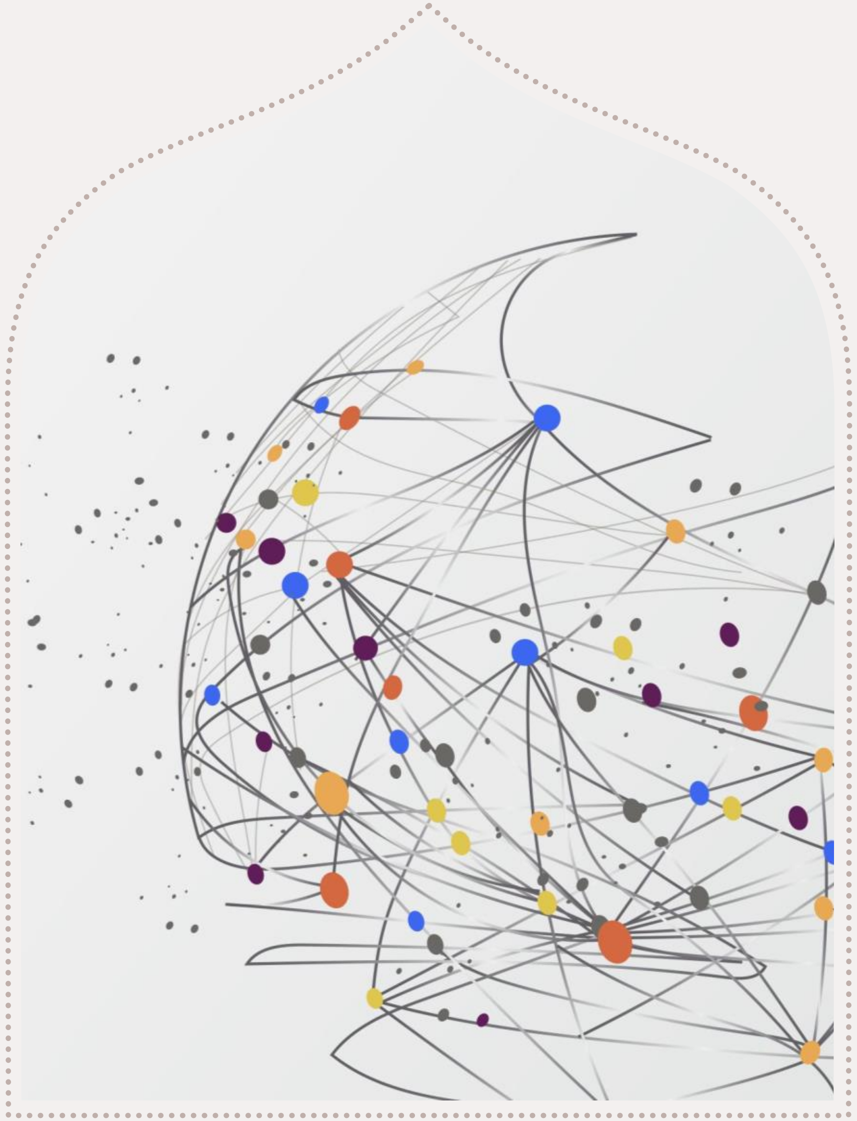
What Factors Predict Dropout Rates for Inpatient Substance Treatment Programs?

Marjete Vucinaj

Data Science, CUNY School of Professional Studies

Data 698: Master's Research Project



Introduction

Substance Use Disorder (SUD) Crisis:

- **46.3 million** people, or 16.5% of individuals (aged 12+) in the U.S., had SUD in 2022 (SAMHSA, 2022).

Overdose Epidemic, 2022 Data:

- Over **107,000 overdose deaths** - one death every 5 minutes (CDC, 2024).

Harm Reduction Efforts and their Challenges:

- **Methadone Maintenance Therapy:** While effective, it carries the potential for misuse and abuse.
- **Access to Clean Needles and Narcan:** Programs promote safer use but face challenges with inconsistent adherence and application (Crapanzano, 2018)

Identifying effective treatment options and taking steps toward recovery is the solution to reducing active SUD and overdose deaths (CDC, 2024).

Study Objective:

Focus: Predict factors for inpatient treatment dropout using predictive models in R.

Why is this important?

- Completion is a vital step toward recovery.
- Dropping out increases the likelihood of continued addiction.

Identifying the factors contributing to dropout rates can help address these issues, potentially improving treatment completion rates and supporting sustained recovery outcomes.

Literature Review

- **Gautam & Singh (2020):**
  - Analyzed TEDS-D 2017 data, focusing on opioid use.
  - Key factors influencing dropouts: Length of stay, geography, employment status, and age; they found RF model outperformed MLRM in predicting dropouts
- **Acion et al. (2017):**
  - Focused on Hispanic patients using TEDS-D 2006-2011.
  - Identified 10 key variables (e.g., age, substance use frequency, employment level); they found that Super Learning and RF performed best.
- **Baird et al. (2022):**
  - TEDS 2017-2019 data; examined treatment completion disparities.
  - Key factors influencing dropouts: Co-occurring mental health diagnosis and employment; found that the Virtual twins model performed best.
- **Gottlieb (2022):**
  - Analyzed local HEROES dataset (Houston, Texas)
  - Overdose history was a top dropout predictor; the RF model performed the best
- **Andersson et al. (2018):**
  - Norwegian study using Cox regression.
  - Mental distress and ADHD increased dropouts; motivation boosted completion.

Research Question

- Do the following factors contribute to dropout rates of inpatient treatment for addiction?

- **Social determinants of health:** include age, gender, race, ethnicity, marital status, education, employment status, living arrangements, past arrests, region, co-occurring disorder, and DSM diagnosis.
- **Substance use severity**: frequency of use, medication-assisted opioid therapy, primary substance, current IV drug use, age of first use, Previous substance use treatment episodes, Route of administration, Substance use type (drugs only, alcohol only, or both).
- **Choice in treatment:** referral source, length of stay, participation in substance use self-help programs.

Data and Variables:

- **Dataset:** Treatment Episode Dataset: Discharge (TEDS-D) 2021 from the Substance Abuse and Mental Health Services Administration (SAMHSA).
- **Scope:**
  - Covers treatment admissions from **all U.S. states, D.C., and territories**
  - Focuses on **publicly funded treatment programs** (limited private treatment data).
- **Challenges:**
  - Data variability due to state differences in collection, funding, and practices.
  - Coercion by criminal justice systems influences referrals in some regions.
- **Confidentiality Measures:**
  - **Data swapping:** Protects PHI by altering variables like census region.
  - **Age recording:** Ensures anonymity for smaller subsets (e.g., very young or old).

Data and Variables

Data Cleaning

- Missing Data Handling: Unknown values (-9) were replaced with NA and omitted
- Target Variable: ‘Outcome’ recoded as binary: Dropout (1), Completion (0)
- Filtered to include only **inpatient service settings** (rehab/residential short-term and long-term)
- After data cleaning the dataset had about 32,000 observations

Feature Engineering

- Length of Stay (LOS): Binned into ranges (e.g., 0–7, 8–14 days).
- DSM Diagnosis (DSMCRIT): Binned into five categories based on severity

Data Preprocessing:

- **Data Splitting:** 70% for training, 30% for testing using createDataPartition.
- **Data Balancing:** Used step_smote() on the training set to balance 2:1 class distribution.
- **Scaling and Centering:** Normalized predictors with step_normalize() for standardization.

Data and Variables

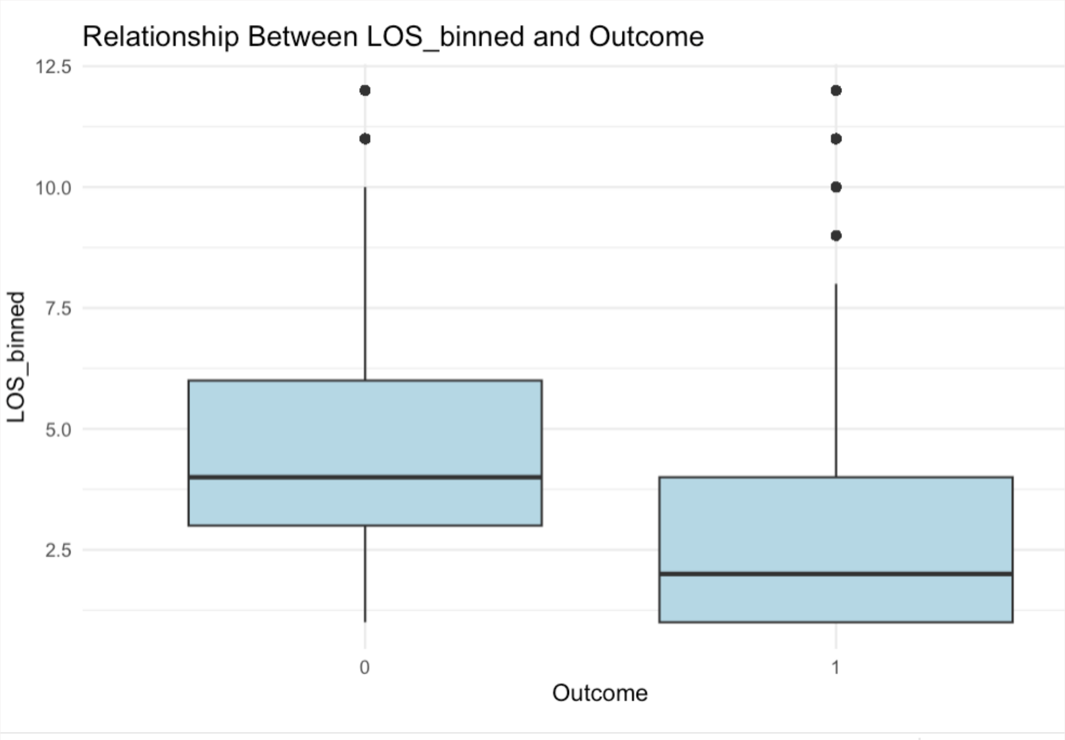
The Random Forest model, which is robust to noise and irrelevant features, was used to determine variable importance.

Inclusion criteria were based on the selecting variables related to the hypothesis and dimension reduction for round 1 of modeling:

Social determinants of health: age, living arrangements, region, and DSM diagnosis.

Substance use severity: frequency of use, age of first use

Choice in treatment: referral source, **length of stay,** participation in substance use self-help programs.



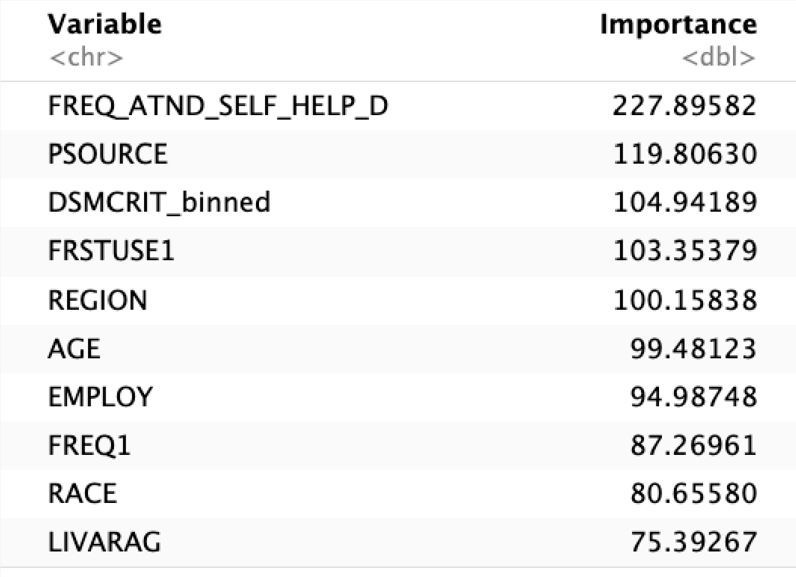
Data and Variables

Round 2 of modeling

Use RF for variable importance with the removal of Length of stay variable

- Frequency of attendance at self - help programs substance use doubled in importance

- A new variable, race, was added as for modelling



Variable	Importance
<chr>	<dbl>
FREQ_ATND_SELF_HELP_D	227.89582
PSOURCE	119.80630
DSMCRIT_binned	104.94189
FRSTUSE1	103.35379
REGION	100.15838
AGE	99.48123
EMPLOY	94.98748
FREQ1	87.26961
RACE	80.65580
LIVARAG	75.39267

Statistical Methods:

Parallel Computing reduces computation time; it divides tasks into smaller sub-stacks, executed simultaneously and aggregated to produce final results.

Random Forest Classification (RF) - ensemble learning technique where multiple decision trees are trained using bagging and final prediction is determined by majority voting across trees.

- 1,000 trees, 3 features per split, 5-node leaf size, max tree depth of 50.

- **Pros:** Reduces overfitting and handles complex, nonlinear relationships.
- **Cons:** Less interpretable than simpler models; computationally expensive.

Extreme Gradient Boosting (XGBoost) - Enhanced gradient boosting algorithm for scalability and efficiency, preventing overfitting through L1 (lasso) and L2 (ridge) regularization.

- Converted to xgb, Max tree depth: 3, Learning rate: 0.05 (slower but better generalization), Row and feature sampling: 70% each, Regularization: L2 (lambda = 2), L1 (alpha = 1), Minimum child node weight: 5, Gamma: 1 (reduce overfitting), 500 boosting rounds with early stopping after 20 non-improving rounds.

- **Pros:** Highly effective for large datasets with customizable, accurate optimization using first and second-order gradient
- **Cons:** Less interpretable than simpler models; computationally expensive.

Statistical Methods:

Lasso Logistic Regression (LR) - Linear model for binary classification, converting log odds into probabilities using the sigmoid function. prediction is determined by majority voting across trees.

- Used glmnet() function; Alpha set to 1 to apply L1 regularization; Cross-validation applied to optimize lambda, controlling the penalty strength on coefficients.

- **Pros:** Simple and interpretable model, effective at reducing overfitting
- **Cons:** struggles with multicollinearity, outliers, and nonlinear relationships.

Neural Networks (NN) - Modeled after the brain, NNs consist of layers of interconnected neurons (input, hidden, and output layers), and uses activation functions and weight optimization to detect complex patterns.

- Hidden layers nodes set 5, decay = 0.2(L2 regularization), training iterations limited to 150 maxit to reduce noise; used balanced training data with classification mode (lineout = false).

- **Pros:** Highly flexible and handles complex, nonlinear relationships, effective for large data
- **Cons:** Lacks transparency in decision-making, reducing interpretability, computationally expensive .

Results

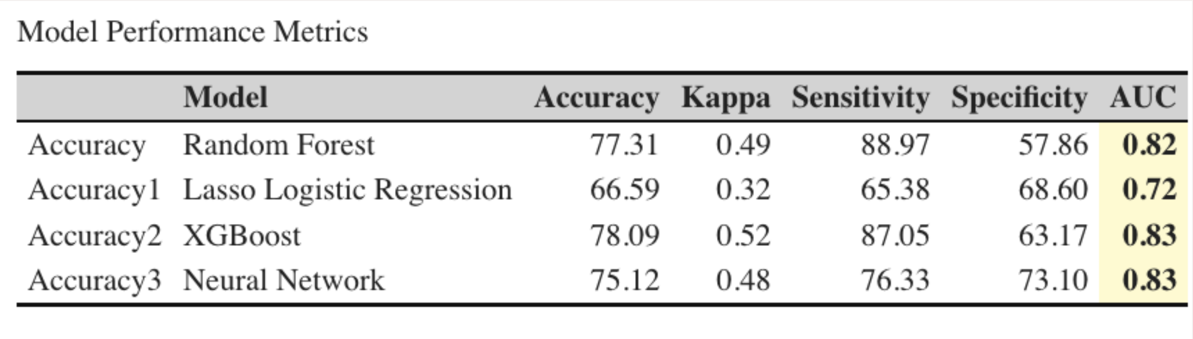
This **first round of modeling** examined how the **length of stay** variable affects the models

RF, XGBoost, and NN performed well

The Length of service variable seems more influential than other features.

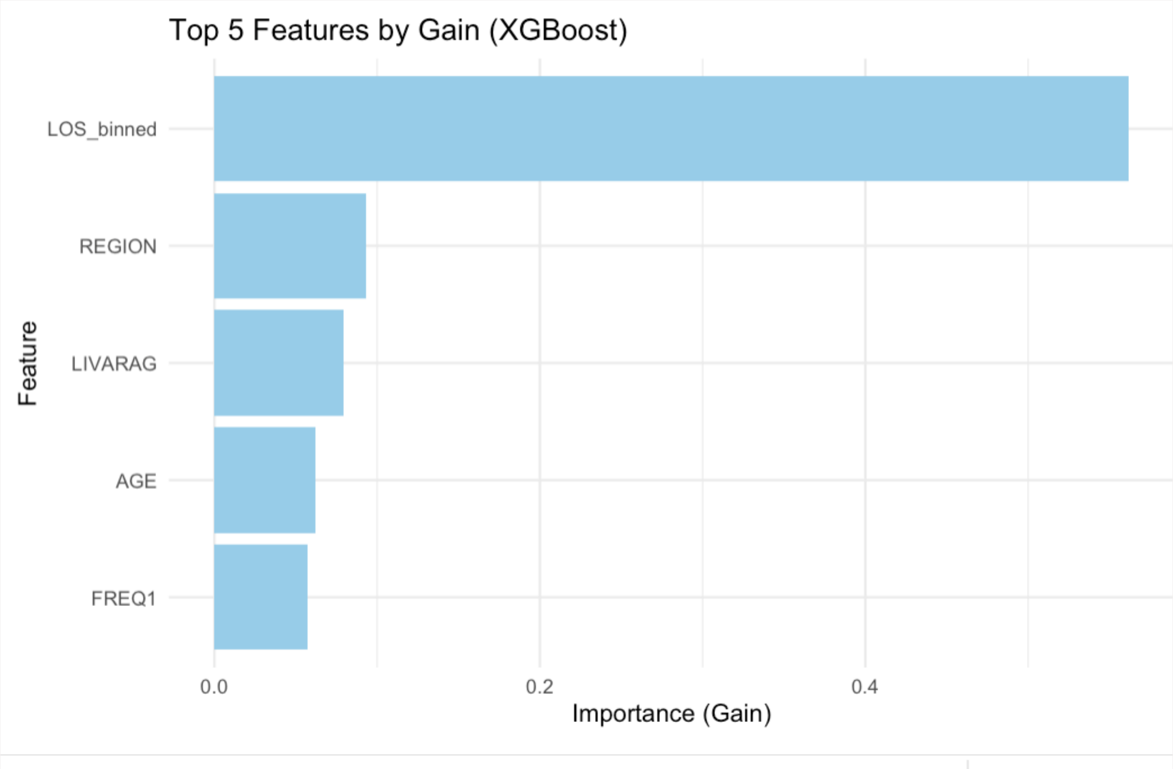
The dominance of this feature might reduce the influence of different variables in the model.

This variable is calculated simultaneously with the treatment outcome; excluding it from the next modeling round is best



Model Performance Metrics

	Model	Accuracy	Kappa	Sensitivity	Specificity	AUC
Accuracy	Random Forest	77.31	0.49	88.97	57.86	**0.82**
Accuracy1	Lasso Logistic Regression	66.59	0.32	65.38	68.60	**0.72**
Accuracy2	XGBoost	78.09	0.52	87.05	63.17	**0.83**
Accuracy3	Neural Network	75.12	0.48	76.33	73.10	**0.83**



Results

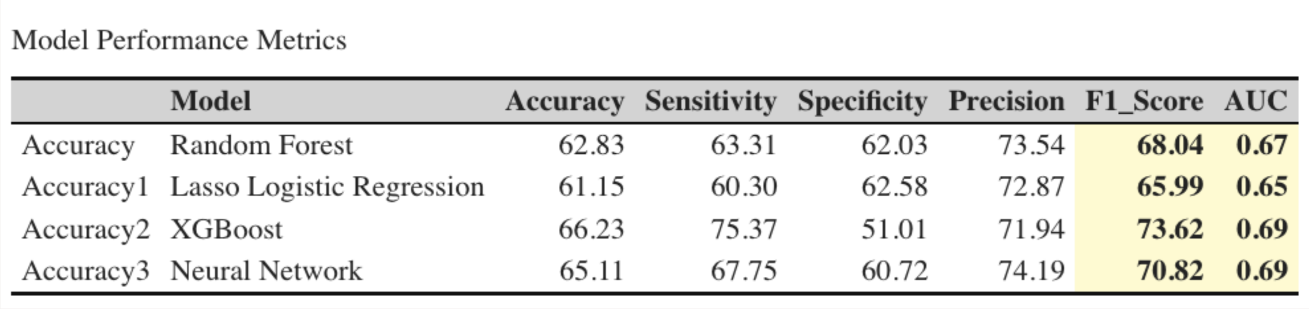
The **second round of modeling** removing the **length of stay** variable

All the models have comparable AUC score.

XGBoost has the highest F1 score of 73.62, highlighting its ability to balance predicting true positives and minimizing false negatives.

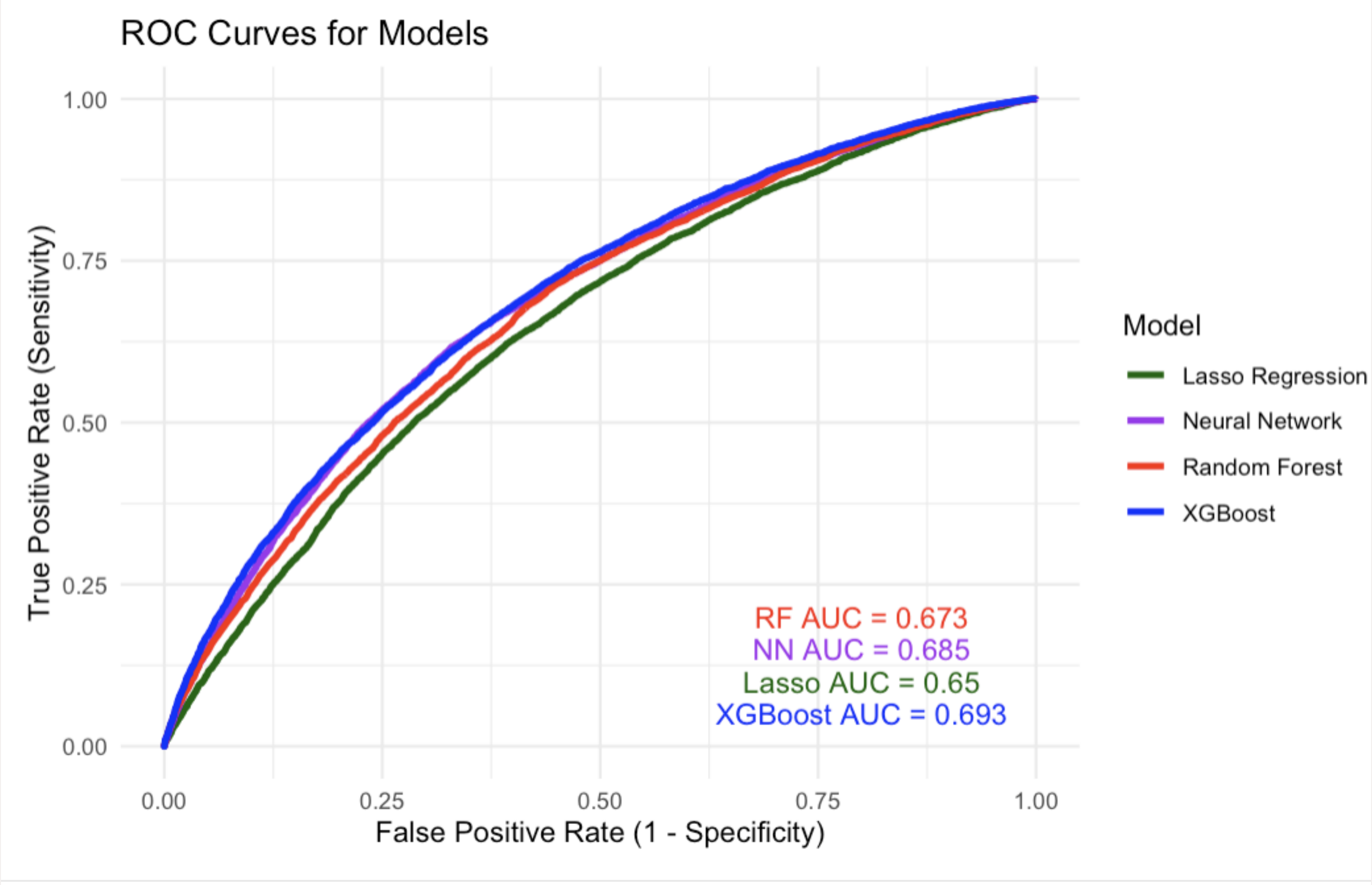
All models have relatively close curves

- The closer the curve is to the top left corner the better the model perform
- XGBoost has a slightly higher AUC metric.



Model Performance Metrics

	Model	Accuracy	Sensitivity	Specificity	Precision	F1_Score	AUC
Accuracy	Random Forest	62.83	63.31	62.03	73.54	**68.04**	**0.67**
Accuracy1	Lasso Logistic Regression	61.15	60.30	62.58	72.87	**65.99**	**0.65**
Accuracy2	XGBoost	66.23	75.37	51.01	71.94	**73.62**	**0.69**
Accuracy3	Neural Network	65.11	67.75	60.72	74.19	**70.82**	**0.69**



Results

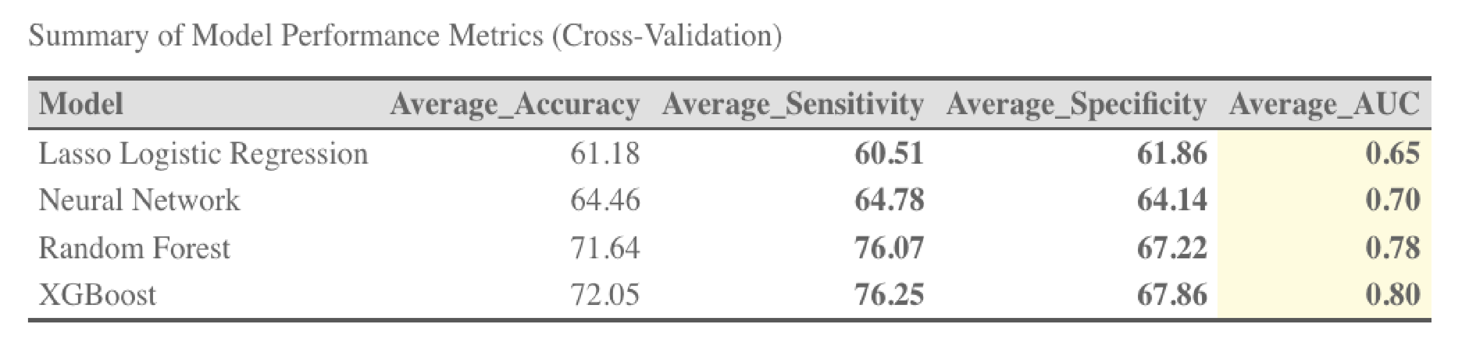
Second round of modeling

Use **cross -validation** to confirm the models' stability and generalization.

- The results show that XGBoost and RF have similar metrics for sensitivity and specificity.
- However, XGBoost's accuracy (72.05) and AUC (.80) suggest outperforming the other models across multiple validation folds.

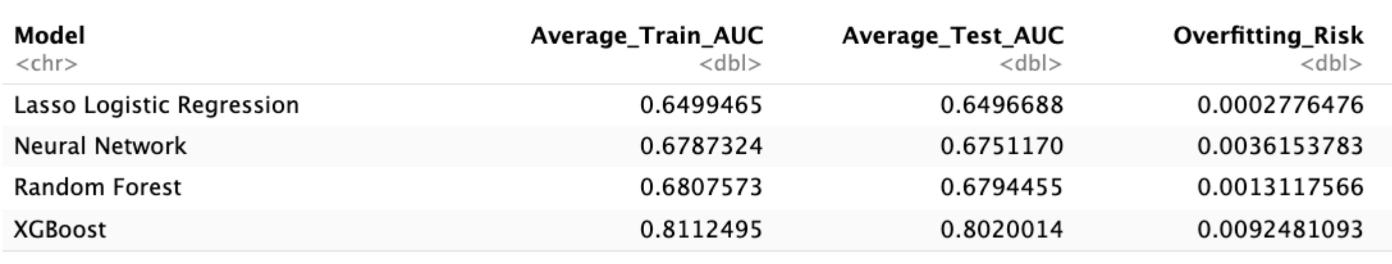
Confirm **internal validity**

- Minimal overfitting is shown in the bottom table, risk, is less than 0.05 for all models,



Summary of Model Performance Metrics (Cross-Validation)

Model	Average_Accuracy	Average_Sensitivity	Average_Specificity	Average_AUC
Lasso Logistic Regression	61.18	**60.51**	**61.86**	**0.65**
Neural Network	64.46	**64.78**	**64.14**	**0.70**
Random Forest	71.64	**76.07**	**67.22**	**0.78**
XGBoost	72.05	**76.25**	**67.86**	**0.80**



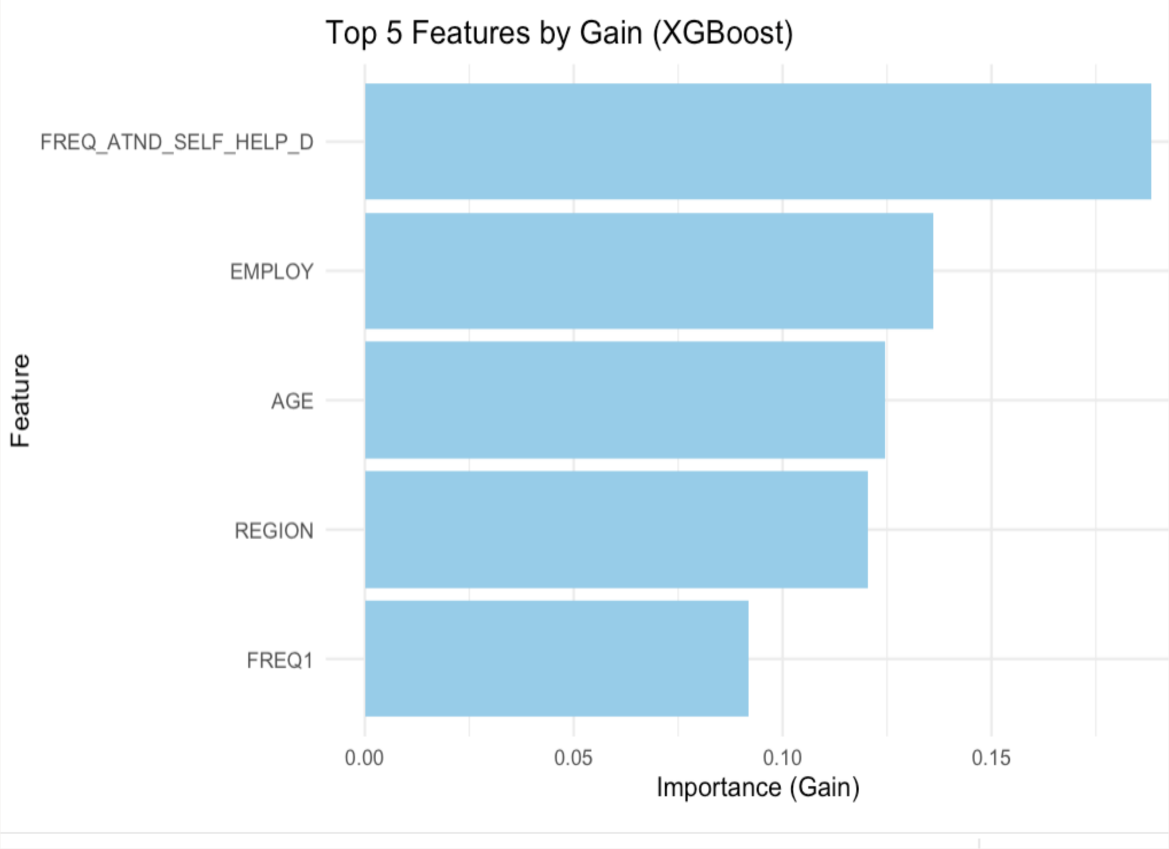
Model <chr>	Average_Train_AUC <dbl>	Average_Test_AUC <dbl>	Overfitting_Risk <dbl>
Lasso Logistic Regression	0.6499465	0.6496688	0.0002776476
Neural Network	0.6787324	0.6751170	0.0036153783
Random Forest	0.6807573	0.6794455	0.0013117566
XGBoost	0.8112495	0.8020014	0.0092481093

Results

Second Round of modelling: XGBoost

The bar chart below shows the top five variables in the XGBoost model

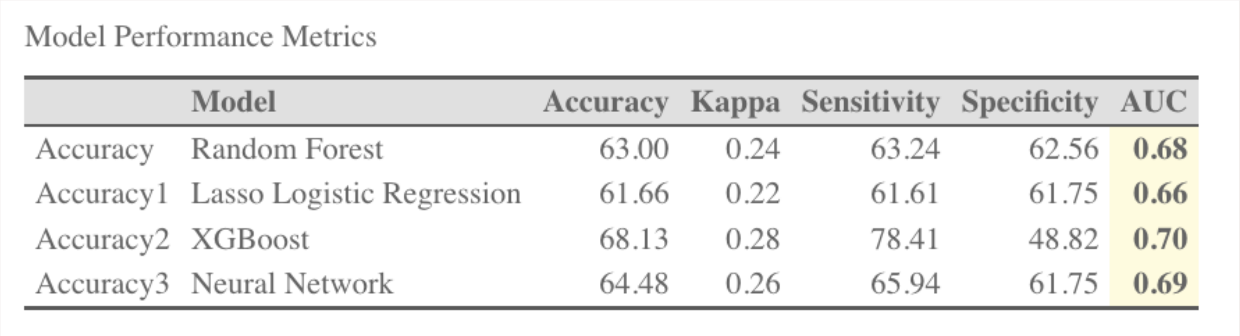
- Uses a gain metric that represents how much a feature contributes to the model's performance.
- The performance metrics on the x-axis are scaled in importance
- the frequency of attendance at self-help programs is the most important
- followed by employment, age, region, and frequency of use.



Results

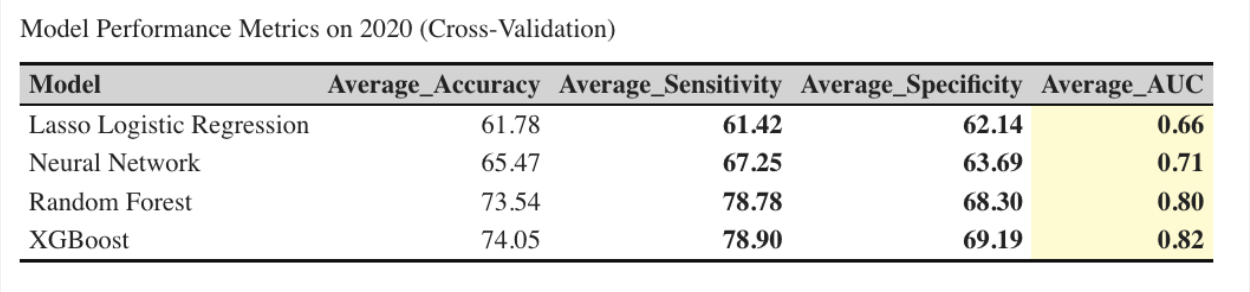
Round 3 of modeling using 2020 data:

- This round aims to prove external validity.
- The same ten predictive variables were used to develop these models.
- The models in round 2 are generalizable to new data from the previous year.
- All models were comparable, with XGBoost slightly outperforming the others with an AUC of 0.70
- CV also confirmed that the XGBoost model's performance was consistent across many folds.



Model Performance Metrics

	Model	Accuracy	Kappa	Sensitivity	Specificity	AUC
Accuracy	Random Forest	63.00	0.24	63.24	62.56	**0.68**
Accuracy1	Lasso Logistic Regression	61.66	0.22	61.61	61.75	**0.66**
Accuracy2	XGBoost	68.13	0.28	78.41	48.82	**0.70**
Accuracy3	Neural Network	64.48	0.26	65.94	61.75	**0.69**



Model Performance Metrics on 2020 (Cross-Validation)

Model	Average_Accuracy	Average_Sensitivity	Average_Specificity	Average_AUC
Lasso Logistic Regression	61.78	**61.42**	**62.14**	**0.66**
Neural Network	65.47	**67.25**	**63.69**	**0.71**
Random Forest	73.54	**78.78**	**68.30**	**0.80**
XGBoost	74.05	**78.90**	**69.19**	**0.82**

Discussion of Results

- Removing the Length of Service variable (from round 1 of modeling) allows the models to rely on other predictive variables; AUC decreased over 10% on the second round of modeling, meaning the models relied too heavily on that LOS variable.
- In their predictive modeling, Gautam and Singh (2020) kept the length of service variable and found it to be the most crucial variable for predicting dropout.
- The XGBoost performed the best in this project, and the results from the 2021 data were validated using the 2020 data (which also found XGBoost to be a top model).
- The top variables found for XGBoost in 2020 were similar to those in the 2021 data. Frequency of attendance at self-help programs was the most important variable, followed by employment, age, and region.
- The results from the 2020 data also showed that referral source was a crucial feature in the modeling; Acion et al. (2017) also found this to be the case.

Discussion of Results

Consistent Predictors Across Studies:

- **Employment Status:** Unemployed individuals had higher dropout rates, consistent with findings by Gautam & Singh (2020), Acion et al. (2017), and Baird et al. (2022).
- **Region:** Rural areas often lack resources, aligning with Gautam & Singh's (2020) findings on regional disparities in outcomes.
- **Age:** Found as a significant predictor in this project and in Gautam & Singh (2020) and Baird et al. (2022).
- **Frequency of Use:** Acion et al. (2017) identified it as a key predictor of treatment outcomes.

XGBoost Usage in Literature: Rarely modeled; Baird, Cheng, and Xia (2022) noted it underperformed compared to their virtual twins model; most studies found **Random Forest (RF)** to perform best, aligning with this project's results.

Inpatient Rehab Treatment Gap: No prior research focused specifically on inpatient rehab settings.

- This study highlights the importance of **self-help program attendance frequency** as a strong predictor, unlike previous studies that only used admission frequency.

Conclusion

Model Comparison: Evaluated Random Forest, Neural Networks, Logistic Regression, and XGBoost.

- **XGBoost performed best**, achieving the highest F1, test AUC (0.70), and cross-validation AUC (0.80), with minimal overfitting (although other models were comparable)
- Results were validated with data from the previous year, confirming external validity.
- **Critical Predictors:**
  - **Social Determinants of Health:** Employment, region, age.
  - **Substance Use Severity:** Frequency of use.
  - **Treatment Choice:** Participation in self-help programs.
- Self-help program attendance identified as the **most significant predictor** in inpatient settings

Conclusion

Implications: Focus efforts on improving treatment outcomes for groups at higher risk of dropout.

- Examples:
  - Increase funding for regions with high dropout rates.
  - Develop age-appropriate interventions.
  - Provide structured care and counseling for individuals with higher use frequency.
  - Encourage participation in **self-help programs** (e.g., Narcotics Anonymous).

Limitations:

- TEDS dataset only includes publicly funded programs and not private treatment data.
- Observations represent admission episodes, not unique individuals.
- No post-treatment data to assess long-term success or abstinence.

Conclusion

Call to Action:

- Addiction remains a public health issue affecting millions.
- Targeted funding and improvements in identified predictors could increase treatment completion rates.
- Societal benefits include reduced healthcare and legal costs.

References*

Acion, L., Kelmansky, D., van der Laan, M., Sahker, E., Jones, D., & Arndt, S. (2017). Use of a machine learning framework to predict substance use disorder treatment success. *PLoS ONE, 12*(4), e0175383. https://doi.org/10.1371/journal.pone.0175383

Andersson, H. W., Steinsbekk, A., Walderhaug, E., Otterholt, E., & Nordfjærn, T. (2018). Predictors of dropout from inpatient substance use treatment: A prospective cohort study. *Substance Abuse: Research and Treatment, 12*, 1–10. https://doi.org/10.1177/1177872718786108

Baird, A., Cheng, Y., & Xia, Y. (2022). Use of machine learning to examine disparities in completion of substance use disorder treatment. *PLoS ONE, 17*(9), e0275054. https://doi.org/10.1371/journal.pone.0275054

Centers for Disease Control and Prevention. (2024, August 12). *Opioid use disorder: Treatment and prevention of opioid overdose.* https://www.cdc.gov/overdose-prevention/treatment/opioid-use-disorder.html

*more listed in full paper