Predicting Youth Risk Behaviors:

Modeling the YRBS for Relevant Discussions with Youth

Becky Peters, Data Scientist; September 2021

Topics for Discussion

- **Part 1**: Background and Motivation
- **Part 2**: Exploring and Modeling the Data
- **Part 3**: Predictions and Dashboard
- **Part 4**: Future Work

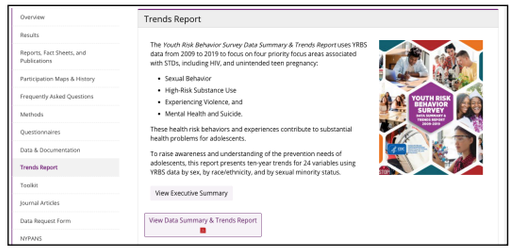
Part 1:

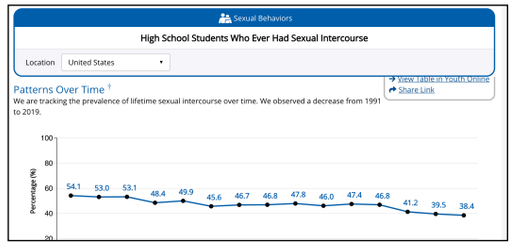
Background and Motivation

The YRBSS tracks risky behaviors for youth health, covering topics such as texting while driving, drinking and drugs, having sex, and bullying. To me, these topics are all like jellyfish. When I was in school the lesson was: Just say no. There was nothing to be discussed, it was just ‘don’t be the person who does ‘x”. This approach has been shown (??) To be detrimental to the better way, having conversations with teens about the dangers of their choices. We can’t help people make choices based solely on fear. Yes, some healthy fear is good, but it causes mistrust in adults when the answer to everything is ‘avoid it’. Teenagers should be aware enough of their surroundings to understand where the jellyfish are and how to avoid being hurt. SO they can make informed decisions for themselves about the direction they want for their lives. Health care professionals, teachers, counselors are all amazing resources to have these conversations, but if we’re talking about the wrong jellyfish with the wrong kids, it turns into a lecture instead of a conversation. Hoping that this model / dashboard can inform the appropriate times for different conversations, and hopefully give youth some control in how they respond to their environment.

Youth Risk Behavior Survey

- CDC Youth Risk Behavior Surveillance System, 1997-2019
  - Violence, sexual behaviors, alcohol / drug / tobacco use, diet, physical activity
- cdc.gov





Part of CDC Youth Risk Behavior surveillance system
to identify and address behaviors that pose risk to youth health
From CDC site:
Do students tell the truth on the YRBS?
“Research indicates data of this nature may be gathered as credibly from adolescents as from adults. Internal reliability checks help identify the small percentage of students who falsify their answers. To obtain truthful answers, students must perceive the survey as important and know procedures have been developed to protect their privacy and allow for anonymous participation.”
Source: https://www.cdc.gov/healthyyouth/data/yrbs/faq.htm
Beautiful media toolkit at https://www.cdc.gov/healthyyouth/data/yrbs/toolkit.htm

https://www.cdc.gov/healthyyouth/healthservices/pdf/OneonOnetime_FactSheet.pdf

Project Inquiries

- Can we use 10 y of aggregated data to predict youth risk behaviors on an individual basis? *(Supervised Multi-Label Classification Algorithms)*
  - *Q58: Have you ever had sex?*
  - *Q25: Have you felt sad or hopeless for 2 weeks in a row?*
  - *Q30: Have you ever tried cigarettes?*
- Which negative behaviors are most related? *(Graph Network)*

Potential Users

- Health Care Providers
  - 2018 data from Statista reports that the majority of pediatricians spend **13-16 minutes** with each patient
- Parents / Youth
  - Provide more targeted information for **relevant conversations**
  - (CO is one of 26 states without a sex ed mandate)
- Education Agencies
  - Implement outreach programs; **individual counseling**

Beautiful media toolkit at https://www.cdc.gov/healthyyouth/data/yrbs/toolkit.htm

https://www.cdc.gov/healthyyouth/healthservices/pdf/OneonOnetime_FactSheet.pdf

Other Important Considerations

- Insights over Answers
- Discussions over Discipline
- Reality over Determinism

Insights over answers: Certainly a high probability of a person engaging in an activity does not mean that the individual has indeed made that choice.

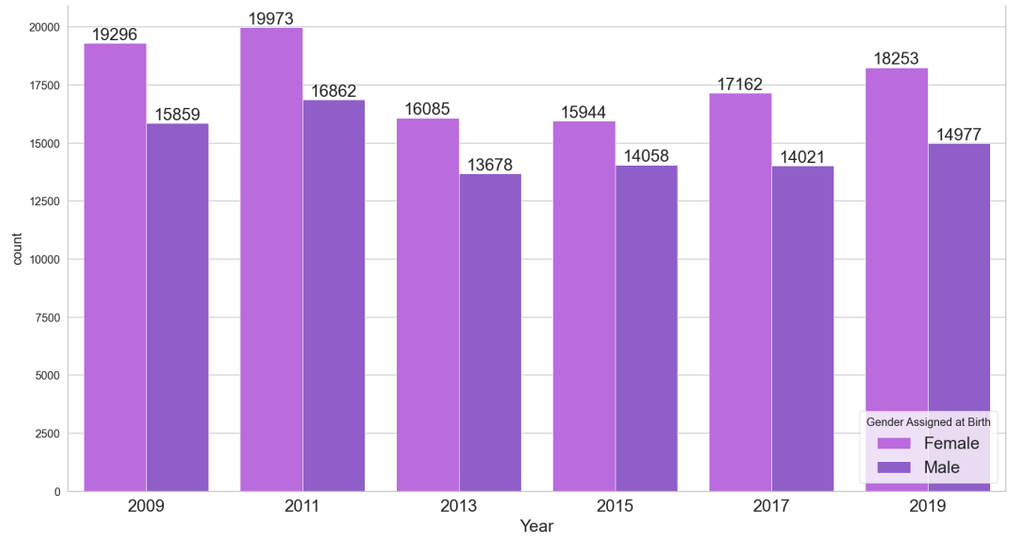
Discussions over discipline: The hope is that adult knowledge of the prevalence of these risk behaviors will lead to more appropriate, relevant conversations with young people about their own health, not punishment for assumptions made about their behavior.

Reality over Determinism: Even if the model were 100 percent accurate, actual choices are impossible to predict for individual people.

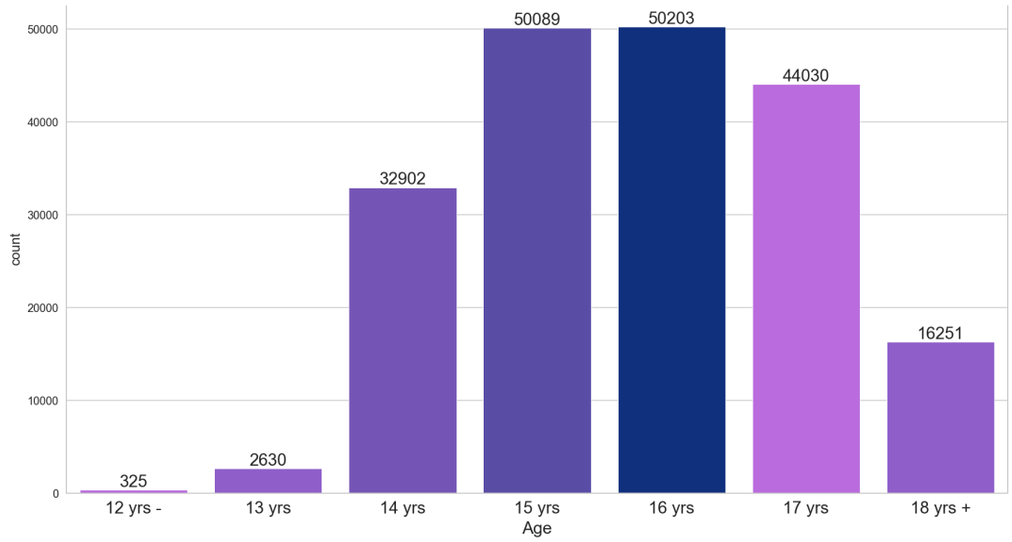
Part 2:

Exploring and Modeling the Data

Dataset: Number of Surveys per Year

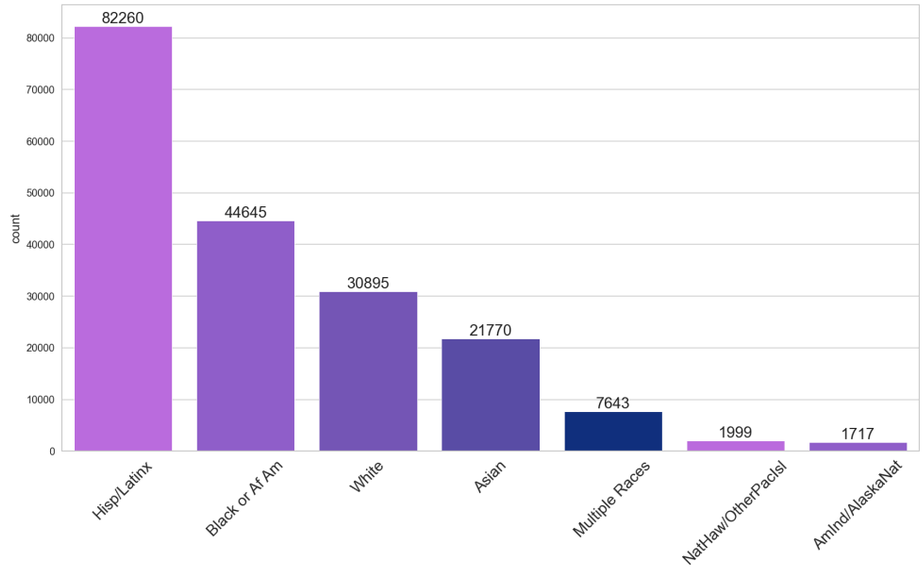


Dataset: Ages of Respondents

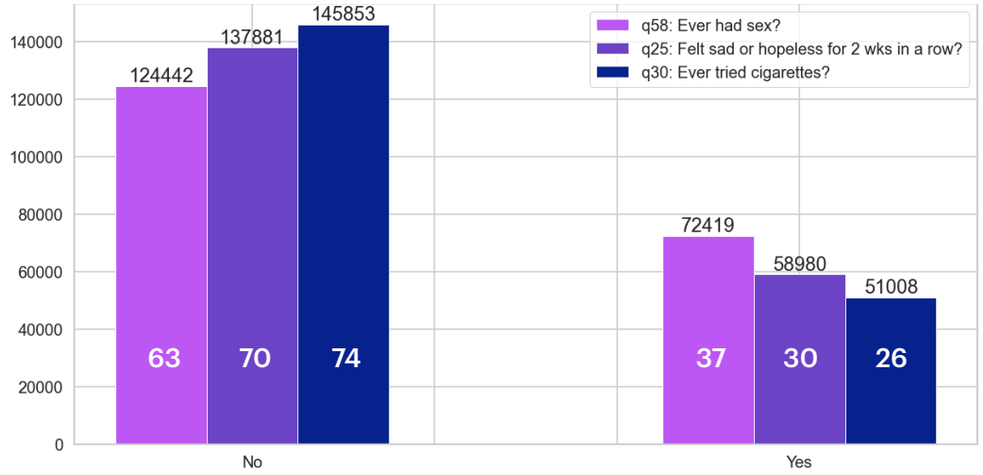


Represents students in grades 9 - 12

Dataset: Race / Ethnicity of Respondents



Distribution of Three Classification Targets



Baseline Accuracy: Percentages listed above.

Baseline Precision: 0%

This will be our baseline for the accuracy metric… we should be able to do better than guessing everyone said no (and get 63%, 70%, and 74% respectively)

Experimenting with Algorithms

scikit-multilearn

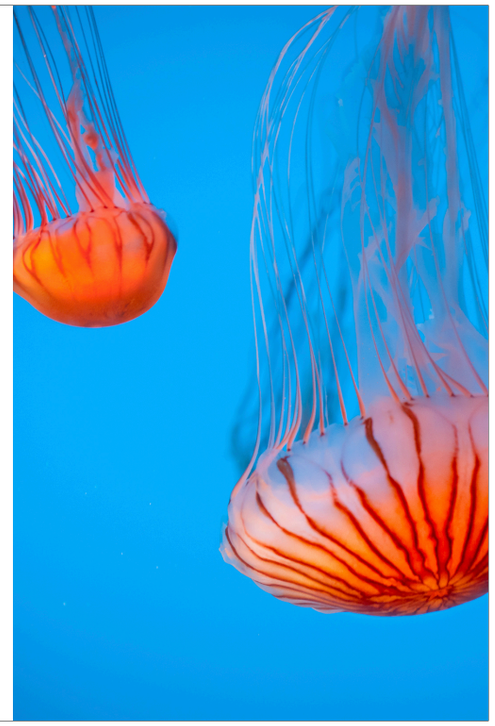
Trained and Compared different Multi-Label Classification Algorithms

- Logistic Regression
- Multi-label kNN
- Binary Relevance
- Classifier Chain
- Gaussian Naive Bayes
- Neural Network
- **Random Forest**

	Q58 (Sex)	Q25 (Sad)	Q30 (Cigarettes)
Precision (TP / TP + FP)	81.2%	83.9%	82.1%
Accuracy (TP + TN) / ALL	84.7%	83.2%	87.1%
Hamming Loss (TP + TN) / ALL	0.14 *(fraction of labels incorrectly predicted)*		

Part 3:

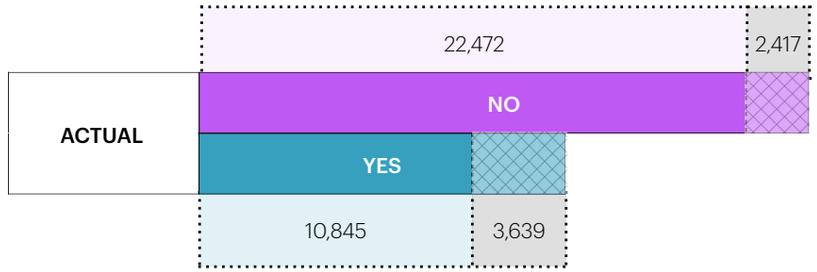
Predictions and Dashboard



How did we do?

Size of Test Set: 39,373

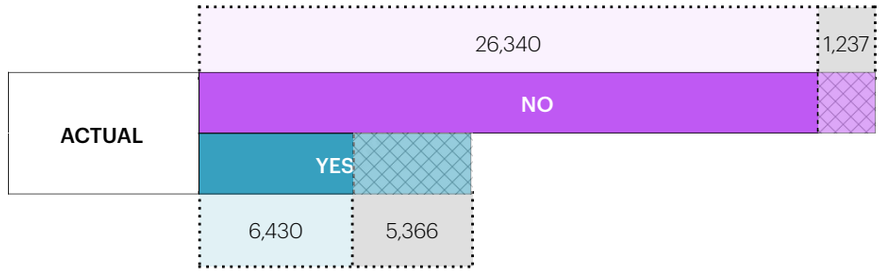
Q58: 'Have you ever had sexual intercourse?'
Baseline Precision = 0%
Model Precision = 81.2%



How did we do?

Size of Test Set: 39,373

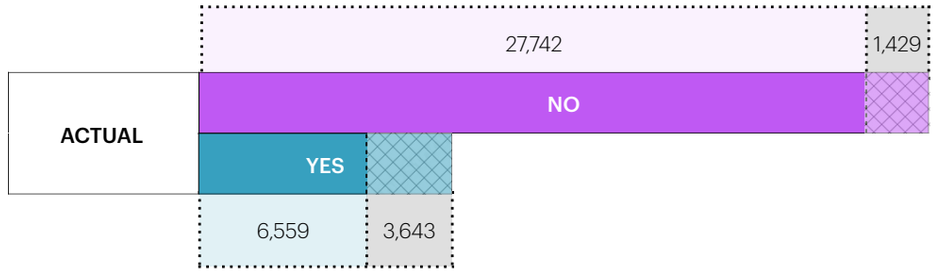
Q25: 'Felt sad or hopeless for 2 weeks in a row?'
Baseline Precision = 0%
Model Precision =83.9%



How did we do?

Size of Test Set: 39,373

Q30: 'Have you ever tried cigarettes?'
Baseline Precision = 0%
Model Precision =82.1%



Dashboard Demonstration (Local Development)

Predicting Youth Risk Behavior

Helping parents and professionals have relevant discussions with youth.

About the Youth Risk Behavior Survey

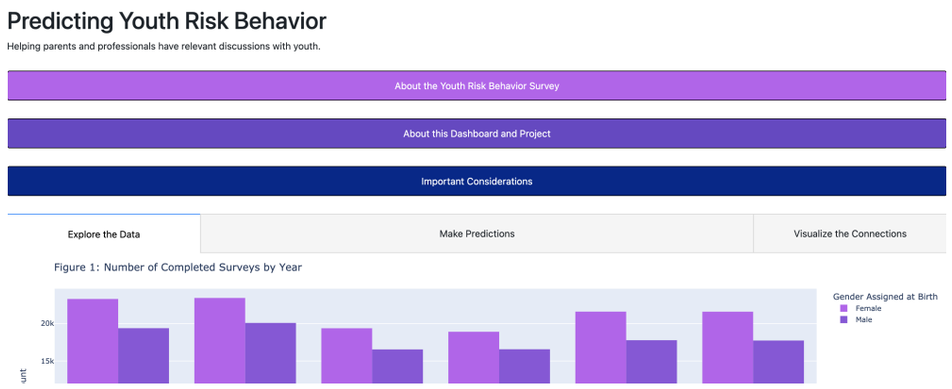
About this Dashboard and Project

Important Considerations

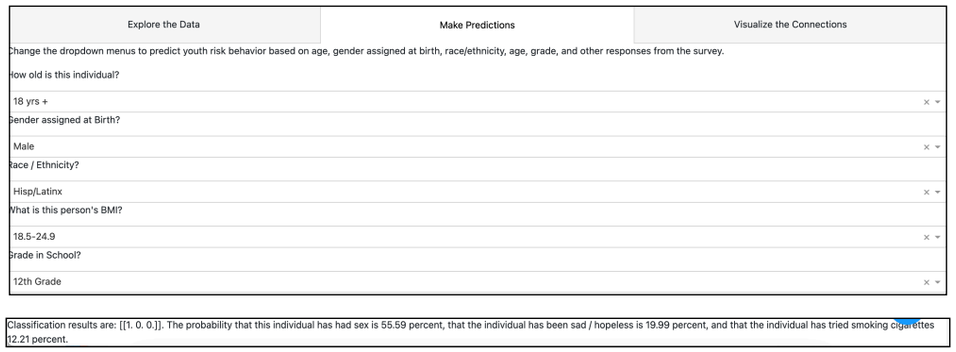
Explore the Data | Make Predictions | Visualize the Connections

The graph network shown below is based on a subsample of 10,000 surveys from the original dataset, maintaining the proportions of target classifications. Each node is a negative risk behavior and the directional edges between the nodes represent the proportion of survey respondents who answered affirmatively to any of those questions. For example, 31.5 percent of respondents who reported feeling sad or hopeless for more than 2 weeks in a row also reported getting less than 8 hours of sleep on an average school night. This graph may provide some further direction to dashboard users about potentially relevant discussions with youth. Numbers listed are percentages.

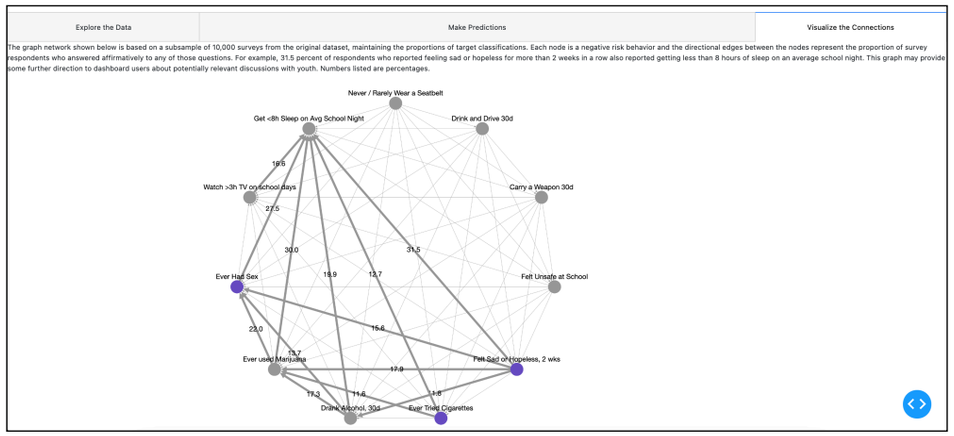
Dashboard Screenshot 1: Explore the Data



Dashboard Screenshot 2: Make Predictions



Dashboard Screenshot 3: Visualize the Connections



Part 4:

Future Work

Future Work

- Dashboard option to “Find a metro area like mine” (possibly match by size, population, political leanings, etc. to possibly refine predictions with geography)
- Consultation with experts in clinical settings for utility of predictions; dashboard creation for specific clinics based on their own survey data
- Tailor the inputs on the dashboard to different audiences

Thank you

Becky Peters, Data Scientist

becky.e.peters@gmail.com

linkedin.com/in/beckyepeters

GitHub.com/beckyepeters

twitter.com/beckyepeters