Predicting the Opioid Crisis in American Communities

STAT 471
Jake Smolenski, Benjamin Liu, Arjun Govind
May 5, 2019

Executive Summary

History

Over the past 10 years, the United States has seen the rate of drug-related deaths skyrocket. Fueled by Purdue Pharma's infamous OxyContin, the rate of opioid-related deaths has led the charge. OxyContin, whose active ingredient is Oxycodone, is twice as powerful as morphine. The drug releases tremendous amounts of dopamine into the body and overloads the opiate receptors, effectively blocking all pain and giving immense pleasure. As the painkillers became widely prescribed in part due to aggressive and misleading advertising by big pharmaceutical companies, addiction took hold in many patients and areas. After doctors cut off prescriptions for those who had recovered from their medical issue, many turned instead to the cheaper and more powerful street drug, heroin, to experience the feeling that they got on OxyContin. With the sudden spike in demand, the heroin supply could not handle the magnitude of new customers, and many dealers turned to methods to "cut" their drugs. This technique involves introducing foreign substances, sometimes psychoactive, sometimes not, into the supply, effectively increasing the volume that the seller can distribute. In order to keep their "product" high quality, dealers had, and continue, to use substances with similar intensity and effect. Here is where the drug fentanyl enters the picture. Massive quantities of this drug, shipped from overseas, flooded the United States. Fentanyl, which is 80-100 times as powerful as morphine and has no medical use, is lethal at almost any non-miniscule dose. Infusing heroin with fentanyl, if dosed right, gives the user a high that is intensely addictive and/or lethal. This past year, fentanyl has become the leading killer in the country, rising rapidly since its introduction. With this epidemic, government resources have been utilized to try to stem the problem, with little success. Cities have deployed task forces to limit the entrance of the drug, but have failed, and have begun to stock stations with clean needles, Narcan (a treatment drug), and have installed dirty needle collection kiosks to mitigate the non-drug issues that are related to this epidemic.

Landscape Today

According to the National Institute on Drug Abuse, opioid overdoses result in more than 130 deaths in the United States every day, and American life expectancy is declining at an alarming rate from a combination of drug overdoses and increased suicide. Communities all across the nation are feeling the effects of the crisis which is straining public health, criminal justice, law enforcement, and social work resources, costing the country an estimated $78.5 billion per year. Our project was inspired by the 2019 Silfen Forum where leading political and medical figures including former Vice President Joe Biden, former Governor Jeb Bush, President Amy Gutmann, Mayor James Kenney of Philadelphia, Professor Bertha Madras, and Professor Jeanmarie Perrone discussed solutions to tackle the issue nationwide. In addition, one of our group members frequents a gym in the Philadelphia Kensington neighborhood where the effects of the opioid epidemic can be seen firsthand and have opened his eyes to the problem that the country faces. We think that some of the causes of the opioid crisis is socioeconomic and demographic; as a result, we strive to identify important factors associated with higher drug-related deaths in hopes that understanding where resources can be deployed to alter conditions and potentially deter people from opioids. Because the dataset used does not have all counties in the United States, this project can be applied to communities outside of our initial data set.

Our data is taken from the Centers for Disease Control WONDER database regarding drug-related deaths from 2012-2017. We merged this dataset with American Community Survey data from the Census Bureau for the same time period. Our final data set has 3,774 observations of 43 variables which include information about ethnicity, gender, income, households, infrastructure, and of course, drug related deaths. We classified all counties with drug-related deaths in the upper 25th percentile as "Needs Resources" which is our response variable.

We conducted exploratory data analysis on a variety of interesting variables across the entire country, plotting their quantitative distributions as well as their geographic distributions. The county with the highest rate of drug-related deaths is Cabell County, West Virginia with more than 150 drug-related deaths per 100,000 people. The county with the lowest rate of drug-related deaths is Hidalgo County, Texas with just around 3 drug-related deaths per 100,000 people. The distribution of economic-based variables such as household income, unemployment, and food stamps is worst in the Rust Belt area of Ohio, Pennsylvania, and West Virginia while the highest home values, percent with both parents, and income can be found in the Northeast and Pacific Northwest.

To determine which predictors are most valuable to determine drug-related deaths, we conducted a variety of data analysis methods including an elastic net, simple tree, random forest, and boosting. We evaluated the misclassification error for each method with a penalty ratio of 1.5:1 for false negatives to false positives. We concluded that the best model was the elastic net model which produced a validation weighted misclassification error of 0.2638 and testing weighted misclassification error of 0.3432. The most important predictors revolved around economic, infrastructure, and social factors.

While we are confident in our conclusions and the predictors of opioid-related deaths in America, we are cognizant that our conclusions are inherently limited by the underlying data of which we have made a number of important assumptions. We assume that the selection of ACS counties is not confounded with any demographic or drug-related deaths variables. In addition, because it is nearly impossible to disentangle specific opioid-related death data on a county level from other drug-related deaths such as cocaine or methamphetamine, we assume that the proportion of opioid related deaths of all drug related deaths are constant for each county which may not ultimately be accurate. In spite of these limitations, we are confident that our classifier does a quality job of identifying communities in need of resources because the predictors make strong intuitive sense.

Data Analysis

Our data was combined from the CDC WONDER database as well as US Census data, queried from the tidyCensus API. We used the CDC Wonder database to gather drug-related death data for counties across the United States and the US Census American Community Survey to capture demographic data about those counties. Demographic data included data on ethnicity, poverty, infrastructure, veterans, households, etc. Our original merged data set had 4,362 observations of 49 variables. Our data spanned from 2012-2017, and each county-year combination represented a new observation in our data. The number of counties in each American Community Survey differed every year, so the number and list of counties from 2012 to 2017 may not be constant. We opted to split by year and county because we wanted to investigate the progression of the opioid epidemic over time and if there were associated demographic changes. Importantly, not every county in the United States is included in the CDC WONDER database or the American Community Survey. The American Community Survey only samples counties with more than 65,000 residents; while this may introduce some bias in the data, it is the best we can do to have updated data. The last Census was only run in 2010 and would not be useful. In addition, we believe that the problem of opioids and drug related deaths does not vary between more populated counties and less

populated counties. Lastly, since some areas draw county lines based on population rather than geography, we think this still has the potential to capture data on a large segment of the population and regions in America.

After merging our data sets together, we had to perform some basic data cleaning. Firstly, because the American Community Survey variables are output as absolute numbers, they are obviously inflated or deflated by the population (or sub-population) of each county. As such, we normalized all of these demographic indicators by dividing by the total county population (or sub-population) to generate proportions. We also transformed the building age variable in the ACS data set; originally, ACS provided the median year that buildings were constructed, and we subtracted the year from 2018 to find the age of each building which we felt was more descriptive of the variable. In addition, we created a variable named "Years since Oxycontin" to identify the chronological effect of the introduction of OxyContin in the market on communities. The development and widespread prescription of OxyContin was the precursor to today's drug epidemic, and the variable "Years since Oxycontin", calculated as the data year minus the year Oxycontin was developed explores if there is a trend or cycle in communities related to the time since OxyContin was introduced. We also created a predictor named households to explore how many households were in each county; this was calculated as the population divided by the average household size. To explore the effect of housing vacancies and other deserted buildings, we also created a total housing to households ratio. Economically depleted areas would theoretically have more vacant residential buildings. Moreover, many counties did not report overall population data, and there were many NAs in this column. We reconstructed the population variable for observations with NAs by summing population male and population female, which were more widely reported.
Although we were able to correct for the NAs in the population variable, there were still other variables that had isolated NA issues, including ethnicity variables such as pct.asian, household variables such as percent without a phone, and percent of multifamily units. Because there was no way to re-calculate this data or extrapolate it from other counties in the data (every county is different), we opted to remove rows with NAs from the data. We felt that this was a reasonable assumption because 1) we have no reason to believe that rows with NAs have a higher or lower rate of drug related deaths and 2) the number of rows removed with NA was still relatively low compared to the total number of rows in the data. After removing rows with NAs, we also removed other variables that we felt had little to no predictive ability including the county name (not possible to extrapolate to new counties outside the data set), percent of structured housing (no variation in the data), and average household income (already represented by median household income). After this data cleaning, we were left with 3,774 observations of 41 predictor variables.

Although the natural response variable with the initial CDC WONDER data was the rate of drug-related deaths per 100,000, we opted to convert this into a binary variable "Needs Resources". In essence, counties with a higher rate of drug-related deaths were considered to "Needs Resources" from the federal government or other policymakers. We chose to use a binary response variable because we felt this would be more practical for policymakers. Whether there are 100 or 110 drug related deaths is not important since both are bad; rather, policymakers need to know where to develop public health and social support infrastructure to stem the opioid epidemic. The literature suggests that somewhere between 60% and 75% of all deaths nationwide in recent years were related to opioids, but it is hard to identify the exact number of opioid-related deaths on a county level. Moreover, the proportion of opioid-related deaths changes year to year and has especially become more prevalent in recent years. In addition, the estimates are also not officially certified by federal health organizations. Because of these uncertainties which we felt could bias our data analysis, we chose to assume that all were opioid-related, and we felt this was reasonable because counties with a higher overall rate of drug-related deaths also probably had the highest rates of

opioid-related deaths. We constructed our binary response variable, "Needs Resources", by identifying the top 25% of counties in every year in drug-related deaths and classified them as "Needs Resources". The remainder 75% were classified as 0. We chose 25% because we felt that federal government resources could realistically only be applied to the top 25% neediest areas and that selecting areas less affected would produce less informative and impactful results.

Our 41 predictor variables in the final data set used for statistical analysis include demographic based variables (such as percent of single mom, percent insured, percent in high school), economic based variables (such as median household income, percent of people on food stamps, percent without a phone, median building age, median home value), and ethnicity based variables (such as percent white, percent black, etc.)

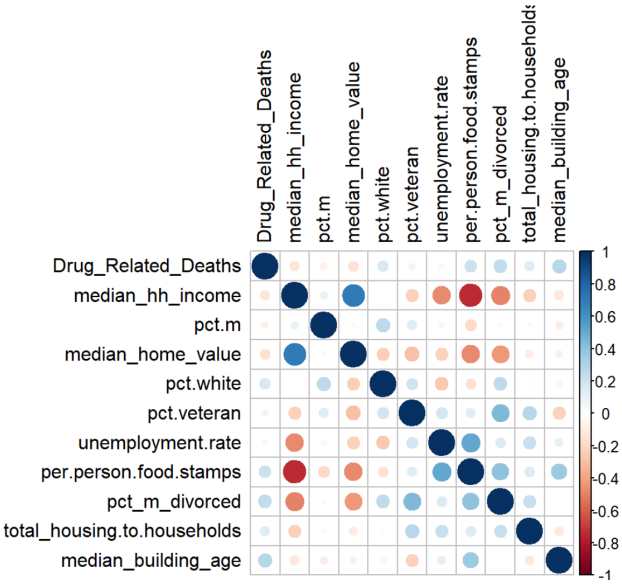
Variable Definitions

Variable	Full Name	Description
County.Code	County Code	Geographic Code (Used only to split data)
Population	Population	Population of the County, ACS only has data for counties with more than 65,000 people
median_hh_income	Median Household Income	Median Household Income, in Dollars, of the County
avg_fam_inc	Average Family Income	Average Income per Family, in Dollars, of the County
average_household_size	Average Household Size	Average People Per Household in the County
households	Households	Number of Households in the County
pct_m_divorced	Percent of Males Divorced	Percentage of Males in the County that are Divorced
pct_m_nvr_married	Percent of Males Never Married	Percentage of Males in the County that have never Married
pct_f_divorced	Percent of Females Divorced	Percentage of Females in the County that are Divorced
pct_f_nvr_married	Percent of Females Never Married	Percentage of Females in the County that have never Married
pct_hh_less_than_10k	Percentage of Households Less Than 10,000	Percentage of Households with Household Income below $10,000 in the County
median_fam_inc	Median Family Income	Median Income per Family, in Dollars, of the County
per_cap_inc	Income Per Capita	Income Per Person in the County
total_housing.to.households	Total Housing To Households	Total Livable Residences Divided by Total Households of People in the County
median_home_value	Median Home Value	Median Home Value, in Dollars, in the County
median_rent	Median Rent	Median Rent, in Dollars, in the County
median_building_age	Medain Building Age	Median Building Age in Years (base year: 2018)
pct.black	Percent Black	Percentage of the Population that is Black or African American
pct.white	Percent White	Percentage of the Population that is White
pct.asian	Percent Asian	Percentage of the Population that is Asian
pct.multi	Percent Multiracial	Percentage of the Population that is Two or More Races
pct.lack.full.kitchen	Percent Lacking Full Kitchen	Percentage of Residences in the County that Lack Full Kitchen
pct.lack.full.plumb	Percent Lacking Full Plumbing	Percentage of Residences in the County that Lack Full Plumbing Services
pct.no.phone	Percent without a Phone	Percentage of Residences in the County that Lack a Phone
pct.50plus	Percent 50 Plus	Percentage of Structures in the County that Contain 50 or More Residences
pct.occupied	Percent Occupied	Percentage of Residences Occupied
pct.veteran	Percent Veteran	Percentage of Population that are Military Veterans
pct.enrolled	Percent Enrolled	Percentage of the Population that is Enrolled in Some Type of School
pct.enrolled.hs	Percent Enrolled in High School	Percentage of the Population that is Enrolled in High School
pct.m	Percent Male	Percentage of the Population that is Male
pct.single.dad	Percent Single Dad	Percentage of the Population that are Single Fathers
pct.single.mom	Percent Single Mom	Percentage of the Population that are Single Mothers
pct.single.parent	Percent Single Parent	Percentage of the Population that are Single Parents
pct.white.pov	Percent White Poverty	Percentage of the Population that is White and Below the Poverty Line
pct.both.parents	Percent Both Parents	Percentage of the Population that are in Households with Both Parents
labor.participation	Labor Participation Rate	Percentage of the Population that is in the Labor Force
unemployment.rate	Unemployment Rate	Percentage of the Population that is Unemployed
pct.unemployment	Percent Unemployment	Percentage of the Labor Force that is Unemployed
pct.insured	Percent Insured	Percentage of the Population that has Any Type of Health Insurance
per.person.food.stamps	Per Person Food Stamps	Percentage of the Population that receives Food Stamps
years_since_OxyContin	Years Since OxyContin Introduced	Years Since Purdue Pharma Brought OxyContin to Market (Base Year: 1996)

To explore the data, we created correlation maps, pairwise plots, histograms/scatterplots, and identified extreme values. First, we created a correlation map of variables we thought were most informative and important in line with our hypothesis and news media that the opioid epidemic was concentrated among young, white males in towns with depressed economic opportunities: Drug Related Deaths, median household income, percent male, median home value, percent white, percent veteran, unemployment rate, food stamps per person, percent of divorced male, total housing to households, and median building age. The correlation map does not show any

associations that pop out immediately, but there are still some interesting associations between variables. For example, median household income is negatively correlated with unemployment rate (-0.49), food stamps (-0.74), and percent of divorced males (-0.50). This makes sense since areas with lower household income have more unemployed citizens who don't earn income and need more food stamps. In addition, more divorced males indicate that more families only have 1 income earner instead of 2 which would increase household income. In addition, the percent of males divorced is associated positively (0.44) with percent of veteran, which indicates the difficulty of veterans re-integrating back into civilian life.

Correlation Map

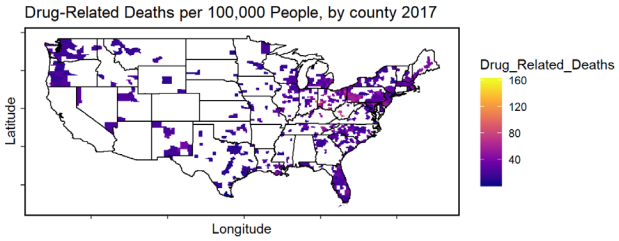


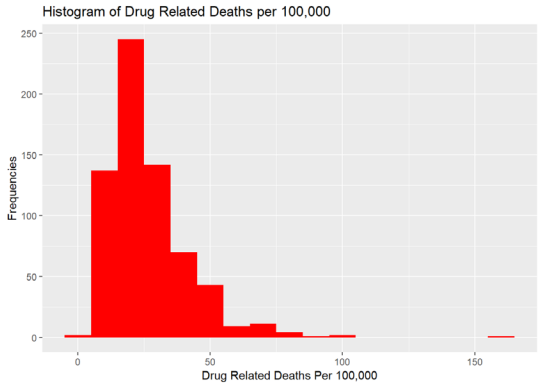
We also created a pairwise plot with five variables: drug related deaths, median building age, unemployment rate, median household income, and population. We see that there is a positive association between median building age and unemployment rate which indicates a county's overall economic prosperity. We see almost no correlation between population and drug related deaths, suggesting that drug related deaths affects counties of all sizes. In addition, there is a slight positive association between drug related deaths and median building age.

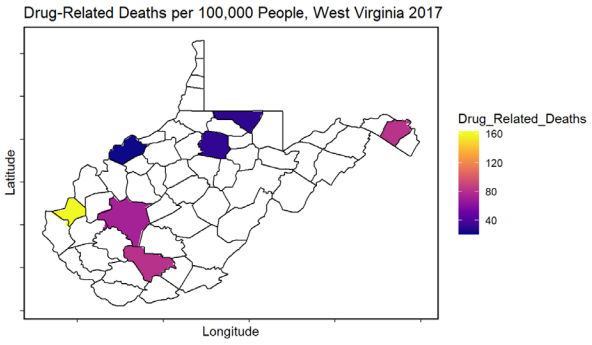
We dove deeper into our drug related deaths variable because it is the variable of interest, looking at distributions geographically as well as within states. Looking at a heat map of drug related deaths across the United States, we see that drug related deaths is concentrated in the Midwest (Rust Belt), Florida, and Pennsylvania, and the worst affected states are Pennsylvania, Ohio, West Virginia, and New Jersey. It is not surprising to see such a concentration in the Rust Belt as these counties used to house the industrial powerhouses of the country. In recent years, many of these jobs have since left the country, leading to economic depression. In addition, the physically-demanding nature of these jobs predisposed the populations to higher levels of injury and the prescription of painkillers, a precursor to the opioid epidemic. Looking at the distribution across Pennsylvania, the worst affected areas are in SW Pennsylvania (Rust Belt) and in Philadelphia, which we are all too familiar with. As remarked earlier, the Kensington neighborhood in Philadelphia is the center of our region's heroin epidemic, but not all neighborhoods are immune as we have seen overdose problems near and on Penn's campus as well. Looking at the West Virginia heat map, you can see that the opioid epidemic is centered around Huntington, WV and surrounding Cabell County. In Ohio, the worst areas are centered in cities such as Cincinnati and Cleveland. The histogram of drug related deaths shows that the distribution is skewed to the right with one large outlier with over 150 drug

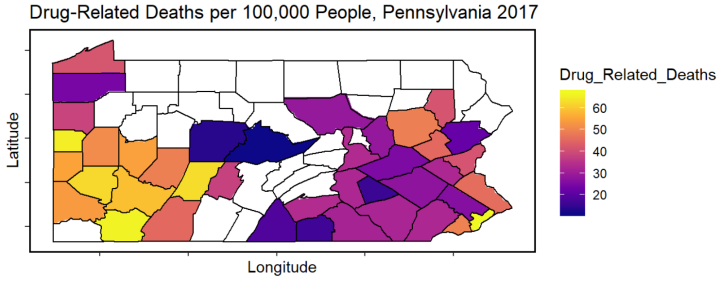
related deaths per 100,000 people. This county is Cabell County in West Virginia. For background, it is estimated that 10,000 of the region's 100,000 residents are addicted, and 1 in every 5 babies born have already been exposed to drugs[1]. West Virginia is really the prime example of the over-prescription of painkillers by doctors and encouraged by pharmaceutical companies; over the past six years, an estimated 780 million hydrocodone and oxycodone doses have been shipped to the state, a staggering average of 430 pills per person.

Summary Graphs of Drug Related Deaths









[1] Sreenivasan, Hari, and Jason Kane. "A Community Overwhelmed by Opioids." PBS. October 02, 2017. Accessed May 05, 2019. https://www.pbs.org/newshour/show/community-overwhelmed-opioids.

We also analyzed household income across the entire country. The heat map shows that the highest median household income is in the Northeast, as expected, and that the lowest is in Florida, West, and Midwest. The lowest household income in the entire country is in Robeson County, North Carolina at $32,729. The household income for Cabell County is $33,750 which is not much higher than the minimum, indicating the overall economic despair of West Virginia and the Rust Belt. Especially in Huntingdon, WV, where the economy centers around coal mining, a rapidly declining industry, the citizens are feeling the pain.

For percent veteran, the largest proportion of veterans live in Florida, which makes sense given the retiree-focused nature of the state for its sun, warm-weather, lower taxes, and higher median age (Figure 1). The highest percent veteran county in the country is in Okaloosa County, Florida at 17.46%. For Cabell County, the percent veteran is 6.49% which is far off the national maximum and is in-line with the nationwide average.

Our percent unemployed is the total amount of unemployed divided by the total county population which is different from the unemployment rate which is divided by the labor force size. For percent unemployed, we see that the highest rates of unemployment lie in the Pacific Northwest as well as across the Rust Belt (Figure 2). The county with the highest unemployment percentage in our sample lies in Harrison County, Mississippi with 5.15%. Cabell County's unemployment percent is very low at only 1.42%. While this theoretically does not align with Cabell County's high drug related deaths, in fact, the majority of jobs in Cabell County are in the coal industry which is low-paying. If the underlying driver behind the opioid crisis is economic problems, then a poorly-paying, painful job is almost as bad as no job, if not worse.

We also mapped percent insured across the entire county (Figure 3). With the introduction of The Affordable Care Act (Obamacare) in the early 2010s, the majority of the country is well-insured with high levels of insurance. Insurance can be related to drug related deaths as people with insurance have more access to rehab facilities and treatment plans. The county with the worst insurance rate is Onslow County, North Carolina with an insured rate of only 80.05%. Cabell County has a high insurance rate of 98.27%. However, given the economic despair and lower median household income of the area, most insurance plans are probably in the bottom tier. As a result, they may not cover expensive rehabilitation facilities. Moreover, the usage of opioid treatment plans such as those involving suboxone and methadone have been supply-limited by the federal government approval process; as a result, insurance may not be a perfect remedy for the opioid epidemic. The Cabell County insurance percentage is also inflated upward because West Virginia was one of the few Republican states to accept federal Affordable Care Act funding to expand Medicaid access.

Lastly, we evaluated the usage of food stamps across the country since it could be an indicator of economic downturn. The areas with highest food stamp usage are in Oregon, Ohio, Pennsylvania, and New York, and the highest county of food stamps is Bronx County, New York at 12.88% of the total population (Figure 4). Cabell County's food stamp percentage is 8.05%. However, the availability of food stamps is not constant across the United States as there are both federal and state level food stamp programs. States with more liberal governments have more expansive food stamp programs which may be the case in New York. West Virginia is controlled by conservative lawmakers who are stingier with food stamps.

Statistical Analysis

For model building, we evaluated 5 different statistical methods – simple logistic regression, elastic net, simple tree, random forest, and boosting – to find the optimal set of predictive variables for communities that need resources for drug related deaths. Each of these models produced a different set of variables with different coefficients. We utilized a loss function that penalized a false

negative 1.5x as harshly as a false positive, resulting in a probability cutoff of 0.4. This reflects our perception of the tradeoff between ignoring a community actually in need with the understanding that government resources are limited. While we think it would be impossible to quantify how bad it would be for a community to suffer an opioid epidemic, resulting in overdoses, community unrest, and broken families, we are also cognizant that wasting resources on too many communities not actually at-risk of an opioid epidemic would leave less for those that actually need federal aid. Especially in light of the current federal budget situation and tightening of government resources, we take into account that Departments of Health Services need to be somewhat selective in funding various opioid diversion or recovery programs around the country. These are all factors that influenced our decision for a 1.5:1 loss ratio of cost of false negatives to the cost of false positives and why our loss ratio is not more extreme for false negatives.

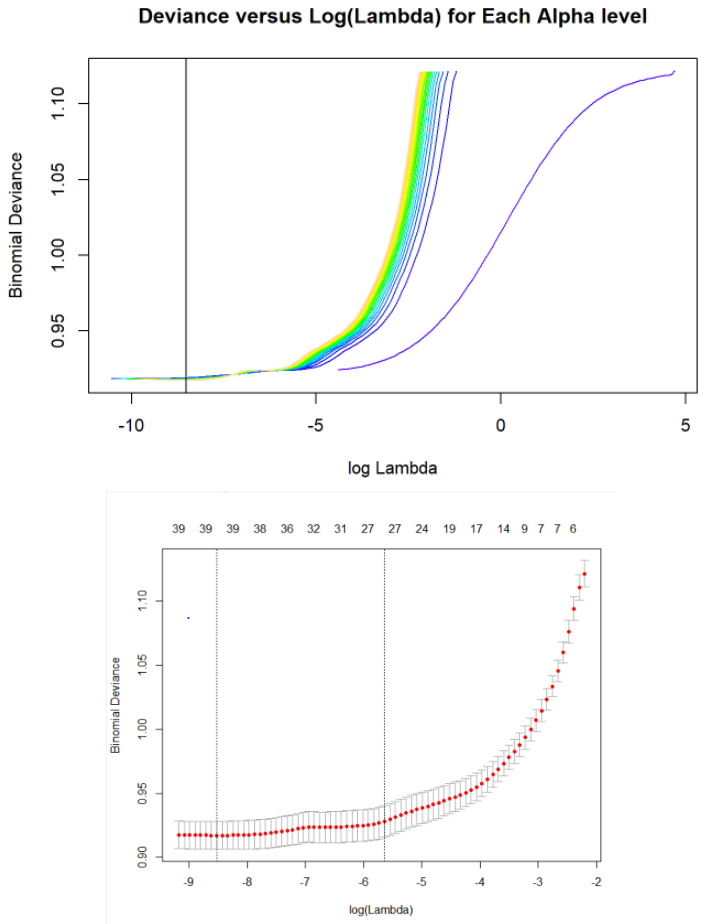
As previously mentioned, we had the choice of using a regression to predict the rate of drug related deaths per 100,000 people or using a classifier such as "Need Resources." Although the data initially was structured in terms of numbers of drug related deaths per 100,000 people, we thought it would be more useful for policymakers to identify areas where resources were needed as opposed to pinpointing the exact number of drug related deaths. We split our data into three different sets (training, validation, and testing) by county to conduct our statistical analysis. Because our data includes data per county per year, we chose to split by county to ensure that our validation and testing data were truly independent from the training data. If we simply had done a split of the entire data set in a purely random fashion, the training, validation and testing data could have had the same counties in them, and as such, naturally overfit the "out of sample" data, as demographic metrics do not frequently change rapidly. We sacrifice our in-sample data, as we realize that the model seeing the same county over six years may lead to overfitting, in order to preserve the independence of the out of sample data. If we had not split the way we did, we would have inflated our performance metrics out of sample and leave us with little ability to extrapolate to counties outside of our data set. For example, our split method meant that all years of Philadelphia County would be in the training data set while all years of Montgomery County may be in the testing data set and all years of Bucks County may be in the validation data set. Out of 770 total counties in the American Community Survey data, we randomly selected 520 for the training data set, 125 for the validation data set, and 125 for the testing data set.

We first used a simple logistic model to gauge a baseline for the remainder of our analysis. This "big fit" was truly a big fit; we included all variables in the logistic, whether they were significant or not. We did not exclude any variables other than those previously mentioned that were removed from the dataset. While we did not intend to use this "big fit" as our final model, we wanted to use it as a comparable to other models to see if it was possible to accurately identify which predictors should be removed or emphasized. This "big fit" produced an in-sample AUC of .804 and a validation AUC of 0.8031. The weighted misclassification error for the training data was 0.2857 and for the validation data was 0.2763. Due to its nature, it makes sense that the "big fit" would perform well on the training data, but its performance on the validation data seems to be a coincidence of the split we have.

We then attempted to use an elastic net model to penalize over complex and overfit models and identify variables that contribute the most to reducing the residual errors and L-1 and L-2 norms. We cross-validated alpha to evaluate whether a full LASSO, full Ridge Regression, or elastic net model would be optimal. After cross-validation, we found that the optimal alpha is 1, indicating a model that only utilizes the L-1 Norm. For the resulting LASSO model, we evaluated it at the lambda value that produced the lowest binomial deviance. This method produced a model with a large amount of predictors (39); however, many were not significant at the 0.05 level. We re-fitted the glm model and conducted Type II Anovas to determine if each variable was truly necessary. We

systematically removed predictors with the largest p-value, re-fit the glm, conducted the Type II Anova, and continued this process until all predictors were significant at the 0.05 level. The final predictors used in the logistic model were: median household income, average household size, number of households, percent males and females divorced, percent male never married, percent household income less than $10,000, total housing to households, median home value, median building age, percent black, percent Asian, percent missing a full kitchen, percent aged 50+, percent veteran, percent enrolled in school, percent enrolled in high school, percent male, percent families with single parent, percent of white in poverty, and percent employment. This is a high number of predictors, and it's not immediately clear which is the most important. However, reducing the model to this set of predictors is already an improvement over the hundreds of demographic data points available. It's clear that opioids and drug-related deaths in general are complex issues across American communities. The training AUC and weighted misclassification error were 0.7947 and 0.2929, respectively, while the validation AUC and weighted misclassification error were, 0.7992 and 0.2638 respectively (Figure 5). This weighted misclassification error compares favorably to the initial "big fit" and was also the best classifier among all methods. Again, it appears that our random split of data contributed to the superior performance of the model on the validation set in comparison to the training data.

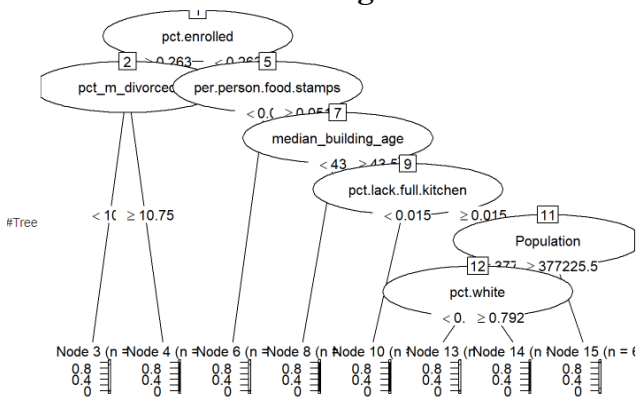
Cross Validation of Alpha and Lambda



We then ran a simple tree on the response variable, Need Resources, using the rpart tree function with max depth = 7. Even after adjusting the minsplits and cp variables, we found that the tree was extremely prone to overfitting, resulting in an in-sample weighted misclassification error of

0. As such, we decided to force the tree to limit its depth, and we felt that 7 was reasonable given the amount of predictors we had as well as the variability in the data. The classification tree is greedy and evaluates all possible variables and split points to determine the best structure to minimize the overall deviance. For our tree, it first split on percent enrolled, and other splits included percent male divorced, foot stamps per person, median building age, percent missing a full kitchen, population, and percent white. It appears that percent enrolled is a proxy for the age of the population, as in general, younger people are in school. There were 8 leaves on the tree. The tree's training AUC and weighted misclassification error were 0.7302 and 0.3040, respectively, and the validation AUC and weighted misclassification error were 0.7451 and 0.2746, respectively (Figure 6).

Tree Diagram

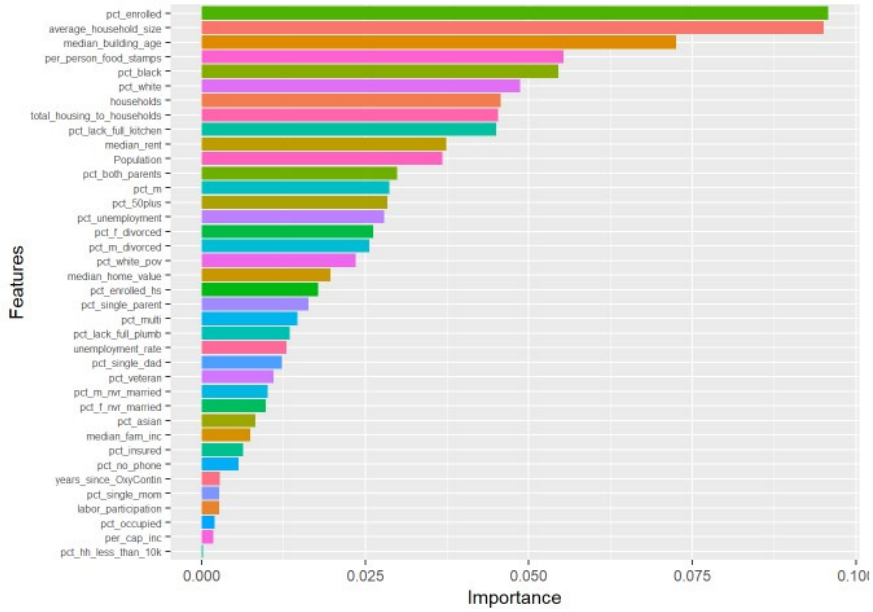


We then adapted the simple tree and ran a random forest to see if this would improve our classifier. The random forest utilizes bootstrapped samples and samples random sets of variables (via the mtry parameter) to build trees split on misclassification error. The final result is an average of all trees built. Similarly to the single tree built above, we ran into problems where the random forest would frequently overfit the training data, resulting in an in-sample AUC of 1 and weighted misclassification error of 0. To rectify this problem, we set the number of trees to be built to only 5 (Figure 7). Although this would limit the overall effectiveness of the random forest method, we were more concerned about overfitting and felt that an average of 5 trees was better than a perfect in-sample average of more trees. In addition, we set the mtry parameter to 7 which is an approximation of the recommended mtry for classification trees ($\sqrt{p}$). After cleaning and removing unnecessary variables from our data, we had a total of 41 possible predictors for the random forest. The random forest did not perform as well as other methods out of sample. It had a training AUC and weighted misclassification error of 0.9927 and 0.04788, respectively. The validation AUC and weighted misclassification error were 0.6987 and 0.3172, respectively (Figure 8). The drastic change from training to testing shows the impact of overfitting.

Lastly, we attempted to boost the model using the xgboost package and function. Boosting takes a variety of different classifiers and assigns different weights to them depending on their performance repeatedly, essentially turning many mediocre classifiers into one theoretically stronger classifier. For our data, we found that boosting did not produce the best results compared to other methods. For our boosting model, we used 5 rounds of boosting and 5 folds of cross-validation. Again, we had to significantly adjust the parameters to prevent overfitting in the data. The boosting algorithm had an in-sample AUC and misclassification error of 0.9491 and 0.1325, respectively. The validation AUC and misclassification error were 0.7607 and 0.3055, respectively (Figure 9). The much worse performance on the validation data shows the overfitting challenge. The boosting algorithm also identified some of the most important predictors for our response variable; the top 5

meaningful predictors were median household size, percent enrolled in school, median building age, percent white, and food stamps per person. The importance chart is included below.

Importance Graph



Our metric to determine the best classifier is the weighted misclassification error because it allows us to take into account the different consequences for generating a false negative and false positive. Out of all of our attempted methods, the elastic net (LASSO) performs the best with weighted misclassification error of 0.2638 followed by simple tree (0.2746), boosting (0.3055), and random forest (0.3172). Although usually boosting and random forest perform better than a simple tree, we think that it performs worse in this case because of its propensity to overfit in our data. As mentioned previously, we had trouble preventing random forest and boosting from overfitting the training data, and this can be seen in its poor performance on the validation data set. The LASSO elastic net uses the L-1 norm to cut extraneous variables from the logistic regression to minimize the overfitting, which we can see reflected in the validation data results. Our final model is the LASSO model with the following parameters and coefficients.

LASSO Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.048e+00	6.505e+00	-0.315	0.752839	
median_hh_income	5.008e-05	9.663e-06	5.182	2.19e-07	***
average_household_size	-1.904e+00	4.565e-01	-4.171	3.03e-05	***
households	9.146e-07	3.496e-07	2.616	0.008897	**
pct_m_divorced	1.044e-01	3.683e-02	2.835	0.004581	**
pct_m_nvr_married	5.651e-02	1.833e-02	3.082	0.002055	**
pct_f_divorced	8.126e-02	3.365e-02	2.415	0.015734	*
pct_hh_less_than_10k	8.144e-02	3.552e-02	2.293	0.021863	*
total_housing.to.households	1.334e+00	5.643e-01	2.365	0.018043	*
median_home_value	-4.005e-06	1.038e-06	-3.857	0.000115	***
median_building_age	3.613e-02	7.226e-03	5.000	5.74e-07	***
pct.black	-2.852e+00	6.530e-01	-4.368	1.25e-05	***
pct.asian	-8.619e+00	3.064e+00	-2.813	0.004905	**
pct.lack.full.kitchen	1.534e+01	7.584e+00	2.022	0.043144	*
pct.50plus	-1.293e+01	3.128e+00	-4.132	3.59e-05	***
pct.veteran	-9.972e+00	3.708e+00	-2.690	0.007154	**
pct.enrolled	-1.114e+01	2.832e+00	-3.931	8.45e-05	***
pct.enrolled.hs	-2.365e+01	1.026e+01	-2.304	0.021219	*
pct.m	-2.200e+01	6.855e+00	-3.209	0.001331	**
pct.single.parent	1.473e+01	6.750e+00	2.183	0.029067	*
pct.white.pov	1.139e+01	5.013e+00	2.271	0.023145	*
pct.unemployment	2.420e+01	5.184e+00	4.669	3.03e-06	***

The positive factors that are associated with an increase in the probability of needing resources (higher drug related deaths) include median household income, percent divorced, percent of household with less than $10,000 income, total housing to households, median building age,

percent single parent, and percent unemployment. A positive coefficient indicates that an increase in these factors is associated with an increase in the probability of a county needing additional resources for the opioid epidemic. These positive factors can all be explained intuitively by the following:

Median Household Income: Although this can be expected to be negatively correlated, the positive correlation can be explained through income inequality. Median household income is higher in cities which also display a higher level of income inequality. Although median income may be higher in a city such as Philadelphia, there are also many at the opposite end of the economic spectrum who may be suffering more from the opioid crisis.

Percent Divorced: Percent divorced is an indication of a broken family situation which may cause people to turn to opioids as a coping mechanism. This could also be an after-the-fact response to opioids as families may be broken because of drug addiction.

Percent of Households with less than $10,000 Income, Median Building Age, Percent Unemployment: These are all indicators of economic despair which can cause people to turn to drugs.

Total Housing to Households and Median Building Age: This is a tracker of old, vacant, and dilapidated buildings. Higher number of vacant buildings is a signal of economic despair, but drug-users also seek out these buildings to use drugs inconspicuously. If they overdose, this could lead to death as there is no one available to revive them with Narcan.

Percent Single Parent: This could be an indicator of broken families, a poor social support system, and levels of stress which could cause drug problems.

The negative factors that are associated with a decrease in the probability of needing resources (higher drug related deaths) include average household size, median home value, percent black, percent enrolled and in high school, and percent male. A negative coefficient indicates that an increase in these factors is associated with a decrease in the probability of a county needing additional resources for the opioid epidemic. These negative coefficients can all be explained intuitively by the following:

Average Household Size: Larger households can be an indicator of more complete families and stronger support systems, which can deter individuals from drugs.

Median Home Value: This is an indicator of county-wide economic prosperity. If economic depression can cause people to turn to drugs, then higher levels of economic prosperity produce lower incentives to turn to drugs as a coping mechanism.

Percent Black: This predictor and other non-white ethnicity distributions are negatively correlated with opioids overdoses. This is in line with the theory that the majority of opioid addicts are white and male. However, this does not indicate that ethnic minorities in America use drugs at a lower proportion than whites; this is merely an indication of drugs of choice and their relative overdose fatality potential.

Percent Enrolled in School: If economic despair is a cause of the opioid epidemic, then a higher percent of the proportion enrolled in education indicates a chance of economic revival. This is also highly correlated with the average age in each county; younger students probably have not been exposed to opioids yet. Lastly, schools also provide a support system away from drugs.

Percent Male: We expected this variable to have a positive correlation. However, the effects of male may already be captured in percent male divorced and never married which both have positive coefficients. In areas with higher opioid overdoses, there may be a higher percent of male who are divorced and never married since these are representative of the despair that plagues drug users. In other words, there is an interaction effect that is not fully extracted from our model.

Conclusion

We found that the elastic net (LASSO) model generates the lowest weighted misclassification error on the validation data (0.2638) for the "Need Resources" response variable because it controls best for overfitting. Run on the testing data, the best LASSO model produces a weighted misclassification error of 0.3432 and AUC of 0.7758 (Figure 10). This is slightly worse performance than the validation data and training data though that can be expected due to some degree of overfitting. Moreover, the structure of our data also lends itself to poor performance; our training, testing, and validation data is split based on county, and there may be differences between counties that are not accounted for in our census demographic data.

Our LASSO model shows that there are a number of economic, infrastructure, and social factors that are related to the probability that a county "Needs Resources" for its opioid epidemic. We show that variables associated with economic despair (median household income, percent of households with income less than $10,000, median home value) lead to higher rates of drug-related deaths. This makes sense because economic pressure and depression can cause people to turn to drugs. In addition, in areas where there is not much legal economic opportunity, people may turn to illegal methods, such as drug dealing, which then increases the supply of illicit drugs in the area. However, it's important to note that both variables – drug proliferation and economic despair – build upon each other. Once addicts are hooked onto drugs, they participate less in the overall economy as most of their money goes to drug dealers and overseas cartels. In addition, they begin to neglect their jobs, and the overall community worsens. This causes employers to leave the community, causing more to turn to drugs, and the cycle continues.

There are also variables related to a county's infrastructure (median building age and total housing to households) that reflect the overall availability of locations for drug users to overdose. While these are also somewhat correlated with economic wellbeing, the presence of older, unused, and vacant buildings provides opportunities for users who prefer private locations to use drugs. Because these buildings are not frequently used, overdoses go undiscovered, leading to a higher rate of death.

Lastly, there are important social factors such as percent divorced, percent single parent, percent enrolled in school, and percent black that are correlated with "Needs Resources". Families that are broken and missing a parent provide less of a support structure for both the parent and a child, which can lead to increased rates of drug usage as a coping mechanism. Schools also provide a social deterrent to drugs. The percent enrolled in school and percent black also provide useful indicators of the demographics (middle-aged whites) that are more inclined towards dangerous opioids.

Our analysis shows that there are three main contributors to the prevalence of opioids in a community and identifies to which counties federal resources should be directed to maximize their utility. One can consider policy recommendations in light of the worst-hit areas of America (Rust Belt) and the worst-hit county (Cabell County, WV). Firstly, the dying industries of these areas lead to higher levels of economic despair. Federal policymakers should revitalize these areas either though public infrastructure investments (creating jobs and important drug-testing requirements), vocational training (such as in IT or computer science to help them shift to higher-paying industries), or offer tax credits to bring industries and better economic opportunities back to those areas. With regards to infrastructure, state and local authorities should tear down old, vacant buildings and replace them with drug clinics or homeless shelters. In this way, the homeless population and drug users would still have a covered place to shelter, yet they would be staffed to prevent drug use and overdoses. Lastly, from the local to the federal level, there should be an increased emphasis on the social work profession by boosting salaries or increasing hiring. Social

workers have the potential to provide support to broken families and direct individuals towards government programs to help them get back on their feet and away from the scourge of drugs. It is important to note that because the opioid epidemic is a multi-pronged problem, there needs to be a coordinated, multi-pronged response and solution to fully address all of the root causes.

Because the American Community Survey and CDC WONDER dataset only include a subset of American counties, our model can be applied to other American counties to identify which counties have conditions that are ripe for opioid proliferation. Although drugs can now be illicitly purchased by mail and online, the vast majority of the supply is neighborhood-based with individual drug dealers. Because of this, some areas of the country have not had opioids reach them, yet our model can forecast which should prepare for their introduction because it would lead to a high number of overdoses and deaths.

While we are confident in our conclusions regarding the social, economic, and infrastructure predictors of drug related deaths and our LASSO model, we are also cognizant of key assumptions we made in our data analysis. The American Community Survey data may be skewed since it only surveys counties with more than 65,000 residents. In addition, we only have data for the total number of drug-related deaths which may not be a perfectly proportional substitute for opioid-related deaths as opioids have become more popular over time (and thus a larger percent of the deaths). Moreover, we encountered significant problems of overfitting in our data analysis that prevented us from fully utilizing more robust techniques. Lastly, our model is only trained and tested on data from the past; changes in the national sentiment, political policy, or drug prevention techniques may limit the usefulness of our model going forward.

Appendix:

Figure 1:

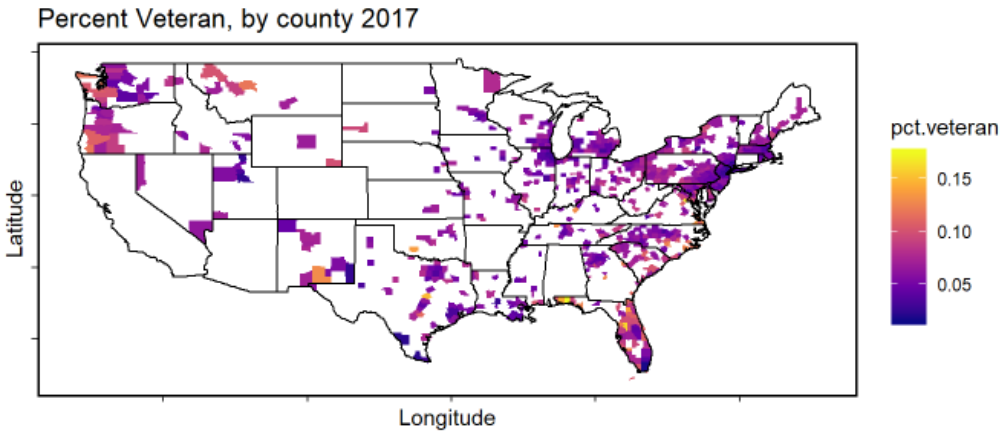


Figure 2:

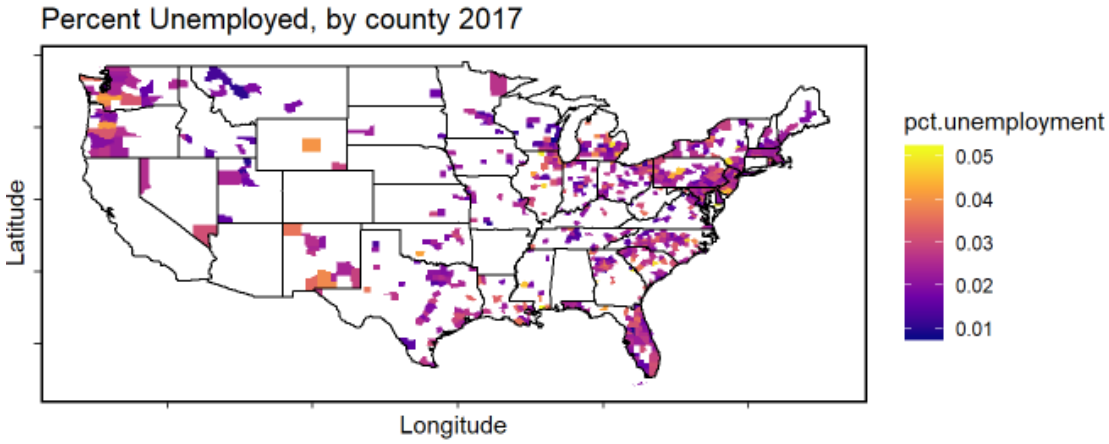


Figure 3:

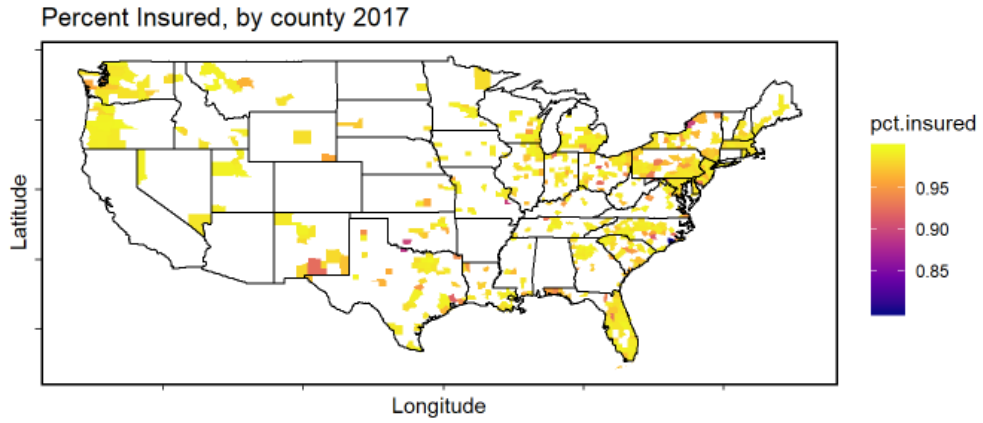


Figure 4:

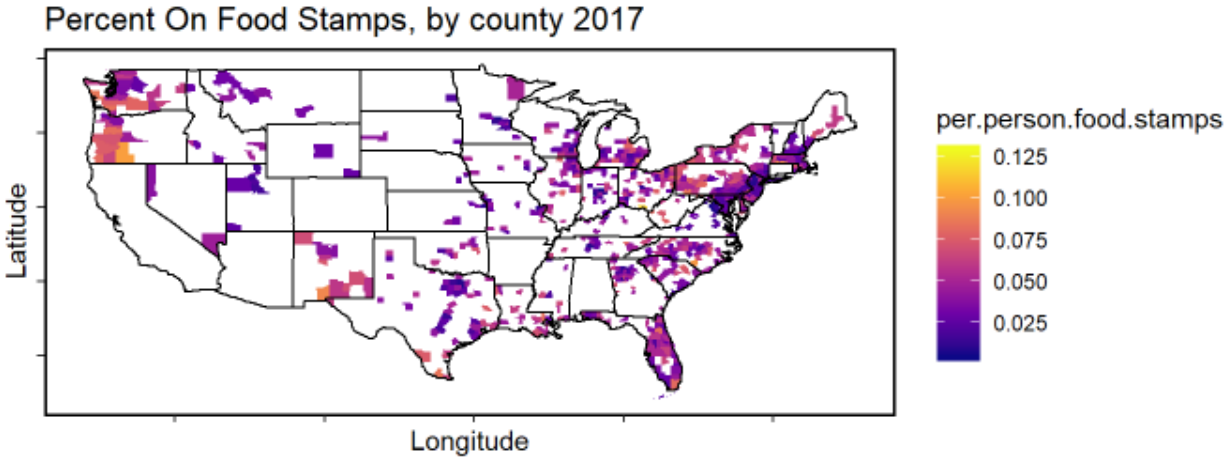


Figure 5: Validation ROC for Elastic Net

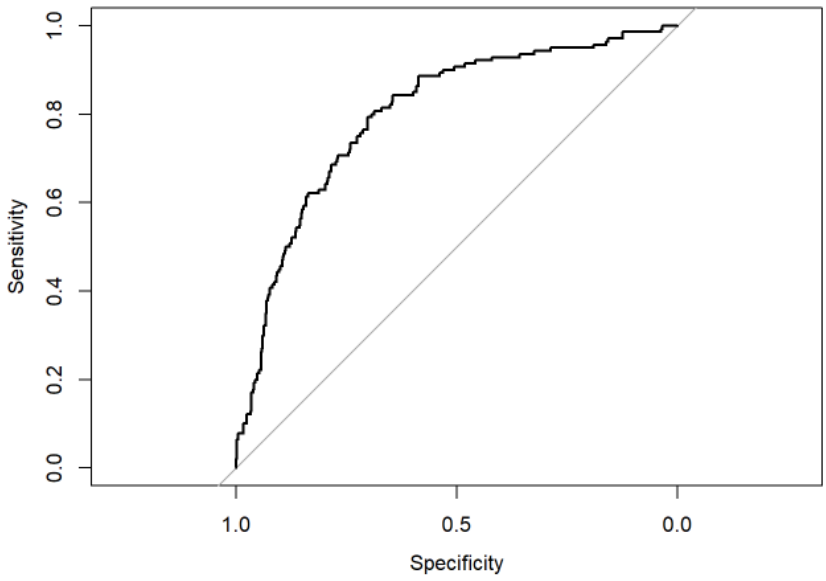


Figure 6: Validation ROC for Simple Tree

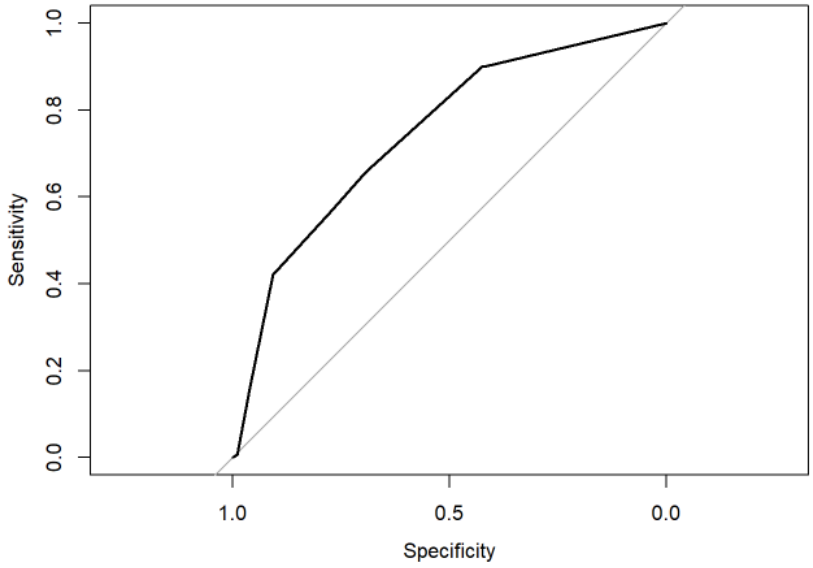


Figure 7: OOB Error vs. Number of Trees for Random Forset

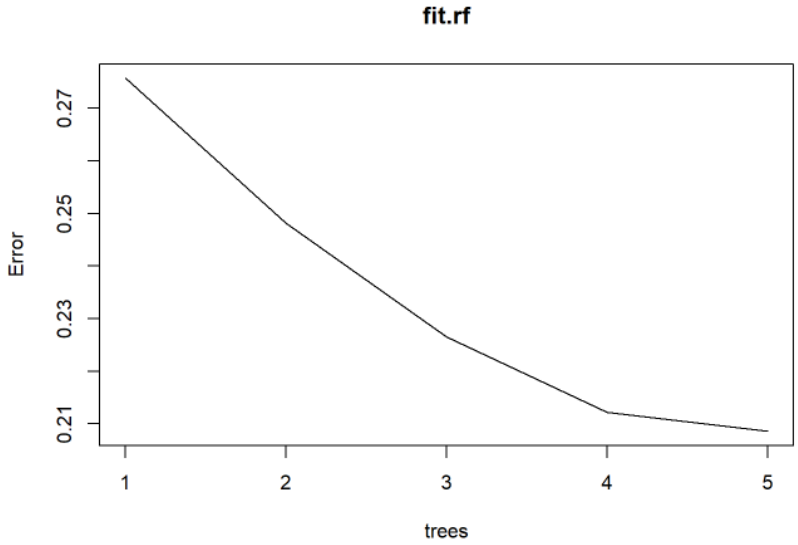


Figure 8: Validation ROC for Random Forest

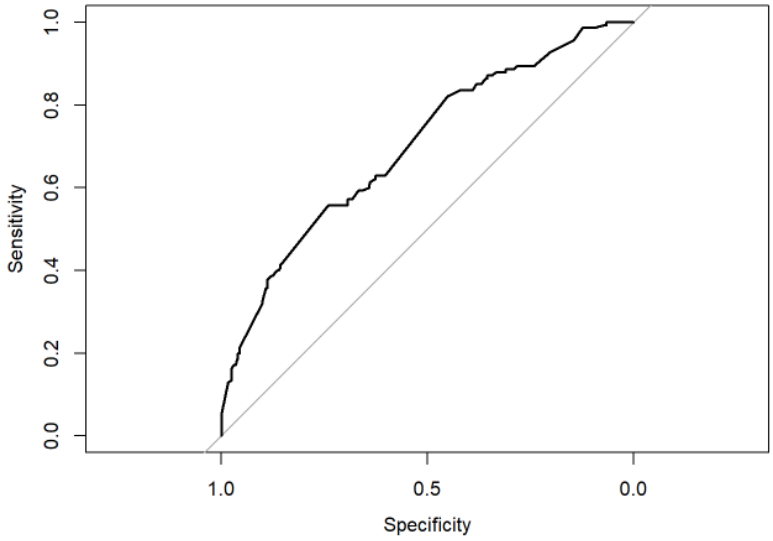


Figure 9: Validation ROC for Boosted Tree

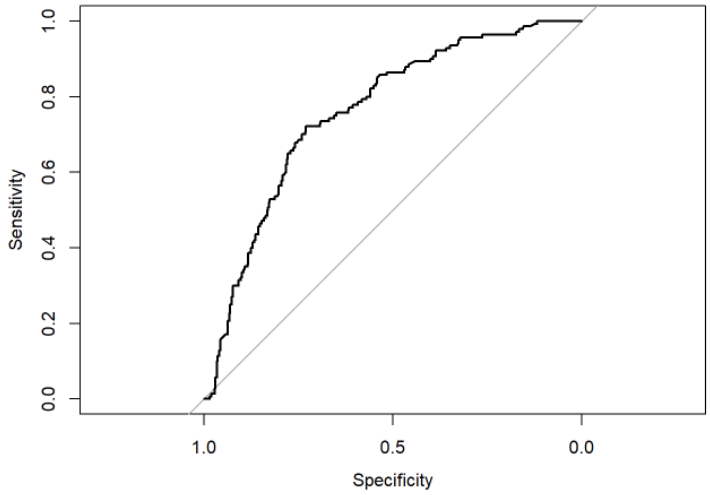


Figure 10: Testing ROC for Elastic Net