Udacity MLND Capstone: Investigating the risk and preventative factors of Lung Cancer

Ryder Bergerud

June 24, 2017

1 Overview

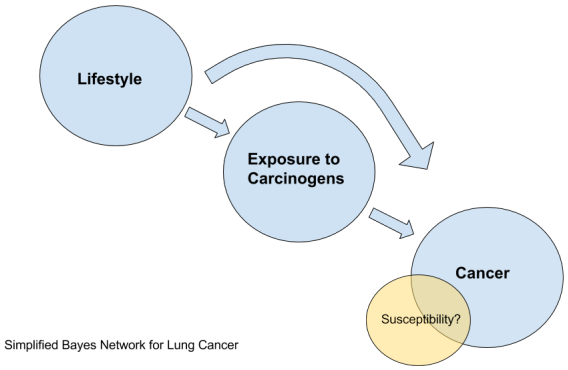
1.1 Problem Domain

I attempt to predict lung cancer rates by county in the United States with a focus on environmental factors. Environmental factors can take on a variety of definitions, the broadest being all events that are external to the body. Here we take environmental factors to mean exposure to carcinogenic substances, and lifestyles factors to mean other choices, such as physical activity and diet.

At least 70-90% of cancers are caused by extrinsic factors [1]. That is, cancer that could occur by naturally-occurring mutations cannot explain the current observed levels of cancer. Understanding the causes of cancer has been a difficult task, largely because being exposed to risk factors doesn't guarantee tumour growth, and because of the delay between exposure and tumour growth and discovery.

1.2 Problem Statement

My goal is to both try to build a model predicting lung cancer incidences by county. I will use machine learning algorithms available in the scikit learn (sklearn) library to build a predictor. I will attempt to build a model that is first of all interpretable, that is one that can give evidence to how and to what degree features relate in their prediction of lung cancer. For instance, do smokers increase their risk in an additive or multiplicative way when exposed to other air pollution? Secondly, I will try and also build a model that predicts well (see metrics). Even though the model will not imply causation between any of the features and lung cancer, epidemiological investigations like this are useful in informing and directing controlled experiments that would happen in a lab that could actually prove causation [2].



Bayes Network.png

Cancer is an environmentally-caused disease [1]. Some substances and viruses are known to be carcinogenic (cancer-causing). As well, some lifestyle choices such as physical activity can be protective factors. What data should we use to predict lung cancer? According the National Cancer Institute, the following are risk factors for cancer-diseases:

- Age
- Alcohol
- Cancer-Causing Substances
- Chronic Inflammation
- Diet
- Hormones
- Immunosuppression
- Infectious Agents
- Obesity
- Radiation
- Sunlight
- Tobacco

[3]

Not all of these are necessarily factors for lung cancer. Specific to lung cancer are

- Tobacco smoke

- Cigar smoking
- Secondhand smoke
- Radon exposure
- Asbestos
- Various inhaled carcinogens
- Arsenic in drinking water
- Air Pollution

[4]

For skin cancer, the gap between exposure to ionizing radiation and detection averages about 20 to 40 years [5]. This suggests that datasets with factors for predicting current cancer incidences should be taken from an earlier periods where possible. Where this hasn't been possible (CHR data), I assume some continuity between current and historical data by using current the data anyways.

Because of privacy concerns, it is difficult to obtain data that with records at the person-level, and so aggregated data is much more available. Since there are over 3000 counties in the US, I thought there might be enough data points to build a predictor, though only with a limited number of features.

Additionally, the US Census Bureau, through 10-year census data and other surveys, makes available hundreds of variables aggregated by county. This gives the opportunity to both manually search for features that correspond to existing identified risk factors, as well as cast a wide net to consider new ones.

Since age is a large and better-understood risk-factor that is well known, we use age-adjusted rates. Age-adjusted rates are calculated by taking the county cancer-incidence rate for each age group, and them summing them up weighted by the proportion of the total US population each age-group takes [6].

This will hopefully simplify the analysis in not having to (a) include age composition as features, and (b) not having to examine environmental causes relative to age.

1.3 Data sources

US Census data is available grouped by county. Datasets used included

- Housing (`HSG02.csv`, `HSG02_H.csv`)
- Labour Force (`LFE02.csv`)

- Manufacturing (Manufacturing_Sheet1.csv, BP_2007_00A1steel/BP_2007_00A1_with_ann.csv, BP_2007_00A1rubberandplastics/BP_2007_00A1_with_ann.csv)
- Business (BZN_01.csv)
- Population (POP01_sheet1.csv,POP02_sheet2.csv)

They key for the column names used in these files (except for files named BP_2007_00A1_with_ann.csv) is found in USCensusMastdata.xls.

Housing data was selecting to survey the age of the existing housing stock, which could be related to indoor airspace and asbestos exposure. Labour Force, Manufacturing, and Business was mainly included to examine occupational hazards that were cancer-causing, including industries such as coal and chemical manufacturing. Population was included because some data was not given as an average or rate, but this figure might allow our learning algorithm to understand the data as such. Statistics here were not averaged on a per-capita basis, though we do make per-capita copies of all features in the data-preprocessing stage. My thought was that exposure of one person might not limit exposure of another in many situations, so it might be useful to have both depending on the feature. Since it would be too tedious to examine each feature for density-dependence, I've let the feature selection stage take care of this.
https://www.census.gov/support/USACdataDownloads.html

County Health Rankings (CHR) data is maintained by the Robert Wood foundation, and collected from US Census and USDA.
- http://www.countyhealthrankings.org/rankings/

TRI (Toxic Release Inventory) data is from the Environmental Protection Agency. - File: EPA_Toxic_Air_Pollutant_Cancer_Risk_by_County.csv
- Source url: https://iaspub.epa.gov/triexplorer/release_chem?p_view=USCH&trilib=TRIQ1&sort=_VIEW_&sort_fmt=1&state=All+states&county=All+counties&chemical=OSHA_IND&industry=ALL&year=1990&tab_rpt=1&fld=AIRLBY&fld=E1&fld=E2&fld=E3&fld=E4&fld=E5&fld=E52&fld=E53&fld=E54&fld=E51&fld=TSFDSP&fld=TSFDSP&fld=m10&fld=m41&fld=m62&fld=potwmetl&fld=m71&fld=m72&fld=m73&fld=m79&fld=m90&fld=m94&fld=m99&fld=RELLBY&fld=on&fld=CAS

The earliest measures of smoking rates by county I could find were from the year 2000. Smoking by County.
Source: Zigler, Cory, 2017, "County-Level Smoking Data", doi:10.7910/DVN/VZ21KD, Harvard Dataverse, V1, UNF:6:L7kVxoDhAmjyTwP0BzYVkQ==

National Cancer Institute maintains county and state-wise estimates on incidences of different cancer diseases. Files:
- lung_incd.csv
- https://www.statecancerprofiles.cancer.gov/incidencerates/

1.4 Metric

Since I will be using regression classifiers, I will use the R2 score to value the model's prediction. R2 is useful since it allows us interpret the score as "how much better (or worse) than a baseline predictor that just predicts the mean, is our model?". The score also is normalized over the total variation over the sample, $\sum_{i=0}^{n_{samples}-1}(y_i - \bar{y})^2$, unlike the mean-squared error.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1}(y_i - \bar{y})^2}$$

It is useful to have a score that considers "outliers", unlike the median-absolute-error, since outliers might have some cause we're attempting to understand in this data.

Another metric I considered was to weight examples by size of their 95% confidence interval for lung cancer incidence rates, which is included in the original data. Unfortunately, this would undervalue rural counties, and bias factors related to the urban environment.

[2]: http://scikit-learn.org/stable/modules/model$_e$*valuation.htmlmedia* − *absolute*−*error*[3] : *http* : *//scikit*−*learn.org/stable/modules/model*$_e$*valuation.htmlmean*− *squared* − *error*

2 Analysis

2.1 Data Exploration

I have begun by collecting a large feature set for this data. The TRI data alone had 1461 features, or about 10 per carcinogen.

The following tables are an example of the data for the chemical 1,2-BUTYLENE OXIDE.

It seemed fair to consider that human exposure is often not as much about volume released as it is how it is dispersed. Fugitive air are emissions that are unintended emissions, or irregular emissions, such as flaring or leaks in pressurized equipment. This could be a greater cause of occupational exposures as compared to the more controlled Stack Air emissions.

One issue with the TRI data was that about half of the counties didn't have any records for emissions of any of the OSHA carcinogens.

2.2 Exploratory Visualization

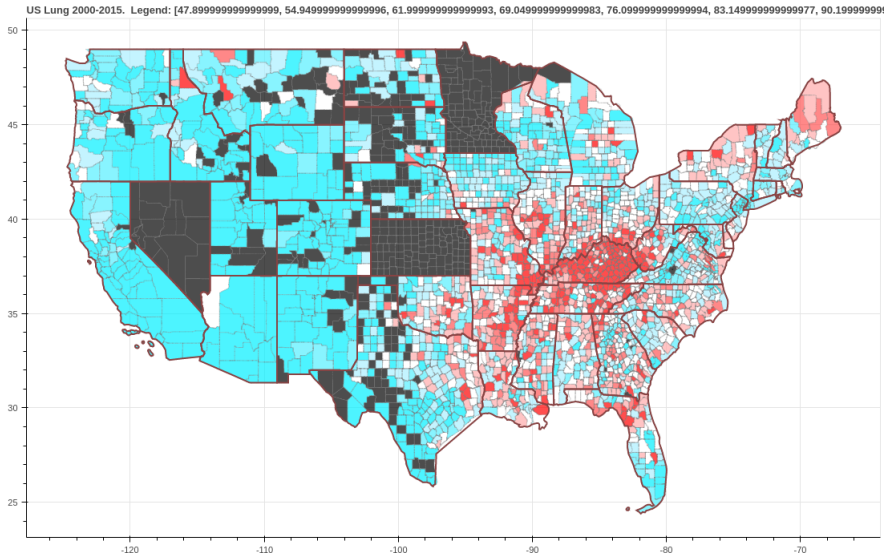
Right off the bat, you can see that TRI emissions do not exist for some of the places with the highest risks for lung cancer. This is most noticeable in the state of Kentucky. When comparing the lung cancer map to the smoking rates map of 2000, we can see that areas with higher rates of smoking also have higher lung cancer incidences 10-15 years later. Even counties that are isolated in this tendency (don't have neighbouring states with high smoking rates) exhibit this relation (counties in Idaho, Montana). We can also see that high-rates of smoking seem to drop sharply along several states lines (Illinois). This might reflect different smoking regulations found in different states at that time.

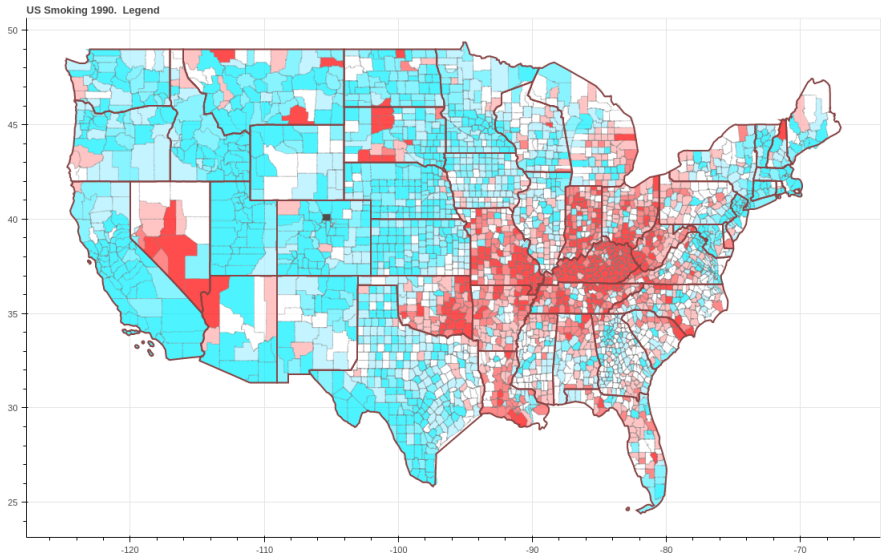
Table 1: 1,2-BUTYLENE OXIDE

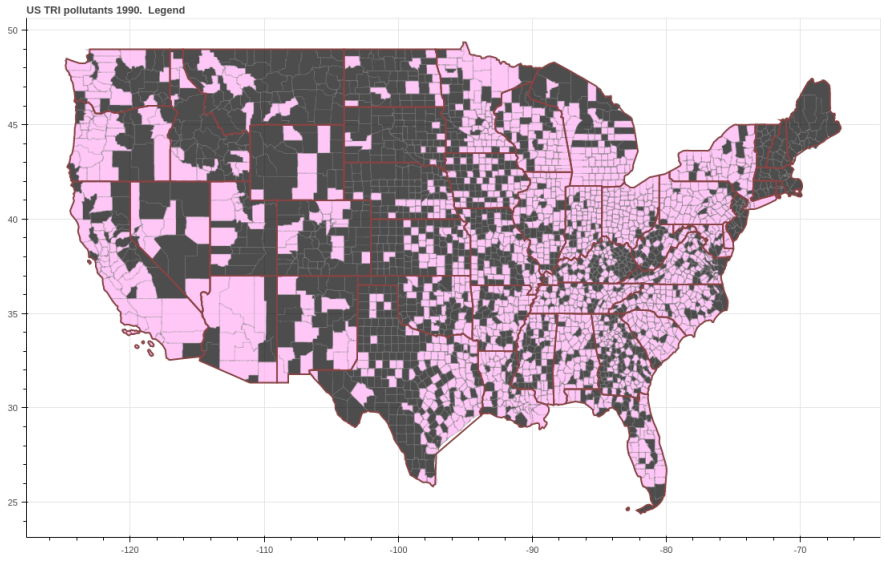
FIPS (index)	Fugitive Air	Stack Air	Total Air	Surface Water	Total Underground Injection
18165	250.0	750.0	1000.0	5.0	0.0
22005	56.0	1100.0	1156.0	0.0	0.0
22019	6908.0	324.0	7232.0	0.0	0.0

Table 2: 1,2-BUTYLENE OXIDE cont.

Total	Land Treatment Application Farming	Total Surface Impoundments	Other Land Disposal	Total On-site Releases to Land	Total On-site or Other Releases	Total Off-site	Total On- and Off-Site
ZIP							
0.0	0.0	0.0	5.0	5.0	1010.0	0.0	1010.0
47842							
0.0	0.0	0.0	0.0	0.0	1156.0	0.0	1156.0
70734							
0.0	0.0	0.0	0.0	0.0	7232.0	0.0	7232.0
70669							


US Lung 2000-2015. Legend: [47.899999999999999, 54.949999999999996, 61.999999999999993, 69.049999999999983, 76.099999999999994, 83.149999999999977, 90.199999999
50
45
40
35
30
25
-120
-110
-100
-90
-80
-70


US Smoking 1990. Legend
50
45
40
35
30
25
-120
-110
-100
-90
-80
-70



2.3 Algorithms and Techniques

Because the sparsity of the data, and the lack of data relating emissions to exposure, I was interested to see how well the TRI data could predict various incidences of cancer on its own.

I will use Linear Regression and Decision Tree Regressor learning algorithms for my interpretable models. The Decision Tree Regressor works when passed an excess of features, since the depth or maximum size of leaf node can be set. However, Linear Regression seems to overfit easily when given too many features.

2.4 Benchmark

I could not find a project that tried to predict lung cancer rates using geographical (non-patient) data. As a benchmark, I will take the success of predicting lung cancer rates using on smoking data, since it is the most obvious cause/factor.

The best score for predictions of lung cancer based on smoking along is 0.47 using Random Forest Regressor, and 0.45 using Decision Tree Regressor.

3 Methodology

3.1 Data Preprocessing

Because our choices of algorithms (Random Forest, Decision Tree, Linear Regression, Gradient Boost) are independent of feature-scaling (unlike KNN or SVR), we omit scaling and normalization from data preprocessing.

In the data preprocessing, because of the sparsity emissions recorded for most carcinogens, we only take data from carcinogens that have positive emissions for at least 1600 counties.

First, we combine features into either `carcinogen_features`, or features that measure carcinogenic substances in some way, or `life_factor_features`, or features that are protective or might correlate with exposure to carcinogens.

For every feature, we then create an additional feature which describes the value of that feature per capita according to the county's 2010 population estimates. This is because many features in the US Census datasets are total counts of some value, such as employees in a certain industry, for the county. For the learning algorithm to make use of this to predict a rates instead of total cases of lung cancer, it makes sense to transform this data to rates per capita.

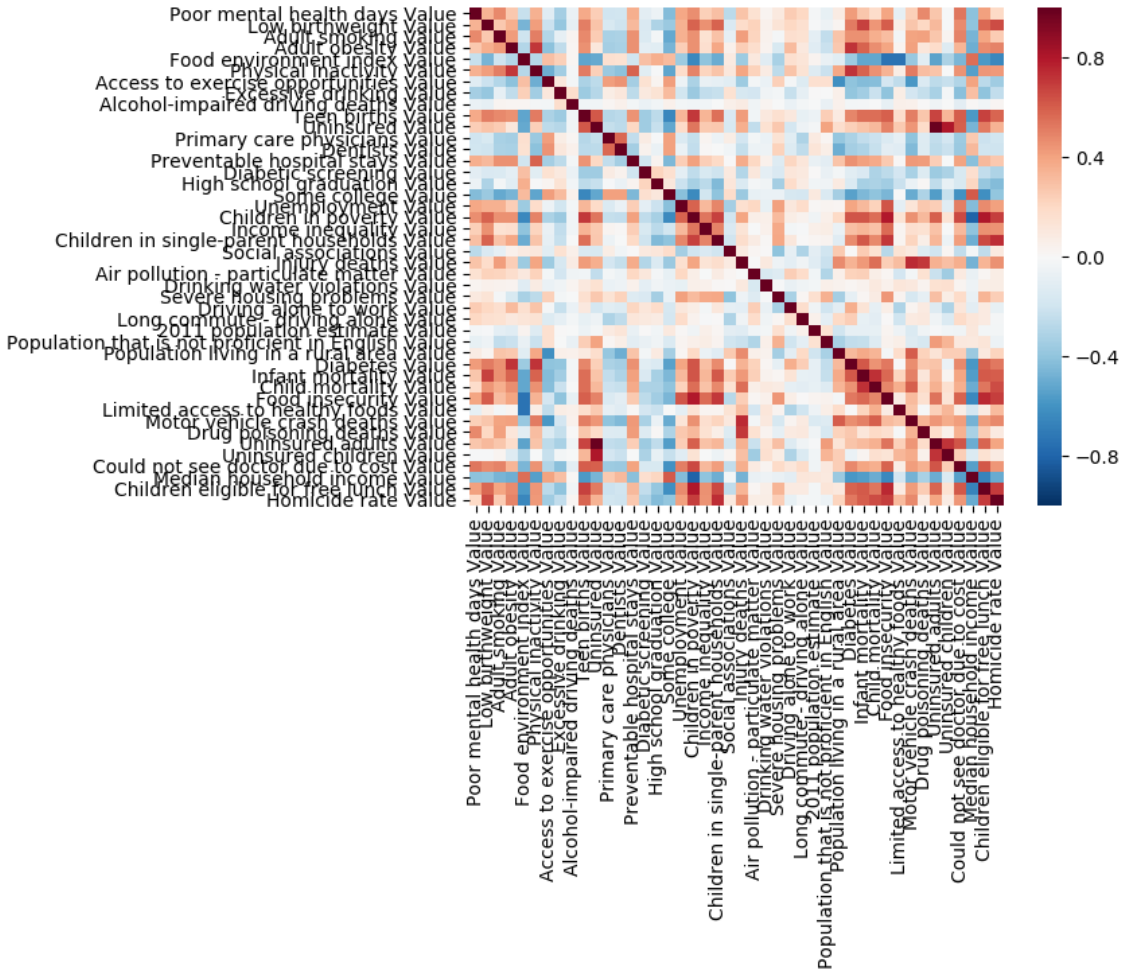
Because of the compounding effects of carcinogens and other cancer factors as shown in research [5], I reduced features using recursive feature elimination with cross validation. Other methods such as KBestSelect will consider variables in relation to the target in isolation, we might miss this compounding effects when choosing our features.

At first, I tried implementing RFECV in two phases. First, since there are 686 possible features to choose from, we first run RFECV with a large elimination step. RFECV eliminates 20 features at each step. The goal here is to quickly remove as many unrelated features as possible, and then with a finer step, consider a smaller list of most relevant features.

Running RFECV this way, I found sometimes the feature set choosen could change significantly, and features that should be included like smokerate2000 were left out. It seems that the feature set becomes more unstable if certain features were sifted out early in the elimination process.

I decided for this reason to try out a possibly more stable feature selection method that would work for a large number of features.

L1 Recovery of features works well for features that don't have strong correlations between each other [7]. L1 will randomly pick among highly-correlated features. However, randomized L1 runs L1 several times giving randomized preweights to features. Randomized L1 will then rank features based on how often they were selected between trials, and so will select correlated features. This is appropriate for this situation where there are many correlated features (see correlation matrix for CHR data).



We can therefore select relevant and correlated features, and then look again at the correlation within the selected group to see whether any further dimension reduction should be done via PCA or IDA.

The downside to this method is that we might miss any features that are interaction features (they predict in a non-additive way), but their main effect (linear relation with target) is insignificant or 0. This is an unlikely situation, but we know ahead of time that there are interaction features (asbestos and lung cancer) [8]. However, in this example, both features (asbestos and smoking) have a main effect.

We can combine using Lasso as a coarse filter, and then using RFECV as a finer sieve that considers interaction effects. Doing this selects the following features:

```
'smokerate2000_x', 'Physical inactivity Value_x', 'Unemployment Value_x',
'Air pollution - particulate matter Value_x', 'Uninsured  Value_x',
'Uninsured children Value_x', 'MAN110202D_y', 'LFE330200D_y',
```

`'LFE305200D_x', 'LFE330209D_y', 'HSG170209D_y', 'HSG305200D_y', Physical inactivity Value_y`

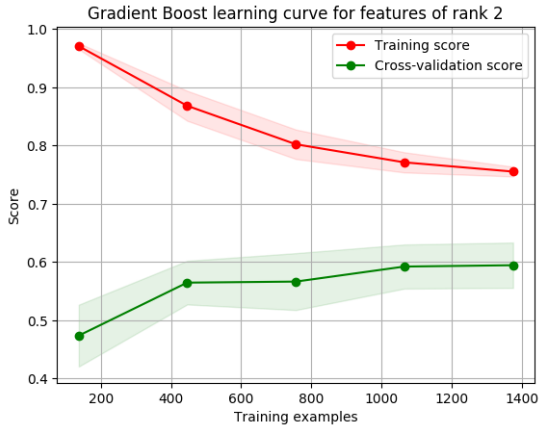
The features with description are:

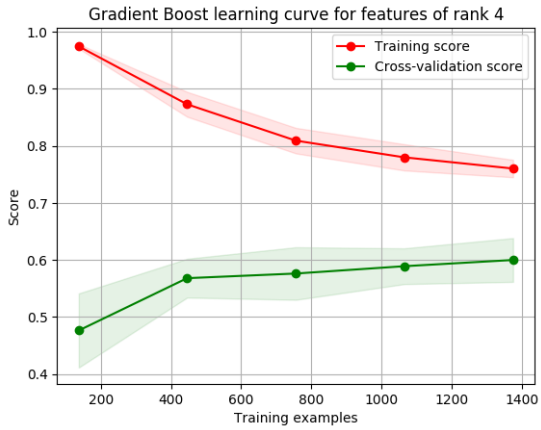
- Smoking rate 2000
- Manufacturing: total (NAICS 31-33) - all (MAN130207D) establishments with 20 or more employees 2007
- Physical inactivity Value
- Air pollution - particulate matter Value
- Uninsured Value (Health insurance)
- Manufacturing: total (NAICS 31-33) - all establishments 2002 (MAN110202D_y)
- Employed persons by industry (NAICS) - agriculture, forestry, fishing and hunting, and mining 2000 /per capita (LFE330200D_y)
- Average travel time to work for workers 16 years and over who did not work at home 2000 (LFE305200D_x)
- Employed persons by industry (NAICS) - agriculture, forestry, fishing and hunting, and mining 2005-2009 /per capita (LFE330209D_y)
- Housing units by year structure built 1939 or earlier, 2005-2009 /per capita (HSG170209D_y)
- Occupied housing units with 1.01 or more persons per room lacking complete plumbing facilities 2000 (sample) /per capita (HSG305200D_y)

Here the postfix y implies that the statistic has been divided by the population of the state, where `x` is unaltered.

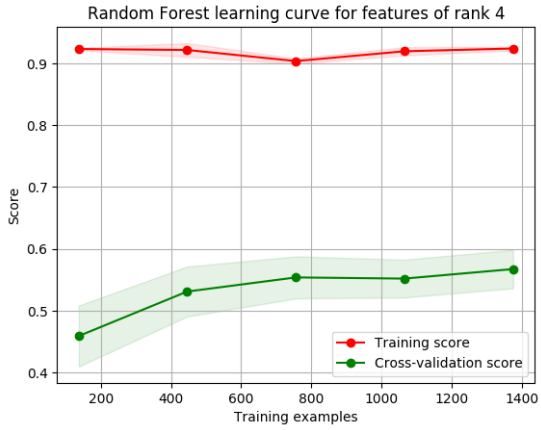
We drop `Physical inactivity Value_y` from the set of features, since its meansing is not clear (it is already a sort of per capita rate) and so it doesn't seem to add any additional information beyond `Physical inactivity Value_y`. We also drop `'Uninsured children Value_x`, since it likely occurs exactly in the same situations as `'Uninsured Value_x'`.

The plot shows that adding additional features does not seem to be altering the model's bias. That is, the (asymptotic) difference between the training data and testing data in CV is the same for features ranked < 2 by RFECV as ranked < 4. This shows evidence that we are likely instead missing features.





The random forest shows much stronger variance, likely due to the model overfitting the training data. This can likely be controlled by tuning parameters later on.



3.2 Implementation

First I examine models that are built around the TRI database. The goal is to see if there emissions of carcinogens from point sources are useful to know own their own for predicting lung cancer. Then we build a model of our selected features from the previous stage.

Several simple, untuned models show that TRI data alone cannot cannot inform predictions of lung, bladder, leukemia, breast, childhood, or lympahatic cancers. In all these cases, this data wasn't able to perform better than the mean, hence hovering around R2 scores of 0. This doesn't mean necessarily that these features are not factors in cancer. For instance, if their joint distribution with major causes like smoking are not independent, a negative correlation between the two could show no relation between just TRI data and cancer incidences.

Using the reduced feature set, training models without tuned parameters shows a better than benchmark score for most ensemble methods, but not so much for the Decision Tree Regressor. One reason might be that without parameter-tuning, a decision tree is prone to over-fitting (default max-depth is none).

3.3 Refinement

To refine the models, I use grid search with cross validation, but in several stages so as to reduce search costs. For ensemble methods, I search among number of estimators (`n_estimators`), since models are most sensitive to this parameter.

4 Results

4.1 Model Evaluation and Validation

Looking at the residuals, 80% of the data falls between -14.94 and 11.57. Looking at the largest residuals (points where the model's error was greater than 2-), their mean smoking rate was a full standard deviation away from training data's mean. So our model has the most difficulty predicting lung cancer incidences when smoking rates are higher. If we were to restrict the data to just counties with low smoking rates, they would likely have lower rates of lung cancer, and hence estimating the mean among these would have less variance, as opposed to the same exercise among counties with high smoking rates. This effect might explain why we see larger absolute error in predicting lung cancer in counties with more smoking.

One question to consider at this point is: how does this model scale? Are the variables scale-free? That is, how well does this model perform if we ran it

at the state level? It is possible that for some features have values that would be less homogeneous at larger scales, such as air quality. Therefore the model may not apply as well with a state-wise value for air quality that varies greatly within the state, as opposed to a measurement that's more consistent at the county level.

Why in general does our TRI data not add significantly to better scores or good predictions on its own?

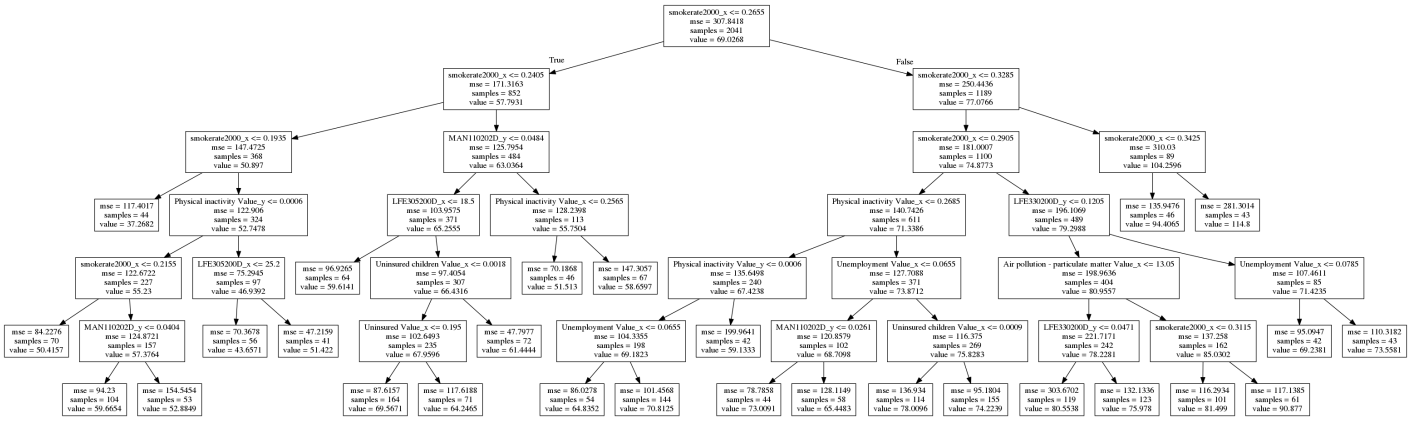
This could be for a few reasons. While these chemicals are known carcinogens, whether and how much they are dispersed from these point sources is not captured in the TRI data. Also, the nature of workplace exposure is not clear from the data.

With many carcinogens, accumulation over time has a linear effect on the probabilities of developing a tumor [9]. Though an inspection of other data from the 1990s shows similar sites for toxic releases, the time period for exposure (1990 until 2009-2013 incidence data) might be too short to develop enough exposure and diagnosis, especially for low-concentrations of dispersed carcinogens.

4.2 Justification

Compared to the benchmark, the inclusion of additional features and training gives a better performing model than our benchmark.

Running the whole modelling process several times shows that the score assigned is fairly robust to the test-train split.



The tree regressor doesn't perform much better than the benchmark tree regressor (0.52 to 0.45). The decision tree for the tuned tree trained on our selected features is shown above.

The score for the gradient boosting predictor varies between 62 and 64 after running it several times. This is likely because our model is still slightly sensitive

to changes in the test/train split, and possibly because of some lack of stability in the feature selection.

Scored feature importances were surprising since smoking along was so successful at predicting lung cancer incidences.

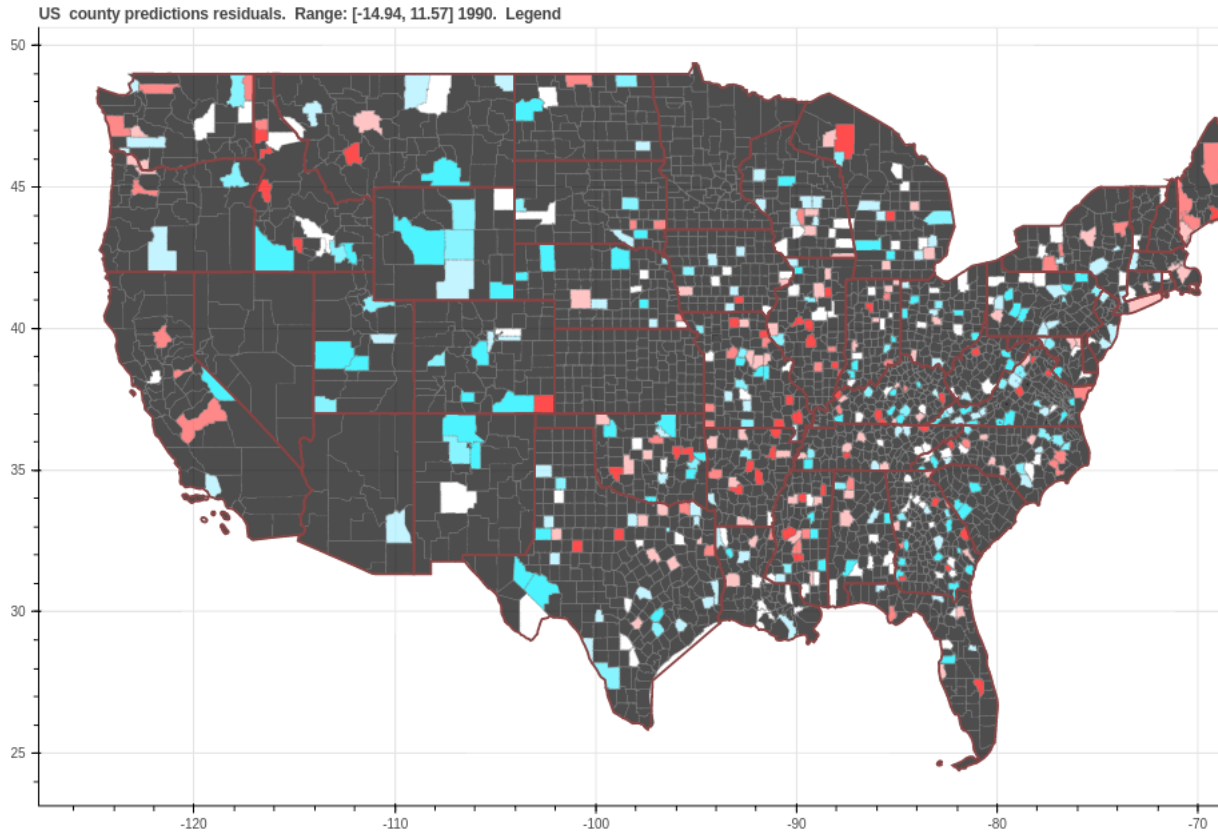
```
(0.090090950383653667, 'LFE305200D_x')
(0.2035934518879296, 'smokerate2000_x')
(0.080709635610441699, 'Physical inactivity Value_x')
(0.0699130453549901, 'Uninsured Value_x')
(0.084411164459612525, 'Unemployment Value_x')
(0.089119589184111767, 'Air pollution - particulate matter Value_x')
(0.080280349560267986, 'Uninsured children Value_x')
(0.039634892420213792, 'MAN110202D_y')
(0.10224914041321241, 'LFE330200D_y')
(0.11627696420339799, 'HSG170209D_y')
(0.043720816522168547, 'HSG305200D_y')
```

It turns out living in an old housing and working in one of Agriculture, Hunting, Fishing, or Mining are the second highest factors of importance in best-scoring model. There are many occupational exposures to carcinogens that happen with mining. Pesticides might also be a factor among agricultural workers. Its possible that older houses are candidates for renovations/demolitions, which could cause harmful exposure to asbestos.

5 Conclusion

5.1 Free-form Visualization

The following visualization is useful to see if their are geographically-related features that might be missing in the model. That is, if some larger areas are being under-predicted, then we might be missing a risk factor in our feature set that is better represented over that area.



It seems there might be underestimation of lung cancer incidences around states the border the Mississippi river. However, it might be plausible to see this much clustering in a uniformly random distribution of errors.

5.2 Reflection

Finding enough data to include a variety of known cancer-causing factors in the analysis was especially challenging, since so much data is collected at the state level. This project started out as an attempt to try to predict cancer incidences, but after having poor success early on training a model, I became more aware that there are often more distinct (primary) causes among different cancers than common ones, and it would be best to focus on modelling just one.

5.3 Improvement

Investigating spatial auto-correlation in the data could be useful, especially in looking at the residual, as a technique to consider whether there is a missing feature [10].

It would have been best to find data over intervals through the past few decades since there is lag between exposure to carcinogens and development of cancer. Unfortunately, this wasn't available. However, if it was available, we could have trained the same model on statistics from earlier decades, and have a predictor for lung cancer rates in the future given current census statistics.

References

[1] Song Wu et al. Substanial contribution of extrinsic risk factors to cancer development. *Nature*, 529(43), 2016.

[2] Extension Toxicology Network. Toxicology information brief. http://pmep.cce.cornell.edu/profiles/extoxnet/TIB/epidemiology.htm.

[3] National cancer agency - causes prevention. https://www.cancer.gov/about-cancer/causes-prevention/risk.

[4] National cancer agency - causes prevention non small cell lung cancer. https://www.cancer.org/cancer/non-small-cell-lung-cancer/causes-risks-prevention/risk-factors.html.

[5] Weichselbaum RR Bast RC Gansler TS Holland JF Frei E. Kufe DW, Pollock RE. *Cancer Medicine, 6th ed.* 2000.

[6] National Cancer Institute. Seer*stat tutorials: Calculating age-adjusted rates. https://seer.cancer.gov/seerstat/tutorials/aarates/definition.html.

[7] Sparse recovery, feature selection for sparse linear models. http://scikit-learn.org/stable/auto_examples/linear_model/plot_sparse_recovery.html#sphx-glr-auto-examples-linear-model-plot-sparse-recovery-py.

[8] R. Saracci. Asbestos and lung cancer: An analysis of the epidemiological evidence on the asbestos—smoking interaction. *International Journal of Cancer*, 20(3):323–331, 1977.

[9] Epa risk assessment of carcinogenic effects. https://www.epa.gov/fera/risk-assessment-carcinogenic-effects.

[10] Sara Gale Michael Jerrett and Caitlin Kontgis. Spatial modeling in environmental and public health research. *Int J Environ Res Public Health*, 7(4):1302–1329, 2010.