Using Machine Learning to Predict Substance Use Treatment Success

Patrico Tyrell, MPH

Contact: patrico_tyrell@yahoo.com

October 2023

Table of Contents

Introduction

Purpose

This project aims to leverage machine learning techniques to investigate disparities in substance use disorder (SUD) treatment success. The primary objective is to develop predictive models capable of identifying individuals who are likely to successfully complete substance use treatment programs. The project's investigation is motivated by the disparities in SUD treatment services and outcomes. By identifying key predictors of successful treatment and revealing disparities, this project seeks to illuminate strengths and weaknesses in service delivery. This analysis has the potential to boost treatment success rates, ultimately addressing unmet treatment needs.

The Problem

According to the Centers for Disease Control and Prevention (CDC), **more than one million people have died since 1999** from a drug overdose. In 2021, **106,699** drug overdose deaths occurred in the United States. The age-adjusted rate of overdose deaths increased by **14%** from 2020 (28.3 per 100,000) to 2021 (32.4 per 100,000).

Opioids, particularly synthetic opioids (excluding methadone), have emerged as the leading cause of drug overdose fatalities. A staggering **88%** of opioid-related overdose deaths involve synthetic opioids.

In 2021, opioids played a role in **80,411** overdose deaths, accounting for a substantial **75.4%** of all drug overdose fatalities. Drug overdose deaths involving psychostimulants such as methamphetamine are also increasing with and without synthetic opioid involvement.

The Data

The project utilized data from the Treatment Episode Data Set: Discharge (TEDS-D), sourced from the Substance Abuse and Mental Health Services Administration (SAMHSA). TEDS-D serves as a comprehensive national data repository, containing annual discharge records from various substance use treatment facilities. These records encompass a wide array of information on admissions for individuals aged 12 and above. Pertinent details include admission demographics such as age, sex, race/ethnicity, and employment status. Additionally, the dataset provides insights into substance use characteristics, including details about the types of substances used, age at first use, routes of use, frequency of use, and prior admissions. Data regarding the number of treatment facilities and response rates was extracted from SAMHSA's PDF reports. Additionally, population statistics and geographical details were sourced from Wikipedia.

Data Wrangling

In the process of preparing the TEDS-D dataset for analysis, I had a few data-wrangling challenges. An overview of the main issues is presented below:

1

The TEDS-D included all admissions and discharges, rather than individual cases.

Solution: To tackle this, a filtering approach was applied to retain only those records corresponding to individuals without prior substance use disorder (SUD) treatment. This transformation aligned the dataset with the goal of studying unique individual instances and resulted in a dataset containing 503,107 distinct individuals.

2

Significant amount of missing data

Solution: While this might not be ideal as some important variables may have been lost, I excluded records with missing values in any of the predictors and outcome variables. This ensured the integrity of the data used for analysis.

3

No single source providing information on the total treatment facility count by state.

Solution: To overcome this challenge, data was extracted from PDF reports to acquire details on the number of treatment facilities surveyed by The Substance Abuse and Mental Health Services Administration and response rates. By combining response rates with the count of facilities surveyed, the total number of treatment facilities for each state was accurately computed.

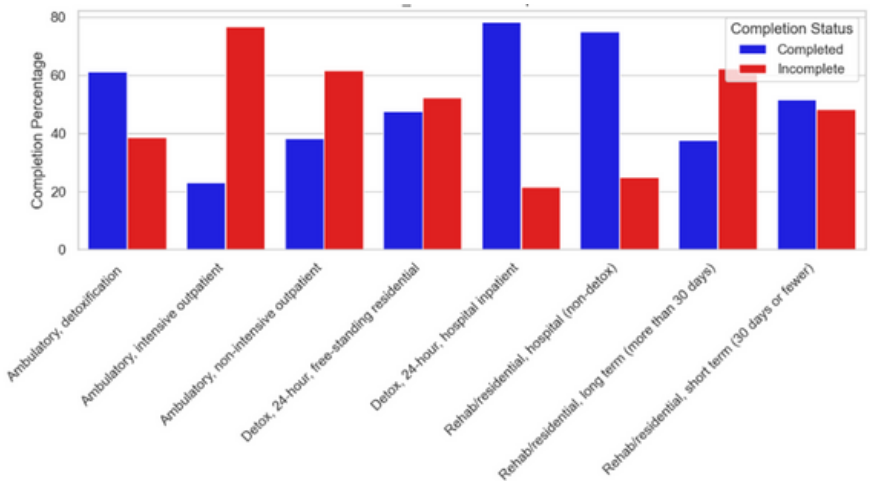
4

Misspelled state names and special characters within the dataset

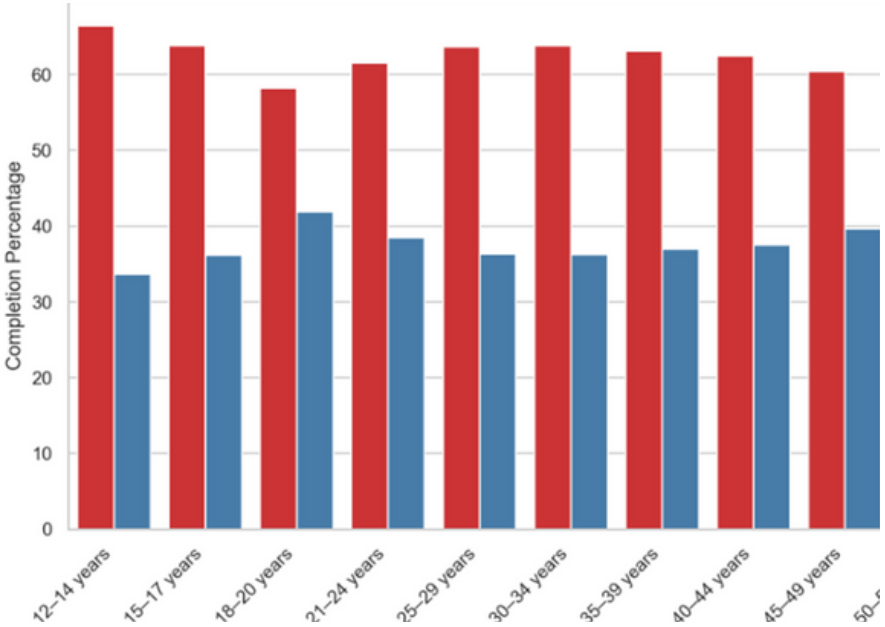
Solution: To facilitate data merging by state, all special characters were removed, and state name spellings were corrected.

Exploratory Data Analysis

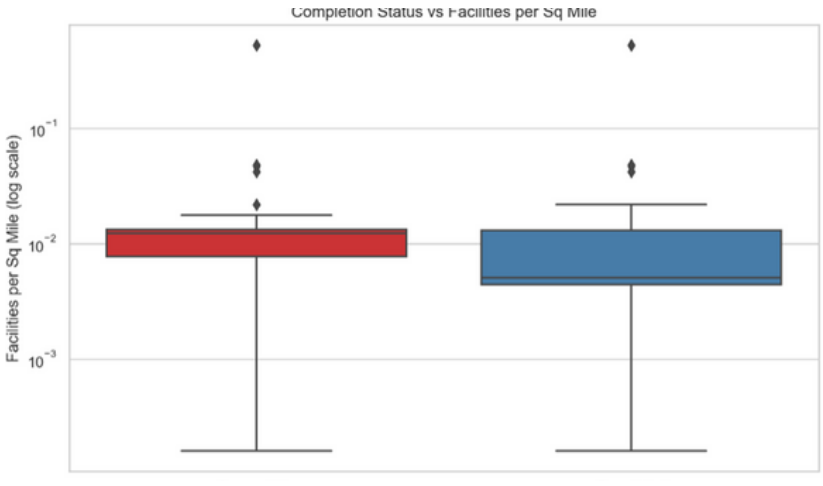
Initial exploratory analysis revealed that treatment success/completion varies based on the type of treatment service or setting they are enrolled in during admission or transfer. Detox and 24-hour free-standing residential services exhibit the highest completion rates, which is in line with the rapid transitions typically observed in these settings. This highlighted that success criteria and durations may differ significantly between outpatient services and other service types like 24-hour inpatient and detoxification.



The data indicates an improvement in completion rates beyond the age of 44, with the 12-14 age group displaying the lowest completion rate.



Boxplot analysis suggests that there could be disparities in the distribution of treatment facilities per area between cases of completed and incomplete treatment. Notably, regions with higher facilities per square mile tend to have higher treatment completion rates, aligning with expectations. This observation highlights potential challenges or limitations in terms of treatment facility accessibility, especially in densely populated or geographically constrained states.



Preprocessing

In the data preprocessing phase, several essential steps were taken to prepare the dataset for machine learning modeling.

These steps included:

- Calculating the population per square mile. The population per square mile was introduced as a new feature, providing information on population density.
- Selecting relevant columns and creating dummy variables for categorical features. Categorical variables were transformed into a numerical format, ensuring compatibility with machine learning algorithms.
- Standardizing numeric attributes, and encoding the target variable. Standardization was applied to numeric features to ensure consistent scaling.
- The label encoding of the target variable 'Completion_Status' facilitated binary classification, with 'Incomplete' represented as 1 and 'Complete' as 0.
- The dataset was then split into training and testing sets, and both sets were saved as CSV files for subsequent model training and evaluation. This comprehensive preprocessing pipeline ensures that the data is well-structured and ready for predictive modeling while preserving critical information for analysis.

Modelling & Evaluation

This is a classification problem as we are trying to predict treatment outcomes (complete or incomplete). I explored these conventional machine learning models to build the model and compare performances:

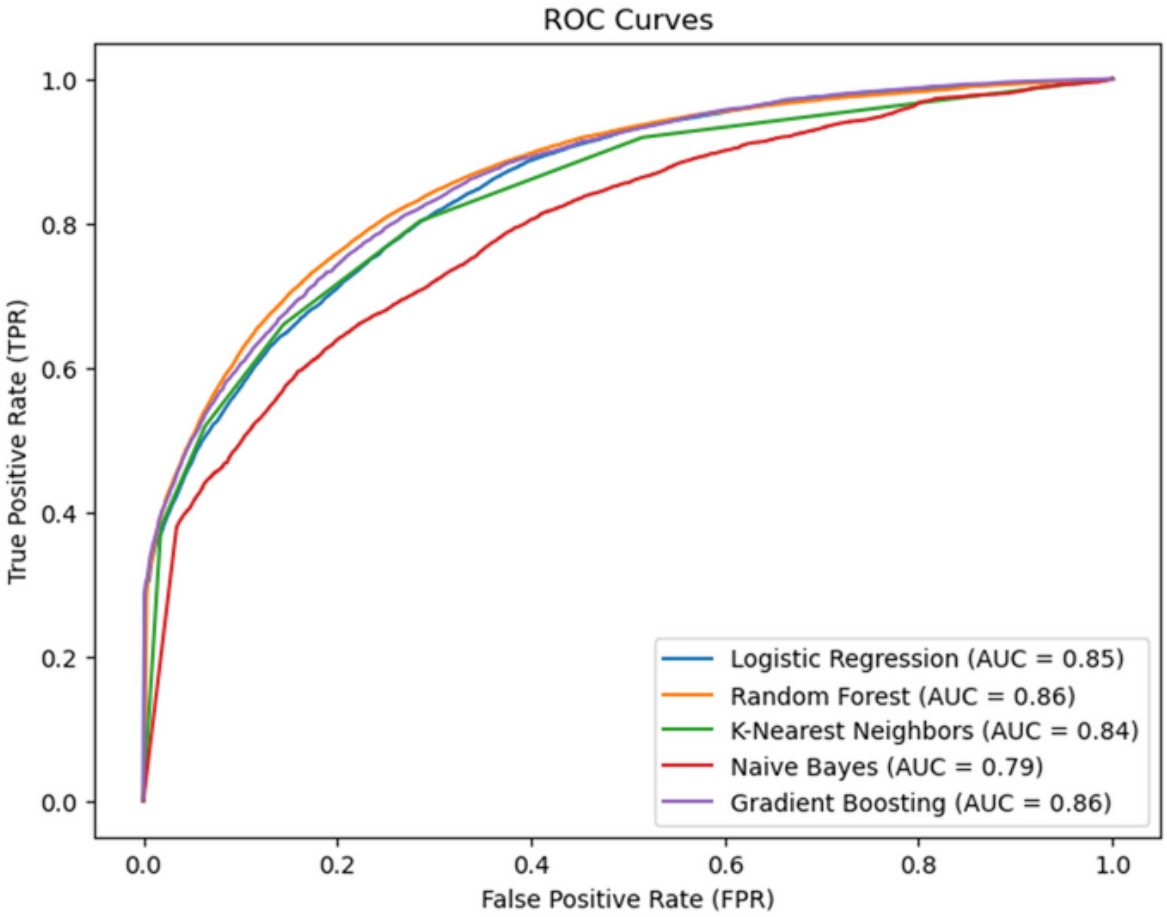
Logistic Regression	The logistic regression model achieved an accuracy of approximately 0.78. It exhibited a balanced precision-recall trade-off, with a higher precision for class 0 (complete) but a better recall for class 1 (incomplete). The model's F1-score for class 0 is notably higher than that for class 1, indicating that it's better at correctly classifying instances as complete. However, it has relatively lower recall for class 1.
Random Forrest	The random forest model performed slightly better with an accuracy of about 0.79. It exhibited a better balance between precision and recall for both classes compared to logistic regression. The F1-scores for both classes are competitive, making it a solid choice. The random forest model also had fewer false negatives (class 1) than logistic regression.
K-Nearest Neighbor (KNN)	The KNN model achieved an accuracy of around 0.78, similar to logistic regression. It displayed a balanced precision-recall trade-off but slightly lower F1-scores compared to random forest. It is a decent performer, but its recall for class 1 is not as high as random forest.
Naive Bayes	The Naive Bayes model had the lowest accuracy among the four models, around 0.74. It showed the lowest precision for class 1 and relatively lower recall for class 1 as well. While it has a balanced precision-recall trade-off for class 0, it struggles with classifying instances as incomplete (class 1).
Gradient Boosting	The gradient boosting model achieved an accuracy of about 0.79, similar to random forest. It displayed a good balance between precision and recall for both classes and competitive F1-scores. It performed well in correctly classifying both complete and incomplete cases, with a balanced precision-recall trade-off.

In summary, the Random Forest and Gradient Boosting models outperformed the others in terms of accuracy and balanced performance for both classes.

Modelling & Evaluation

In addition to accuracy, I examined evaluation metrics such as ROC and AUC. These metrics were important as the model appears to be classifying class 0 (incomplete) better because of class imbalance. Imbalanced datasets can heavily influence the performance of machine learning models, especially in classification tasks.

Consistent with previous metrics, In this case, both Random Forest and Gradient Boosting have the highest AUC scores of 0.86, suggesting that they perform similarly well in terms of ROC curve analysis. Since the metrics suggested Random Forest and Gradient Boosting, I chose to proceed with hyperparameter tuning the Random Forest model.



Hyperparameter Tuning Random Forest Model

I utilized RandomizedSearchCV to tune the model with a randomized search over hyperparameters to reduce the computational cost compared to using grid search.

While accuracy increased slightly, ROC-AUC Score decreased after hyper tuning using the random search. The decrease in the ROC-AUC score after hyperparameter tuning using random search could be due to the randomness involved in the search process. Randomized search explores a random subset of the hyperparameter space, and in some cases, it may not find hyperparameters that improve the model's performance on your specific dataset.

Hence, I performed cross-validation to get a more stable estimate of the model's performance. This helped ensure that the hyperparameters were chosen based on more robust performance estimates.

I performed 5-fold cross-validation on your Random Forest model, and the results indicate that the mean ROC-AUC score of approximately 0.879 suggests that your Random Forest model is performing well on the cross-validated subsets of your training data. The standard deviation is relatively small, indicating that the model's performance is consistent across different folds. This information provides more confidence in the model's predictive ability and its stability when applied to unseen data.

Feature Importance

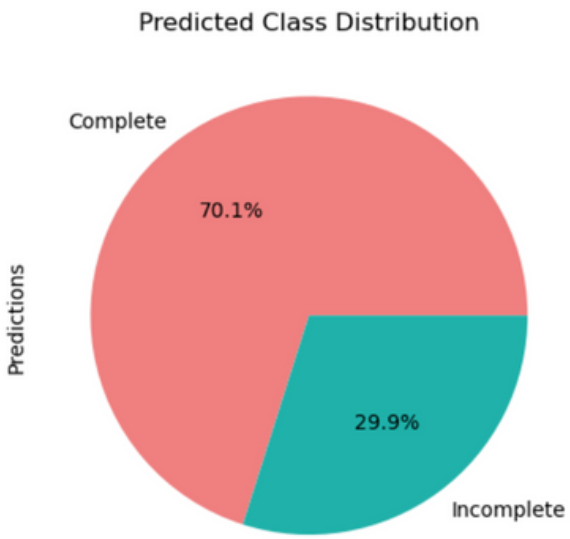
The features that were most influential in predicting treatment success were:

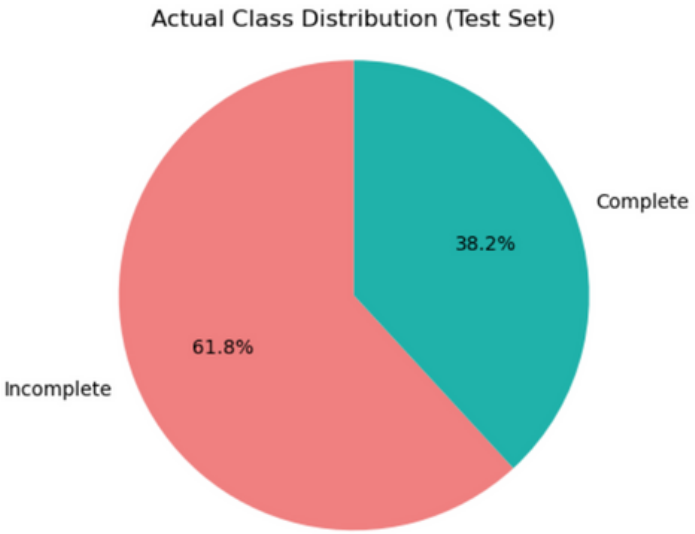
- The **Census division**, specifically "Mountain," appears to be highly influential. It suggests that treatment outcomes may vary significantly by region, with the Mountain division having a strong positive impact on treatment success.
- **Population density per square mile** is the second most important feature. It indicates that areas with higher population density may have a positive influence on treatment success. This could be due to better access to resources and support services.
- The **state of Arizona** seems to be a crucial predictor, suggesting that individuals receiving treatment in Arizona may have a higher likelihood of successful treatment completion.
- **The total number of treatment facilities** in the state is a significant factor. More treatment facilities could mean better access to care, increasing the chances of successful treatment outcomes.
- **Individuals who do not report alcohol at admission** have a positive impact on treatment success. This feature suggests that those without alcohol use issues are more likely to successfully complete treatment.
- Conversely, **individuals who report using only other drugs** (not alcohol) at admission also contribute positively to treatment success.
- **Kentucky's state**-specific impact on treatment outcomes is relatively smaller than Arizona, but it still plays a role in predicting success.
- Longer lengths of stay within the range of **61-90 days** have a positive influence on treatment success. This suggests that extended treatment durations may be more effective.
- ·Similarly, very long stays (**181-365 days**) also contribute positively to successful treatment outcomes.
- ·Stays within the range of **91-120 days** are another important factor, indicating that moderate-duration treatments have a favorable impact.

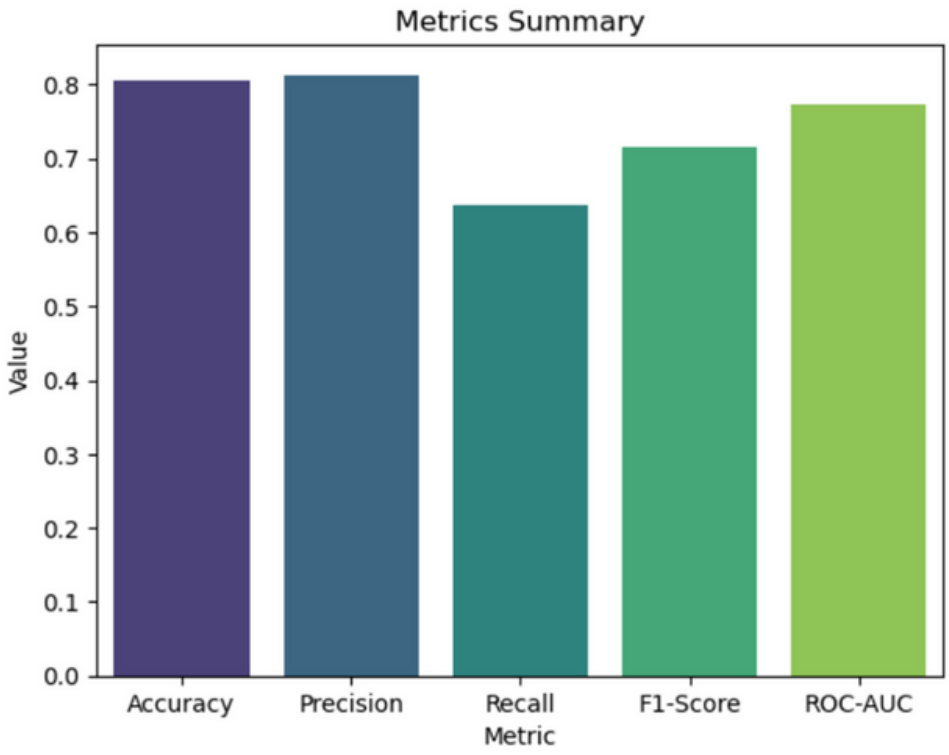
Final Predictions

My Random Forest model has achieved notable success in predicting treatment outcomes, as indicated by an accuracy of approximately 80.56%. This means that roughly 80.56% of the cases were correctly classified by the model. Additionally, the precision score of about 81.30% demonstrates that, among the cases predicted as successful treatment completions, the majority were indeed accurate predictions.

However, there are some downsides to consider. The model's recall, at around 63.70%, suggests that it missed identifying a substantial portion of actual successful treatment completions. This means there is room for improvement in capturing more positive cases. The F1-Score, which balances precision and recall, stands at about 71.43%, indicating a reasonably good balance, but further optimization may enhance overall performance. Lastly, the ROC-AUC score, approximately 77.33%, suggests the model's ability to distinguish between the two classes is moderately effective, but fine-tuning could boost this aspect of the model.







Future Directions

While my journey through model selection and hyperparameter tuning has yielded encouraging results, there are several avenues for future exploration and improvement:

1. **Feature Engineering**: Continue to explore feature engineering techniques to create new variables or transform existing ones, potentially uncovering additional patterns that can enhance predictive performance.
2. **Feature Importance**: Dig deeper into the feature importances generated by the Random Forest model.
3. **Add more Variables**: Explore the potential impact of socioeconomic and demographic factors on treatment success, such as **income**, **education**, **race**, **ethnicity**, **sex**, and **employment status**.
4. **Ensemble Methods**: Experiment with other ensemble methods like XGBoost, LightGBM, or AdaBoost to determine if any of them can surpass the performance of Random Forest.
5. **Imbalanced Data Handling**: Since the dataset is highly imbalanced (more incomplete cases than complete cases), I can explore techniques such as oversampling, undersampling, or the use of synthetic data generation methods to mitigate class imbalance.
6. **Model Interpretability**: Investigate techniques for improving the interpretability of Random Forest models, as they can sometimes be seen as "black boxes." Techniques like SHAP (SHapley Additive exPlanations) values or partial dependence plots can shed light on model predictions.
7. **Deployment**: Since I would like to use the model in practice, I will have to consider the deployment pipeline, model monitoring, and integration with existing systems.
8. **Ethical Considerations**: I have to ensure that the model is used in an ethical and responsible manner, avoiding biases and unintended consequences.

Acknowledgements

Thank you for reading my report to the end!

I would like to express my gratitude to the Schenectady Public Health Services for generously funding my course. Special thanks go to my Data Science mentor, Upom Malik, for his invaluable guidance and advice. I am also deeply appreciative of my wife, Andrea Glasgow, for her unwavering support throughout this journey.