EPA Reporting and Cancer Trends

Created by Manlai, August, and Michael

Table of Contents

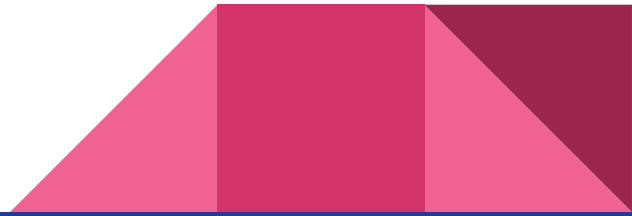
Objective

The Environmental Protection Agency makes many datasets available to the public based on records and reporting across industry and the environment.

Given a variety of factors including demographic data, water quality, and industrial chemical reporting, is there a relationship to area cancer trends?

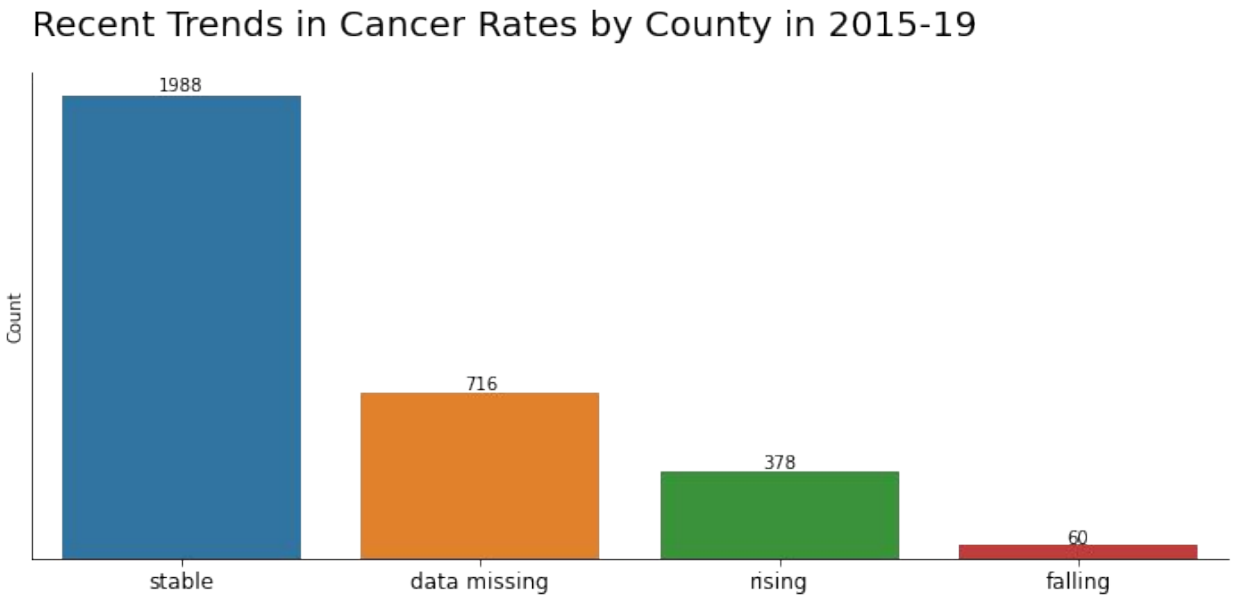
Public Good

An effective model could potentially forecast trends in health using leading environmental indicators, potentially leading to more environmentally-friendly policies and improved health.

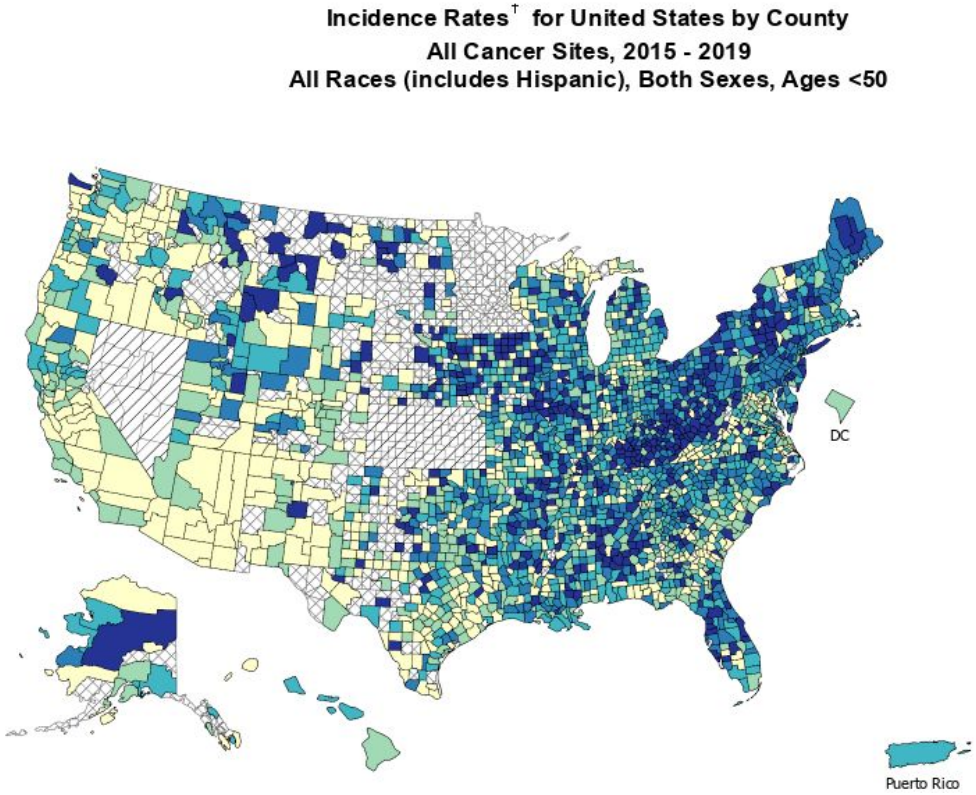


Challenges

- Imbalanced classes
- Merging datasets from different sources, different reporting years, with missing counties
- Complex subject matter with many, many variables



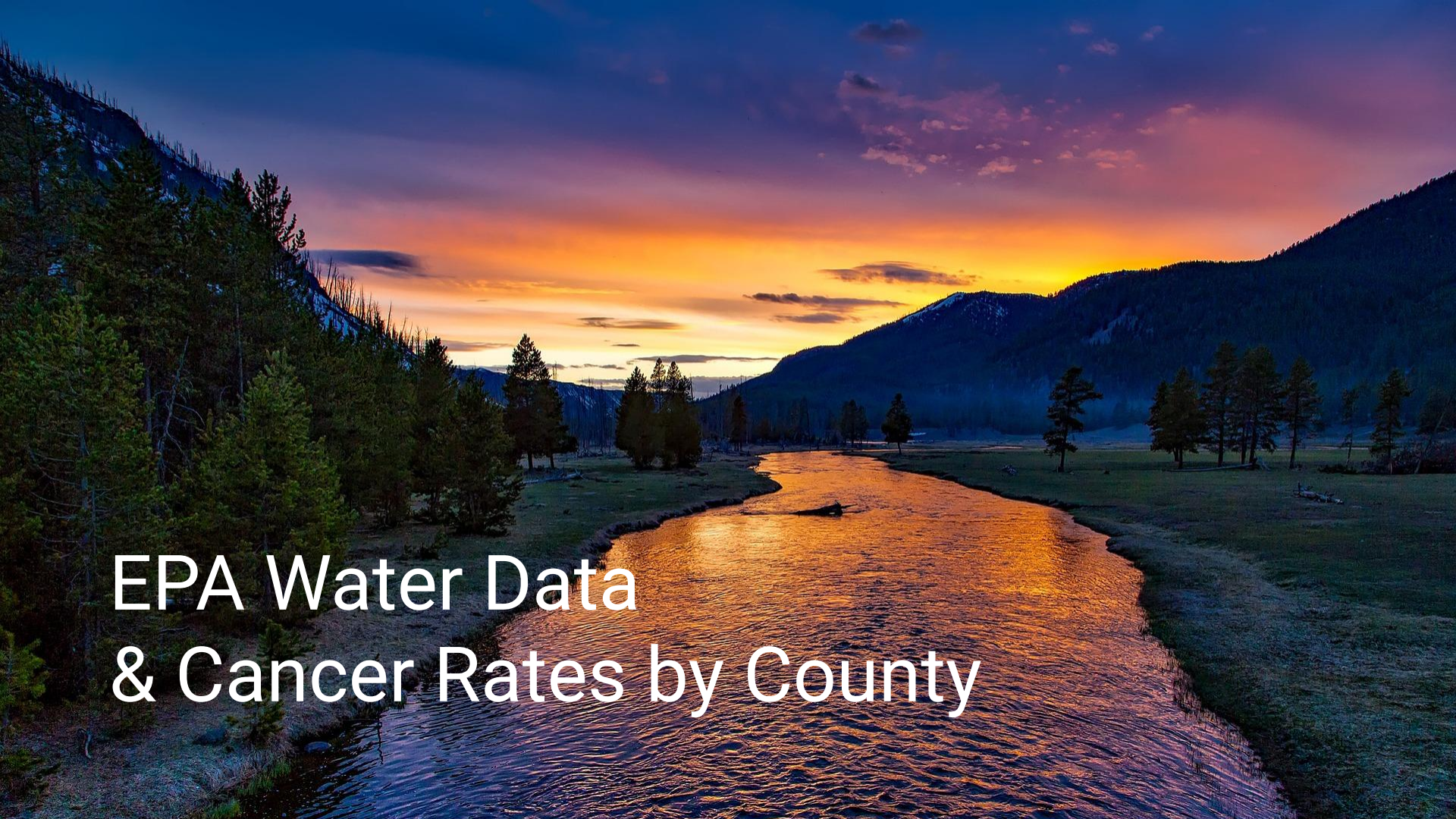
NIH Cancer Trends by County



Rising: **16%**
Falling: **3%**
Stable: **81%**

(Where county data is available. Baselines for models differ because of other county-level data).

Source: https://www.statecancerprofiles.cancer.gov/map/map.withimage.php?00&county&009&001&00&0&01&0&1&5&2#results

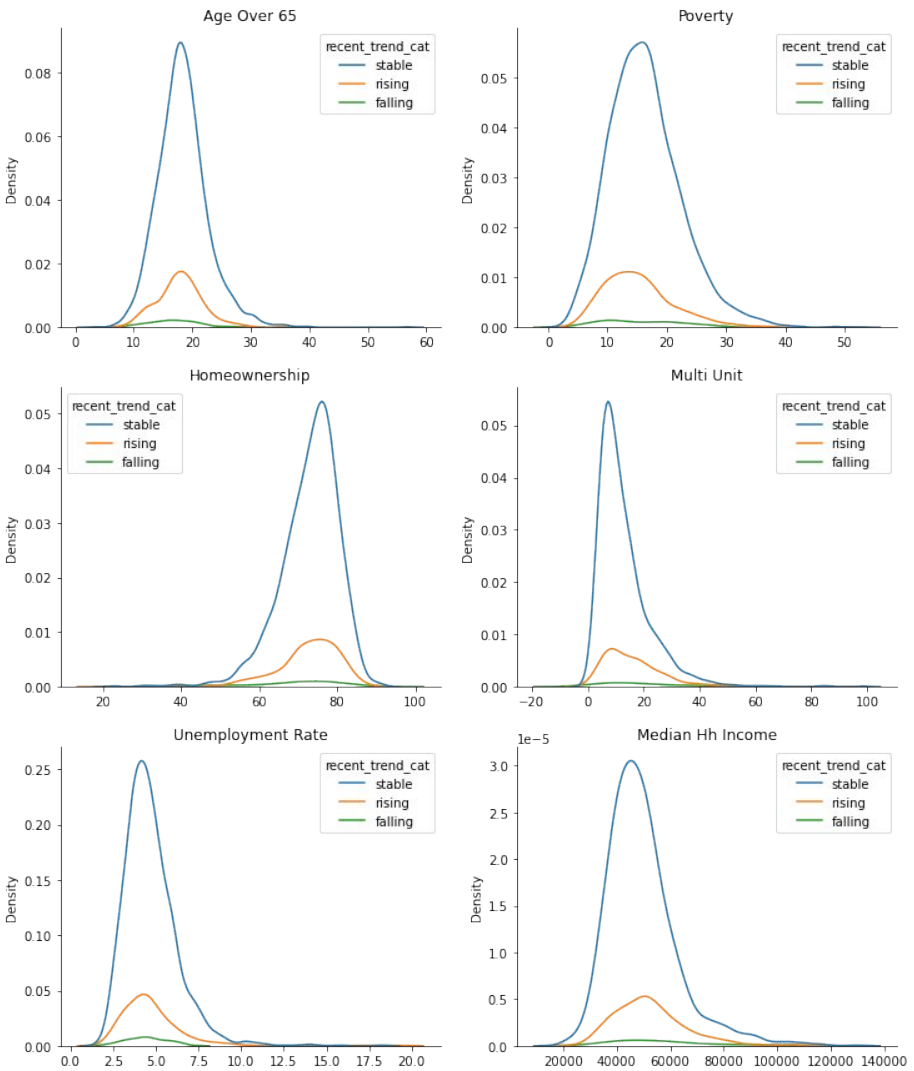


EPA Water Data & Cancer Rates by County

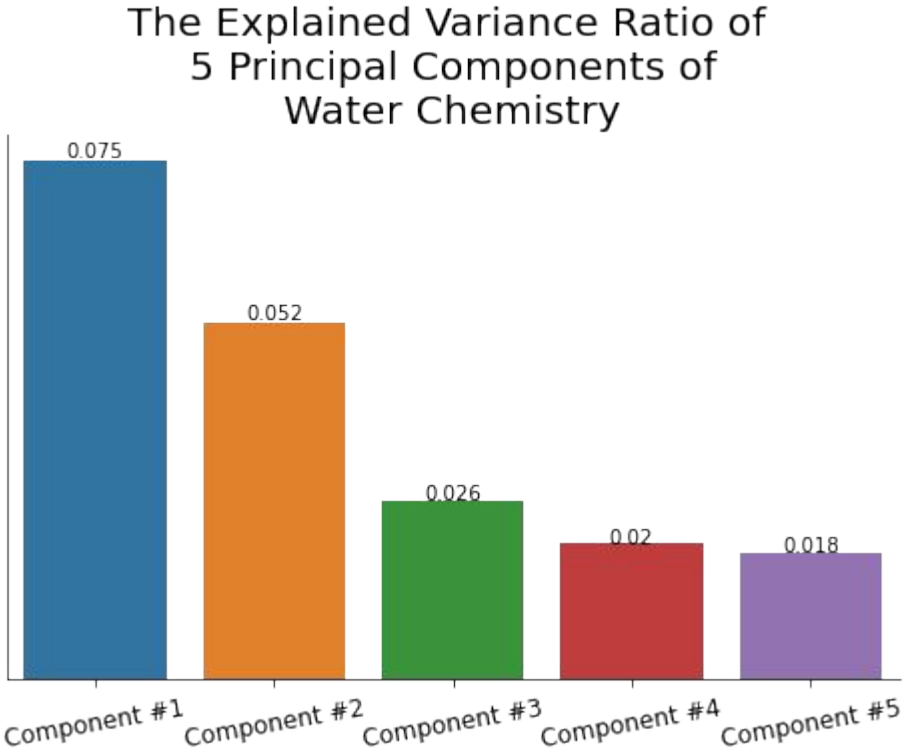
Demographic Data

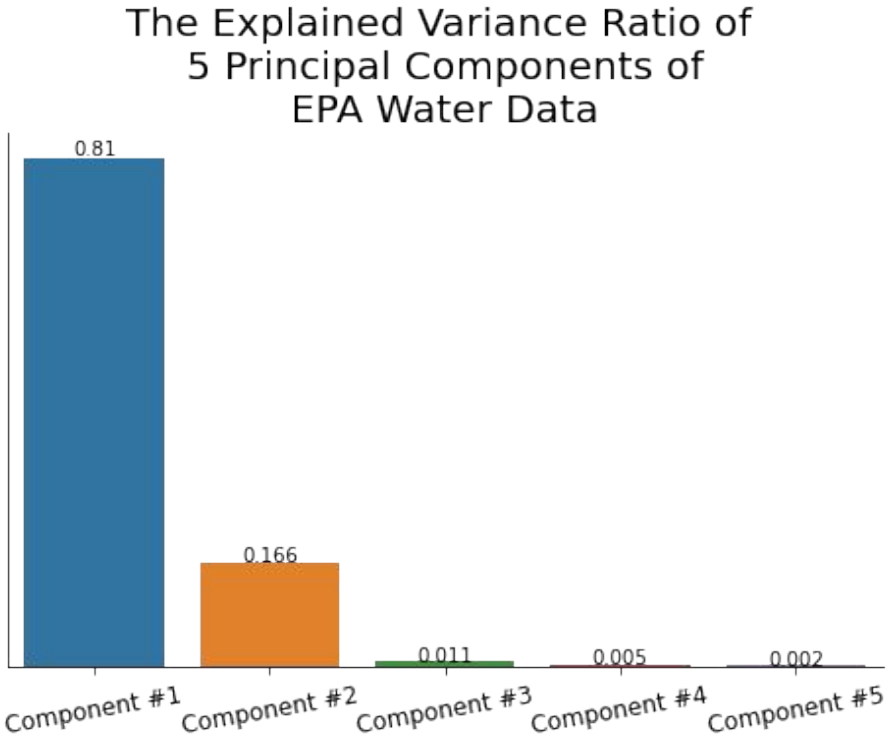
Distribution of cancer rate categories are not distinct across demographic dimensions:

- Age
- Poverty Status
- Homeownership
- Dwelling Type
- Local Unemployment
- Median Household Income

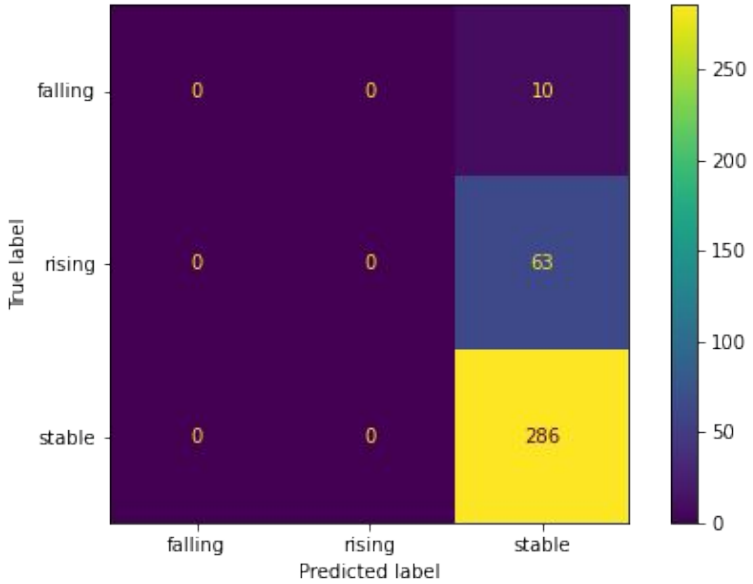


PCA on Dummy Variables





Baseline Model: Water Chemistry vs Recent Trend



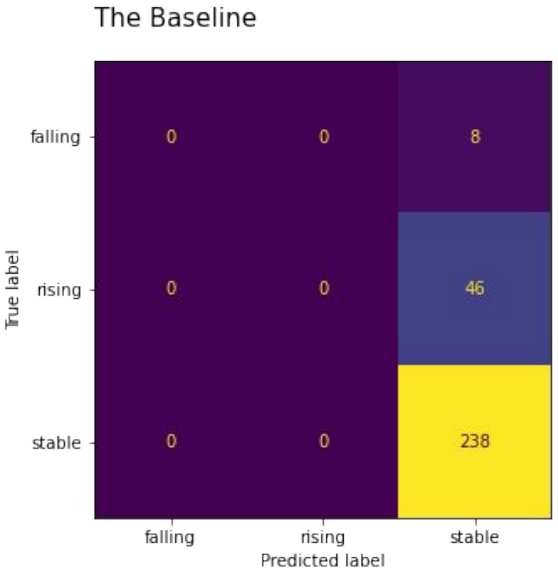
Models Compared:

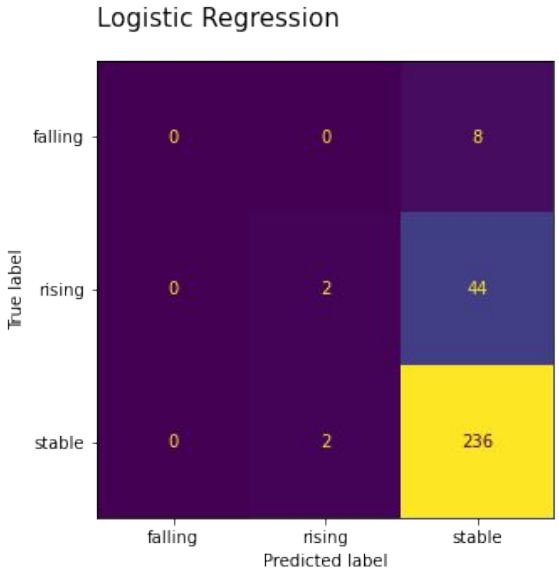
- The baseline
- Logistic Regression (with L1 Penalty)
- Random Forest (300 estimators, Depth 1)

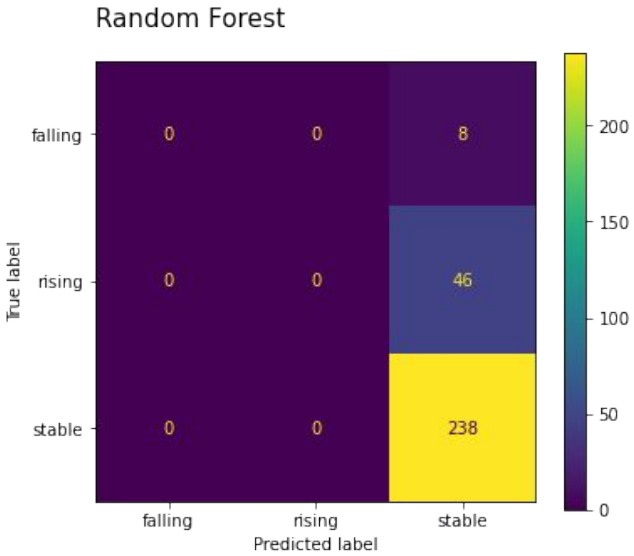
Each of these approaches performed no better than the baseline.

Baseline model: EPA Water Data vs Recent Trend

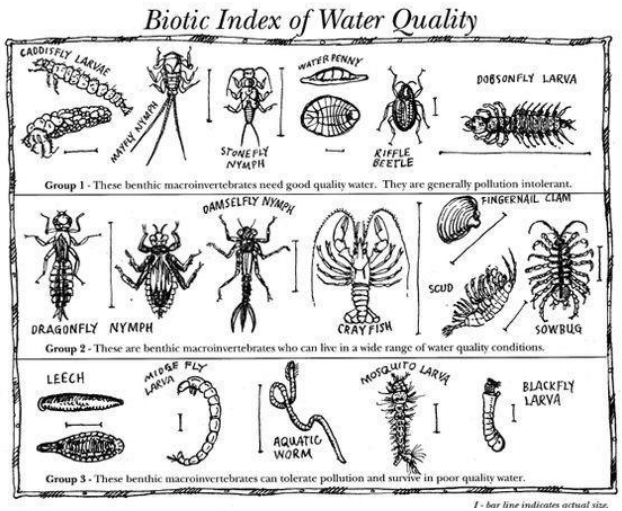
Confusion Matrices







Candidate Model: DNN on Water Data



Img source: epa.gov

The EPA publishes a large number of features measured in lakes, rivers, and streams across the US.

This dataset covered 1,168 counties with hundreds of environmental and biological factors related water quality where NIH cancer trend data was also available.

Once again, our model was at parity with the baseline.

Baseline Model:

81.5%

Validation Score:

81.6%

EPA Chemical Data Reports & Cancer Rates by County

What is EPA Chemical Data Reporting

From the EPA Website:

The Chemical Data Reporting (CDR) rule, under the Toxic Substances Control Act (TSCA), requires manufacturers (including importers) to provide EPA with information on the production and use of chemicals in commerce…

The information is collected every four years from manufacturers (including importers) of certain chemicals in commerce generally when production volumes for the chemical are 25,000 lbs or greater for a specific reporting year.

Chemical Data Reports

- 2012 Reporting
- 30,000 Individual Chemical Reports
- Merged with NIH Cancer Data by County

Example:

Chemical	Site	County	Count	Cancer Trend
Diazotized substituted benzenamines	Dow Chemicals	Fake, FL	60	Rising
Methane	Shell Chemical	Example, TX	1780	Stable
Borated reaction product of polybutenyl	United Carbide	Instance, CA	20	Falling

Chemical Data Reports

- 2012 Reporting
- 30,000 Individual Chemical Reports
- Merged with NIH Cancer Data by County

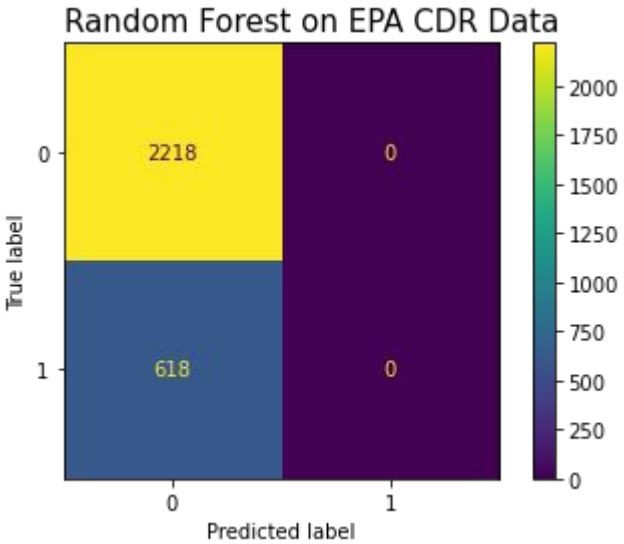
Engineered feature to capture a county's total reports with each individual report.

Example:

Chemical	Site	County	Count	Cancer Trend
Diazotized substituted benzenamines	Dow Chemicals	Fake, FL	60	ising
Methane	Shell Chemical	Example, TX	1780	table
Borated reaction product of polybutenyl	United Carbide	Instance, CA	20	alling

Candidate Model: Random Forest Classifier

This approach was almost perfectly equivalent to the baseline, so it did not represent a great model.



Baseline Model:

78.2%

Validation Score:

78.2%

Candidate Model: RF Classifier & Remove Outlier

One county, Harris County, TX, had vastly more reports than anywhere else, and it's cancer incidence was stable, so we removed this county to see if our predictions were more accurate. They weren't.

The baseline here shifted because a large number of the stable class were removed.

Baseline Model:

76.7%

Validation Score:

76.7%

Candidate Model: DNN on Reshaped Data

After several modeling attempts wherein rows were individual reports (counties have many reports), the data was reshaped to have each row represent a single county and each reported chemical as a binary feature, along with demographic data for each county.

Here again, the model was not substantially better than the baseline. This model also only predicted **seven** counties as having rising rates, out of **253.**

Baseline Model:

79.7%

Validation Score:

79.9%

Candidate Model: DNN with Duplicate Data

Since the imbalanced classes seemed to be the source of much of our difficulty, we duplicated the ‘rising’ data. This lowered the baseline (counties where cancer is not on the rise) and lifted our test accuracy slightly.

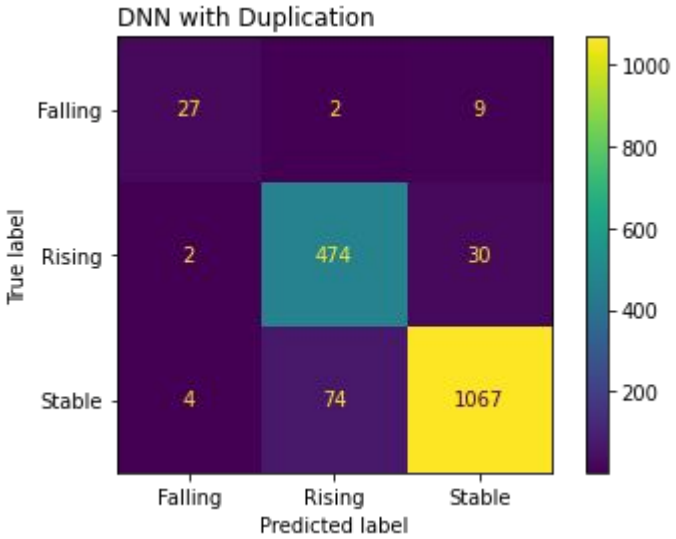
Interesting, duplicating the ‘rising’ rows also led the model to make predictions in the even-smaller ‘falling’ class, which the previous model did not.

Baseline Model:

67.7%

Validation Score:

80.7%

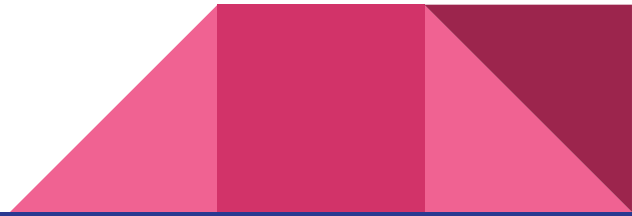


Conclusions

Improving on baseline at predicting a rare class proved difficult despite trying a number of approaches and a variety of data points. With only a few thousand counties (depending on the availability of other data) it may not be enough to train an effective model with this type of complex data.

Looking at historical data and trends over time would be one approach, and that is supported by the improved performance in the model with duplication.

Additional analysis of which features are actually important might also be a good next step.



Thank You!

Images from Pixabay except where noted.
Data from public datasets maintained by the NIH, EPA, and CDC.