# Survey of Non-Medical Use of Prescription Drugs Program

## United States 19Q1 Launch

Siddharth Bowgal | Arjun Putcha

# Meet the Team

Siddharth Bowgal
B.S. Statistics & Analytics, Class of 2022

Arjun Putcha
B.S. Biomedical Engineering, Class of 2022

# Drug Misuse is a major problem in the US

## Opioid Crisis

In 2019 alone, about 50,000 people died from overdose [1]

## Misuse of prescriptions

21-29% of patients misuse their prescribed opioids [2]

## Drug abuse is only growing

Need to keep improving ways to stop drug abuse

Sources: National Institutes of Health [1,2]

# How we approached the data

## Subset the Data

- Features: Variables not explicitly related to questions about drug use. Included variables about alcohol, tobacco, and cannabis use

## Setting up for Machine Learning

- One-hot encoding of categorical variables, compare classification models
- Models compared: SVM, Decision Tree (with AdaBoost), Bagging, Logistic Regression, Random Forest

## Stratify Data and Run Models

- Stratification on: Income, Age, Region
- Models used: Logistic Regression, PCA + Logistic Regression, Random Forest

## Evaluate Model Performance

- Confusion matrices
- Accuracy scores, Features_importances

# Model Performance

| | | Demographics + Alcohol, Tobacco, Marijuana, and OTC use | | | | |
|---|---|---|---|---|---|---|
| | | **US** | **US_NE** | **US_Midwest** | **US_South** | **US_West** |
| **PCA + Logistic Regression** | Accuracy | 0.94 | 0.944 | 0.953 | 0.929 | 0.931 |
| | Specificity | 0.01 | 0.03 | 0.01 | 0.01 | 0.02 |
| | Sensitivity | 1 | 1 | 1 | 1 | 1 |
| **Random Forest Tree** | Accuracy | 0.94 | 0.945 | 0.952 | 0.928 | 0.933 |
| | Specificity | 0.02 | 0.03 | 0.01 | 0.01 | 0.02 |
| | Sensitivity | 1 | 1 | 1 | 1 | 1 |

# Model Performance

**Demographics + Alcohol, Tobacco, Marijuana, and OTC use**

| | | Age | | | | | | Income | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **18-24** | **25-34** | **35-44** | **45-54** | **55-64** | **65+** | **< $25K** | **[$25K, $50K)** | **[$50K, $75K)** | **[$75K, $100K)** | **> $100K** |
| **Logistic Regression** | Accuracy | 0.879 | 0.874 | 0.891 | 0.933 | 0.962 | 0.983 | 0.931 | 0.93 | 0.941 | 0.941 | 0.947 |
| | Specificity | 0.12 | 0.07 | 0.06 | 0.03 | 0 | 0.03 | 0.05 | 0.03 | 0.08 | 0.18 | 0.14 |
| | Sensitivity | 0.98 | 0.98 | 0.98 | 1 | 1 | 1 | 0.99 | 0.99 | 1 | 0.99 | 0.99 |
| **Random Forest Tree** | Accuracy | 0.884 | 0.88 | 0.898 | 0.934 | 0.963 | 0.983 | 0.933 | 0.934 | 0.939 | 0.937 | 0.95 |
| | Specificity | 0.04 | 0.07 | 0.05 | 0 | 0 | 0 | 0.02 | 0.03 | 0.05 | 0.03 | 0.03 |
| | Sensitivity | 1 | 0.99 | 0.99 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Logistic Regression specificity has the highest average across Income -> fewer false positives

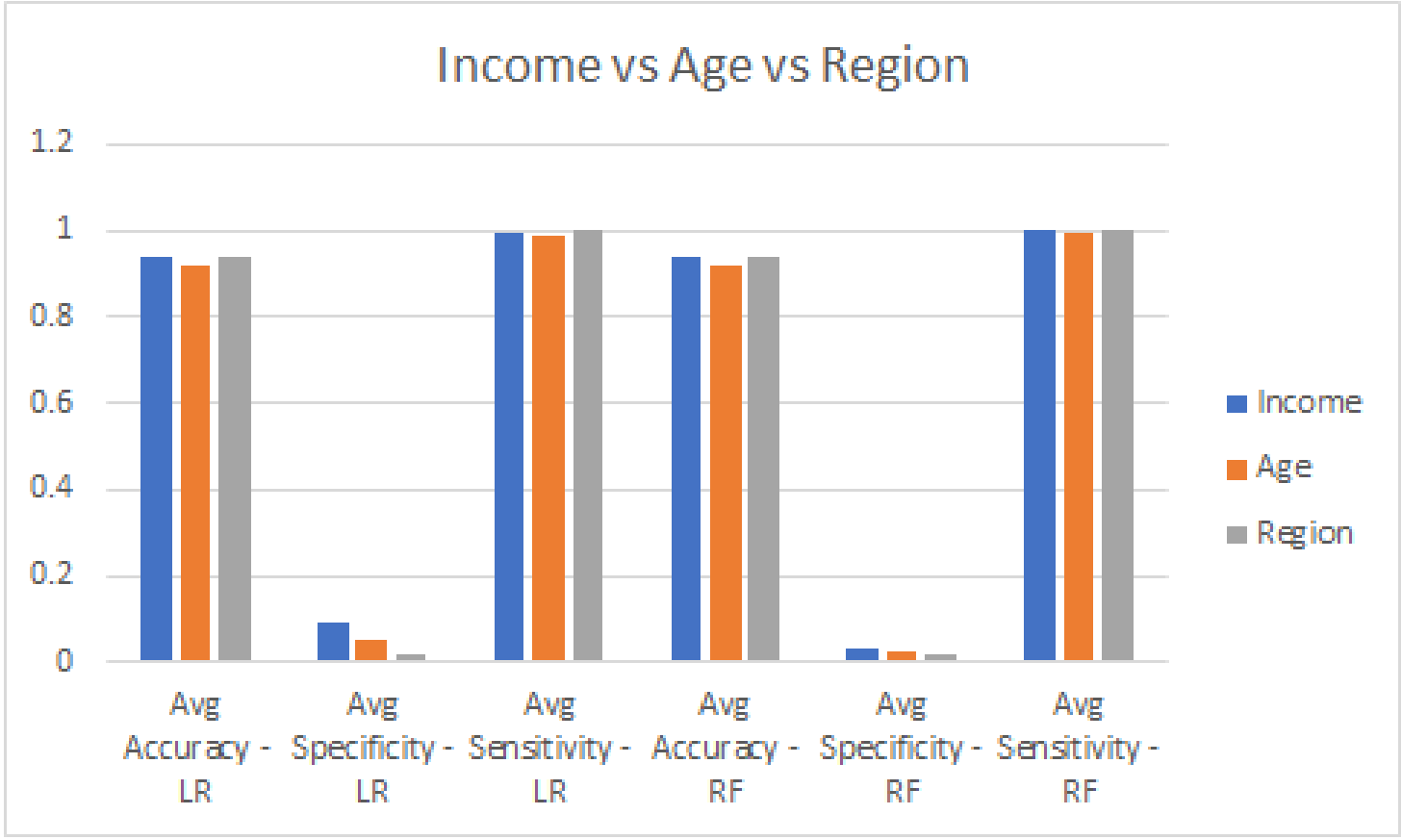
# Most important features

# Our approach achieves high accuracy

**88%**

minimum accuracy achieved by our Logistic Regression and Random Forest Models

- Low specificity (high false positives), High sensitivity (low false negative) across models



# Type I and Type II errors

- With relatively higher occurrences of false positives, we run a greater risk of predicting someone to safely use their medically prescribed opioid, when they actually will use it for a non-medical purpose
    - There is room for improvement here

- However, with low occurrences of false negatives, our model is less likely to falsely assume someone will abuse their prescribed opioids

# Outlook

- Part of our analysis included feature variables related to specific drug questions. This model achieved nearly 100% accuracy, so further research can be done to see which questions about drug use should be included in a survey

| | | **Demographics + Drug Data** |
|---|---|---|
| **PCA+ Logistic Regression** | Accuracy | 0.998 |
| | Specificity | 0.98 |
| | Sensitivity | 1 |
| **Logistic Regression** | Accuracy | 0.995 |
| | Specificity | 0.96 |
| | Sensitivity | 1 |
| **Random Forest Tree** | Accuracy | 0.994 |
| | Specificity | 0.96 |
| | Sensitivity | 1 |