Identifying Key Predictors of High-risk Drug-Overdose Mortality: A Machine Learning Analysis of U.S. Counties

Nandita Vaidyanathan
Claremont Graduate University

The Drug Overdose Crisis

- Drug overdose mortality is one of the leading causes of injuries in the US
- The Drug overdose crisis stems from:
  - Historical opioid over-prescription for pain medication
  - The proliferation of low-cost, high-potency synthetic drugs
  - Increasing accessibility to illicit drugs
- Traditional drug overdose mortality surveillance rely on hindsight data
- Need for forecasting overdose rates, even if it is for the short term, is crucial in directing limited public health resources

Current Research and Gaps

- Studies identified individual-level risk factors affecting risk of overdose mortality that include demographic (eg. older non-hispanic white males with low levels of education) and behavioural factors (eg. regular smoking, history of mental illness)
- Many of these studies use traditional statistical methods to determine causality
- Recent research using ML techniques using individual-level data have demonstrated the value of integrating socioeconomic and criminal justice variables in predictive models
- Need for:
    - Focus on identifying community patterns rather than individual-level analysis for generalizability
    - County-level analysis to account for heterogeneity in US economic, and social structures
    - Exploring non-linear relationships between county-level characteristics and overdose mortality using ML techniques

Study Objectives

- Identify county-level predictors of high-risk drug overdose
- Compare traditional vs Machine Learning approaches
- Provide insights for targeted interventions

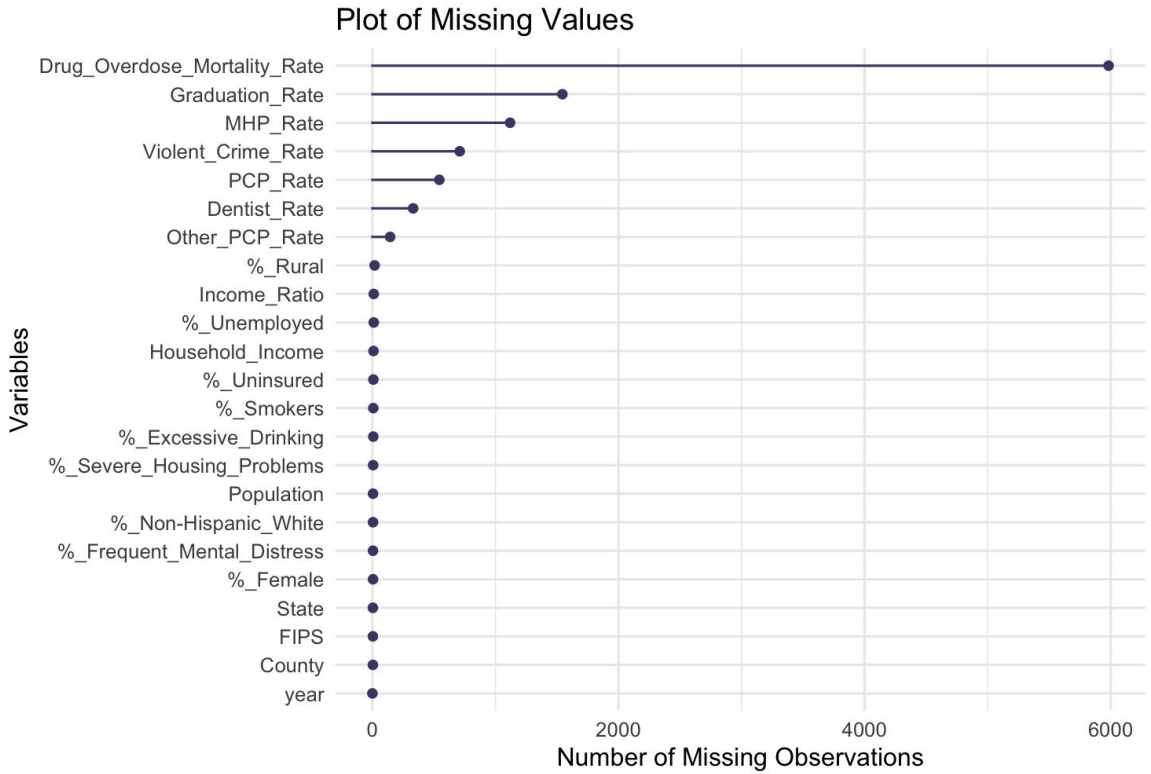
Research Question: Which county-level characteristics best predict high-risk drug overdose mortality in the US?

1. How do health system factors impact drug overdose mortality?
2. What socioeconomic factors are most predictive?
3. How do different ML approaches compare in prediction accuracy?

Data Sources and Variables

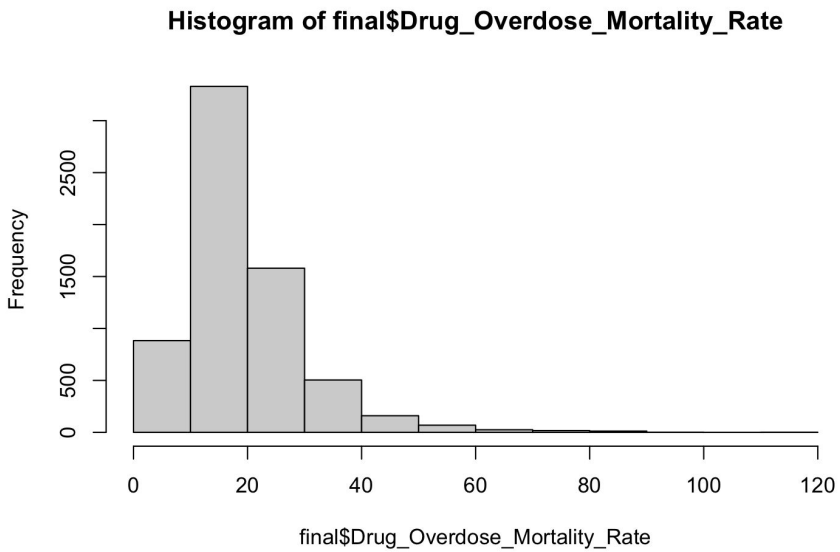
- Data Source (2016-2019)
  - County Health Rankings & Roadmap program from Roberts Wood Johnson Foundation
  - Includes from multiple authoritative sources including the Behavioral Risk Factor Surveillance System, American Community Survey, CDC Wonder Mortality Data etc.
- Categories of variables
  - Health-related and Healthcare Access
  - Community safety
  - Sociodemographic
- Total: 23 variables across 1,856 counties (12,563 observations)

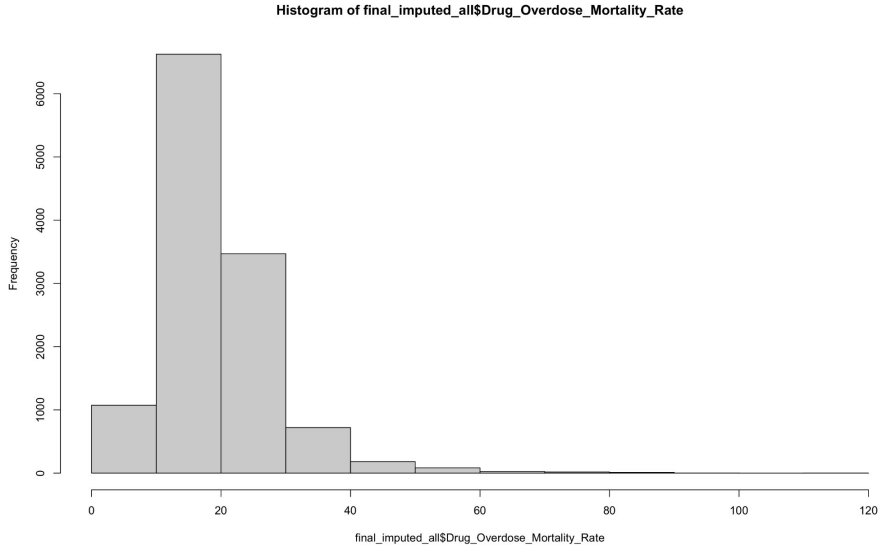
The issue of Missingness and how it was addressed



- 5,983 missing values in mortality rates
- CDC suppression rule (death counts < 10)
- Identified independent variables with missingness

- Tested for Missing Completely at Random (MCAR)
  - Rejected null hypothesis
- Used Multiple Imputation method using MICE package to handle missing values in the independent variables
  - Used predictive mean matching method, a semi-parametric imputation method
  - it might be more appropriate than the regression method since the normality assumption is violated
- For the dependent variable, K-nearest neighbor method
  - It fills in missing values based on the similarity of observations, which is particularly helpful in preserving relationships between neighboring observations in the feature space
  - K was set to 5
  - KNN assumes that observations that are similar in the feature space will have similar outcomes, which is a reasonable assumption in many cases but could introduce bias if the data have a more complex missingness pattern, which I haven't fully accounted for





Distribution of dependent variable before and after imputation using KNN

Methodology

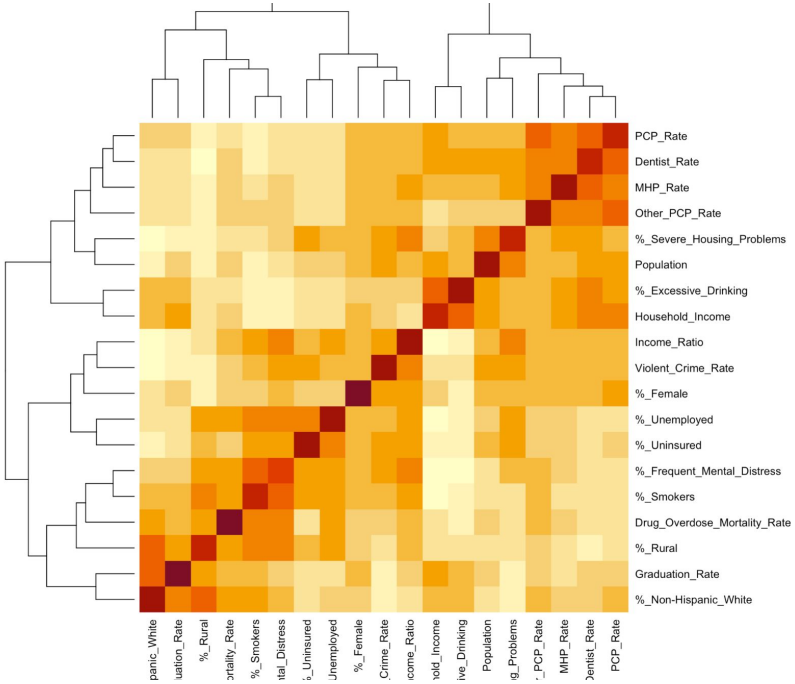
- High risk county is defined as counties with drug overdose mortality rate in the top 25% percentile of drug overdose mortality rate (High risk = 1)
- Set up baseline probit model
    - $P(Y=1|X)= \Phi(\beta0+\beta1(\text{Socio-economic Vulnerability Variables})+\beta2(\text{Health-related and Health Access Variables})+\beta3(\text{Community Safety Variables})+\epsilon)$
- Run LASSO regression
- Run Random Forest
- Run XGBoost
- Model Comparison

Descriptive Statistics

	n	mean	median	max	min	sd	skew	range
percent_Frequent_M	12560	11.83	11.73	22.21	6.60	1.98	0.34	15.61
Drug_Overdose_Mortality_Ra	6582	19.26	16.93	111.54	2.54	10.43	2.04	109.00
percent_Uninsured	12557	16.46	15.74	51.26	2.54	6.98	0.57	48.72
Other_PCP_Rate	12421	68.64	58.55	1433.89	0.00	52.55	6.04	1433.89
Household_Income	12556	49081.27	47168.00	136191.00	21658.00	12802.85	1.41	114533.00
Homicide_Rate	4961	6.18	5.00	44.80	0.60	4.68	2.40	44.20
percent_Non_Hispar	12560	76.70	84.08	98.61	2.76	20.01	-1.20	95.85
percent_Female	12560	49.91	50.35	57.00	26.57	2.27	-3.18	30.43
percent_Rural	12547	58.59	59.49	100.00	0.00	31.48	-0.16	100.00
Population	12560	102635.30	25770.00	10170292.00	86.00	329754.50	13.90	10170206.00
percent_Smokers	12558	18.01	17.43	42.75	6.55	3.67	0.86	36.21
percent_Excessive_Drinking	12558	17.00	17.00	29.44	8.40	3.27	0.15	21.04
PCP_Rate	12021	55.07	49.29	477.22	0.00	35.00	2.30	477.22
Dentist_Rate	12234	43.92	39.45	725.13	0.00	29.62	3.43	725.13
MHP_Rate	11446	135.90	93.78	2002.66	0.00	148.81	3.29	2002.66
Graduation_Rate	11021	86.29	87.70	100.00	2.50	8.35	-1.58	97.50
percent_Unemployed	12554	5.42	5.09	24.01	1.19	2.06	1.56	22.82
Income_Ratio	12554	4.52	4.42	10.66	2.56	0.73	1.18	8.10
Violent_Crime_Rate	11855	247.88	199.79	1885.30	0.00	192.46	1.92	1885.30
percent_Severe_Housing_Pro	12559	14.41	13.92	71.26	0.99	4.77	1.88	70.27

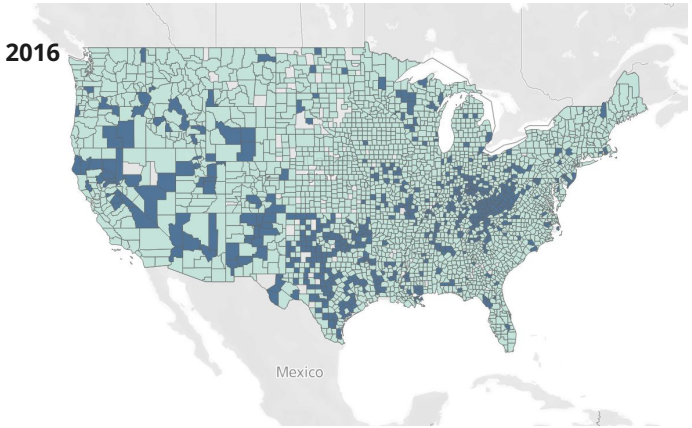
- Drug overdose mortality rates: Mean 19.3, max 111.5 per 100,000
- Mental health providers range: 0 to 2000+ per 100,000, showing severe disparities
- Demographics: 77% Non-Hispanic White mean, 59% rural population
- Economic factors: Income range $22k-$136k, mean unemployment 5.4%

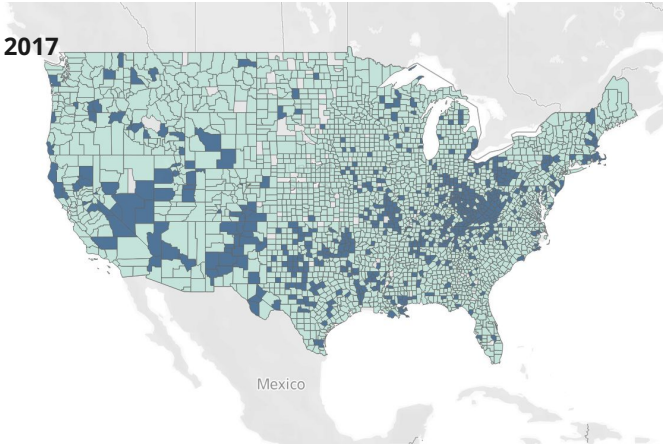
Correlation Heatmap

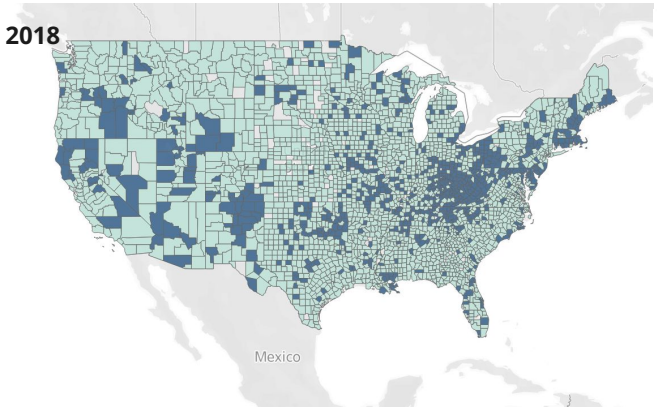


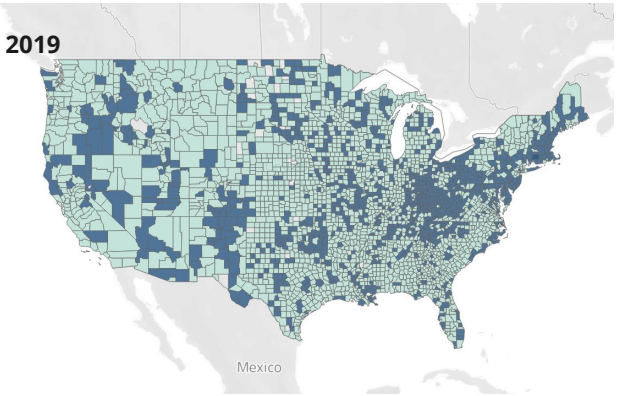
- Strongest correlation appears among the healthcare provider metrics, suggesting these resources tend to cluster together in certain counties.
- HH income is correlated with the healthcare provider rates, suggesting that counties with higher income levels tend to have better access to healthcare resources across all provider types
- HH income has a strong positive correlation with excessive drinking
- Interestingly, drug overdose mortality rates show positive correlations with smoking rates and frequent mental distress , while having a negative correlation with excessive drinking

County-level distribution of High Risk Drug Overdose Mortality Rate



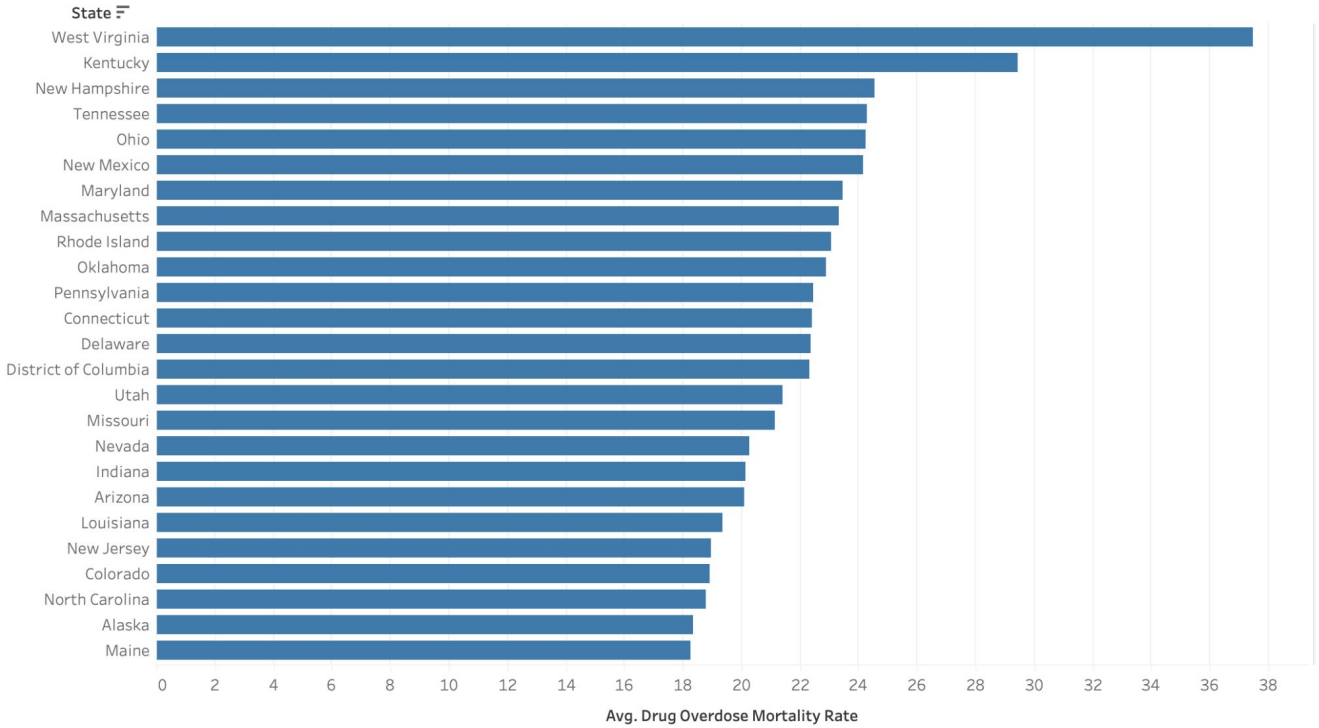






- We see a clear temporal trend– from 2016, the national average increased from 17.4 deaths per 100,000 to 21.0 in 2019.
- There are consistent geographic patterns, with rates concentrated in Appalachian regions, particularly West Virginia
- Rural eastern counties show consistently high risk, notable spots in New Mexico (Rio Arriba county)
- We see an evolution from rural to urban county spread in 2019
- Persistent state-level disparities in West Virginia, Kentucky, and Ohio

Top States with Drug Overdose Mortality Rate



Results

Variable	Coefficient	P-value	AME
% Frequent Mental Distress	0.225***	3.77e-11	0.063
% Uninsured	0.088***	3.23e-05	0.025
Other PCP Rate	0.038**	0.012	0.011
Household Income	0.027	0.262	0.008
% Non-Hispanic White	0.369***	¡2e-16	0.104
% Female	-0.073***	5.32e-07	-0.021
% Rural	0.079***	1.34e-04	0.022
Population	0.190***	¡2e-16	0.053
% Smokers	-0.022	0.375	-0.006
% Excessive Drinking	-0.173***	¡2e-16	-0.049
PCP Rate	-0.073***	5.89e-05	-0.021
Dentist Rate	0.022	0.179	0.006
MHP Rate	0.073***	3.19e-05	0.021
Graduation Rate	0.068***	2.16e-05	0.019
% Unemployed	0.135***	1.78e-13	0.038
Income Ratio	0.106***	4.71e-10	0.030
Violent Crime Rate	-0.013	0.402	-0.004
% Severe Housing Problems	-0.101***	3.29e-07	-0.028
Year	0.191***	¡2e-16	0.054

Notes: *** p¡0.01, ** p¡0.05, * p¡0.1

Probit model key results:

Strongest predictors (marginal effects):

- Non-Hispanic White: +0.104
- Frequent Mental distress: +0.063
- Population size: +0.053
- Year effect: +0.053

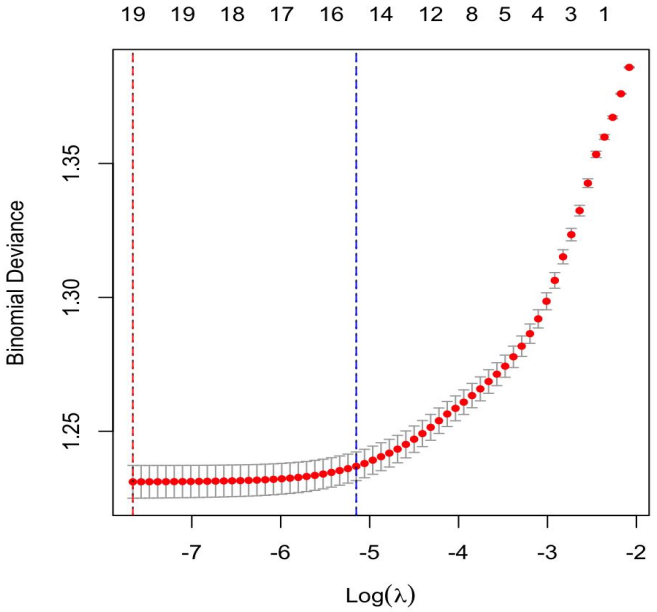
Healthcare effects:

- Primary care physicians: -0.021
- Mental health providers: +0.021
- Excessive drinking: -0.049 (protective)

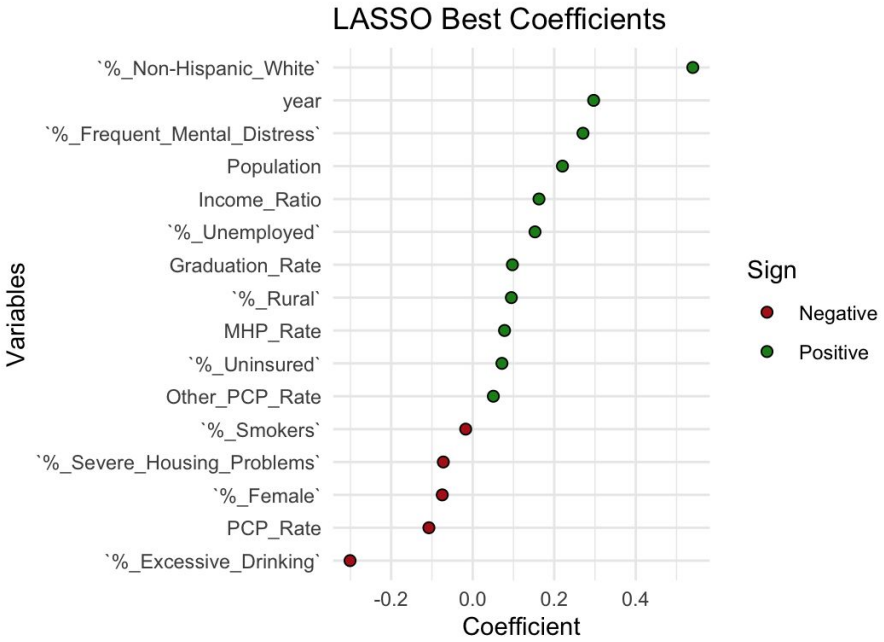
Socioeconomic impacts:

- Unemployment: +0.038
- Income inequality: +0.030
- Rural population: +0.022

Non-significant: dentist rates, smoking rates, violent crime



- Cross-validation (k=10):
  - Very low lambda values
  - Most variables highly informative
  - Minimal regularization needed
- Model comparison:
  - Lambda.min (red line): retained all variables
  - Lambda.1se (blue line): dropped 3 variables
    - Household Income
    - Dentist Rate
    - Violent Crime Rate



Based on the LASSO model with lambda at lambda.1se:

Strongest positive associations:

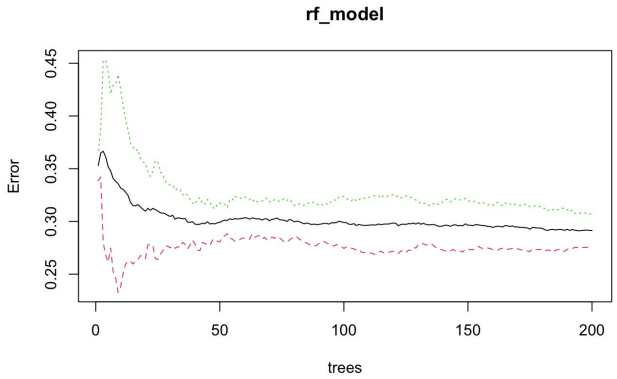
- Non-Hispanic White (+0.4)
- Frequent Mental Distress (+0.3)
- Population size (+0.2)
- Income Ratio (+0.2)

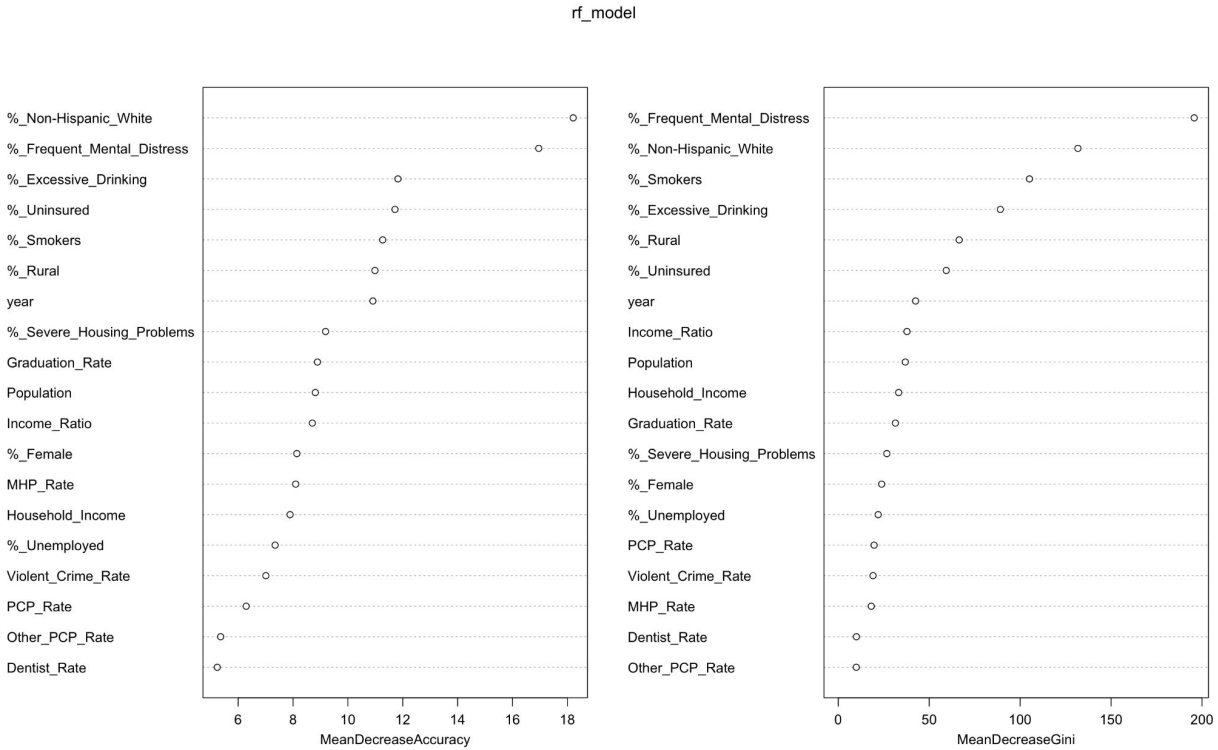
Negative associations:

- Excessive drinking (-0.2)
- Primary care physicians (-0.15)
- Female percentage (-0.1)

Random Forest

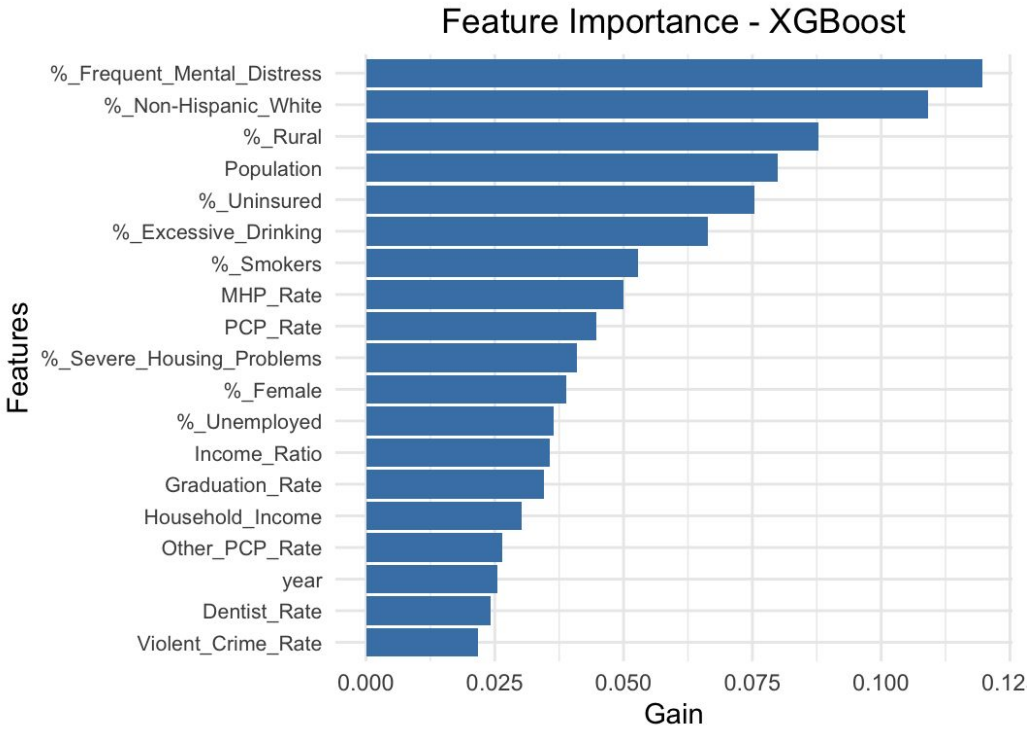
- RF is good at handling complex relationships and non-linearity
- Default hyperparameters showed robust performance
    - Maximum leaf-nodes set to 20
    - Minimum node-size set to 1
    - The error rate stabilises after adding about 150 trees





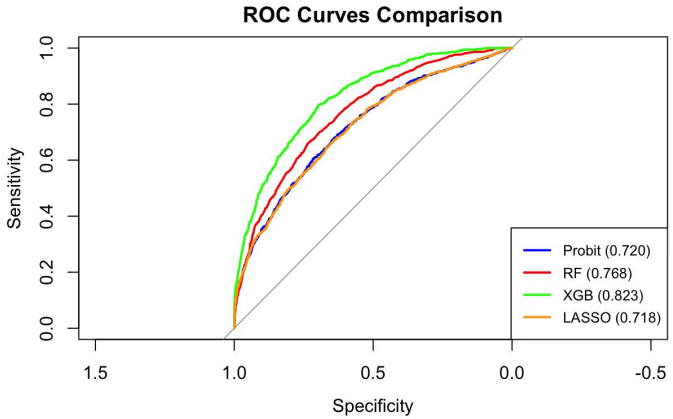
- The Mean Decrease Accuracy measure, shown in the left panel, indicates how model accuracy declines when each variable is randomly permuted.
- It confirms the prominence of the key demographic and mental health variables, while also highlighting the relatively lower importance of healthcare supply measures (Dentist Rate and Other PCP Rate)

XGBoost

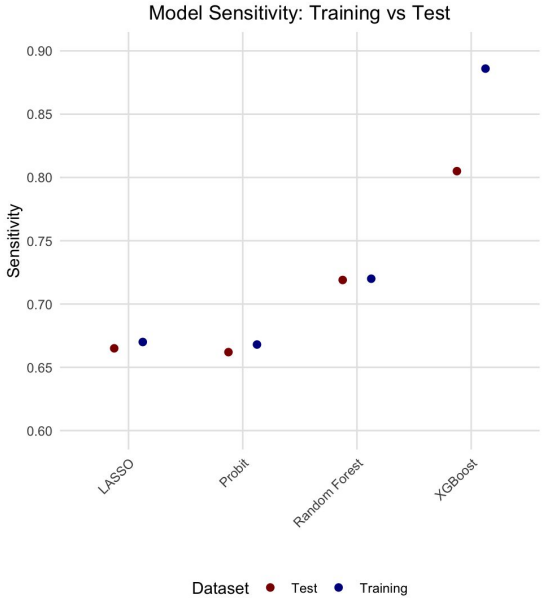


- EXtreme Gradient Boosting is characterised by its sequential approach in learning from previous errors and builds trees iteratively
- Conservative settings for generalization:
  - Learning rate: 0.1 (lower than default)
  - Max tree depth: 5
  - Min child weight: 2
  - Early stopping: 20 rounds (5-fold cross-validation)
- Feature importance plot shows gain in accuracy of the model brought by the features to the model
- Frequent mental distress emerges as the top variable suggesting it is relatively more important for prediction, closely followed by other demographic variables, similar to previous models

Model Comparison



The ROC curves illustrate the superior discriminative ability of the ensemble methods compared to the linear models. We can see that the ROC curves of the linear models are overlapping, while we see a clear dominance of the XGBoost model.



- Sensitivity measures our ability to correctly identify high risk areas
- XGBoost has the highest sensitivity score, however, has the largest test-train gap
- Random Forest model has the most stable performance for sensitivity analysis, while the traditional models show consistent but low performance

Model	Accuracy	Sensitivity	Specificity	Precision	NPV	F1_Score
Probit Model (Train)	0.6631	0.6721	0.6542	0.6602	0.6661	0.6661
Probit (Test)	0.6614	0.6672	0.6439	0.8505	0.3891	0.7478
Lasso (Train)	0.6610	0.6744	0.6476	0.6568	0.6654	0.6655
Lasso (Test)	0.6606	0.6653	0.6461	0.8510	0.3886	0.7468
Random Forest (Train)	0.7220	0.7283	0.7158	0.7193	0.7249	0.7238
Random Forest (Test)	0.7073	0.7187	0.6725	0.8696	0.4404	0.7870
XGBoost (Train)	0.8862	0.8617	0.9106	0.9060	0.8682	0.8833
XGBoost (Test)	0.7687	0.8047	0.6593	0.8777	0.5264	0.8396

- The Random Forest and XGBoost demonstrated superior overall performance compared to the traditional linear models, Probit and LASSO
- Although XGBoost achieved higher accuracy and sensitivity on the test set, it showed a notable drop in specificity, suggesting potential overfitting despite its high training performance
- Random Forest's more balanced specificity suggests more reliable overall predictions

Key Findings

- Demographics
  - Non-Hispanic White percentage emerged as #1 predictor
    - Reflects historical prescribing disparities
    - Highlights systemic healthcare access issues
  - Rural communities particularly vulnerable
- Health-related and Healthcare Access
  - Frequent Mental distress: consistently top predictor across all models
    - Strong link to substance abuse disorders
    - Creates cycle of vulnerability
  - The Healthcare Paradox:
    - Fewer primary care doctors = higher risk
    - More mental health providers linked to higher risk
      - Likely reflects reactive resource allocation
      - Better reporting in these areas
  - Protective effect of excessive drinking
  - No impact from smoking rate
- Results suggest need for targeted interventions focusing on mental health integration and rural healthcare access
  - Strengthen integration of mental health and substance use treatment services in high-risk counties
  - Develop early intervention programs in counties with high mental distress rates
  - Focus on creating innovative healthcare delivery models for rural counties

Limitations and Future Directions

- ML models can be improved using extensive hyperparameter tuning and cross-validation strategies
- Exploring different thresholds for high-risk and evaluate impact on predictions
- Incorporating spatial analysis to account for spatial autocorrelation and examine regional clusters
- Exploring complex ML techniques, like Deep Neural Networks, which has been established as the strongest ML technique for overdose risk-stratification by recent studies