Visualization of the Opioid Crisis in the USA

Bathla, Rajat (GT) | Davey, Mark (GT) | Kumar, Deepak (GT) | Mahajan, Amey (GT) | Pardo, Adrian (GT) | Ymeri, Genci (GT)

Introduction

Within the United States, **opioid abuse has doubled in prevalence since 2004**. This alarming reality begs three questions: what are the **underlying factors** of this "opioid crisis"; are opioid transactions truly, **postively assoiciated with deaths**; and is there anything that we can do to **mitigate this nationwide issue**? By analyzing socioeconomic factors across multiple data sources, we developed a **county-level holistic analysis** to better understand the location and degree of this crisis. We created an **interactive website** to visually identify local opioid trends and explore correlation data. Website: **https://kariato.github.io/opioiddrugsanalysis.github.io**

Opioid Dataset

88+ GB of data was downloaded from various government agency websites and Kaggle. Python along with Dask and Pandas libraries was utilized to process these large datasets. Although drug transaction data was not complete for every year, the final merged dataset was comprised of the following components with dates ranging from **2006** to **2018**.

Opioid transactions	179 million rows
Mortality	25 million rows
OSHA fatalities	~9,000 rows
Area deprivation index	~250,000 rows
BLS unemployment data	~60,000 rows

Overview: Machine Learning

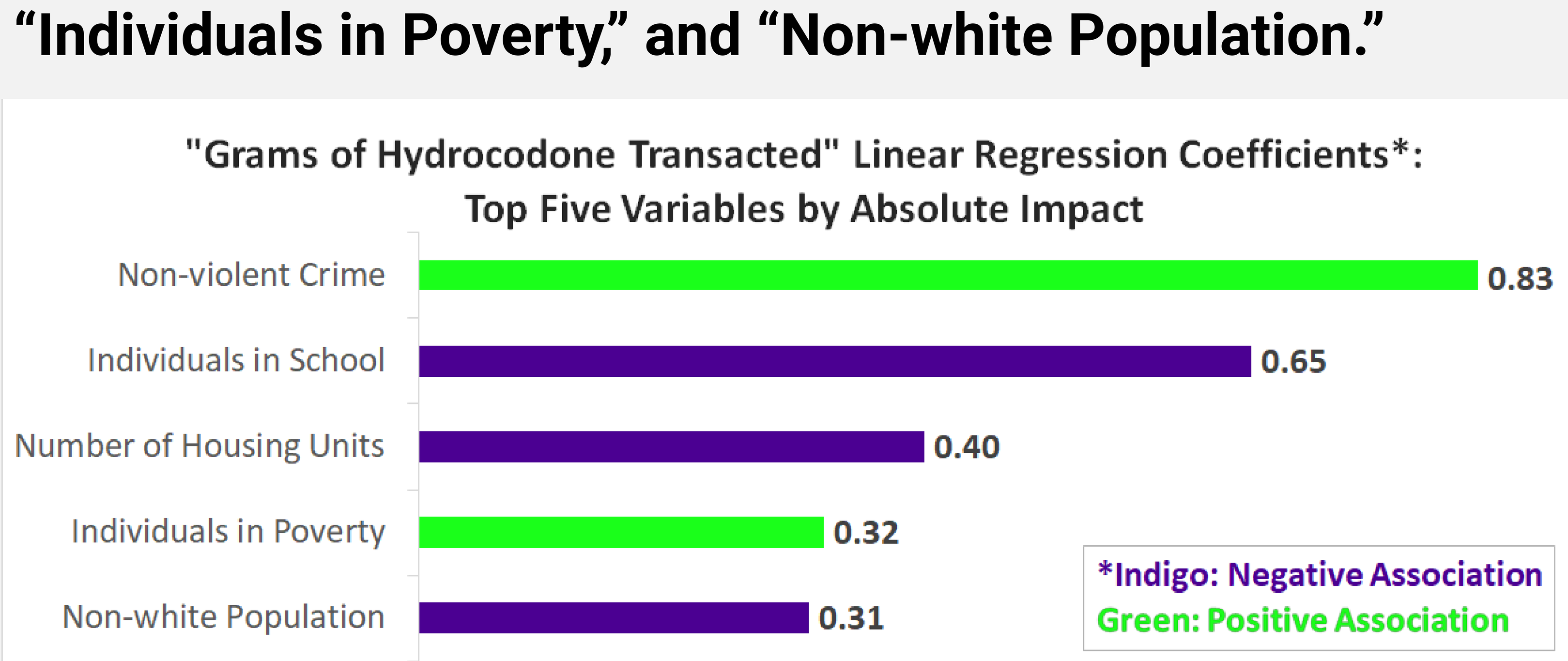
We employed **linear regression**, **LASSO regression**, and **random forests** for two prediction tasks on a per-county basis:

(1) Prediction of **grams of opioids transacted** (hydrocodone and oxycodone)
(2) Prediction of **deaths related to opioids** (both with and without opioids included in the model)

To explain opioid usage, we opted to use Linear Regression for coefficients are more interpretable than RF importance scores. For deaths, we opted to focus on RF results because they were **~50% more accurate** than linear and LASSO regression when comparing hold-out set RMSE.

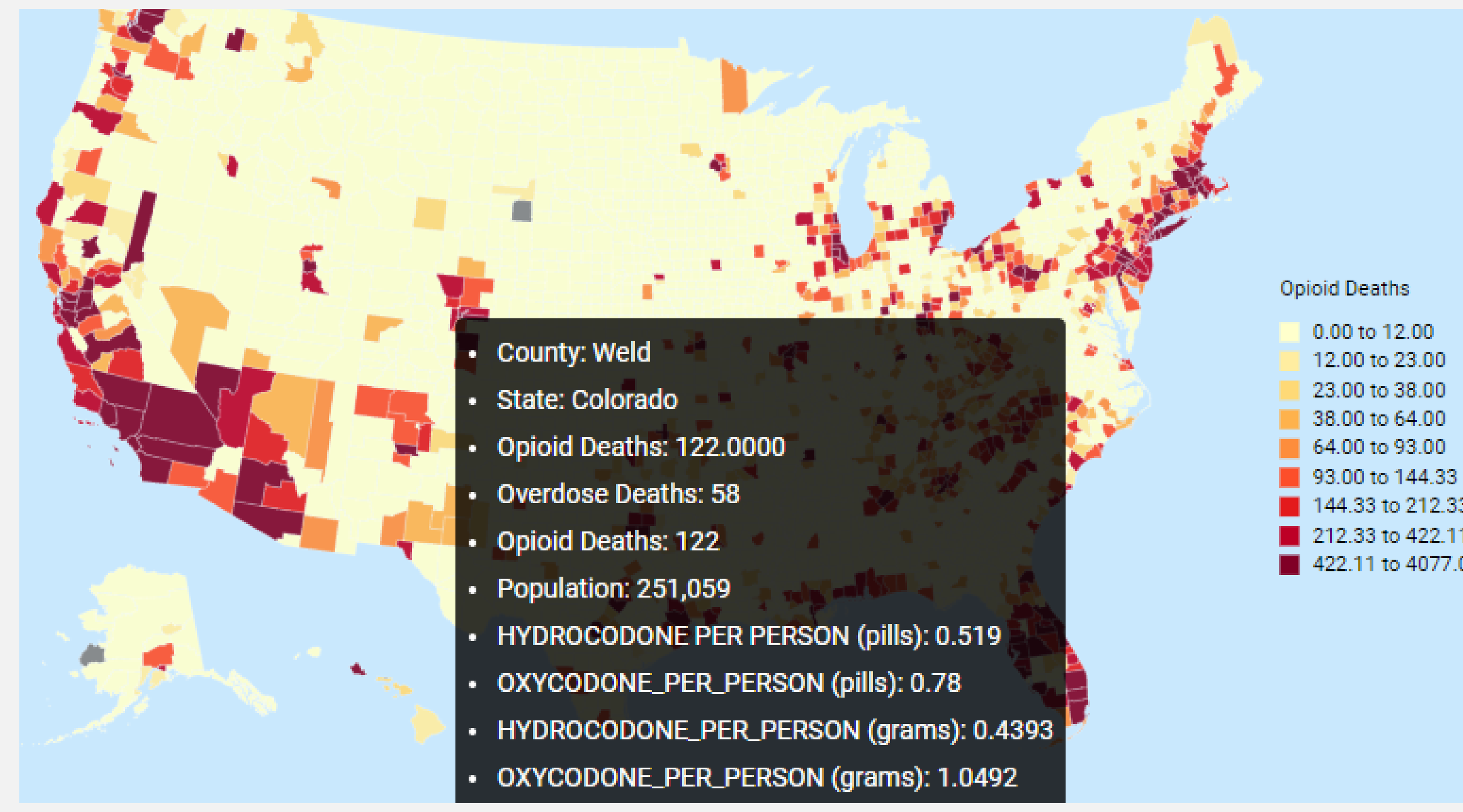
What explains opioid usage?

Linear Regression indicated that "Grams of Hydrocodone transacted" is **best explained by "Non-violent Crime," "Individuals in School," "Number of Housing Units," "Individuals in Poverty," and "Non-white Population."**



This specific model suggests that we would expect to see a higher amount of Hydrocodone transacted in counties where non-violent crime is prevalent, school attendance is lower, housing units are scarce, poverty is common, and the non-white population is lower.

Opioid Deaths in USA: 2006-2012

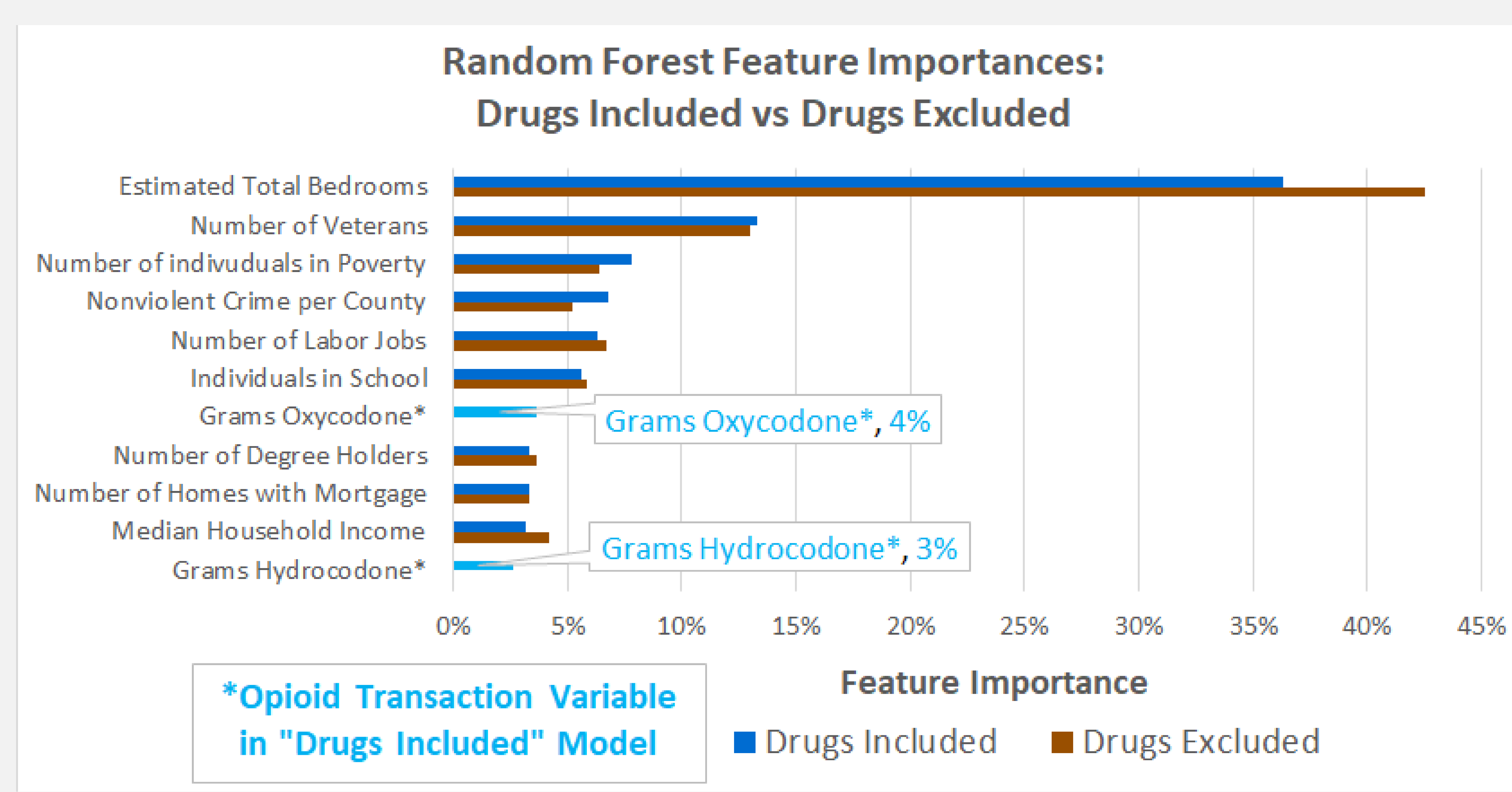


To what extent do prescriptions explain opioid-related deaths?

It appears that opioid usage is not as associated with mortality as other socioeconomic factors. We compared the feature importances of two Random Forest models:

(1) Prediction of deaths related to opioids **with Opioid usage as input**
(2) Prediction of deaths related to opioids **without Opioid as input**

Feature importance scores from RF-Mortality Model indicate that, on a relative basis, opioid transaction features ("Grams Oxycodone" and "Grams Hydrocodone") are not as important as "Estimated Total Bedrooms," "Number of Veterans," or "Number of Individuals in Poverty."



Discussion

We believe that our project can help preemptively **identify counties** vulnerable to opioid abuse and allow for national, state, or local intervention. Our data suggest that it would be wrongheaded to insist on financial remedies to the opioid crisis. School systems, controlling the crime rate, and housing availability seem to be much more impactful areas for possible local government investment.

Future considerations that could improve this project include:

(1) **More comprehensive dataset** (Despite using an 88GB dataset, we only had 12 years worth of data.)
(2) **Inclusion of restricted health data points** (e.g. comorbidities)
(3) **Further exploration of illegal markets** (Very little data currently exists.)