Substance Abuse Prediction Model

Data by Professor Jordan P. Davis

By Rupali Bahl, Megan D’Souza, Sarah Okamoto, Prathik Rao

2019

Agenda

Mission

Data

Regression

Survival

Future Work

Team Weed

Our Process

HOW WE CAME TO OUR IDEA

01
Current Issues

Vaping Crises & Deaths, Opioid Crisis - how can we study these

02
Data Available

Datasets we found for opioids had data about marijuana

03
Marijuana Abuse

We finally reached our topic because of the way the data lead us

The Problem

WHY WE CHOSE TO FOCUS ON SUBSTANCE ABUSE

- In 2005, ~¼ million emergency room visits in the US involved marijuana
- Recognizing which patients are more likely to relapse can help treatment centers reallocate resources to patients that need it
- Marijuana is the most popular illicit drug in the US (~24 million current users)
  - ~4 million of those people experienced significant issues related to their usage of the substance
- Marijuana often leads to use of and experimentation with other, harder drugs
- Most people who abuse marijuana do not seek treatment, but for **those who check into rehabilitation centers, about 60% will relapse**

Prof. Jordan P. Davis

- Research addressing substance use and the developmental needs of marginalized and vulnerable populations
- Intervention work on Mindfulness-Based Relapse Prevention

The Data Set:

20,000 Individuals checked into treatment centers

- 13,000 marijuana abusers
- 0-3 months: Treatment
- 3-12 months: Post-treatment
- 2 hr interview (once in 3 months): psych, mental, physical health
- Intervention
- Cognitive behavioral therapy

Measures demographics such as gender and age, trauma, and mental health conditions

The Goal:

- Help those who abuse marijuana and are seeking treatment
- Predict the likelihood of relapse for patients checked into rehabilitation centers
- Allocate resources and plan interventions according to these predictions

Key Features

IN-DEPTH PREDICTION

01

Trauma

Do people who’ve experienced trauma, stratified by type of trauma, have higher chances of relapse?

02

Ethnicity

Do people of different ethnicities have significantly different chances of relapse?

03

Gender

Do men and women have significantly different chances of relapse, and can we train our model based on this?

Related Work

What's already out there?

01

Individualized relapse prediction: Personality measures and striatal and insular activity during reward-processing robustly predict relapse

02

Use of a Machine Learning Framework to Predict Substance Use Disorder Treatment Success

03

Neural Activation Patterns of Methamphetamine-Dependent Subjects During Decision Making Predict Relapse

Preprocessing

01
Marijuana Days
Filtered data only to patients being treated for marijuana abuse

02
Trim Predictors
Only keep predictors relevant to marijuana abuse

03
Missing Data
Filled in missing data with mean/mode for features that had less than 25% missing values

Classification
(Logistic Regression)

	Predicted: NO	Predicted: YES
Actual: NO	2705	1008
Actual: YES	1345	1565

people who relapsed in the 1st 3 months

	Predicted: NO	Predicted: YES
Actual: NO	1698	1291
Actual: YES	1050	2584

people who didn't relapse in the 1st 3 months

	Predicted: NO	Predicted: YES
Actual: NO	3858	407
Actual: YES	1734	624

people who didn't relapse in the 1st 6 months

	Predicted: NO	Predicted: YES
Actual: NO	5633	25
Actual: YES	948	17

people who didn't relapse in 1 year

Linear Regression

Metrics for Each Model

R^2

Determination Coefficient

a difference of the total variance and the variance still not explained by your model

MAE

Median Absolute Error

the median difference between the approximated value and the true value

EV

Explained Variance

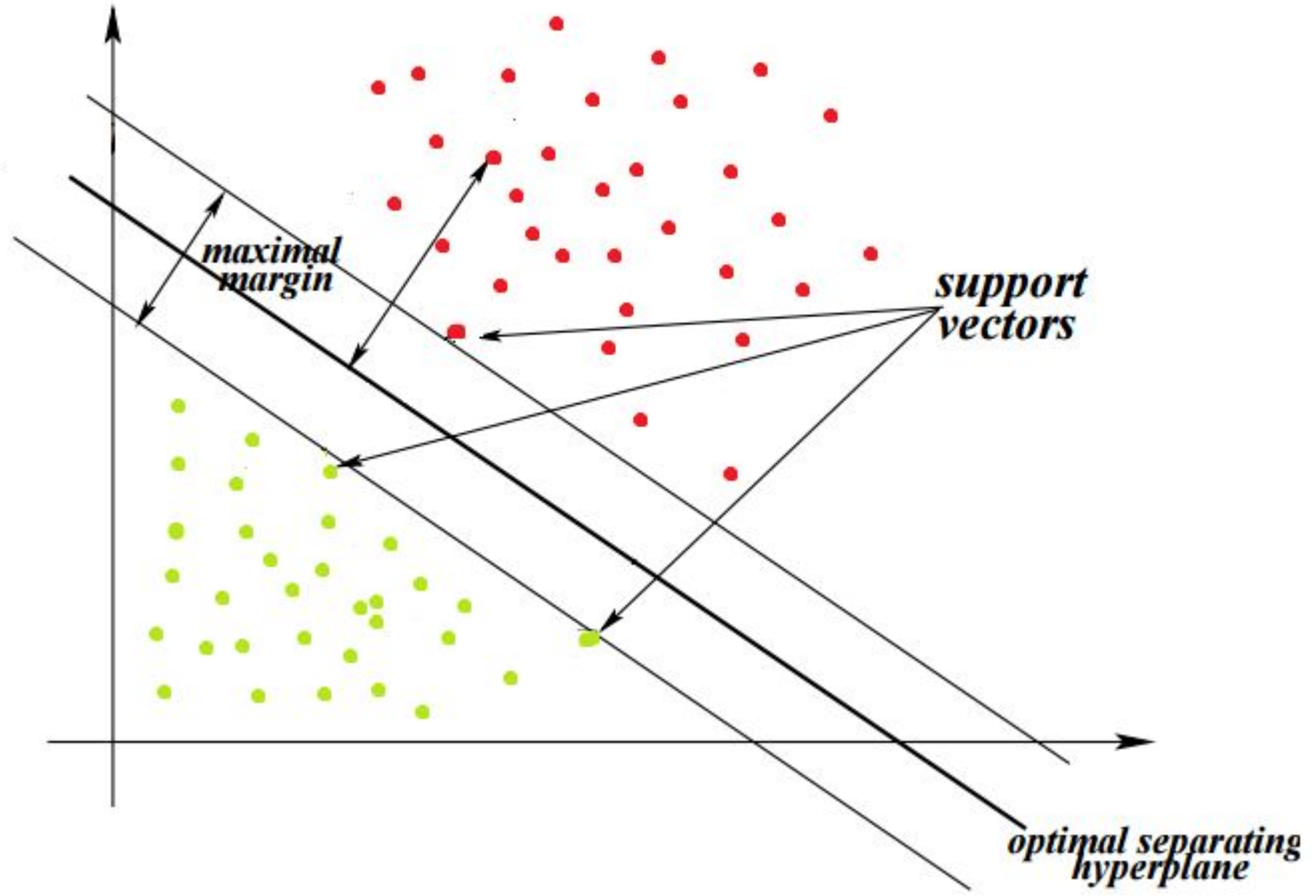
the total variance is explained by factors that are actually present and is not due to error variance.

Attempted Regression Models

Linear Regression	XGBoost	Lasso	Random Forest	SVM
R^2: 0.068365	**R^2:** 0.105686	**R^2:** 0.063992	**R^2:** -0.007847	**R^2:** -0.036086
EV: 0.069254	**EV:** 0.106588	**EV:** 0.064907	**EV:** -0.007806	**EV:** 0.057095
MAE: 76.696494	**MAE:** 75.418133	**MAE:** 77.961828	**MAE:** 77.25	**MAE:** 61.872242
~77 days	~76 days	~78 days	~78 days	~62 days

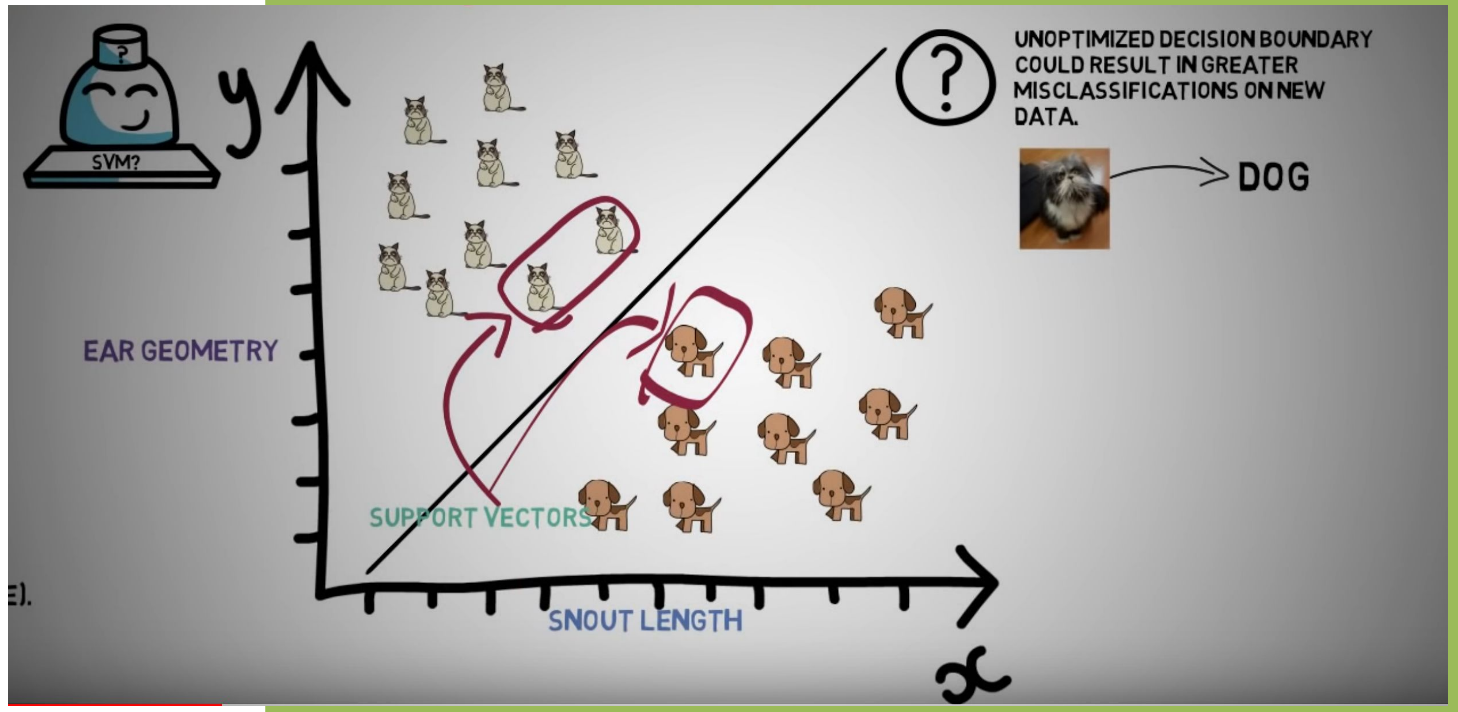
Support Vector Machine (SVM)

How does it work?



Linear SVM

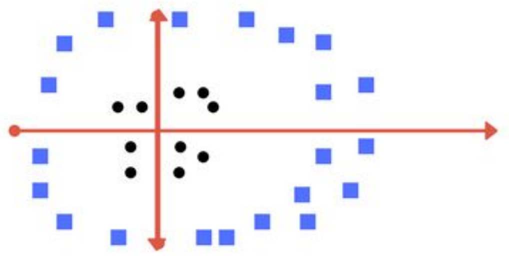
- A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane
- Given labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane that categorizes new examples
- Works by selecting the extreme points (all points that are close to the opposing class) and creating support vectors from them
- Get hyperplane (a line that divides the two classes) by finding the line between the support vectors

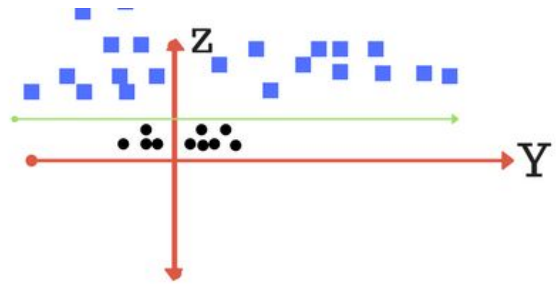


How does it work? (cont.)

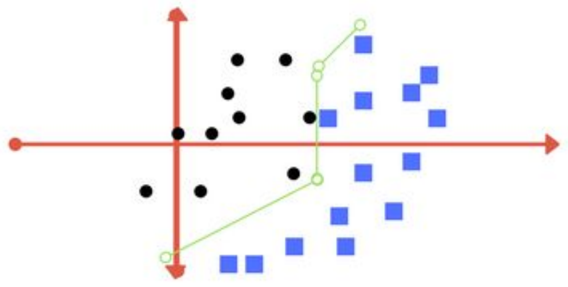
Examples

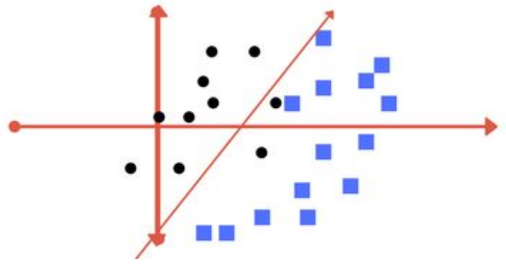
- In a more complex space (non-linear SVM), apply transformation (kernel) adding more dimensions to find a clear separation

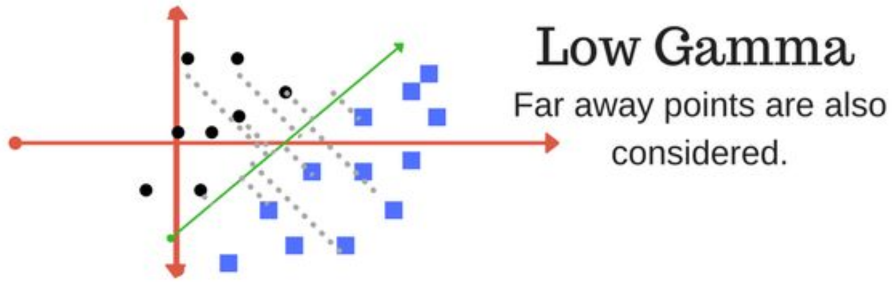


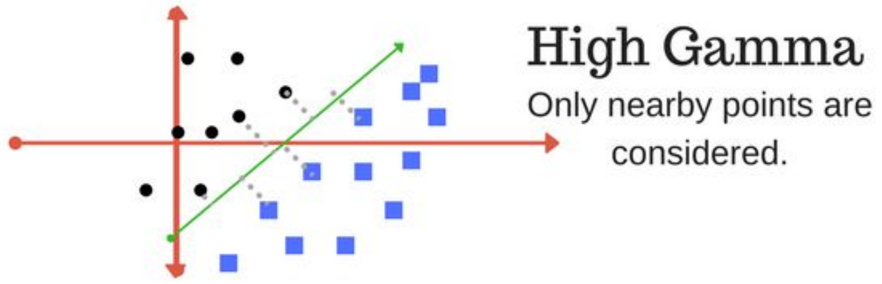


- May also run into overlapping data plots (regularization)









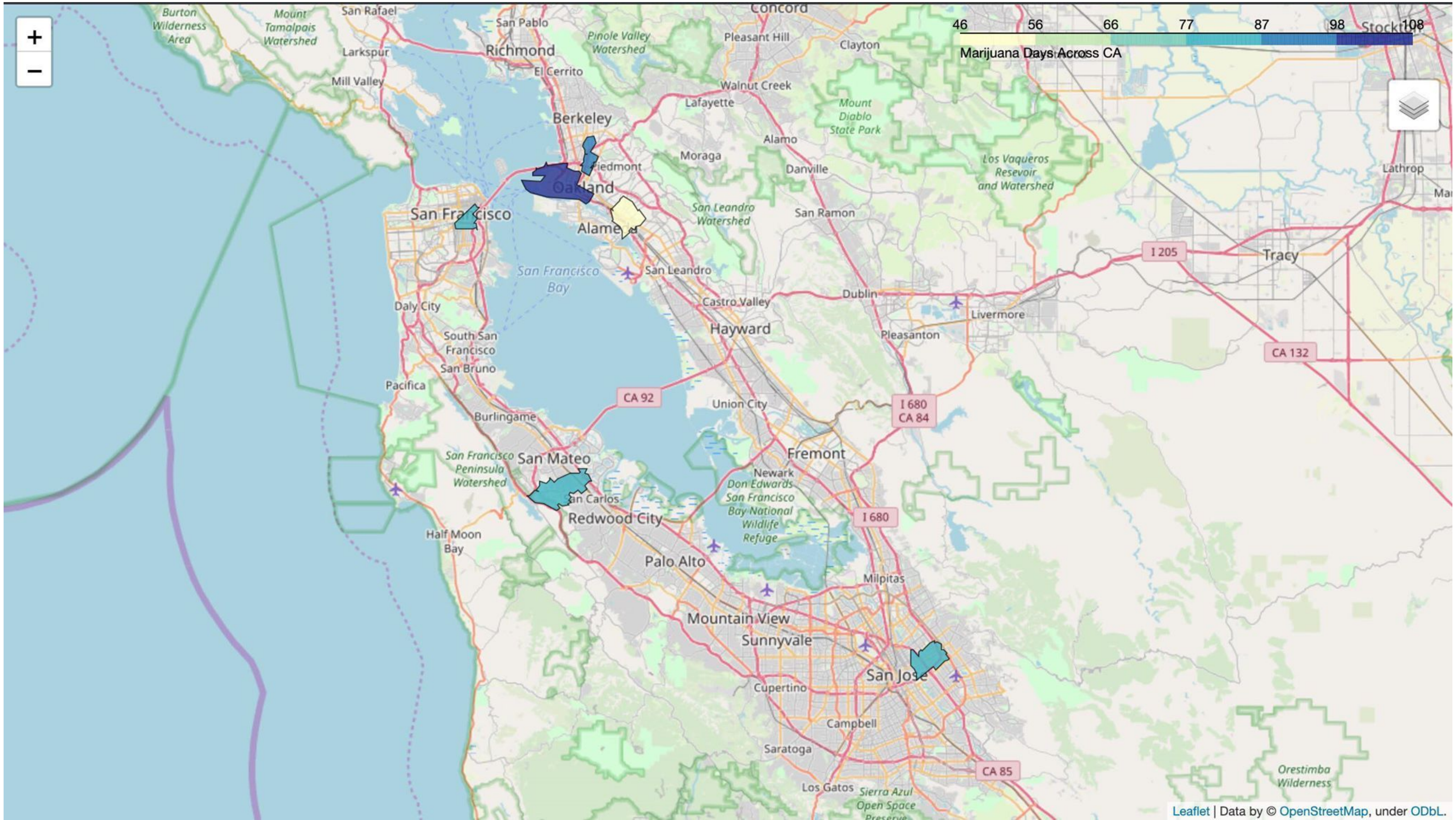
Feature Importance

FEATURES IN ORDER OF HIGHEST IMPACT ON MODEL

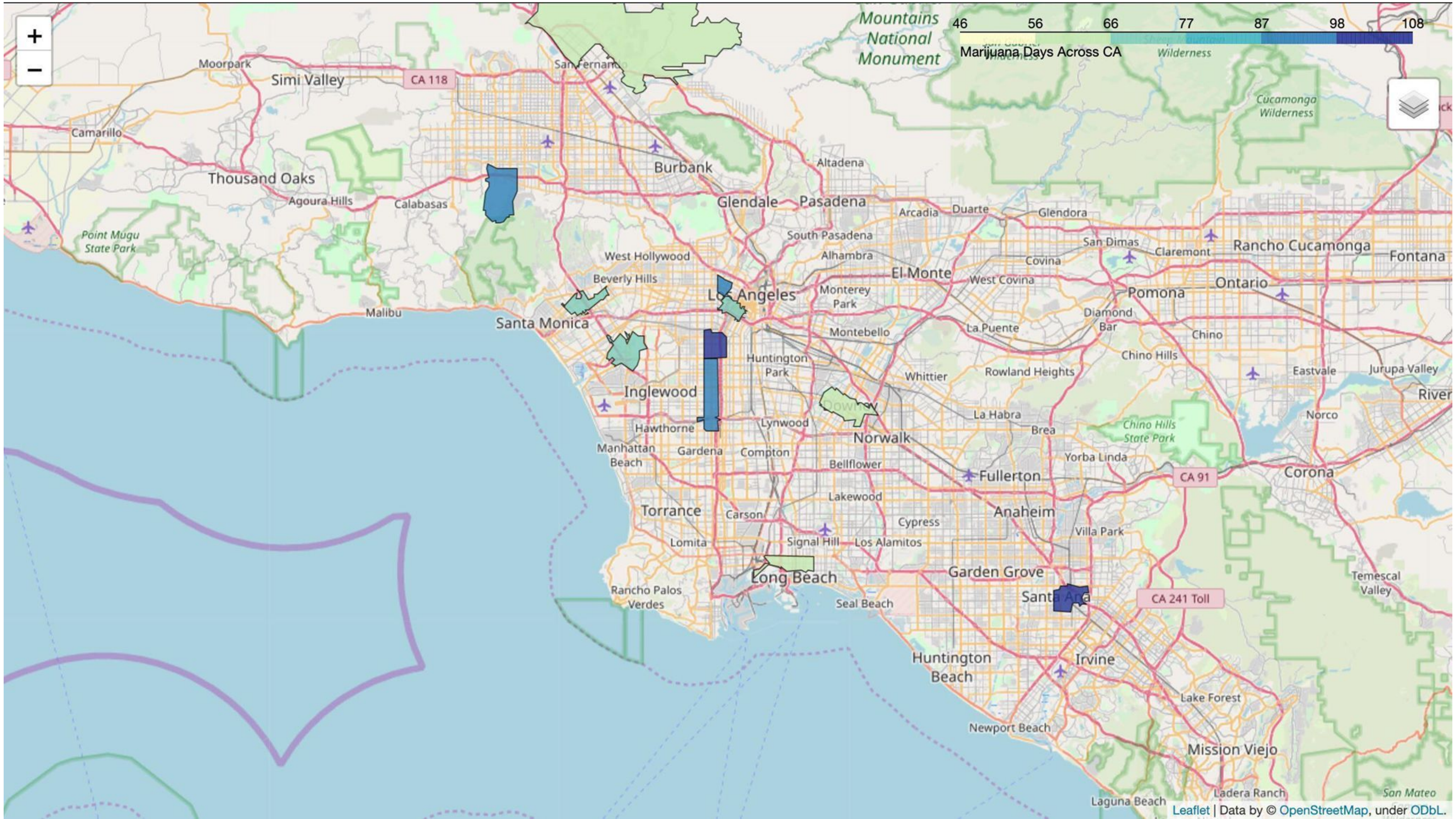
- **SPSm_0** - Substance Problem Scale (Past Month)
- **S2x_0** - P90: Days in controlled environment
- **dldiag** - Dual diagnosis
- **female -** whether or not the patient is female
- **HIVrisk** - HIV risk Scale across NPS, SxRS and GVS items
- **ncar** - Participant is not close to anyone in recovery [E5g, E6g, and E7g=4 or skipped]
- **tottxp4** - Total number of treatment planning needs endorsed-per LaVerne
    - **examples of treatment:** medicare, job placement, etc.
- **prsatx** - Any prior substance abuse treatment

	Coefficients	column_name
31	-13.159539	SPSm_0
30	11.817768	S2x_0
13	-11.300838	dldiag
0	11.123025	female
27	-10.883599	HIVrisk
24	9.775970	ncar
5	-8.836512	tottxp4
4	-8.273132	prsatx

Map of the Bay Area



Map of the Greater Los Angeles Area



Survival Analysis

Thanks Aaron!!!

Metrics

Concordance Index

Trauma

None: 0.581
Experienced: 0.575

Gender

Male: 0.591
Female: 0.577

Race

White: 0.591
Non-White: 0.586

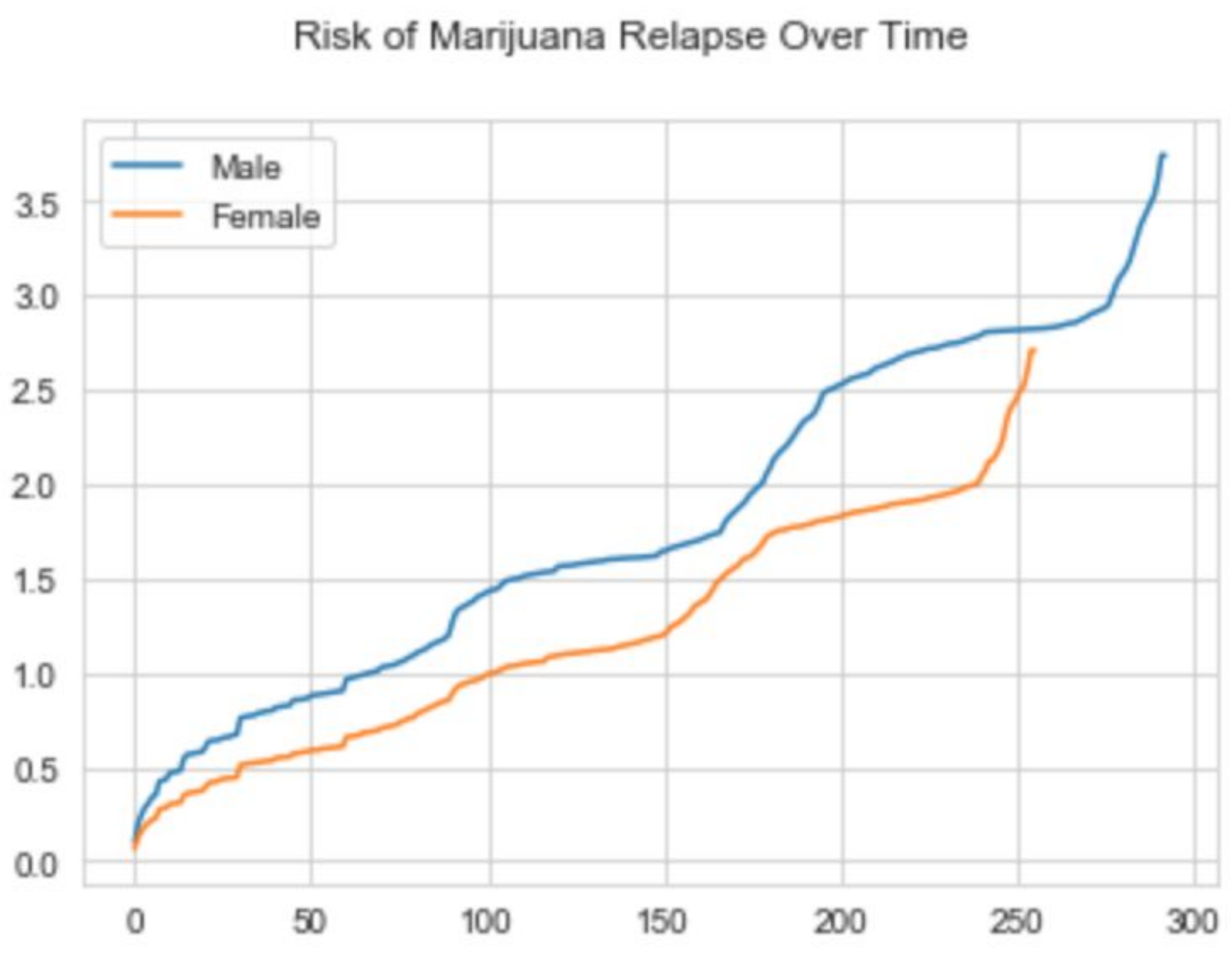
HUMANITIES GOLD STANDARD: 0.2-0.4

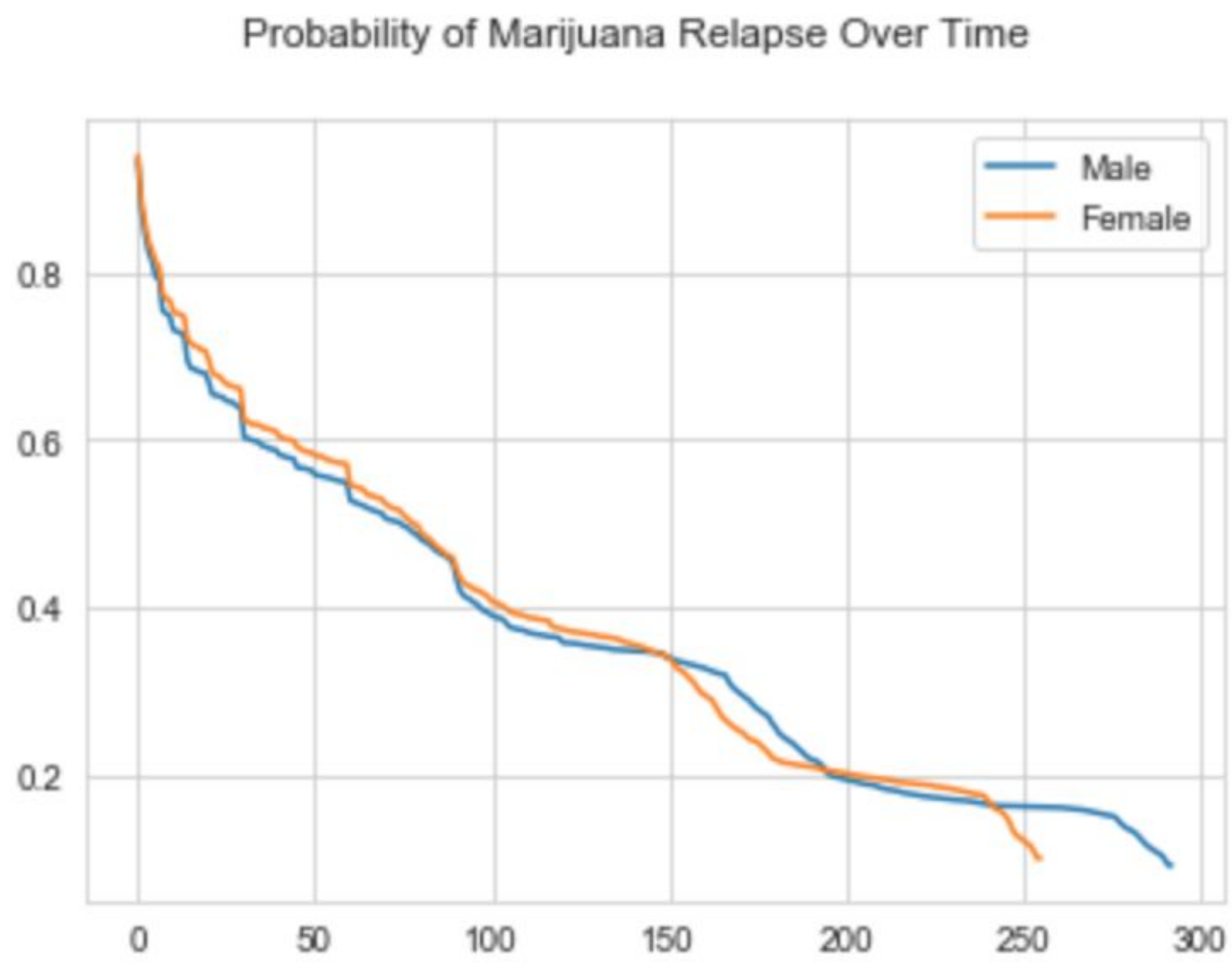
Web App - Interpretability

- Interface makes our models more readable and interpretable for people, like Jordan, who work in social work
    - Also important for people working in the rehabilitation centers

[Web App Demonstration]

Gender Plots

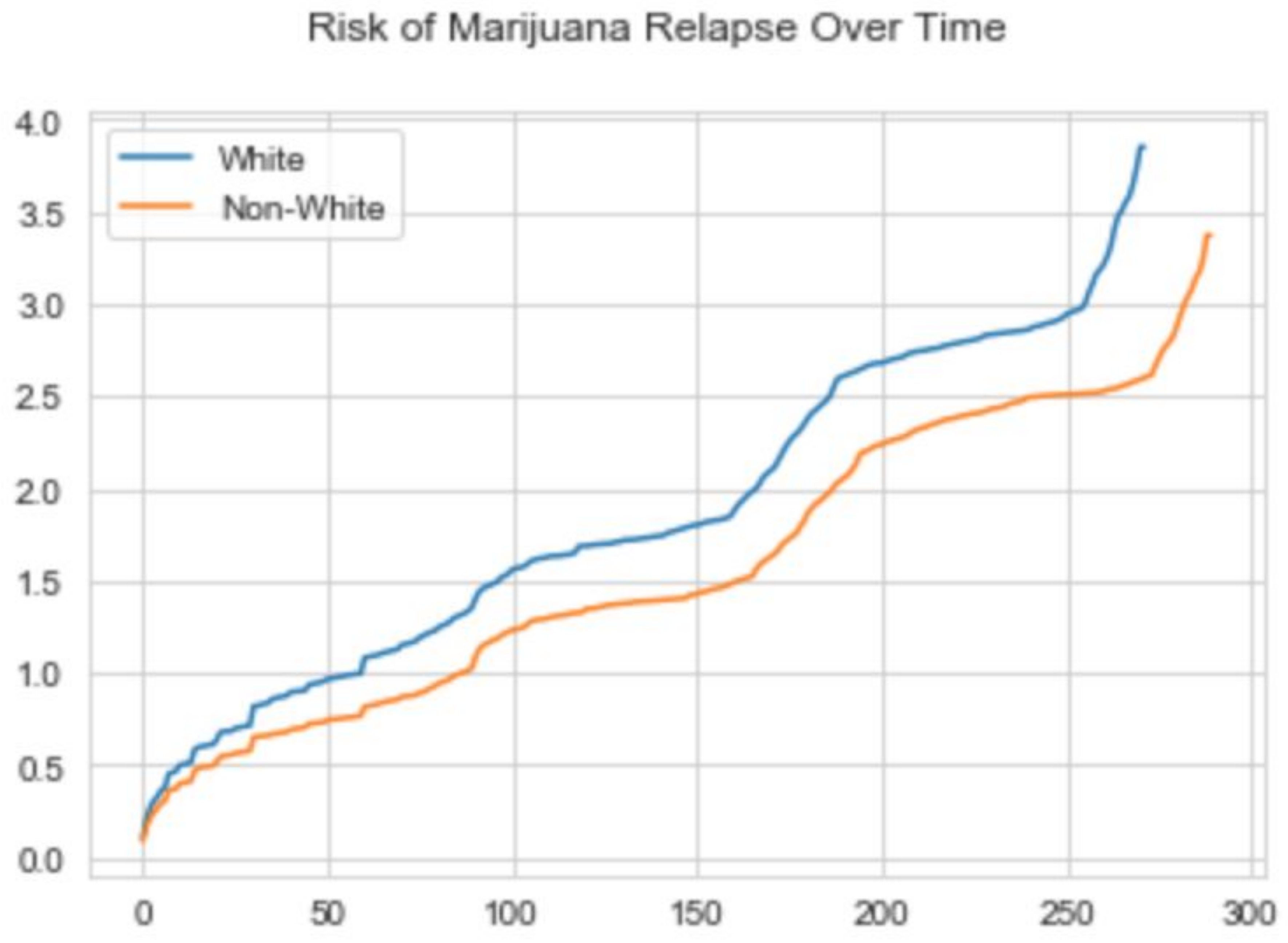




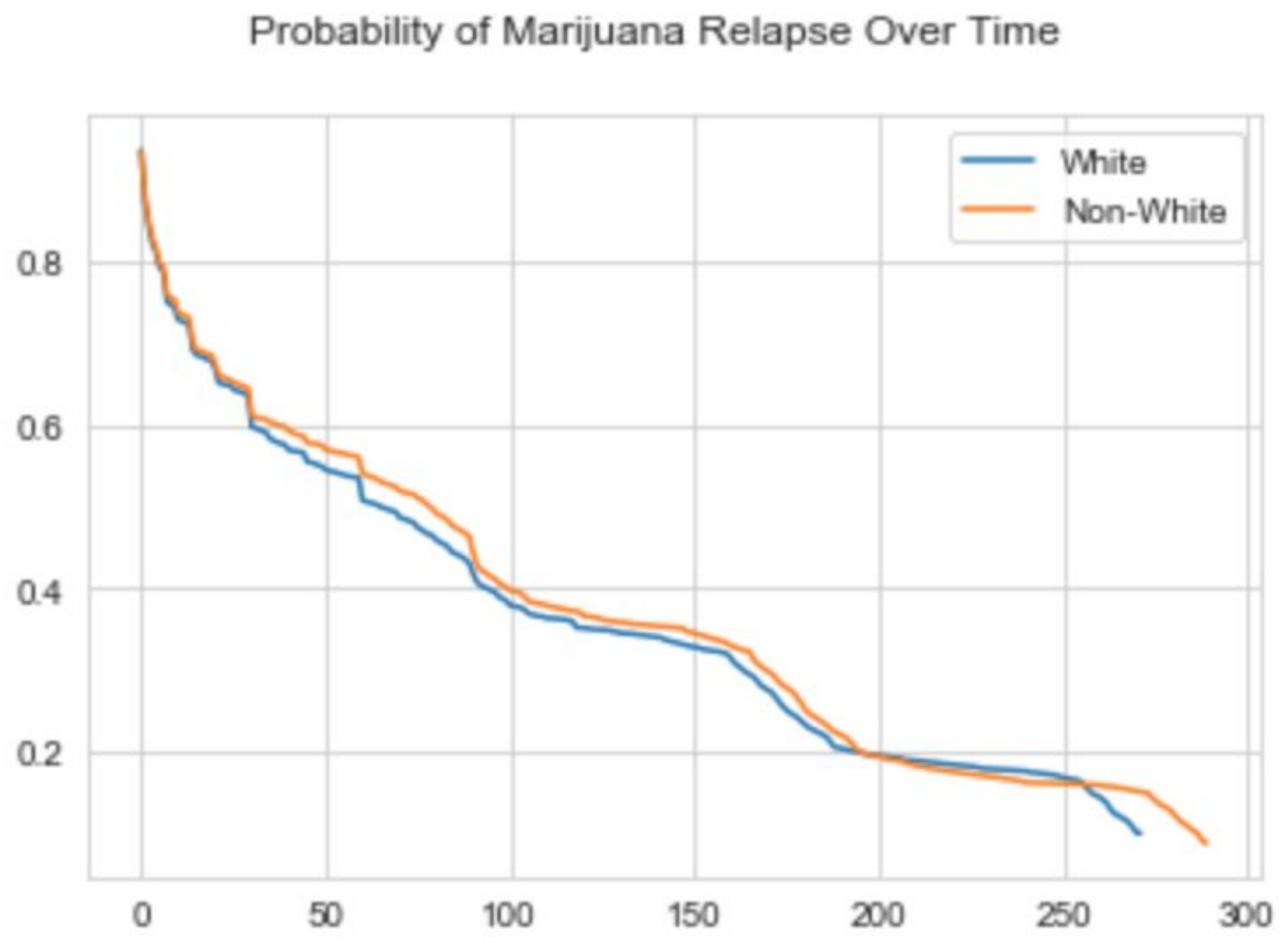
Gender Feature Importance

	Coefficient	Males	Females
0	nonwhite	-0.001659	0.077357
1	unemplmt	-0.065102	-0.143643
2	B2a_0	-0.019487	-0.011622
3	prsatx	0.098602	-0.031836
4	tottxp4	0.012020	0.058482
5	TRI_0	0.038297	0.055626
6	GVS	-0.025235	-0.019732
7	tsd_0	-0.112690	-0.071935
8	und15	0.046469	0.097521
9	CWS_0	-0.014230	-0.000077
10	IPI	0.001308	0.021452
11	S9y10	-0.009030	0.006118
12	dldiag	0.018592	0.014563
13	DSS9_0	-0.008402	-0.017082
14	ADHDs_0	0.030735	-0.036960
15	CDS_0	0.002761	0.025465
16	suicprbs_0	-0.031600	-0.096766
17	CJSI_0	-0.004864	-0.018974
18	LRI7_0	0.011775	-0.035244
19	SRI7_0	-0.019816	0.024310
20	ERS21_0	0.080240	0.077044
21	homeless_0	-0.015414	0.065972
22	S6	-0.003445	0.022992
23	ncar	-0.027114	-0.073205
24	engage30	0.033549	-0.050416
25	init	-0.041395	0.029993
26	HIVrisk	0.102601	0.067406
27	totttld	-0.059037	-0.115038
28	POS_0	-0.013149	-0.029020
29	S2x_0	-0.028015	-0.042544
30	SPSm_0	0.102842	0.154341
31	EPS7p_0	0.012088	0.033121

Race Plots



Risk of Marijuana Relapse Over Time
White
Non-White
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
250
300

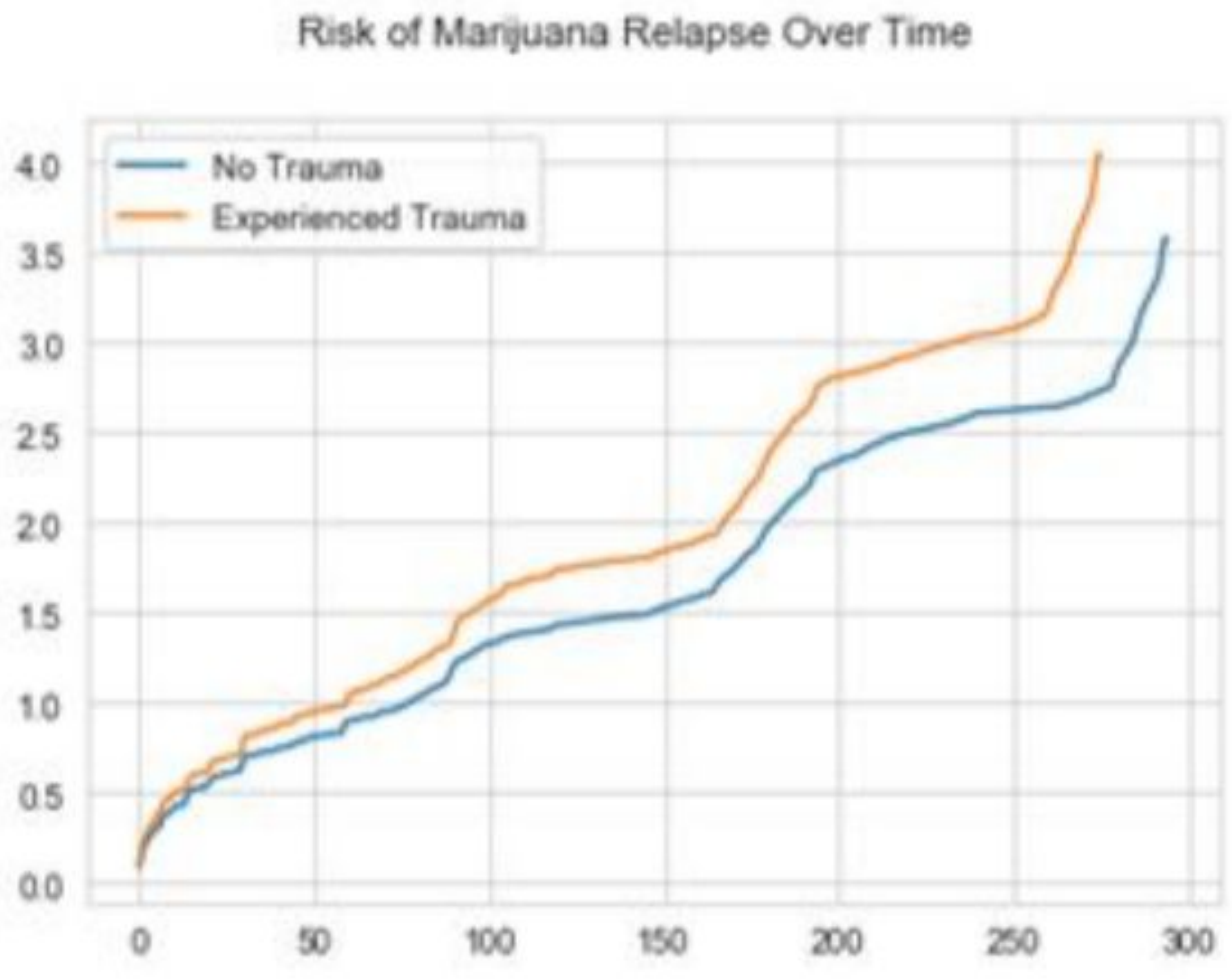


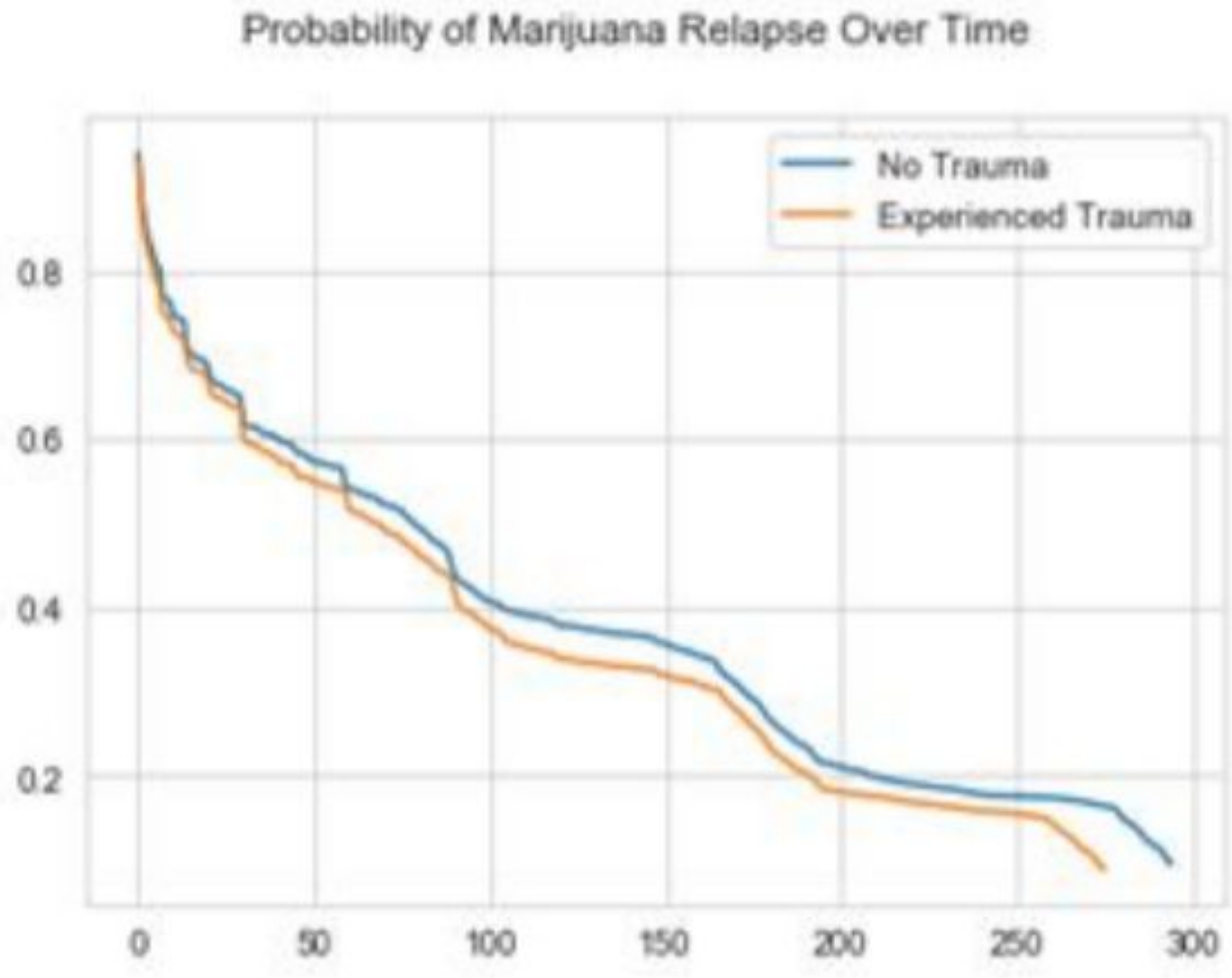
Probability of Marijuana Relapse Over Time
White
Non-White
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300

Race Feature Importance

	Coefficient	White	Non-White
0	female	-0.072901	-0.015640
1	unemplmt	0.000952	-0.066822
2	B2a_0	-0.018040	-0.020140
3	prsatx	0.105581	0.045368
4	tottxp4	-0.015695	0.020157
5	TRI_0	0.068928	0.008180
6	GVS	-0.029742	-0.039498
7	tsd_0	-0.085786	-0.041287
8	und15	0.054036	0.039762
9	CWS_0	0.002176	-0.028679
10	IPI	-0.002156	0.007601
11	S9y10	-0.016232	-0.001418
12	dldiag	0.077237	0.034820
13	DSS9_0	-0.011906	-0.011531
14	ADHDs_0	-0.026304	0.043348
15	CDS_0	0.052800	-0.007989
16	suicprbs_0	0.001791	-0.106738
17	CJSI_0	-0.030512	0.004886
18	LRI7_0	-0.065563	0.022558
19	SRI7_0	-0.013944	-0.005865
20	ERS21_0	0.134650	0.066843
21	homeless_0	-0.022125	0.021429
22	S6	0.030899	0.008466
23	ncar	-0.059395	0.011944
24	engage30	-0.013672	0.049003
25	init	-0.083359	0.033814
26	HIVrisk	0.131783	0.133057
27	totttld	-0.050170	-0.091527
28	POS_0	-0.022152	-0.015301
29	S2x_0	-0.086043	-0.009803
30	SPSm_0	0.059989	0.150322
31	EPS7p_0	0.007770	0.024547

Trauma Plots





Future Work

Augment Dataset

Do environmental factors (such as socioeconomic status) impact the time it takes for someone to relapse?

Approach:

- Augment dataset with other factors from social explorer for each census tract
  - Poverty status, public assistance, unemployment status, age (less than 18), etc.
- Use the address of the rehabilitation center a patient was checked into
  - Add columns for the latitude and longitude of each center
  - Using the latitude and longitude, find the FIPS code (correlates to census tract in the social explorer csv file)
  - Append FIPS code column to original dataset
  - Join the two datasets on the FIPS column

Future Work Cont.

- Fairness
  - False Positive - predict someone will relapse when they do not
  - False Negative - predict someone will not relapse but they do
    - Most important to minimize
  - True Negative - predict someone will not relapse and they do not
  - True Positive - predict someone will relapse and they do

- Prioritize high sensitivity
  - Few false negatives
- High Specificity
  - Few false positives

Future Work Cont.

Extensions

- Look into predicting relapse for other substances
  - i.e. opioids
- Optimization of rehabilitation centers
  - Given the number of clinics in a given area, measure the impact of adding a new facility to that region
- Anything with success/fail outcome

Thanks Prof!!

Any Questions?

Thanks for listening

- Problem (background on marijuana relapse) - MEGAN
    - Original Idea - interested in vaping or opioids → data led us to marijuana
- jordan <3 - Rupali
    - his background
    - Data set and possibly background about Jordan’s work
- High level outline of goals -Rupali
    - Help those who abuse marijuana and are seeking treatment
        - Predicting relapse will be helpful in prevention -- allocate more resources
        - Possibly predict/optimize effective placement for rehabilitation centers
- Related Work (what is already being done) - Rupali
- Our planned approach - Rupali
    - pre-processing
- Attempted Regression Models - Sarah
    - Linear Regression
        - r^2, median, explained variance (?)
    - XGBoost
    - Random Forests
    - SVM
        - all of the features and their coefficients (feature importance)
        - Talk about maps thing
- Classification (lasso and logistic regression) ***if we get it working*** - Rupali/Megan
- Survival Model
    - censoring
    - nx2 vector - the first one predicts if the event actually occurs
    - if a user does not relapse, the model might predict that they will relapse in 500 days, but the data stops at 365, our model would think we are inaccurate, but censoring combats this issue
    - Present survival plots & hazards of demographics we chose, and feature importance, and concordance index
- Web app - interpretability
- Augmented Dataset ***if we get it working*** - Sarah
    - Do socio economic / environmental factors affect a person’s relapse time?
- Future Work - Megan
    - Extensions