Care

GENERAL DASHBOARD LAYOUT WITH EXAMPLES

Dashboard/Story Board Layout

judith.calvo | June 18, 2020

Slicers to create Level of Care, Gender and Ethnic Interactivity

- Findings/story
- level of care
- length of stay
- key behavioral issues at admission
- family strengths

Metrics: Accuracy, precision, F1 Score

Data Cards: sample size, total clients originally sampled, average age, only 3-4 key story points.

ROC Curve Visualization embedded in Dashboard as a .png image

Process

Tabular Format — Logistic Regression Findings from Statsmodel analysis

LIne graph of clusters (for the PCA tab)

Box Plot of Percent Successful by Level of Care (for the PCA tab)
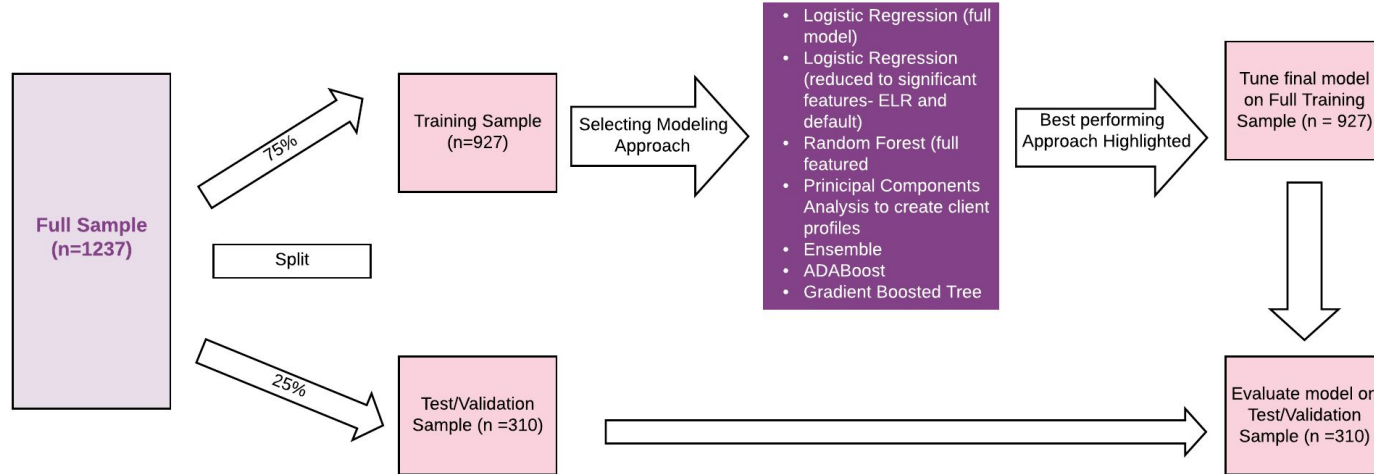
Data Analysis

Example ROC Curve

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

Receiver Operating Characteristic

True Positive Rate

False Positive Rate

Logistic Regression (area = 0.64)
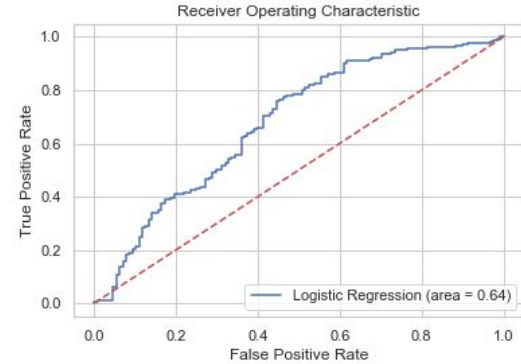
SCHEMATIC OF ANALYSIS

Schematic of Analysis
Pipeline and ML Modeling

Full Sample (n=1237)

75%

→ Training Sample (n=927)

Split

25%

→ Test/Validation Sample (n =310)

Selecting Modeling Approach →

- Logistic Regression (full model)
- Logistic Regression (reduced to significant features- ELR and default)
- Random Forest (full featured
- Prinicipal Components Analysis to create client profiles
- Ensemble
- ADABoost
- Gradient Boosted Tree

Best performing Approach Highlighted →

Tune final model on Full Training Sample (n = 927)

Evaluate model on Test/Validation Sample (n =310)

Drawing adapted from:
Webb, et al. (2020)

Storyboard Elements: How Model was Trained

- Logistic Regression Modeling
- Logistic model fit
  - Default and Elastic Net Regularization
  - Classification Reporting
- Random Forest
  - Classification and Feature Importance
- ADABoost & Ensemble
- Gradient Tree Boosting
- PCA Classification

Accuracy Scores

- Logistic Regression with Elastic Net Regularization yielded an accuracy score of 68%
- Gradient Boosted Tree Yielded the same accuracy (68%)