EXPLAINABLE ENVIRONMENTAL INFLUENCERS:
USING ANALYTICS TO IDENTIFY OUTSIDE INFLUENCES ON LOCAL POLITICAL FIGURES FOR ENVIRONMENTAL POLICIES

Katie Finnegan

Kevin McClenahan

Katelyn Work

Practicum Report

Georgia Institute of Technology
Master of Science in Analytics

Sponsored by the Environmental Defense Fund

April 2023

Abstract

Over the past several months, our team has worked towards providing the Environmental Defense Fund (EDF) with scripts which automate data pulls, aggregate and summarize data, model policy making decision making, and visualize results. Our goal was to provide tools which aid EDF in making steps to further their own internal models while also giving insight into local legislators and into what data sources may be useful for EDF to pursue as they continue to untangle the complex area of local legislative decision making.

This document outlines the methodology we used to solve specific challenges presented to the team by EDF, the challenges we faced, and our recommendations for next steps.

I. Introduction

Despite the significant advancements in advancing policy to address climate change in recent decades, there are still obstacles which may impact the speed of creating new bills or strength of the legislation. When analyzing the potential factors at play, it is important to consider the impact of outside entities on politicians.

The Environmental Defense Fund (EDF) is developing an algorithmic method to model the third party influences impacting environmental policy creation at local levels of government in the United States. Understanding these outside entities in the broader social network of a politician, referred to here as "Explainable Environmental Influencers", can help the EDF team to prioritize actions to address potentially bad players with data-backed, quantitative metrics. Given the usual lack of easily accessible information at lower levels of government, it becomes especially important to apply systematic models at local and state levels to efficiently understand the influential landscape.

For this practicum, the EDF team has provided a selection of potential deliverables, or "challenges", which break apart creation of this influence model into smaller tasks. Each challenge addresses a particular vector of influence on a politician and requires combining and analyzing data sources to model levels of political influence exhibited by outside individuals and organizations. Our team chose to take on Challenge 5 ("Geospatial Intelligence for Environmental Influence"), Challenge 7 ("The Census Taker"), and Challenge 8 ("List Locally").

Challenge 5 sought to understand how to map the political influence of geography to politicians. Particularly, districts in certain geographic locations may contain businesses, organizations, or key individuals who seek to influence their local politicians to act in their favor in creating or barring environmental policies.

The intent of Challenge 7 was to analyze United States Census data and algorithmically create a confidence score for the data quality of each congressional district. After known issues with data collection for the 2020 Census due to the COVID-19 pandemic, the EDF team needed a more thorough understanding of the confidence in census data or any notable inconsistencies when considering the use of the data for creating other models.

Finally, Challenge 8 asked for an aggregation of local legislation including modeling what category the legislation falls under, which city or county it belongs to, and who is supporting or against it. While this type of information is readily available at the federal level, indexing this data at the local level becomes much more difficult.

This report will describe our data sources, methodology, and results for each respective challenge with the intent to summarize our learnings and provide an overview of our generated models built for the EDF organization.

II. Methodology

A. Geospatial Intelligence for Environmental Influence (Challenge 5)

For this challenge, the goal was to 1) provide a dataset to EDF which contains addresses for businesses within a particular geographic area, and 2) provide a geo-spatial visualization of state US legislative districts, along with supporting information about what businesses or business sectors may have influence within that area.

For the first part of the challenge, we created a Python script which will pull the address for any company which has made a filing with the SEC. For the second part, we utilized a natural language processing (NLP) model to find the similarity of legislation being supported by a particular candidate and business sectors in that area.

We then used this as a part of an influence score, which takes into account legislation, as well as the footprint of business sectors within a specific district. The highest scoring sectors were mapped within an interactive visualization. While this approach did not ultimately yield the results we had hoped for, it did still lead to the creation of a rich geographic dataset and concrete next steps to pursue.

i. Data Sources

- US Census Data
  - State ShapeFiles: A ShapeFile is a file which contains geometry which can be used for visualization of geographic data. These ShapeFiles contain boundaries for lower legislative district (i.e. House level equivalent for each state) and state.
  - Business sector information: These files contain the most up-to-date numbers on employment and physical locations by business sector and county within a state.
- OpenStates API
  - We utilized OpenStates to pull down: 1) name of current local representative by district, and 2) currently sponsored legislation per each representative
- SEC Filings
  - We utilize the SEC website to generate a list of all businesses which have created a filing with the SEC, along with their address. While this was not used within our model below, it was requested by EDF as they believe it would be helpful within their own influence model. We were able to scrape a little over 12,000 business locations from this website.
- S&P 500 Data
  - We finally created a lighter-weight version of the SEC filing script, which restricts to pulling data only for companies which are within the S&P 500. This reduced the time to pull information from over an hour to less than a minute, due to the limited scale.

ii. Analysis

Influence Model - Overview

Throughout the semester, we provided EDF with "influencer stories" which highlight the potential influencers for an individual politician. For Challenge 5, we saw a way to automate this process, to provide potential influential business sectors within a particular district.

Ahead of formulating our model we investigated three local politicians and their potential business influences manually, which helped identify the correct data sources as well as become familiar with what results might be expected when we designed our model. In doing so, we identified that the US government classifies business sectors via the North American Industry Classification System (NAICS). This groups businesses into both high-level and more granular categories (See Appendix 1 for full list of codes). It allowed our team to have an identifier across multiple data sources which we could join to so that we could create a more robust data source about industry.

The NAICS data enabled us to manually analyze employment and geographic footprint (i.e. number of physical locations) by business sector at the state and county level. We assumed that a business sector which employed many people or had many physical locations in a given state district, could lead to that sector being influential towards a politician for their policy making decisions.

After this manual analysis, we designed a Python script which pulls the same information from the US Census website, and integrated this information into our model. In total the script generates a geography file containing: district, representative name, party affiliation, and top business sector by number of people employed, total annual payroll, and total number of physical locations.

Beyond the employment and/or physical presence, we wanted to connect this information to policy, to build a stronger case for influence. To do this, we sought to tie the politician to an industry, either through donations or through the politician's policy decisions. Tying a donation to a particular industry proved to be cumbersome, as it can be difficult to relate a business name back to an industry. However, we were able to form a methodology for tying sponsored legislation back to an industry.

Using the title of the sponsored legislation, we found the "similarity" to each business sector name in the NAICS, utilizing a NLP model comparing word vectors of bill titles and business sector name. If the similarity breached a threshold, we would consider the legislation to be related to that sector. We talk about the evaluation and results below, as this particular model did not yield great results. We wouldn't recommend productionalizing this for EDF, but did find that it may be a way to parse through legislation and remove bills which are not related to industry at all.

Influence Model - Evaluation

We utilized the Python package SpaCy to conduct our NLP analysis. SpaCy has several different options which can be chosen to evaluate similarity. The similarity feature works by converting two pieces of text to word vectors, and then comparing those word vectors to each other. Word vectors are created utilizing pre-trained models from SpaCy that can be configured at runtime.

The under-the-hood math is shown below, summarized as the dot product of the two word vectors divided by the product of their norms:

```
result = xp.dot(vector, other.vector) / (self.vector_norm * other.vector_norm)
```

After running our data through SpaCy and comparing bill titles to legislation, we found that SpaCy may not be the right package for this use case. In short, there may not be enough text for SpaCy to evaluate to determine a similarity score. We came to this conclusion as we tried multiple configurations, but all had a lack of variability between the similarity scores for a particular bill. Meaning that, in general, the similarity method will return scores which are close to each other for most to all business sectors. It does not do a good job of distinguishing between specific sectors. An example is shown below, where the first bill does not match well with any sector, and the other has similarity scores which match "well" with many sectors. The sector which should be selected as the most similar is highlighted in Red for the second example.

	Bill Name	
NAICS + Sector Name	Foster, Autherine Juanita Lucy, death mourned	Teachers' Retirement System, employees who served as postdoctoral fellow, allowed to purchase up to two years credit in the system
11: Agriculture, Forestry, Hunting & Fishing	0.066	0.619
21: Mining	0.146	0.705
22: Utilities	0.262	0.624
23: Construction	0.270	0.581
31-33: Manufacturing	0.129	0.582
42: Wholesale Trade	0.063	0.627
44-45: Retail Trade	0.159	0.710
48-49: Transportation & Warehousing	0.163	0.592
51: Information & Media	0.057	0.592
52: Finance & Insurance	0.126	0.636
53: Real Estate Rental & Leasing	0.095	0.578
54: Professional, Scientific, and Technical Services	0.203	0.590
55: Management of Companies & Enterprises	-0.013	0.316
56: Administrative Support & Waste Mgmt.	0.040	0.405
61: Education	0.206	0.681
62: Health Care & Social Assistance	0.161	0.604
71: Arts, Entertainment, and Recreation	-0.001	0.369
72: Accomodation & Food Services	0.167	0.686
81: Other Services	0.028	0.706

Table 1. Sample Similarity Scores for two Bills

We attribute this issue to lack of text to evaluate, which creates a small vector for SpaCy to evaluate. We would propose to utilize a classification model instead. However, a classification model also has drawbacks, because some training set would need to be produced, which is tedious to build. We do think, however, that this model or something similar could be used if EDF wished to download legislation in bulk, and remove bills which may be more administrative in nature. These are bills such as Alabama HJR8: “To Recognize March 2023 as "Chronic Kidney Disease Awareness Month", which simply provide recognition, acknowledgment, or are more administrative or symbolic in nature. Detailed results are discussed in a subsequent section.

Business Location Data Source

After receiving feedback from EDF, they indicated a hope to still be able to locate businesses by address in an automated fashion.

Our first thought was to utilize stock symbols and then somehow attach the address via another website. We were unable to utilize this method, but through this exploration came up with two alternatives, one light-weight, and one more comprehensive. Both methods utilize Security Exchange Commission (SEC) filings to pull address information. When a company files with the SEC, the address of company headquarters is listed on the filing.

Required to pull this filings is the Central Index Key (CIK), which is a unique identifier tied to a company who has completed a filing:

1. **Light-weight - Pull data for companies listed in the S&P 500:** For this method we pull data into a dataframe in Python directly from Wikipedia. This should contain fairly up-to-date information on which companies are in the S&P 500, although we acknowledge it may not alway be completely accurate. This table contains the CIK for each business, making it possible to pull data from the SEC web API. The script is simple and loops through each business and pulls the relevant business data.
2. **More Comprehensive - Pull data for any company who has filed with the SEC:** This method, while still fairly simple in design, takes much longer to execute as it handles a little over 12,000 businesses. There isn’t a bulk operation to pull many CIKs at once at this time from the SEC. This does give a more comprehensive picture of business locations throughout the US, though. One of our learnings from this exercise is that some companies which file with the SEC are international companies, so they end up not being useful for EDF. We weren’t able to solve this issue with the time allotted, but could be a somewhat simple modification to make to the script overall if there were an already made dataset to cross reference the address to and remove any entries with non US addresses.

B. Census Taker (Challenge 7)

The objective of this challenge was to methodically evaluate available 2020 Census data and report an analysis of the quality data, deliver a list of Congressional districts with issues to EDF, and provide confidence scores for each district.

i. Data Sources

- US Census Data (2020)
  - Census demographic data: Population and demographic data from the 2020 Census for each **US county**
  - Quality metric data: Quality data for each **US county** including information such as unresolved housing units which went to count imputation, percentage of unresolved housing units managed via other records or proxy interviewers, and counties which were population counts only (not including demographic data)
  - US county / local district relational database: Table matching **state house congressional districts** to US counties

ii. Analysis

There were three key quality metrics of concern being analyzed to develop the quality score model, where for each of these metrics a higher value indicates lower quality demographic information. All of these metrics relate to how unresolved addresses were later resolved, with the most ideal case being either direct interviews with the household or self-responses from the resident(s).

The first key quality metric of concern was **proxy interviews**, denoting the percentage of unresolved households which were resolved through an interview with a neighbor, landlord, or building manager. The second was **administrative records**, which shows the percentage of unresolved households which were resolved via information the Census Bureau already has access to. Finally, we evaluated the percentage of districts with **population count only**, which shows the percentage of households that did not include any demographic information.

While all of these key metrics were summarized down to the congressional district level and are themselves available for individual analysis, we wanted to create a single “quality score” to describe the overall quality of the district. The intent of this score is to indicate districts that could potentially have a data quality concern, for which the census data should be evaluated under this consideration and further evaluation of each individual quality metric should be conducted if the situation calls for it.

The quality score calculation was accomplished by calculating the aggregated sum of each quality metric together and determining the percentile of that score relative to the entire distribution for all counties. This percentile was subtracted from 100% such that districts with a high percentile net quality sum (indicating a lower quality score) were assigned a lower quality score, while the districts with a low percentile net quality sum

(indicating a higher confidence in data quality) were assigned a higher quality score. Scaling was not required as each metric was of a similar magnitude.

Percentile was considered across all counties in the United States rather than on a state-wide basis. While some states may show a higher average of unresolved follow ups or other indications of poor quality census data compared to the United States average,

C. List Locally (Challenge 8)

In this challenge, we were tasked with providing a more algorithmic approach to collecting and predicting local legislation data. Using Python, it was possible to automate an API call to OpenStates.org and retrieve around 6K rows of data to model with. This local legislation data allows us to predict vote outcomes on a bill and can be narrowed down by modifying the API call to even specify on specific bill sponsors.

ii. Data Sources

- OpenStates API
  - We utilized OpenStates to pull down: 1) name of current local representative by district, and 2) currently sponsored legislation per each representative. Bill data was also collected for several states.
- Legiscan
  - Legiscan was another data source we attempted to use to track open legislation. While it offered valuable data, it required licenses for use which limited our work.
- FollowTheMoney
  - Followthemoney.org tracks financial contributions to politicians or political candidates from third party individuals, organizations, or PACs. This was not used for the project, but the provided resources will soon have interoperability with this website. This will allow for a connection between companies backing legislators as sponsors and specific legislation subjects.

ii. Analysis

Every day the world is getting better about recording and labeling their data. Websites like OpenStates have been hard at work developing an API that can be used for the exact problem presented in Challenge 8: List Locally. While there are websites that offer a more robust dataset such as legiscan.com, these tend to be full of free-license restrictions that make it difficult for utilizing their data sources.

For this project, we opted for OpenStates due to their easy to use API and lack of free-license restrictions. In contrast to Legiscan, they offered 250 calls per API key with a limit of 10 calls per minute. This allowed us to build an auto-API caller program. We used this autocaller to extract nearly 6k rows of bill/legislation data from OpenStates for the states AK, WA, AZ. This data was then exported as a series of JSON files which automatically accumulate. Using the BILLS API call on OpenStates website, we were able to retrieve around 180 rows of legislation data per call.

Data Cleaning

As with all ML projects, models are only as good as their input data. For the returned OpenStates JSON data, there were a series of obstacles that heavily reduced the number of features that could be utilized for any modeling.

- Related_bills: This column intended to represent relation to other similar bills by subject. This field was always blank across all 50 states; a lost opportunity.
- Only around 40% of returned legislation had 'votes' field information, a returned column from OpenStates that had lists of voter names, their vote in support or opposition to the bill, and political party affiliation.
    - The votes were then sorted by most recent vote instance of each bill.
    - Voter names were then categorized by party, counted, and used to determine the yes/no counts for bill result outcome.
- Of that 40% of returned bill data that had voter information, only 10% of those had bill 'subject' information (with varying results between states).
- This remaining bill population provided the bill subjects alongside the names of the supporting legislators, providing the ability to correlate the two in a LR prediction model with bill pass/fail as the binary result column.
- This process was then automated in Python for easier EDF retrieval and provided as a Jupyter notebook with detailed instructions for its usage and future improvement suggestions.

Model Features

Feature Name	Status	Type	Processing
identifier	Input	Category	Dummy encoding
jurisdiction.classification	Rejected	Category	
session	Input	Category	Dummy encoding
classification	Input	Category	Dummy encoding
from_organization.classification	Input	Category	Dummy encoding
bill_result	Target	Numeric	
bill_subject_3	Input	Category	Dummy encoding
bill_subject_4	Input	Category	Dummy encoding
bill_subject_5	Input	Category	Dummy encoding
state_party_affiliation	Input	Category	Dummy encoding
bill_subject_1	Input	Category	Dummy encoding
from_organization.name	Input	Category	Dummy encoding
bill_subject_2	Input	Category	Dummy encoding
jurisdiction.name	Input	Category	Dummy encoding
bill_party_affiliation	Input	Category	Dummy encoding

Table 2. Model features and variable types

With the data retrieved, columns created, and rows cleaned, the next step was to try and predict something meaningful from our data. Logically, the most important aspect of a bill is if it passed or failed in its most recent vote session. The Python code provided in the engine already accounts for this, as only the latest instance of a vote for each returned bill is used to determine the most recent bill outcome. This was then combined into one dataset containing the bills of the three states. The models evaluated were Random Forest, Logistic Regression, and LightGBM.

There were a few challenges with this project. One challenge presented with this project was obtaining enough data to make meaningful predictions off of. Due to differences between legislator websites, a main data source like OpenStates was vital to the success of a ML model. Additionally, some states in the OpenStates database are lacking in 'vote' and 'subject' data. Given that the Challenge 8 model is based on those two parameters, it is difficult to find states that provide meaningful data that can be used for ML predictions. The 6K rows that we returned and used for modeling took many hours of work to retrieve due to this issue. We have now automated the process for EDF, and the function will flag any states that are missing this info as it retrieves data automatically.

In the near future (according to followthemoney.org) there is a join being made between OpenStates and followthemoney.org. Once this is complete, it will be easy to retrieve specific sponsors and companies backing legislators to the source.

III. Results

A. Geospatial Intelligence for Environmental Influence (Challenge 5)

Our results yielded some valuable information on business location, sectors that are present as well as their potential influence on local politicians. While our results from the NLP model did not give us confidence in productionalizing this aspect of the project, we do think we learned some valuable information about trying to use this type of model to determine relationships between two text sources.

For the data generation, we were able to build multiple scripts which create data sources for EDF which were identified to be helpful in their goal of creating a robust influence model. The first generates a shapefile for US state local legislative districts across all 50 states and ties districts to their local representative and party affiliation. The script takes about 30 minutes to run on a local machine, to pull information for all (~4,000) representatives. It does complete the task of combining this data together with data on businesses to create a comprehensive dataset which EDF can use to understand district party affiliation and business sectors which drive employment or have a large footprint in the area. This process, as well as the other script referenced below, could be optimized by employing different technology to run the script. Something like Spark may be more suitable for this type of large scale data generation or another resource which can parallelize some of the work. Taking the time to refine the script to only pull data for data changes would be another strategy to make this process more efficient over time.

The NLP model which we attempted to employ to relate bills to business sector did not ultimately pan out as we had hoped. While the model can identify if a bill could have impact on some business/industry in general (correctly classified bill as being related to any business or industry 78% of the time on our test set), it does not do a good job of identifying the specific business sector that the policy may be related to. Frequently, the model would identify over half of the sectors to be "similar" (score >0.6), with a less than 1% variation between the similarity scores. Due to this, we don't think this is a viable option for use in production to associate granular business sectors, but the model could be used to de-clutter data from bills which are more administrative in nature (e.g. acknowledgment of holiday, mourning after someone has passed, etc.).

Finally, our business address script also was able to provide headquarter addresses for over 12,000 different businesses which have filed with the SEC. This script takes a long time to run (around 2 or more hours), due to the amount of web scraping requests that need to be made. We mitigated this by generating a light-weight version of the script which pulls headquarter addresses for public companies in the S&P 500. This should provide EDF with a more scalable way to pull data. Again, this could be further optimized down the line to only pull data for companies which are not already listed rather than completing the entire data pull again.

B. Census Taker (Challenge 7)

First, the county to district relational database was used to match the county-level census data with congressional districts. One notable limitation observed here is that some districts lie in more than one county, and some counties contain more than one district. While we cannot easily correct for this, we can provide all available data for each district to gain a better understanding of potential influences and overall data confidence.

An example of this phenomenon is shown in Figure 1 below: Alabama state house district 2 lies in both Lauderdale and Limestone counties, but total constituents under this district do not necessarily equal the population sum for both counties.

STNAME	CTYNAME	State District ID	County ID	TOT_POP	TOT_MALE	TOT_FEMALE
Alabama	Lauderdale County	State House District 2	1077	94043.0	45416.0	48627.0
Alabama	Limestone County	State House District 2	1083	107517.0	54050.0	53467.0

Figure 1. Alabama state house district 2 match to county-level data

Next, the underlying distributions and descriptive statistics of each key quality metric under evaluation were assessed to check normality and skew of the data. Results are shown below.

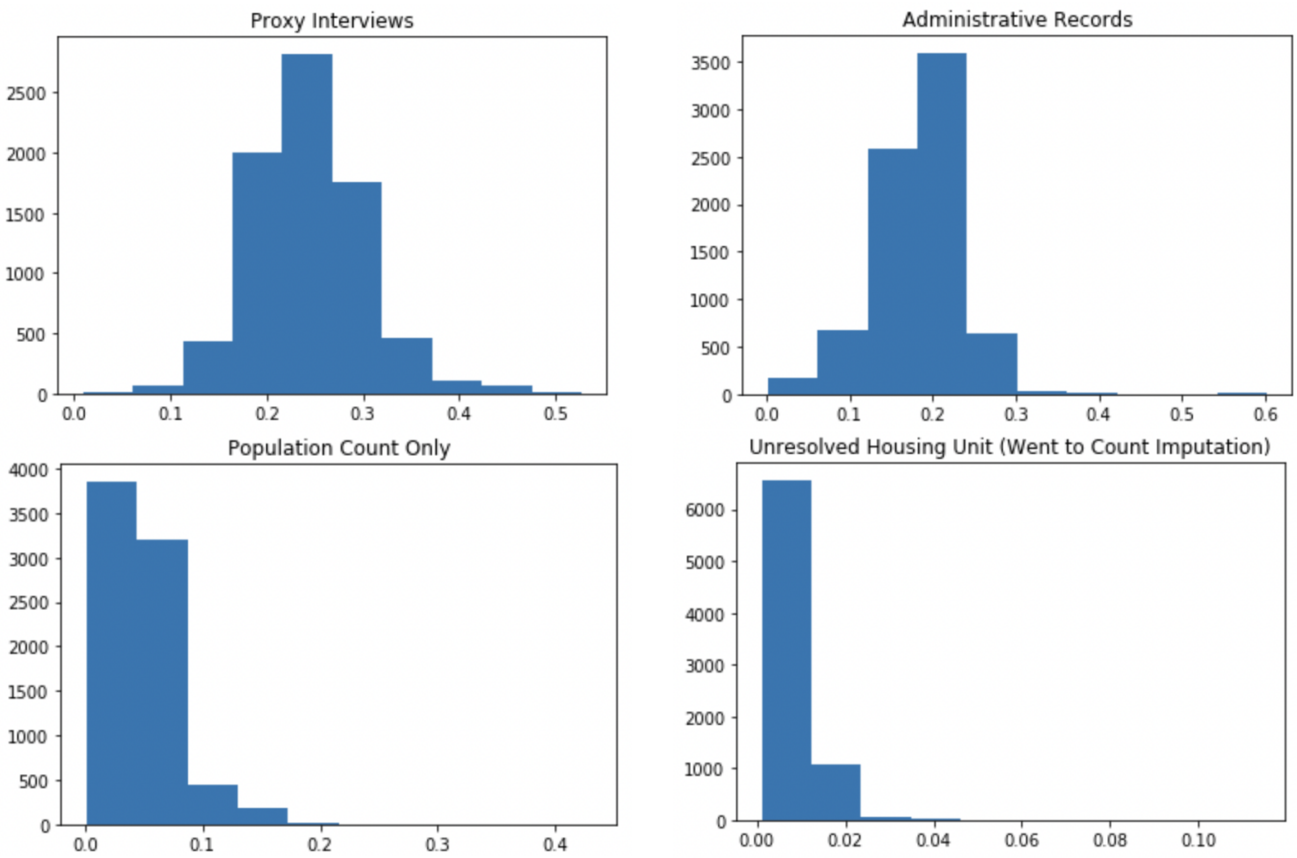

Figure 2. Distributions of the key quality metrics for census data

From the exploratory data analysis, it is clear there are few counties with remaining unresolved housing units. This metric was accordingly excluded from further analysis, and remaining work focused on the proxy interview, administrative record, and population count only methods of data resolution.

	Proxy	Administrative Records	Population Count
count	2684.000000	2684.000000	2684.000000
mean	0.223188	0.172050	0.042329
std	0.058878	0.059231	0.030868
min	0.009000	0.001000	0.001000
25%	0.188000	0.138750	0.023000
50%	0.217000	0.177000	0.037000
75%	0.252000	0.208000	0.054000
max	0.527000	0.602000	0.431000

Figure 3. Descriptive statistics for key quality metrics for census data

Next, these three quality metrics under consideration were summed together to create an aggregate indicator for overall data quality. We again can analyze the distribution of this metric, and find it to approximate a Gaussian fit.

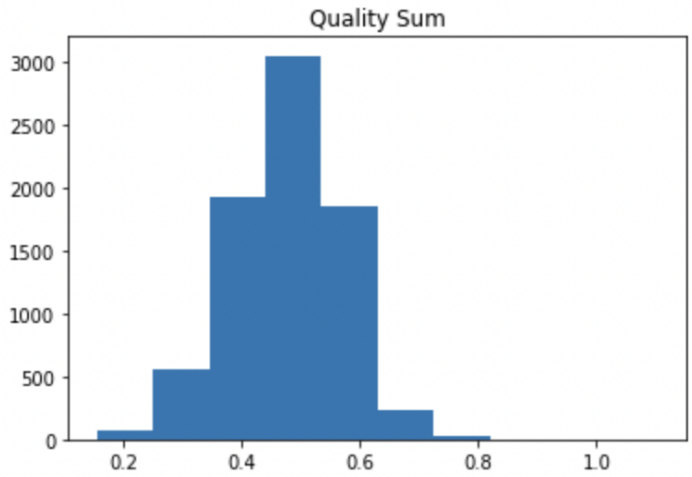

Figure 4. Distribution of the aggregated quality sum metric

The percentile was then analyzed and a reversed percentile score was assigned to each locality. Districts with a high quality sum metric (representing a high proportion of households resolved via imperfect methods) were assigned a low percentile score, and vice versa. This percentile score is the final deliverable to consider as the quality score for each area, where low percentile districts represent areas where the end user should consider a potentially low quality of census data and consider further evaluation of how they use the census data for that district.

From here, we can evaluate a list of districts with the lowest aggregate quality score based on the bottom percentile of districts and evaluate their specific issues. This list can be pulled considering whichever percentile criteria the end user prefers, for example the bottom 5% of all districts; here, we will show the five districts with the lowest percentile quality scores.

State	State District ID	County	% Unresolved Housing Unit	% Proxy	%Administrative Records	% Population Count Only	Quality Metric	Percentile
Idaho	State House District 31	Clark County	0.008	0.386	0.447	0.272	1.105	0.000
North Dakota	State House District 29	Steele County	0.011	0.306	0.363	0.43	1.099	0.013
Montana	State House District 30	Golden Valley County	0.075	0.241	0.222	0.431	0.894	0.026
North Dakota	State House District 28	Logan County	0.044	0.527	0.182	0.139	0.848	0.039
North Dakota	State House District 14	Pierce County	0.015	0.088	0.495	0.288	0.871	0.052

Figure 5. Local districts with the worst quality metric scores

In evaluating these districts, we find a few commonalities:

- All districts except for ND district 14 have above average proxy responses, with ND district 28 representing the max value in the dataset.
- All districts have above average administrative record imputations.
- All districts have an above average population count.

It is clear that the combination of factors sum to yield significantly higher than average quality metric scores.

Since our dataset combines the quality metrics with the actual census data at the county level, we can further study if there are any demographic or population commonalities between these bottom five districts.

State	State District ID	County	Total Population
Idaho	State House District 31	Clark County	792
North Dakota	State House District 29	Steele County	1810
Montana	State House District 30	Golden Valley County	831
North Dakota	State House District 28	Logan County	1883
North Dakota	State House District 14	Pierce County	3953

Figure 6. County-level population data for districts with the worst quality metric scores

Immediately, we observe that every district with the worst quality metric scores has an incredibly low population. This suggests that small counties may have a harder issue with direct household data collection during the US census interview period.

This quality metric and the ability to immediately understand potential issues provides valuable insights when using census data for other studies. End users can understand at a glance the quality percentile of the district at hand compared to the total US population, and use this to further deep dive into which individual quality metrics or population details are worth noting when using the data.

C. List Locally (Challenge 8)

Between the three models, it turned out that random forests fit the data best. Despite this fit, we would refer to the later two models (LR, and LightGBM). They are less susceptible to overfitting in this low data scenario.

Despite RF came out on top with an accuracy of 90.4%, it is likely that there is some possible overfitting happening with the model due to a lack of data. Because of this, we suggest that logistic regression (LR) may provide a better predictor for the data as it cannot account for every scenario as RF can with its many decision trees.

Logistic Regression

Despite its name, Logistic Regression is a classification algorithm, using a linear model (i.e., it computes the target feature as a linear combination of input features). Logistic Regression minimizes a specific cost function (called logit or sigmoid function), which makes it appropriate for classification. A simple Logistic regression algorithm is prone to overfitting and sensitive to errors in the input dataset. To address these issues, it is possible to use a

penalty (or regularization term) to the weights. This implementation can use either 'L1' or 'L2' regularization terms

In building the LR model, 15 features were placed against the bill_result binary column. 1 being a bill that was passed and 0 being a fail.

The evaluation metric used to tune the hyperparameters was ROC AUC computed on the validation dataset. After the best hyperparameter combination was found, the same metric was also computed on the test dataset. The final value was 0.874.

Model Parameters

Search parameters	
Strategy	Random search
Random state (hyperparameter search)	1337
Max number of iterations	24
Max search time	0 (no limit)
Parallelism	4

Table 2. Model search strategy and parameters

Cross-validation	
Cross-validation strategy	K-fold
Number of folds	5
Random state (cross-validation split)	1337
Stratified	Yes

Table 3. Cross-validation parameters

This is a solvable problem with the tools we have provided. By feeding the model more data, it will become more and more accurate in determining bill outcome. While the overfitting is just a suspicion at this time due to the higher than expected accuracy, it is possible that state party affiliations and bill subjects are polarized within the USA. If this is truly the case, then the supplied LR model is as highly effective at predicting bill outcomes as it claims.

As for lightGBM, it is also likely this model will perform well with the supplied data due to the features it is currently weighing on to make predictions on bill outcome. (see fig. 7)

The accuracy results were:

- RF: 90.4%
- LightGBM: 89.0%
- Logistic Regression: 87.4%

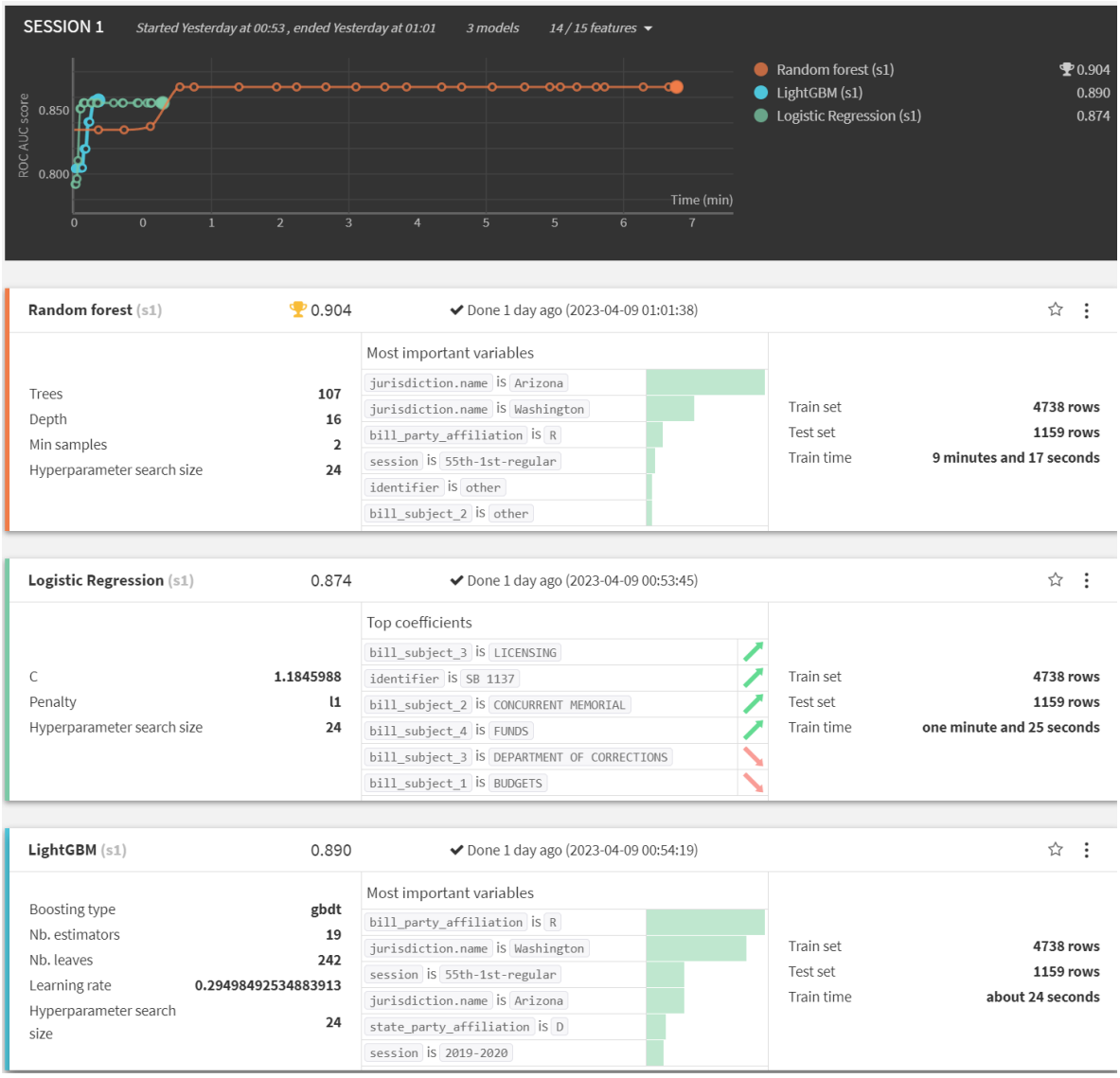

Figure 7. Model specifics, features used, and score trends

The RF model was built using industry standard tool, Dataiku. This model read through nearly 6k rows and used an 80/20 testing split to provide an accuracy outcome of 90%.

While a great deal of more data is needed prior to using this model for real world predictions, it is clear that the features of bill subject and party affiliations are directly related to bill pass/fail outcomes. The model goal is to predict bill_result given a total of 15 features. Using a dataset of 4738 rows, the process led to the selection of the Logistic Regression algorithm.

	Predicted 1	Predicted 0	Total
Actually 1	930	25	955
Actually 0	141	63	204
Total	1071	88	1159

Figure 8. LR Confusion matrix

What this means for prediction of environmental influences: Having a model that can predict vote outcome means that one can predict how bills related to environmental subjects (as can be filtered within the bill subjects column) will turn out in court. This can even be tailored to specific legislators and even specific state or house legislations by altering the provided API call. Doing this would allow for machine learning to determine how a specific legislator/legislation will act based on bill subject.

Essentially, this LR algorithm provides EDF with a tool for predicting the future of a bill based on its subject and the legislation it will go through, which is a powerful tool for expecting what is to come next in terms of environmental influence.

IV. Biases and Ethics

In building these models, it is pertinent to consider the potential biases or ethical concerns of the data sources and use of the data.

Certain data sources have an inherent source of bias with them, particularly considering we were using third-party datasets and information sources. We found several issues where a dataset has a particular political lean as we evaluated information related to politicians or bills. With this, their assumptions and methodologies could skew the provided information in line with their motives. In addition to this, data collection from primarily republican or primarily democratic states presents a situation where the model will obtain that bias and reflect it in its predictions in our modeling, particularly in bill predictions for Challenge 8.

We must also consider how the results of our work may produce cognitive or confirmation biases when in use by human operators, and several times caught ourselves needing to check our own biases while searching for information to build our models. It was erroneous to assume more conservative politicians or groups in right-leaning districts may have a negative approach to environmental legislation, while liberal politicians or groups in left-leaning districts would act positively.

We were careful in ensuring our models were created with a full set of representative data, and we encourage EDF end users and practitioners of our models to be considerate of their own biases. Users who interact with our models should be informed that the data was collected

factually and ethically, but must be used responsibly. There should be disclaimers about the sources of the data and responsible use of our models for any users as a systematic reminder of unconscious bias and drawing conclusions.

Another critical consideration in our analysis is the benefits and risks of personal privacy and civil liberties upon implementing our methodology. We believe our work can provide considerable benefit to the EDF team and customers in evaluating the influence of corporations and individuals on politicians. However, we also must be respectful and ethical in using data regarding potentially influential individuals and political figures under review.

V. Lessons Learned

- State-level results don't indicate national level results. Looking into data across multiple different states, it was quickly easy to see that states behave differently, and have different levels of information available. While it was good to initially test out ideas at the state level, our solutions could have been tested at the national level more quickly to expose those gaps earlier on.
- Not all APIs are easy to work with. OpenStates provides a plethora of data, but lacks complete subject and vote information as one pulls data from different states. While this issue is state side, and not the fault of OpenStates, it does create a difficult scenario for pulling enough data to make valuable insights on.
- Legiscan is a valuable resource that should be investigated by EDF. It provides a near 1:1 solution for tracking legislation data that includes vote sponsors and followthemoney.org connections that would promote the ability to determine bill outcome based on what companies are backing the bill. This was the original datasource for Challenge 8 before license issues prevented further research.

VI. Conclusions

The provided data, code, and documentation provides EDF with a series of powerful tools that can help in the battle of obtaining large amounts of legislation data for higher level analytics.

In Challenge 5, creating a comprehensive dataset at the state level is tedious. State-by-state structure varies for many of the API pulls, and bulk API data pulls (e.g. single pull for all 50 states, or by region) are non-existent. Despite these challenges, we were able to create two comprehensive datasets for EDF to use to further their efforts to model influence towards local politicians. We hope that we also gave some building blocks and lessons learned in our attempt to build an NLP model. This sort of model would be better suited for larger amounts of text, rather than sparse text that we used. If we were to re-do our approach, we would likely try out a clustering model which was trained on pre-classified data.

For Challenge 7, we have aggregated 2020 US census data at the county-level with state congressional district level detail and created a simple model to generate an indication of data quality by considering several key quality metrics. Though the county and district matching was imperfect due to the imperfect matching of districts and counties (for example, a district lying in multiple counties or a county containing multiple districts), the output of this study provides a

hopefully valuable resource for end users to consider the data quality for both county and state congressional level census data after known issues with 2020 US Census data collection.

Finally, in Challenge 8, legislation data was pulled via API, wrangled, and fed into a Logistic Regression model that produced bill prediction outcomes. While overfitting is a potential risk at below 10k rows of data, the framework for more data collection and training more predictive models is provided to EDF in detail for their legislation data collection journey. Recommendations to improve this model would simply be to run the API caller program using a pro license from OpenStates and to provide it with each state abbreviation. This would collect enough data to make even more meaningful results. Additionally, we would advise to use swing states, as their data has a valuable contrast between Republican and Democratic bill outcomes that would help train the model on less biased data.

With the challenges completed, we have provided a foundation of data with algorithmic approaches to the very unstructured world of legislation data. Having this foundation of resources provides EDF with a solid ground to continue the collection and ML models that were provided in our final report submission in their crusade to make the world of legislative tracking a more predictable place.

References

Basseches, J. A., Bromley-Trujillo, R., Boykoff, M. T., Culhane, T., Hall, G., Healy, N., Hess, D. J., Hsu, D., Krause, R. M., Prechel, H., Roberts, J. T., & Stephens, J. C. (2022, February 16). *Climate policy conflict in the U.S. states: A critical review and way forward - climatic change*. SpringerLink. Retrieved April 11, 2023, from https://link.springer.com/article/10.1007/s10584-022-03319-w

Bureau, U. S. C. (n.d.). *Explore Census Data*. Explore Census Data. Retrieved April 11, 2023, from https://data.census.gov/

Everyday AI, Extraordinary People. Dataiku. (2023, February 14). Retrieved April 11, 2023, from https://www.dataiku.com/

Legiscan. LegiScan. (n.d.). Retrieved April 11, 2023, from https://legiscan.com/

Open States: Discover Politics in your State and Congress. Open States: discover politics in your state and Congress. (n.d.). Retrieved April 11, 2023, from https://www.OpenStates.org/

Tools. Home - FollowTheMoney.org. (n.d.). Retrieved April 11, 2023, from https://www.followthemoney.org/

APPENDIX A: NAICS Codes

Code	**Industry Title**
11	Agriculture, Forestry, Fishing and Hunting
21	Mining
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)