SUBSTANCE ABUSE MODEL

Rupali Bahl
USC
rupaliba@usc.edu

Megan D'Souza
USC
mdsouza@usc.edu

Sarah Okamoto
USC
saokamot@usc.edu

Prathik Rao
USC
prathikr@usc.edu

Abstract

In 2005, 242,200 emergency room visits in the United States involved marijuana [5]. With the increase in Marijuana abuse over the last few years, there has also been an increase in resources and government funds allocated in treatments towards substance abuse drug disorders. It is important that these resources are used efficiently to optimize the number of patients treated. One of the biggest problems faced by treatment centers and drug abuse patients is relapsing during or after the treatment. Recognizing which patients are more likely to relapse and at what point in their treatment are they most likely to relapse can help treatment centers reallocate resources to create high impact. Additionally, using the zip codes in the data set, we explore geographical areas where marijuana abuse and relapse is the highest to decide which areas require more rehabilitation centers and other SUD treatment resources. Roughly 13,000 patients that checked into a treatment center for marijuana abuse were studied and 21 feature columns were significant in our predictive model. The dataset tracks their progress through three checkpoints (after 3 months, 6 months, and finally, one year) throughout their 12 month treatment process.

1 Introduction

With the legalization marijuana in some states, it is of utmost importance to ensure that people understand the risks of the substance. According to the National Institute on Drug Abuse, adolescents' perception of the risks of marijuana have declined over the past decade [4]. Marjuana is the most popular illicit drug in the United States with about 24 million current users, and nearly 4 million of those people were either addicted or experienced significant issues related to their usage of the substance [6]. Marijuana often leads to use of and experimentation with other, harder drugs and teenagers who use marijuana are at a significantly greater risk for developing an addiction to something more dangerous [3]. Most people who abuse marjiuana do not seek treatment, but even for those who check into rehabilitation centers, about 60% will relapse [9]. As marijuana abuse continues to rise among adolescents in the United States, and the rates of relapsing remains high, it is necessary to take preventative measures and ensure that rehabilitation resources are effective and available to all communities.

1.1 Problem Statement

One way to help people who abuse marijuana is to look at those that are seeking treatment and predict when a person might relapse. A relapse is defined as the moment a person uses marijuana again for the first time after receiving treatment. Our goal is to track the amount of time to relapse for different kinds of patients and use this data to improve their treatment in order to make it more successful. When people are close to their predicted relapse day, treatment centers can allocate more resources to help prevent the possibility of relapse. This kind of data can be used to increase the effectiveness of treatment centers so that better care and resources can be provided to patients, especially at times when they may be at a higher risk of relapsing. If rehabilitation centers are more effective, more users may seek help, which is incredibly important as the majority of people who abuse marijuana do not seek any form of treatment. Additionally, increasing the number of drug rehabilitation centers would also contribute to reducing the abuse of marijuana. Using the estimated time before a person relapses in a given area could be indicative of where building a rehabilitation center would be most effective. This paper focuses on developing a predictive model that determines the number of days before a person will relapse after checking into the center and explores the average relapse time for specific zip codes to look into optimizing center placement in the future.

1.2 Goals and Approaches

First, in this project, we present a predictive, linear regression model that determines the number of days before a person is expected to relapse within a twelve month time frame following their initial interview at the center. This model uses individual data such as gender, geographic location, number of days before relapse, age, mental health disorders, etc. from previous patients that have checked into rehabilitation centers and then predicts the number of days before relapse for any patient, given their data from the initial interview. This can be used to help rehabilitation centers provide the most effective treatment and allocate an increased number of resources for patients when they are at a higher risk for relapse. Secondly, we plan to augment our dataset and find census tract data on the zip code level to determine which areas may have significantly shorter average relapse times (higher marjiuana usage). Using this geospatial data, we can explore where adding new treatment centers may have the greatest impact as an optimization problem. The predictive model in conjunction with the optimization problem could be used to select the best locations to place a new rehabilitation center. This would ensure that more resources are available to areas that have the fewest days until relapse.

---

2 Background & Related Work

2.1 Background

Drug abuse is a major problem in the United States. Our team's original idea was to create a study on opioid abuse and treatment, and hopefully find a way to help mitigate the opioid crisis. When researching data, we discovered a very extensive dataset on marijuana treatment and relapse, from Professor Davis at USC, and decided to pursue work on marijuana abuse. For some background on marijuana abuse, marijuana is the most popular illicit drug in the country, with over 24 million users. As college students, we see marijuana usage on a regular basis among peers and were interested in learning more about the usage and abuse of this drug. In 2005, approximately a quarter of one million emergency room visits were due to marijuana-related causes. Marijuana usage often leads to use of and experimentation with other drugs. Most people who abuse marijuana do not seek treatment for their abuse, but out of those who do check into rehabilitation centers, about 60% of them will relapse. By recognizing which patients are more likely to relapse, we can help treatment centers reallocate resources to patients that need it, and better help people overcome their abuse.

2.2 Related Work

One project that we found similar to our own model is a study on methamphetamine relapse. This study, titled "Individualized relapse prediction: Personality measures and striatal and insular activity during reward-processing robustly predict relapse," was conducted and developed by researchers Joshua L. Goodwin, Tali M. Ball, Marc Wittmann, Susan F. Tapert, and Martin P. Paulus. They used the random forest classification technique to generate individual predictions for relapse, using fMRI scans, to suggest that neuroimaging can be used to predict relapse. Our model differs from this because we are using linear regression. We tried to build a random forest model, but it was not as accurate as the linear regression. This model also tracks the binary outcome of whether an individual relapses, whereas ours predicts the number of days until the next relapse [1].

Another study we found, "Use of a Machine Learning Framework to Predict Substance Use Disorder Treatment Success" uses a Super Learning methodology to optimize patients' pathways to the best possible treatment outcome. It is similar to our project as we both look at the patients who have signed up for Substance Use Disorder treatments and try to analyze how what kind of factors lead to successful treatments versus relapses. Our study predicts when are certain patients likely to relapse given a SUD treatment for marijuana using Linear Regression, and the other study evaluates what method works best at predicting successful treatment using Super Learning [2].

The final paper we examined that is similar to ours is "Neural Activation Patterns of Methamphetamine-Dependent Subjects During Decision Making Predict Relapse," and was written in July 2005 by researchers Martin P. Paulus, MD, Susan F. Tapert, PhD, and Marc A. Schuckit, MD from USCD. Their study also examined fMRI neuroimaging to predict likelihood of relapse, and their model explored how different parts of the brain were connected to relapse probability. Our study focuses on socioeconomic and life event factors, rather than medical data, and focuses on data past 2005 which could make a significant difference on the data [7].

2.3 Improving Upon Past Work

Our model differs from past work because we are studying marijuana abuse, and past projects have focused on methamphetamine or broadly substance abuse. These projects do track relapse, but they do not focus on treatment centers. They track relapse from other data sources, such as surveys and census. Instead of creating models based on demographics and traits such as trauma, they focus on either treatment methods administered or on fMRI imaging.

3 Data

3.1 Data Sources

The data that our model uses is from Professor Jordan P. Davis's work at the University of Southern California. It maps the recovery and relapse of individuals who checked into a treatment center for aid regarding addiction to various substances, primarily marijuana. In addition to tracking their recovery, the data also records certain factors such as age, gender, address of rehabilitation center, trauma disorders, impulsive personality index, and other relevant features that could affect an individual's potential to relapse. Below is the outline for steps taken to load and preprocess the data, and a summary of how we trained our models.

3.2 Pre-Processing

In order to pre-process the data, we took the following steps: first, we pushed the database provided by Professor David onto Github and linked that url in our code in order to read into pandas dataframe object. Second, we needed to calculate the percentages of null values in each column, and drop columns with >25%. Then we needed to identify the columns that we did not need, and drop the columns irrelevant to the study provided by Professor Davis (other predictor columns and labelling data including state, zip code, etc.) Finally, we needed to fill columns with numerical data filled with NaNs with the mean of the column, for categorical data with the mode of the column.

To train our models, we created a train-test split using sklearn's train_test_split to train the model on one part of the data and test on the other. All of the models were trained using this approach.

4 Predictive Model

The machine learning methods we used to develop the marijuana relapse predictive models included linear regression, linear support vector machine (SVM), random forest, XGBoost, logistic regression, and Cox Proportional-Hazards survival analysis.

4.1 Approaches

4.1.1 Metrics Used to Evaluate the Models

For classification, we used confusion matrices to compare the predicted number of days until relapse with the actual number of days before a patient relapsed.

For regression, we used R^2 , median absolute error, and explained variance to evaluate our models. For survival analysis, we used concordance index to evaluate the predictions made by our model. To evaluate model performance, we chose to primarily consider median absolute error. This metric tells us that our predictions are accurate within this margin of error, which will be helpful to those who might be using our model to estimate when their patients might relapse.

Concordance index is used to validate the predictive ability of a survival model. It is defined by the proportion of pairs divided by the total number of possible evaluation pairs. A concordance index of greater than 0.5 indicates good prediction ability and because the concordance index for each of our subgroups within trauma, gender, and race is approximately 0.6, our model has good predictive ability.

4.1.2 Classification

Classification constructs a model to classify data and predict categorical class labels. For this project, we specifically used Logistic Regression, a supervised classification model, to predict whether or not a given patient is likely to relapse in a given time frame. Logistic Regression is similar to Linear Regression in its underlying technique, but uses a more complex cost function, defined as the Sigmoid Function, instead of the linear function. This function maps predicted values to probabilities between 0 and 1, and the classifier gives us a set of outputs based on this probability and a chosen threshold, defined as the decision boundary.

In our training dataset, we divided the dataset into four sections. The first section only looked at the people that relapsed within the first 90 days, which is the treatment period. These patients were then removed from the dataset for the next analysis that only looked at patients in the post treatment phase (3-12 months). We ran the model on all the patients that did not relapse in the first 3 months. Then the patients who relapsed in the next 3 months were removed, and the model was run on all the patients who did not relapse in the first 6 months. Finally, the patients who relapsed in the last 3 months were removed, and the model was run on all patients who did not relapse in the observed 1 year. To analyze this data, we output confusion matrices to analyze the false positives, false negatives, true positives, and true negatives in each case.

4.1.3 Regression

All of the regression models contain a decision function which allows us to easily find the feature importance for each of our models. In this paper, we compare linear regression, random forests, support vector machine, and XGBoost using median absolute error, R^2 and explained variance to compare the performance of each model. For interpretability, we created a map that colors areas in Southern California based on the average number of days until relapse for each of the regression models.

The features in our regression models were either given a positive or negative coefficient; this number reflects the impact of that particular feature on the predictions. A positive coefficient indicates that the feature prolonged the number of days until relapse whereas a negative coefficient indicates that the feature corresponded to a quicker relapse.

4.1.4 Survival Analysis

Survival analysis works to establish a connection between covariates and the time of an event. This approach is appropriate for our predictive task as survival analysis measures the time in a sample until a specific event occurs, knowing that not everyone or everything in the sample will experience the event. In our model, we predict the number of days until someone that abused marijuana will relapse in our sample of patients, but not every patient will relapse during their time in rehabilitation.

Furthermore, survival analysis takes into consideration training data that can only be partially observed, or censored. In our dataset, after rehabilitation is complete, there is no way to keep track of whether patients relapse later in life. In our model, we used survival analysis to predict the hazard function for the approximate probability that a person does not relapse between 0 and 365 days (the duration of treatment) following their initial interview. For each day following the initial interview, our model provides a probability that the event (relapse) has not already occurred. We also create a plot that reflects the risk of marijuana relapse overtime (probability that the event will occur).

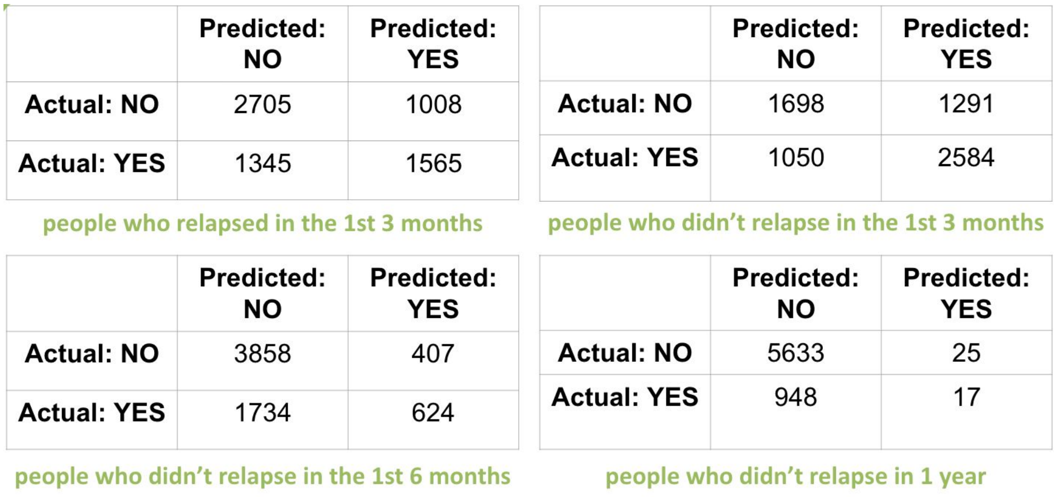
For our model, we chose to use Cox Proportional-Hazards Model (CoxPh) survival analysis, but for future work, we plan to also use survival analysis on random forests and compare the results.

In our model, we looked at three different metrics that are protected by the Equality Act to prevent discrimination: trauma, gender, and race. Using survival analysis, we compared the hazard plots for these subsets.

4.2 Results

4.2.1 Classification Results

The results from our classification model, or logistic regression model, were not especially conclusive. Out of the people that actually relapsed within the first 3 months, the model predicted that 4050 people would not relapse, but only 2705 of these were accurate, and the rest relapsed. Out of those didn't relapse within the first 6 months, the model is a little better: it predicted that 5592 people would not relapse but only 3858 of those were accurate. Our model got better the further out the predictions were, but we realized that this was not the best route to pursue for our model so we moved on to 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

Table 1: Logistic Regression Confusion Matrices

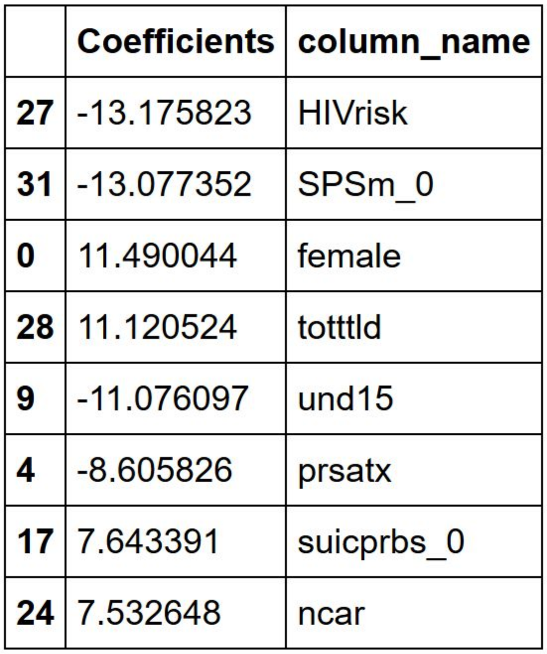
4.2.2 Regression Results

4.2.2.1 Linear Regression

The features that had the highest impact on our linear regression model included HIVrisk, SPSm_0, female, totttId, and und15 (Table 2). The HIVrisk feature was an indicator of whether or not a person was at risk for HIV. This had the highest impact on the model and because its coefficient is negative, our model suggests that a person that is at risk for HIV will have a shorter relapse time. SPSm_0 represents a patient's substance abuse severity at the time of the first interview. This value was also negative, suggesting that SPSm_0 also indicates that a user will have fewer days until relapse. Und15 indicated that a patient was under the age of 15 when they first used marijuana. Both of these seem like reasonable predictors of a faster relapse time because if a person was severely abusing marijuana before seeking treatment, it would make sense that they would be more likely to relapse sooner. Additionally, abusing marijuana at a younger age might also correspond with using the substance over a longer period of time, especially during a critical development stage in life.

Female and totttId were both positive values, indicating that both features corresponded to a prolonged relapse time. TottId represented the number of days that a patient underwent treatment (as some people fail to continue seeking treatment) and female indicated that a patient was female. It also seems reasonable that a patient would have a longer time before relapsing if they were being treated consistently.

To evaluate the effectiveness of our model, we looked at the median absolute error, R^2 and explained variance which were 76.70, 0.07, and 0.07 respectively. The median absolute error indicates that our predictions were accurate within approximately 77 days, or about 2.5 months. Practically, those working in a rehabilitation center could start allocating more resources to a patient within 77 days of the person's projected day of relapse. However, because resources are limited, facilities may not have enough to accommodate every patient or compensate for all 77 days of error.



	Coefficients	column_name
27	-13.175823	HIVrisk
31	-13.077352	SPSm_0
0	11.490044	female
28	11.120524	totttId
9	-11.076097	und15
4	-8.605826	prsatx
17	7.643391	suicprbs_0
24	7.532648	ncar

Table 2: Linear Regression Feature Importance

The predictions for our linear regression model can be visualized in Figure 1 below. In this map of Southern California, some areas are colored according to the scale of average days until relapse for the area. According to this model, in Santa Ana, on average, people have a longer time before relapsing. Whereas in Downey, people on average relapse more quickly. However, this does not necessarily indicate that environmental features have a significant impact on a person's time until relapse (this is discussed further in section 4.2.6 Augmented Dataset).

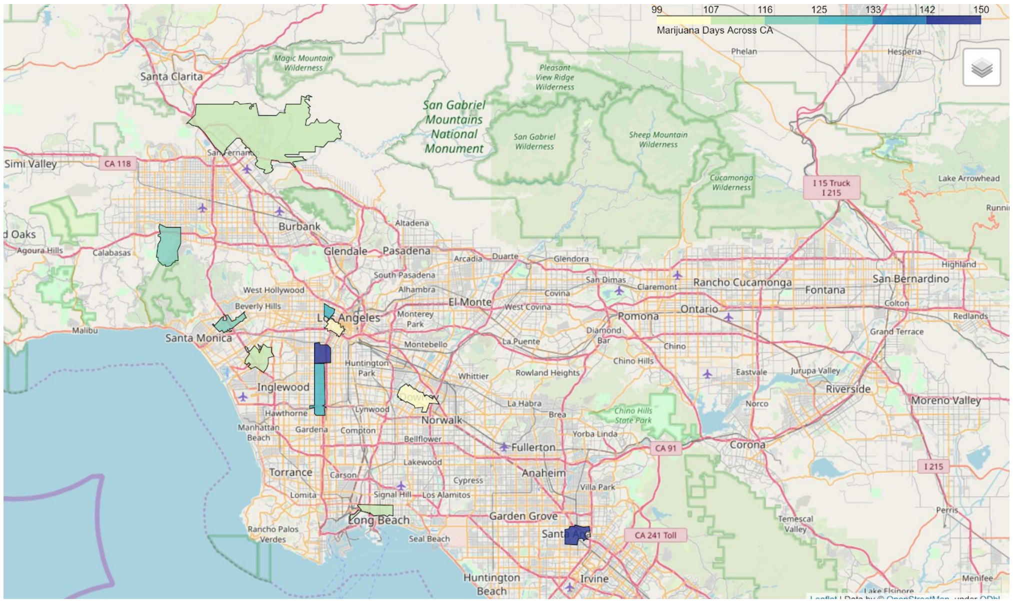
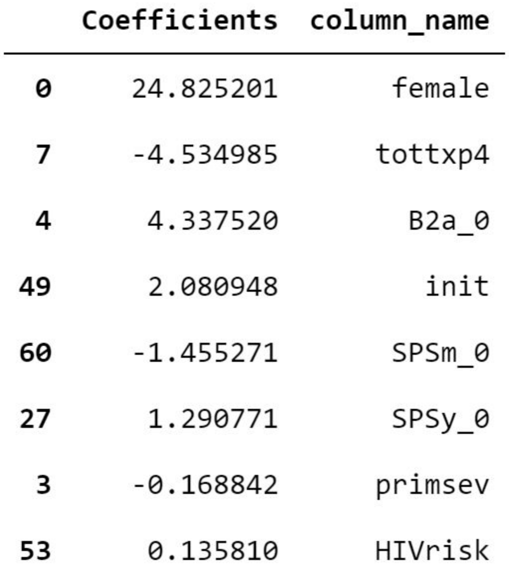


Figure 1: Linear Regression Predictions Map of Southern California

4.2.2.2 Random Forests

Our random forest model had a median absolute error of 77.25, R^2 of -0.09 and an explained variance of -0.06. Similarly to the linear regression, this model would be able to approximately predict the number of days until a patient relapses within about 2.5 months.

The features, SPSm_0, HIVRisk and female, that were impacting the linear regression model were also among the highest feature coefficients on the random forest model. The second highest coefficient, tottxp4, which indicated the number of treatment plans required for a given patient, negatively impacted the model, meaning that if a patient has more treatment plans, their time until relapse would be decreased. If a patient has more treatment plans, it might indicate that a patient had a greater risk of relapsing or perhaps the numerous treatments are more difficult for a person to maintain which would make sense as to why there is a negative correlation.



	Coefficients	column_name
0	24.825201	female
7	-4.534985	tottxp4
4	4.337520	B2a_0
49	2.080948	init
60	-1.455271	SPSm_0
27	1.290771	SPSy_0
3	-0.168842	primsev
53	0.135810	HIVrisk

Table 3: Random Forests Feature Importance

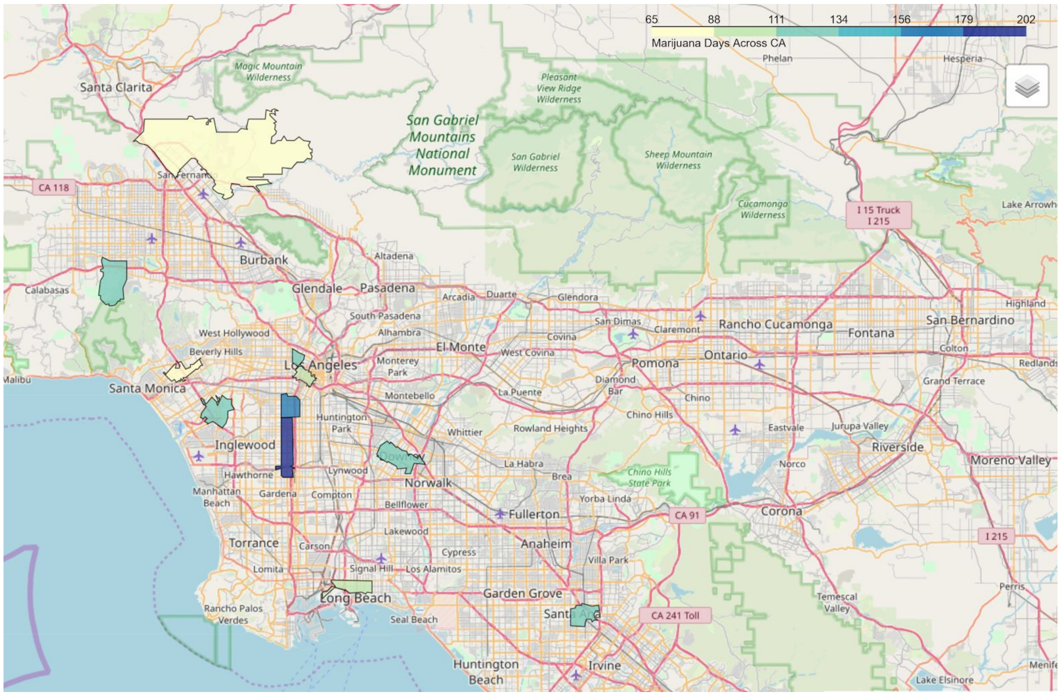


Figure 2: Random Forest Predictions Map of Southern California

4.2.2.3 XGBoost

Our XGBoost model had a median absolute error of 75.42, R^2 of 0.07 and an explained variance of 0.11. Similarly to the previous two regression models, our XGBoost model would be able to approximately predict the number of days until a patient relapses within about 2.5 months.

Also in line with the other models, totttId, SPSm_0 and HIVRisk were among the most impactful features. B2a_0, indicating the patient's age at the time of the first interview, also had a high coefficient in the random forest model.

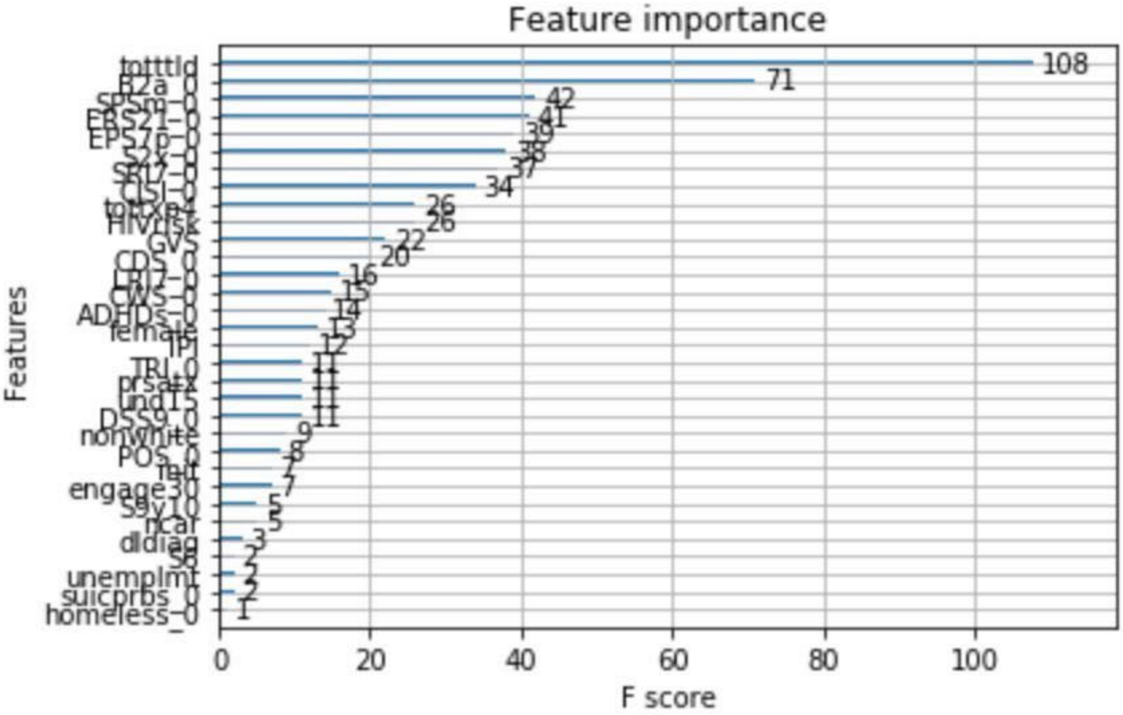


Figure 3: XGBoost Feature Importance

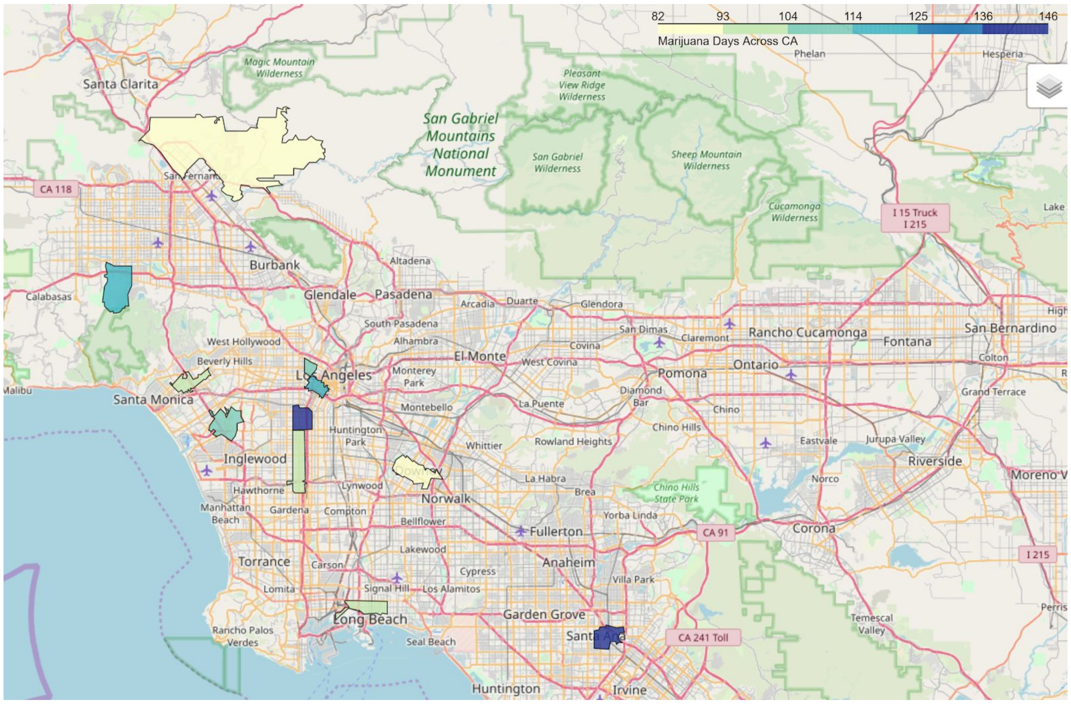
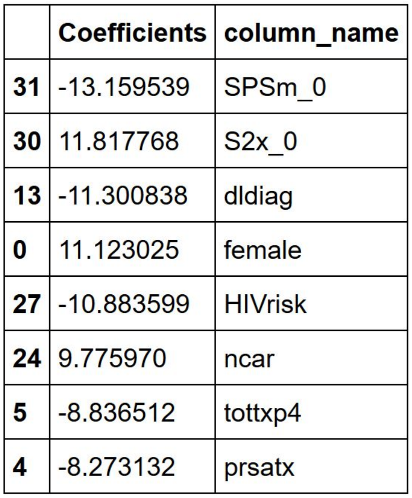


Figure 4: XGBoost Predictions Map of Southern California

4.2.2.4 Support Vector Machine (SVM)

Our XGBoost model had a median absolute error of 61.87, R^2 of 0.11 and an explained variance of 0.57. Based on our SVM model's median absolute error, we would be able to approximately predict the number of days until a patient relapses within 62 days or about 2 months. This number indicated that the predictions using SVM were more accurate than the other regression models.

Despite a reduction in median absolute error, many of the same features also had higher impact on the SVM model including SPSm_0, female, tottxp4, and HIVrisk. Following SPSm_0, the second most impactful feature, Sp2x_0, represented the number of days a patient was in a controlled environment. The positive coefficient indicates that a person that spends more time in a control environment would have an increased amount of time before relapsing, which is reasonable because a person may not have access to marijuana while in the controlled environment. Dldiag had the third highest coefficient and corresponded to whether or not a patient had more than one diagnosis. This feature negatively impacted a person's time until relapse, meaning that if a patient had a dual diagnosis, they were more likely to relapse faster, which is also reasonable because if a person has more than one underlying issue, their substance abuse might be more severe.



	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

Table 4: Support Vector Machine Feature Importance

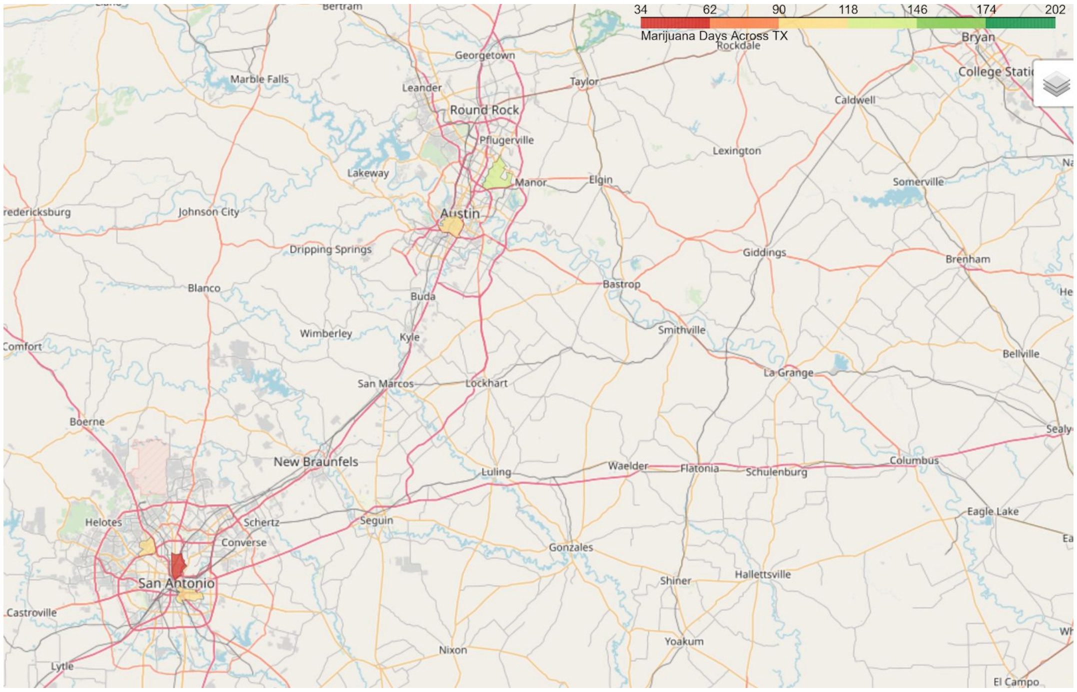


Figure 5: SVM Predictions on Map of Texas

4.2.3 Survival Analysis Results

Gender

Looking at the hazard plots for Risk and Probability for Marijuana Relapse for our gender subset (Figure 6), we can conclude that males are at a higher risk of marijuana relapse over time.

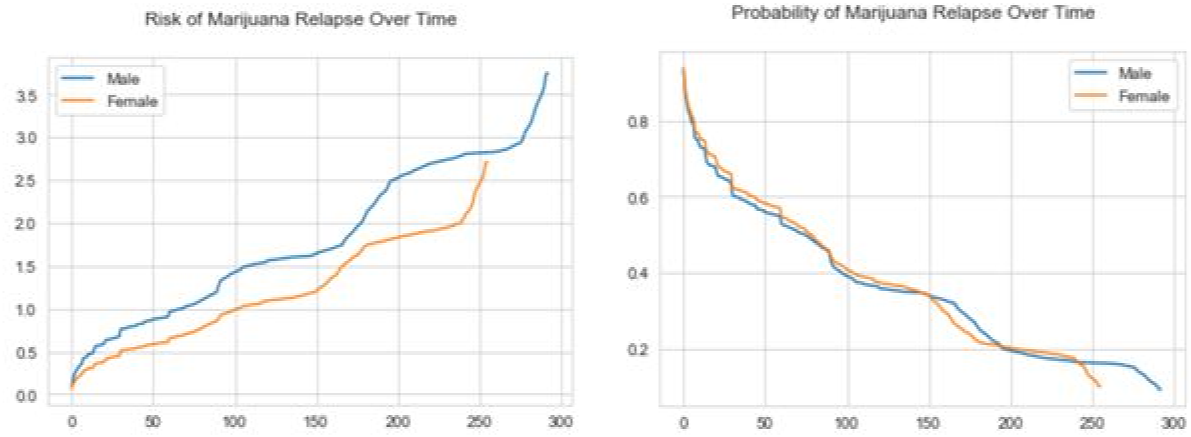


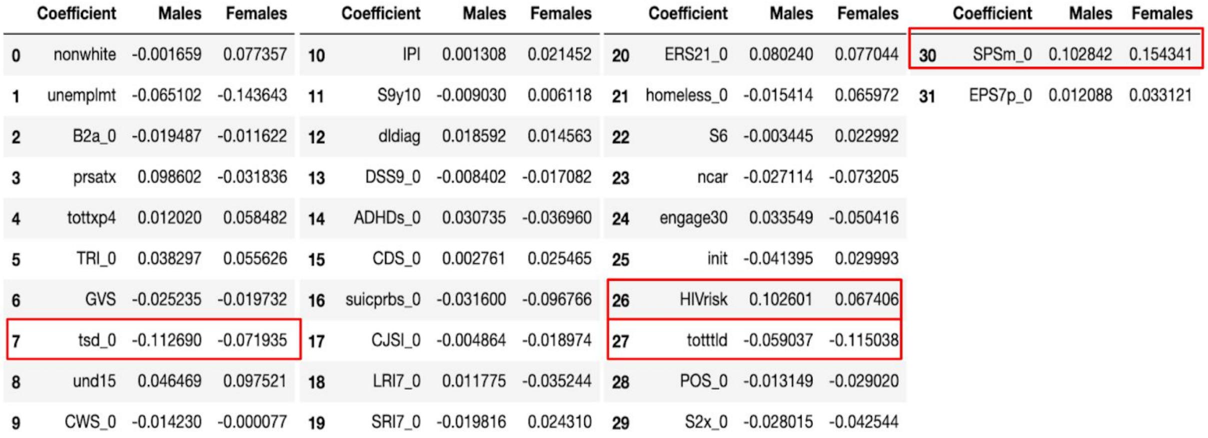
Figure 6: Gender Hazard Plots for Risk and Probability of Marijuana Relapse over time respectively

On performing survival analysis on this subset, we narrowed down to 4 features that gave us significant results - traumatic stress disorder (‘tsd_0’), risk of HIV (‘HIVrisk’), days a patient underwent treatment (‘totttld’), and the substance abuse severity scale at baseline (‘SPSm_0’). Figure 7 highlights the different feature importances in our gender subset when predicting survival, which in our case is, how likely is a patient to not relapse.

From this table, we can see that there is a negative correlation between trauma and survival, which means a patient with traumatic stress disorder is more likely to relapse. Trauma also plays a bigger role in predicting marijuana relapse for males, as compared to females.

HIV risk has a positive correlation, which means patients with higher risk of HIV are less likely to relapse. We speculate this might be due to patients being more wary about consuming drugs when they are aware of this underlying condition. Similar to trauma feature importance, HIV risk feature importance is higher for males than females.

The results for the feature importance of the number of days a patient underwent treatment was an interesting finding. The results give us a negative correlation, which means a patient who underwent treatment for more number of days, is more likely to relapse. This intuitively does not make sense. But a highly possible explanation for this result is that some patients do not complete treatment and fail to show up for their interviews. This throws off the data in favor of people who do not show up for their check-ups, which is most likely to be inaccurate. But if we look at the gender importance, SPSM_0 plays a bigger role in predicting marijuana relapse for females than males.



	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

Figure 7: Gender Feature Importance

Race

The next metric that we looked at in our survival analysis is race. The data was divided between white and non-white patients. The plots for the risk and probability of marijuana relapse for race show that white patients have a higher risk of marijuana relapse as compared to non white patients.

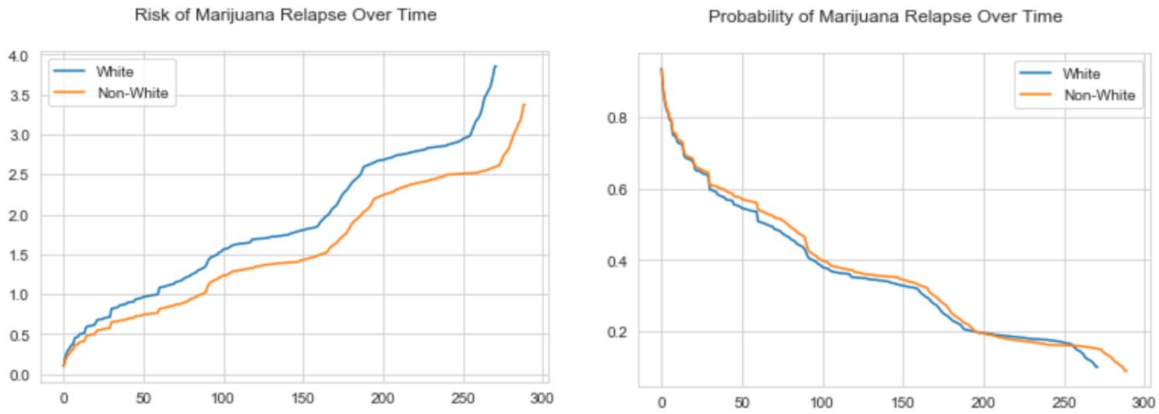
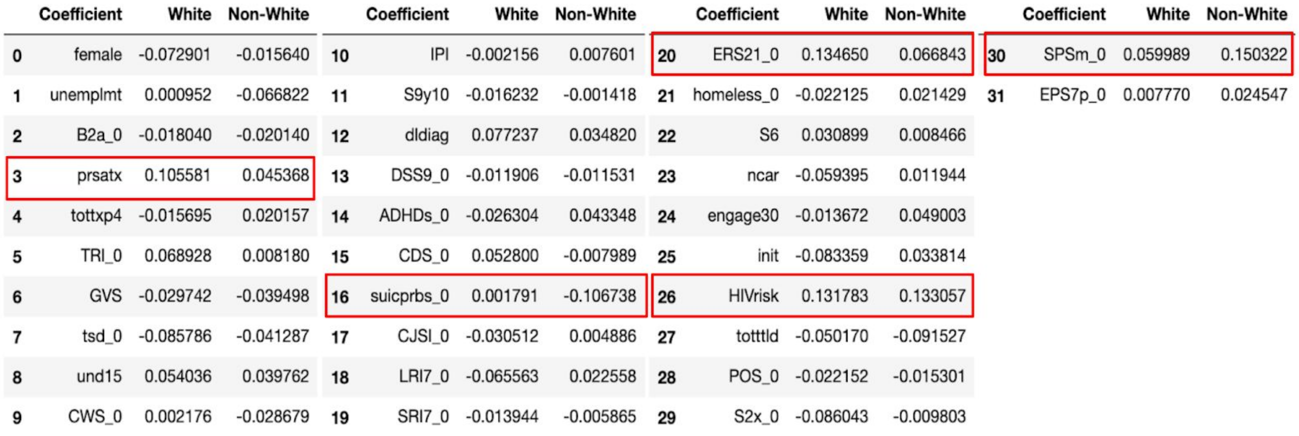


Figure 8: Race Hazard Plots for Risk and Probability of Marijuana Relapse over time respectively

Figure 9 shows the feature importance for the different features in our survival analysis for our race subset. Some significant features in this analysis were - prior substance abuse treatment ('prsatx'), probability of suicide ('suicprbs_0'), environmental risk factor scale ('ERS21_0'), risk of HIV ('HIVrisk'), and substance abuse severity scale at baseline ('SPSm_0').

A positive correlation between prior substance abuse treatment and survival (probability that a patient will not relapse) can be noted, which intuitively makes sense. This correlation is higher amongst white patients. We can speculate that this disparity might exist due to differing socio-economic status and cultural differences, where white patients are more likely to have had, or been able to afford, prior treatment.

Higher probability of suicide correlated to higher probability of marijuana relapse amongst non white patients in our dataset, but had very low feature importance for our non white patients. HIV risk showed similar positive correlation with not relapsing like it did with out gender subset.



	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

Figure 9: Race Feature Importance

Trauma

The final metric that we analysed in our survival analysis is trauma. The 3 kinds of trauma that we looked at were sexual trauma, physical trauma, and psychological trauma. Any patient who experienced any one of those was considered in the analysis. The plots for the risk and probability of marijuana relapse over time show that patients who experience any kind of trauma are at a higher risk of marijuana relapse at every point of their treatment.

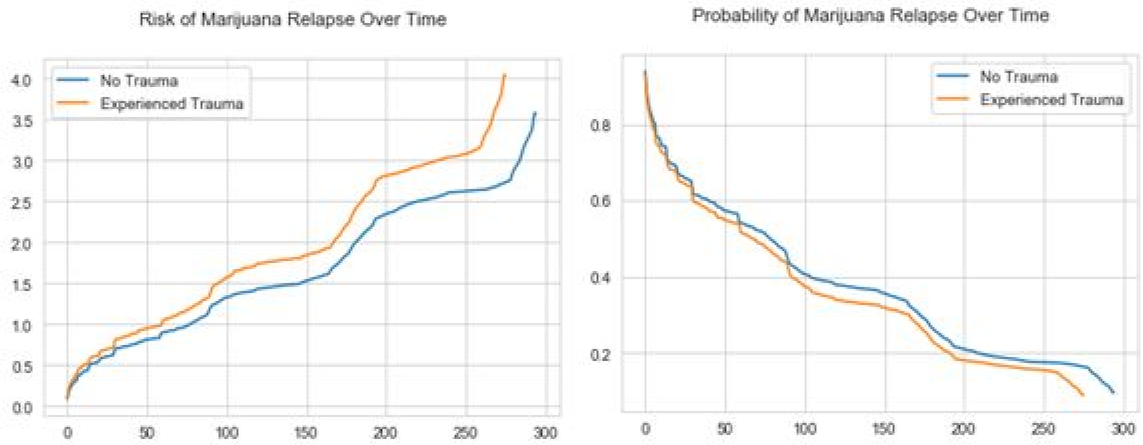
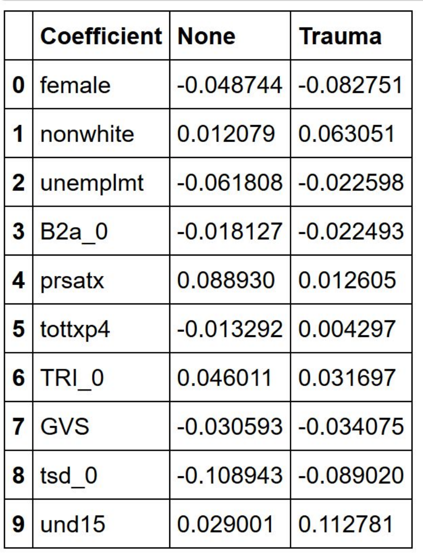


Figure 10: Trauma Hazard Plots for Risk and Probability of Marijuana Relapse over time respectively

Looking at the feature importance of some of the features in our survival analysis in Figure 11, we can see that female patients that underwent trauma had a higher probability of relapsing that those who did not. We can also note that unemployed patients that underwent trauma are a lot more likely to relapse than those who did not undergo trauma.



	Coefficient	None	Trauma
0	female	-0.048744	-0.082751
1	nonwhite	0.012079	0.063051
2	unemplmt	-0.061808	-0.022598
3	B2a_0	-0.018127	-0.022493
4	prsatx	0.088930	0.012605
5	tottxp4	-0.013292	0.004297
6	TRI_0	0.046011	0.031697
7	GVS	-0.030593	-0.034075
8	tsd_0	-0.108943	-0.089020
9	und15	0.029001	0.112781

Figure 11: Trauma Feature Importance

4.2.4 Web App

Since the data will be primarily used by social workers and people at the rehabilitation centers (or those treating marijuana abusers) or even patients themselves, we developed a web application that helps people interpret our model. Interpretability is important, especially in our case, as we want people to be able to use our model without having to understand how it works. In the web app, a person can select some personal characteristics including female/male and white/nonwhite. Using our survival analysis model, the application will produce a plot similar to Figure 10, except that the plot will be more accurate to the individual given their selected features.

4.2.5 Overall Results

Of the models that we tried, survival analysis performed the best. The results of our logistic regression (classification) model were largely inconclusive. While the results for our regression models, particularly SVM were reasonable, the median absolute error was still approximately 62 days, or 2 months, in which our predictions could be inaccurate. Survival analysis had the smallest median squared error of 55 days and also had a concordance index of approximately 0.6 among each subgroup, meaning that our model had good predictive ability.

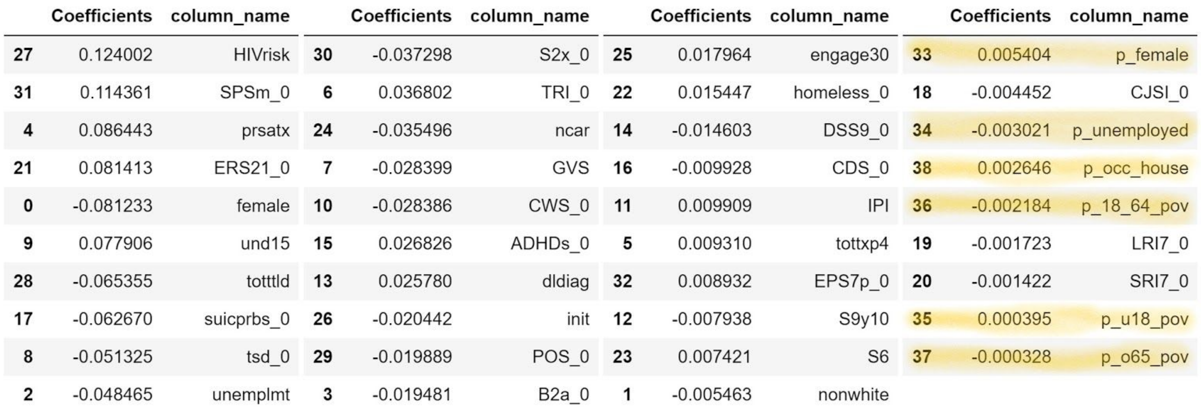
4.2.6 Augmenting Dataset with Census Level Data

Our model used a patient's personal data to approximate the number of days it would take to relapse. We were also interested in seeing if environmental factors based on a patient's location would impact the time it took for a given person to relapse. If our model indicated that living in certain locations had an impact on a person's projected date of relapse, we could potentially use this information to determine effective placement of new rehabilitation centers. In order to answer this question, we augmented our original dataset using census tract level data.

We collected our data from Social Explorer, using the five-year estimate American Community Surveys from 2013-2017 [8]. In our model, we chose to include the following features (all of which are used to calculate concentrated disadvantage, or neighborhoods with high percentages of residents with low socioeconomic status) for each census tract: the percentage of female headed households, the percentage of people under the age of 18 with poverty status, the percentage of people between 18 to 64 years of age with poverty status, the percentage of people over the age of 65 with poverty status, unemployment rate, and percentage of owner occupied housing.

Using the rehabilitation center addresses from the original dataset (198 unique locations), we were able to add columns that corresponded to the latitude and longitude of each facility using a web geocoder API. We then made API calls to the Federal Communications Commission's area API to get each address' corresponding census tract Federal Information Processing Standards (FIPS) code, appending a column to the original dataset. We were able to join the original dataset with a dataset with census tract information from Social Explorer, on the FIPS code column [8].

Using the augmented dataset, we reran our best performing mode, SVM survival analysis. After rerunning the model, it was evident that the census tract features (p_female, p_u18_pov, p_18_64_pov, p_o64_pov, p_unemployed, p_occ_house) had next to no impact on our predictions as demonstrated by Figure 12 below. The results of our model on the augmented dataset refute the idea that environmental factors influence the time it takes for someone to relapse.



	Coefficients	column_name		Coefficients	column_name		Coefficients	column_name		Coefficients	column_name
27	0.124002	HIVrisk	**30**	-0.037298	S2x_0	**25**	0.017964	engage30	**33**	0.005404	p_female
31	0.114361	SPSm_0	**6**	0.036802	TRI_0	**22**	0.015447	homeless_0	**18**	-0.004452	CJSI_0
4	0.086443	prsatx	**24**	-0.035496	ncar	**14**	-0.014603	DSS9_0	**34**	-0.003021	p_unemployed
21	0.081413	ERS21_0	**7**	-0.028399	GVS	**16**	-0.009928	CDS_0	**38**	0.002646	p_occ_house
0	-0.081233	female	**10**	-0.028386	CWS_0	**11**	0.009909	IPI	**36**	-0.002184	p_18_64_pov
9	0.077906	und15	**15**	0.026826	ADHDs_0	**5**	0.009310	tottxp4	**19**	-0.001723	LRI7_0
28	-0.065355	totttld	**13**	0.025780	dldiag	**32**	0.008932	EPS7p_0	**20**	-0.001422	SRI7_0
17	-0.062670	suicprbs_0	**26**	-0.020442	init	**12**	-0.007938	S9y10	**35**	0.000395	p_u18_pov
8	-0.051325	tsd_0	**29**	-0.019889	POS_0	**23**	0.007421	S6	**37**	-0.000328	p_o65_pov
2	-0.048465	unemplmt	**3**	-0.019481	B2a_0	**1**	-0.005463	nonwhite			

Figure 12: Coefficients for Features After Running Survival Analysis on Augmented Dataset

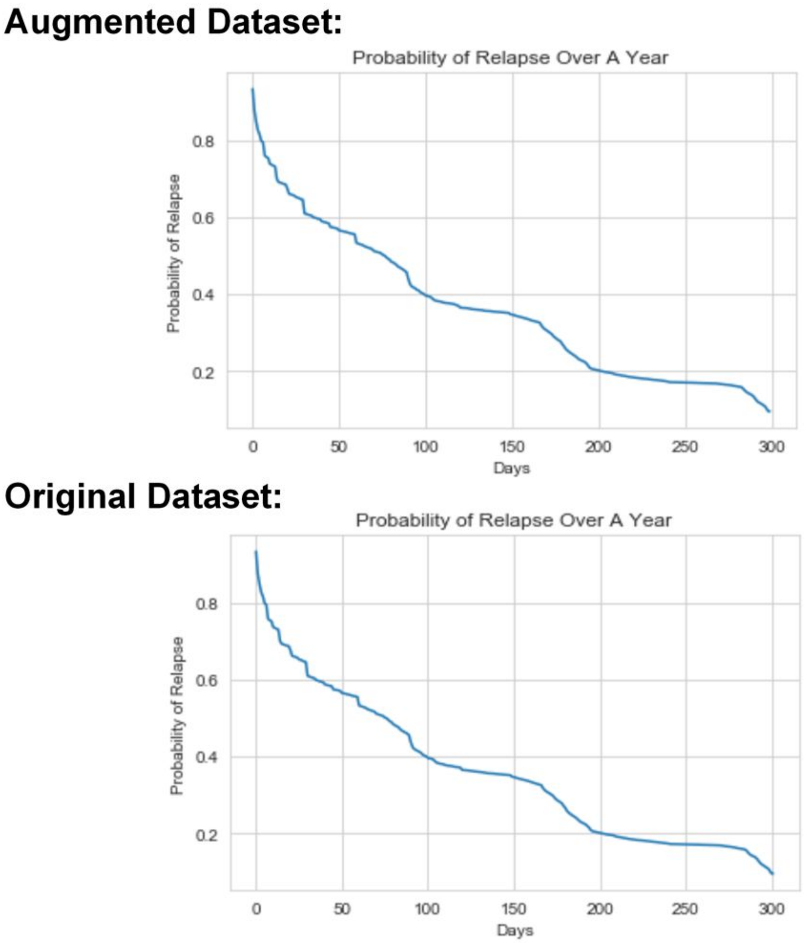


Figure 13: Plot of probability of relapse over a year for the augmented and original datasets respectively

5 Conclusions and Possible Extensions

5.1 Conclusions

In summary, we feel that our project has given us a robust model which can be used to predict relapse in other similar populations with usage of marijuana. By experimenting with different environmental factors to measure in our model, and by using many different models in order to find the one with the best fit, we were able to develop a model with Support Vector Machine with -55 median absolute error accuracy. This means that we were underpredicting relapse by around 55 days. We also developed a Web App for doctors to be able to predict their patients' relapse chances based on their factors, and this can be further developed as our model becomes more specific and we acquire more data. This interface makes our models more readable and interpretable for people, like Professor Davis, who work in social work as well. In the following sections, we will discuss other examples of extensions and future work for continuing our project.

5.2 Possible Extensions

Some possible extensions we would like to pursue, first of all, are the other models such as fairness/optimization. We would like to make our model as accurate as possible, and pursuing newer models will aid us in this. Going a step further, our project can be used to model predict relapse of other substances such as opioids, vape usage, cigarettes, alcohol, and other abusive substances. This can then be used to optimize rehabilitation centers - by analyzing the given number of centers within an area, we can measure what the impact of adding a rehabilitation center there would add, and pursue adding centers appropriately to maximize impact. More broadly, this can be applied to other situations with a fail/success outcome, such as likelihood to fail exams or courses due to environmental factors, or success and failure of a marriage or interview.

5.2.1 Calculating Fairness for Each Model

In our progress toward calculating fairness thus far, we have found that our dataset contains significantly more males than females and a higher number of nonwhite than white people. Additionally, for our survival model, our concordance index is slightly higher for males, white people, and those who have never experienced trauma. Even though this difference is marginal, it demonstrates that our survival model provides more accurate predictions within the subgroups of gender, race, and trauma respectively.

To calculate fairness in the future for our regression models, we could check if our model is under predicting or over predicting for any subgroup of our dataset. For each of the groups, we would get two metrics: “days too late” and “days too early.” “Days too late” would be calculated by subtracting the actual day the patient relapsed from the predicted date (representing over prediction). “Days too early” would be calculated by subtracting the predicted date from the actual day the patient relapsed (representing under prediction). The median absolute error metric that we have calculated for each model would be the sum of the two metrics.

References

[1] J. L. Gowin, T. M. Ball, M. Wittmann, S. F. Tapert, and M. P. Paulus, “Individualized relapse prediction: Personality measures and striatal and insular activity during reward-processing robustly predict relapse,” *Drug and Alcohol Dependence*, 30-Apr-2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0376871615002148. [Accessed: 17-Nov-2019].

[2] L. Acion, D. Kelmansky, M. van der Laan, E. Sahker, D. S. Jones, and S. Arndt, “Use of a machine learning framework to predict substance use disorder treatment success,” *PLOS ONE*, 10-Apr-2017. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175383. [Accessed: 17-Nov-2019].

[3] “Marijuana Addiction and Abuse - Understanding Marijuana Abuse,” *AddictionCenter*. [Online]. Available: http://www.addictioncenter.com/drugs/marijuana/. [Accessed: 17-Nov-2019].

[4] “Marijuana,” *NIDA*. [Online]. Available: https://www.drugabuse.gov/publications/research-reports/marijuana. [Accessed: 17-Nov-2019].

[5] “Marijuana Statistics - Cannabis Use Statistics - Drug-Free World,” *Foundation for a Drug-Free World*, Jun-2005. [Online]. Available: https://www.drugfreeworld.org/drugfacts/marijuana/international-statistics.html. [Accessed: 17-Nov-2019].

[6] M. Gonzales, “Marijuana Statistics and Facts,” *DrugRehab.com*, 21-May-2018. [Online]. Available: http://www.drugrehab.com/addiction/drugs/marijuana/statistics/. [Accessed: 17-Nov-2019].

[7] M. P. Paulus, “Neural Activation Patterns of Methamphetamine-Dependent Subjects During Decision Making Predict Relapse,” *Archives of General Psychiatry*, 01-Jul-2005. [Online]. Available: https://jamanetwork.com/journals/jamapsychiatry/article-abstract/208753. [Accessed: 17-Nov-2019].

[8] Social Explorer. Census Tract Data [Data file]. Retrieved from: https://www.socialexplorer.com/tables/ACS2017_5yr/R12411264

[9] “Understanding Marijuana Addiction Relapse: Relapse Triggers,” *Recovery.org*, 02-Nov-2019. [Online]. Available: https://www.recovery.org/marijuana/relapse/. [Accessed: 17-Nov-2019].