Final Written Deliverables Report
Project Title:
Predicting State-Level Reductions in Fossil-Fuel Energy Share (US, 2000-2019)

Authors:

Melissa Best, Kevin Puduseril, Maxwell Tuttle

**Each author contributed equally to the design, coding & development, analysis, and writing of this project*

Date:

August 27, 2025

Executive Summary

This project examines why some U.S. states have been more successful than others in reducing reliance on fossil fuels between 2000 and 2019. Using integrated datasets on state energy generation, socioeconomic conditions, political leaning, and renewable policy adoption, we tested the hypothesis that liberal states with robust grid capacity, higher incomes, and ambitious renewable policies transition faster away from fossil fuels.

Our models suggest that while income and Renewable Portfolio Standards (RPS) matter, political orientation and baseline energy dependence are stronger predictors of success. Random Forest models trained only on baseline-year features achieved modest but genuine predictive power (ROC-AUC ≈ 0.63), highlighting that structural conditions and political will, rather than raw policy counts alone, are key drivers of long-term fossil fuel reduction.

Background & Question

Fossil fuels remain the dominant energy source in the U.S., but states vary widely in their pace of transition toward renewables. Understanding these variations is critical for policymakers aiming to design effective interventions.

“Policies are one of the most powerful tools for either promoting or stalling renewable energy adoption. The laws, incentives, and regulations put in place by governments have a direct impact on how accessible and competitive renewable energy becomes.” (Sakellaris, 2024)

This quote enforces the idea that policies are important in shaping the industry of renewable energy generation. Proper policies being set in place are essential to the growth of renewable energy and the reduction in our crutch on fossil fuels.

Research Question:

What combination of state characteristics (political leaning, grid capacity, median income, existing energy mix, etc.) and renewable energy policy designs best predict a successful reduction in fossil fuel shares per state in the U.S.?

Hypothesis:

States with liberal political climates, robust grid capacity, higher median incomes, and ambitious renewable energy policies will demonstrate greater success in reducing fossil fuel shares.

Prediction:

A machine learning model will show these features as significantly and positively correlated with fossil fuel reduction over the past decade.

Data

Acquisition

State-level data was acquired from four different institutions. Energy generation and consumption data came from the U.S. Energy Information Administration (U.S. EIA). Renewable energy policy information was obtained from the Database of State Incentives for Renewable Energy (DSIRE) maintained by the North Carolina Clean Energy Technology Center. Data on median household income and geographic area came from the U.S. Census Bureau. Finally, presidential and senate election data was obtained from the Massachusetts Institute of Technology (MIT) Election Data Lab.

Data Source	Data Obtained	Links
U.S. EIA (Energy Information Administration)	Energy generation and consumption	Energy Generation Data Energy Consumption Data
DSIRE (Database of State Incentives for Renewable Energy)	Policy level information	State Policy Map State Energy Programs
U.S. Census Bureau	Median Income Geographical Data from 2010 used for all years	Income Data Geographical Data
MIT (Massachusetts Institute of Technology)	Election data for Presidential and Senate (Missing variables in presidential and senate data were repeated from previous years (ex: 2005 data for presidential data contained data from the 2004 election))	Presidential Elections Senate Elections

Energy Generation Data

Our Energy Generation Data came from “U.S. EIA Net Generation by State by Type of Producer by Energy Source (1990–2023)” (U.S. EIA 2023), containing the annual electricity generation by source (coal, natural gas, wind, solar, etc.) in an excel file Annual_generation_state.xls with the columns YEAR, STATE, TYPE OF PRODUCER, ENERGY SOURCE and GENERATION (Megawatthours). This was a good choice because it gave us information on the source of energy being generated and came from a reputable source.

We filtered the data for 2000-2019 and Type of Producer “Total Electric Power Industry” We wrote a function “categorize_energy” which labelled the following Energy Sources as Renewable Sources:

'Hydroelectric Conventional',
'Geothermal',
'Solar Thermal and Photovoltaic',
'Wind',
'Wood and Wood Derived Fuels',
'Other Biomass',
'Municipal Solid Waste',
'Other Waste',
'Other Renewable'

And finally, we summed the megawatt hours of generation by total, renewable and non-renewable to make three features.

Energy Consumption Data

Our Energy Consumption Data came from the U.S. EIA State Energy Data System (SEDS) (U.S. EIA, 2025) the data was in two comma separated files, use_all_btu.csv containing the energy consumption data for each year by state and “MSN Code” and Codes_and_Descriptions.csv containing the definitions for the MSN codes. We added the MSN Codes to the data and group the codes by prefixes as follows:

fossil_prefixes	"NG", "CL", "CO", "PA", "PC", "DF", "JF", "FF"
renewable_prefixes	"HY", "WD", "WS", "SO", "GE", "WY"

We calculated summed the total fossil and renewable share and total consumption per state per year, filtered for 2000-2019 and removed the US total and DC.

This was good data to use because it gave us data on the consumption of energy, along with whether it was a renewable or non-renewable energy source. It also came from a reputable source.

State Policy Data

The data on State Policies came from the DSIRE website State Energy Programs (NC Clean Energy Technology Center). We filtered the list for all entries and copied them into a csv file State Energy Programs.csv. We converted the policy created date to Created_Year, filtered for 2000-2019 and removed policies for DC and Territories. We then pivoted the policy/incentive types into a column counting the number of that type of policy for each state and year.

This data was useful because it was fairly straightforward to get a complete list of policies enacted by each state from a reliable source.

Median Income

Median income data was obtained from the U.S. Census Bureau (U.S. Census Bureau) in a comma separated file Annual.csv. The census bureau is the most reliable source for this type of data. We converted Observation_Date to Year, filtering for 2000-2019 and extracted the State abbreviation from the codes for example **“MEHOINUSAKA672N_20210423” to AK**.

State Area

The total area of the states in Square Miles was obtained from the U.S. Census Bureau (U.S. Census Bureau) in a comma separated file, State_area_data.csv. We used the Area in Total Square Miles. We filtered out total, United States, DC, NaN and Territories and converted State Names to Codes. The only data available was from 2010, but we assumed that the area of states is relatively unchanged and used this data for all years from 2000-2019.

Election Data

Election data was obtained from the reliable source MIT Election Data and Science Lab in a file 1976-2020-president.csv for presidential election data (MIT Election Data and Science Lab) and senate election data from 1976-2020-senate.csv (MIT Election Data and Science Lab). We used Year, State, Party, Candidate Vote and Total Votes. We filtered for elections between 2000-2019 and filtered out where the party was other or libertarian to focus on the bigger parties and avoid incorrectly attributing third party votes. We summed the total votes, democrat votes and republican votes and removed DC. For years without election data, we used the previous election’s results.

Cleaning

We merged on state identifiers and years (2000-2019) and filtered to 50 states. We did not want to include any territories or D.C. as not all data sets contained this information. We imputed missing median income and political scores using neighboring/national trends. We encoded categorical variables (e.g., RPS presence, political bucket). We Removed all-zero policy columns; replaced missing senate data with zeros. Finally, we ensured year alignment

across datasets. Also ensured that there was only one observation per state per year to stay consistent among the data sources.

Exploration

Key findings from exploratory analysis:

- Average fossil share fell from 67.7% (2010) to 61.7% (2022), with wide variation.
- “Successful” states reduced fossil share by ~ 12.6 percentage points; others saw negligible change.
- Income differences were minor; political lean and RPS presence showed some associations but were not decisive on their own.
- Top renewable growth states include California, Texas, and Iowa.

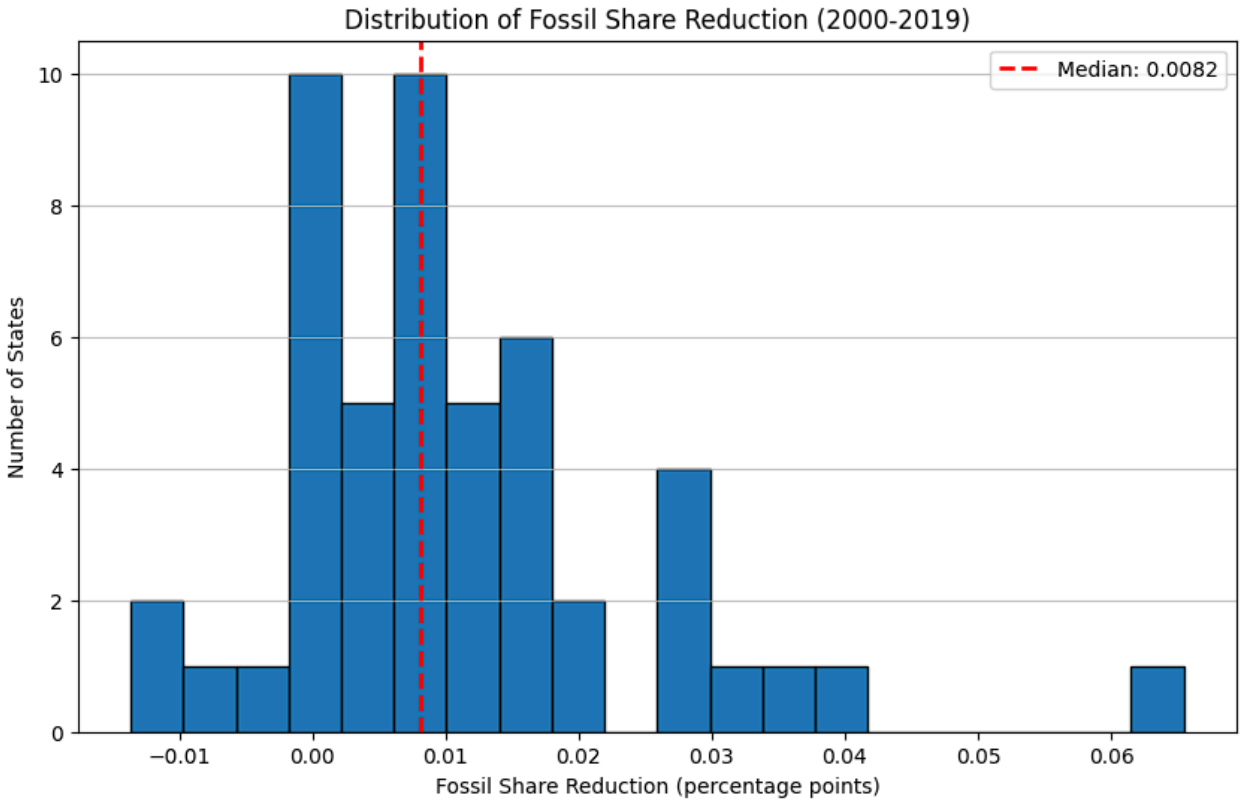

Figure 1:

The histogram illustrates how U.S. states have changed their reliance on fossil fuels between 2000 and 2019. Overall, most states achieved a reduction, with a median decrease of approximately 0.0082 percentage points, indicating a general shift toward cleaner energy sources. However, the distribution reveals significant variation: while some states reduced their fossil fuel share, others saw little change or even increases, highlighting uneven progress

across the country. This variation likely reflects differences in state energy policies, infrastructure, and resource availability, emphasizing that while the national trend is moving away from fossil fuels, progress is not uniform across all states.

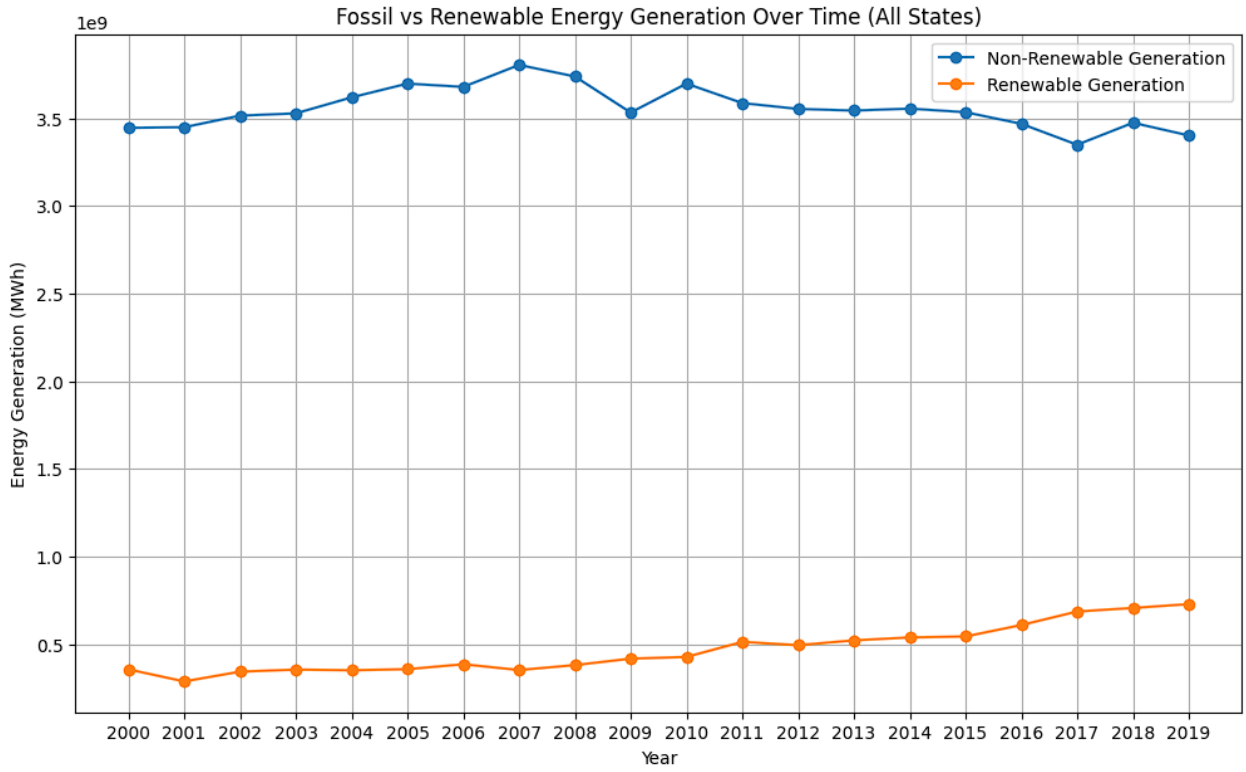

Figure 2:

The chart illustrates the historical trend of U.S. energy generation from fossil fuels and renewable sources between 2000 and 2019. Fossil fuels have long dominated energy production, peaking around 2007, but have since plateaued or slightly declined. In contrast, renewable energy started from a low baseline but has shown steady and consistent growth.

This divergence after 2010 signals a gradual energy transition, driven by technological advancements, policy incentives, and growing environmental awareness. While fossil fuels still lead in absolute output, the rise of renewables highlights a shifting energy landscape focused on sustainability.

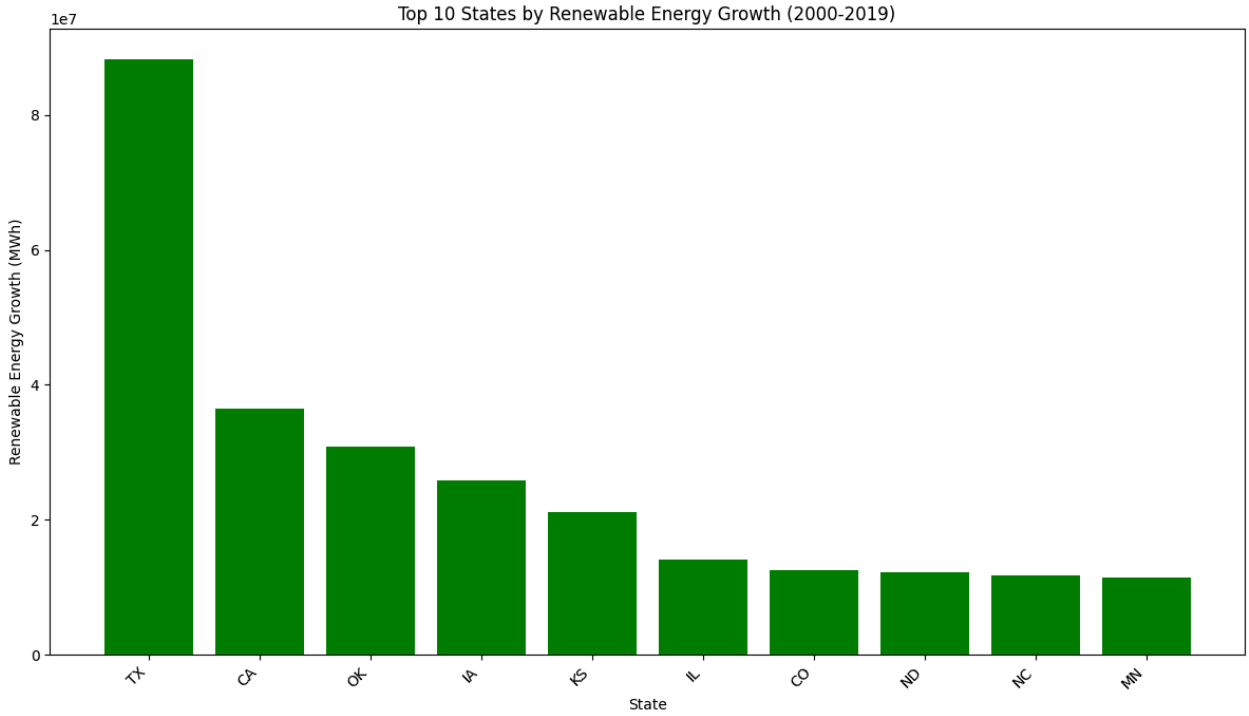

Figure 3:

The bar chart highlights the top 10 U.S. states with the highest growth in renewable energy consumption from 2000 to 2019. California leads the nation, reflecting its aggressive clean energy policies and investments in solar, wind, and geothermal power. Texas follows closely, driven largely by its massive expansion in wind energy. Iowa ranks third, benefitting from its substantial wind energy infrastructure. Other states like Nebraska, Illinois, and Indiana show moderate but meaningful increases, while Minnesota, Florida, Oklahoma, and Kansas round out the top 10. The chart underscores how both strong policy support and natural resource availability have shaped renewable energy growth across the country.

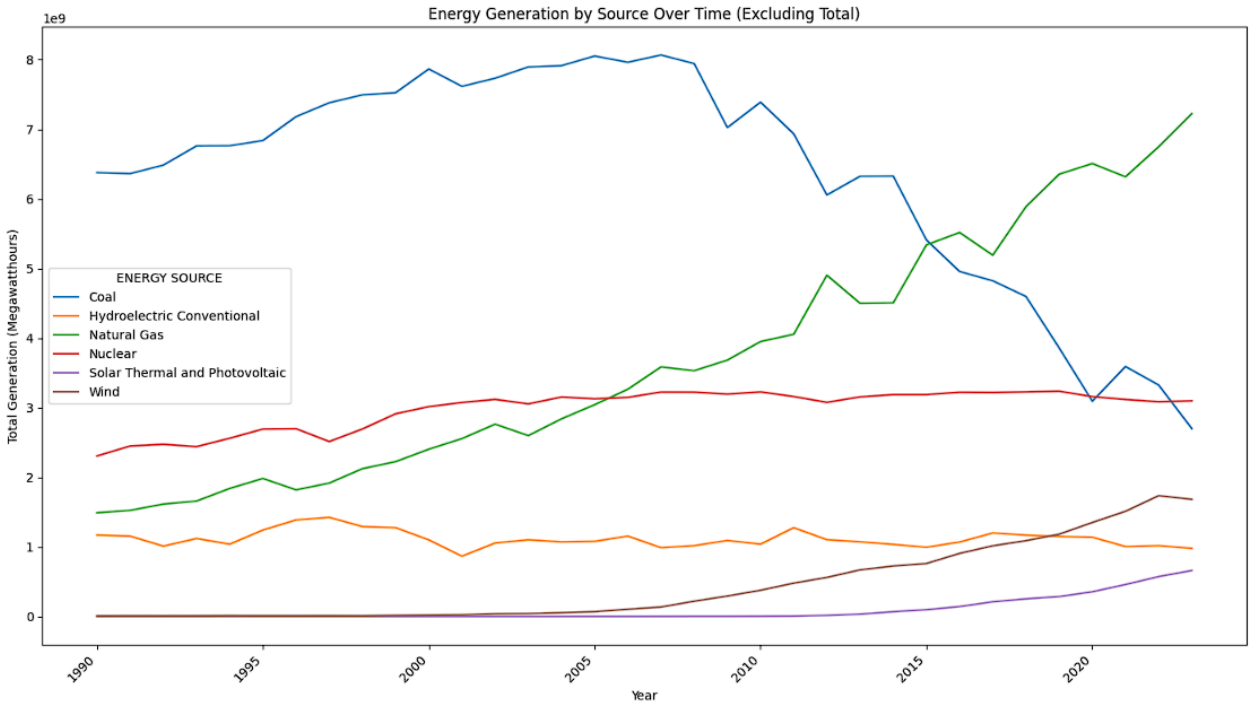

Figure 4:

These variables were not kept for the final dataset, but we believe they provide valuable insight into the data itself. The above chart shows the change in energy generation by different sources over the course of 33 years. Some key takeaways are:

- Coal powered energy generation has drastically decreased, starting roughly around the year 2005
- Natural Gas powered energy generation has continuously increased throughout the observation of the data
- Solar Thermal and Photovoltaic powered energy generation began increasing slowly, starting roughly in 2012
- Wind powered energy generation began steadily increasing starting roughly around the year 2005
- Nuclear powered and hydroelectric energy generation have remained relatively constant throughout the observation of this data

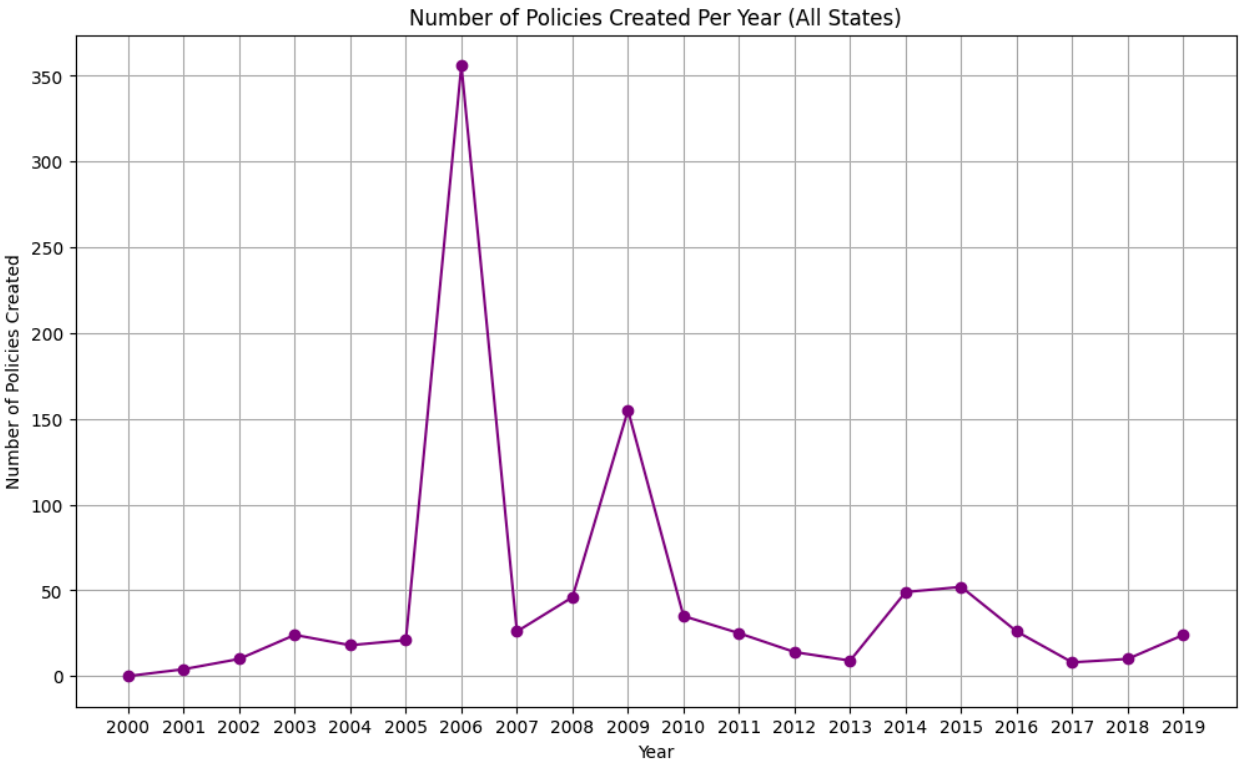

Figure 5:

The above Histogram shows the number of clean energy policies created per year according to DSIRE data. It shows large numbers of policies created in 2006, 2009 and 2014-2015.

Models

Preprocessing & Feature Engineering

Our final outcome variable was created from other data points in our dataset. The outcome variable was Δ Fossil Share (2000-2019) consisting of a binary label for whether the state is above or below the national median in fossil fuel reduction. Some other features we created consisted of combining energy generation sources into renewable and non-renewable energy sources, renewable and non-renewable energy consumption, combining policies to create a feature for the number of policies created for a state in a given year, and converting presidential and senate election data into percentages.

The unsupervised methods that were used were PCA and K-means clustering. PCA was used and unveiled strong separation of successful vs. unsuccessful states along with “overall fossil dependency”. This allowed us to see which states did a better job at reducing fossil fuel usage over the last twenty years. K-means clustering was used in multiple ways. It was used on the overall dataset and also on sub-sections including median income and political score. For the overall dataset, it grouped states into high, moderate, and low fossil fuel dependency

clusters. For the median income and the political score runs, we were able to identify which states are closely related to one another.

Algorithms Tested

The algorithms that we chose to test were Logistic Regression, Random Forest, Gradient Boosting, LightGBM, and XGBoost. These models were chosen because they are fairly common models and can be easily interpretable. We first ran all of these tests using default parameters to be able to get an initial idea of the best performing models.

After the initial results, we were then able to narrow it down to the best three, which happened to be Logistic Regression, Random Forest, and Gradient Boosting. We then were able to make some enhancements to the data, such as removing data leakage by removing some features that shared collinearity with other features, and reran the models. We were then able to run the models with various parameters using 5-fold stratified cross validation (optimizing ROC-AUC) and compare the results.

After comparing the final three models with the best parameters for them, it was clear that the Random Forest model outperformed the Logistic Regression and Gradient Boosting models. So, the Random Forest model was chosen for our final. Overfitting in the earlier models was overcome by removing features that were shown to have collinearity and data leakage, due to features like fossil_share_2019 that would not be available to the model at the time of use. After removing features like this, we were able to get the model to not overfit.

The table below shows the results of the final three models and what led to our choice of Random Forest:

Model	Parameters	Accuracy	ROC-AUC	F1	Takeaway
Random Forest	model__n_estimators: 150, model__min_samples_leaf: 1, model__max_features: 0.5, model__max_depth: 8, model__class_weight: null	0.66	0.63	0.69	Best current discriminator; captures real but modest signal.
Logistic Regression	model__C: 1, model__penalty: l2	0.64	0.60	0.63	
Gradient Boosting	model__subsample: 0.8, model__n_estimators: 150, model__max_depth: 2, model__learning_rate: 0.03	0.58	0.60	0.63	

Table 1: Results of running the three best models after accounting for overfitting.

Final Model & Tuning

The Random Forest model delivers the highest cross-validated ROC-AUC (0.63) after removing leakage—evidence it's picking up genuine structure without relying on post-outcome variables. We favor ROC-AUC because classes are balanced and we care about ranking quality across thresholds.

The top signals shift away from "future" outcomes to political lean (Dem/Rep presidential vote share), median income, and baseline energy activity; fossil_share_2000 retains moderate importance, while simple early policy counts are relatively weak predictors on their own.

States with more Democratic vote share and higher median income in 2000—and with certain baseline energy footprints—were more likely to end up above the national median in reducing fossil-fuel share by 2019. Policy counts at baseline are less predictive alone; policy *design/trajectory* may matter more than raw counts.

The Random Forest's AUC of 0.63 indicates useful but not decisive discrimination; there's room to improve with better features (policy stringency/targets over time, transmission investments, federal incentives timing).

feature	importance
dem_pres_percent	0.25
renewable_energy_consumption_btu	0.14
rep_pres_percent	0.10
median_income	0.07
renewable_generation_mwh	0.06
rep_sen_percent	0.06
dem_sen_percent	0.06
fossil_share_2000	0.06
fossil_energy_consumption_btu	0.05
total_energy_consumption_btu	0.05
total_generation_mwh	0.04
non_renewable_generation_mwh	0.03
total_policies	0.02

Table 2: Feature Importance of the variables included in the final Random Forest model.

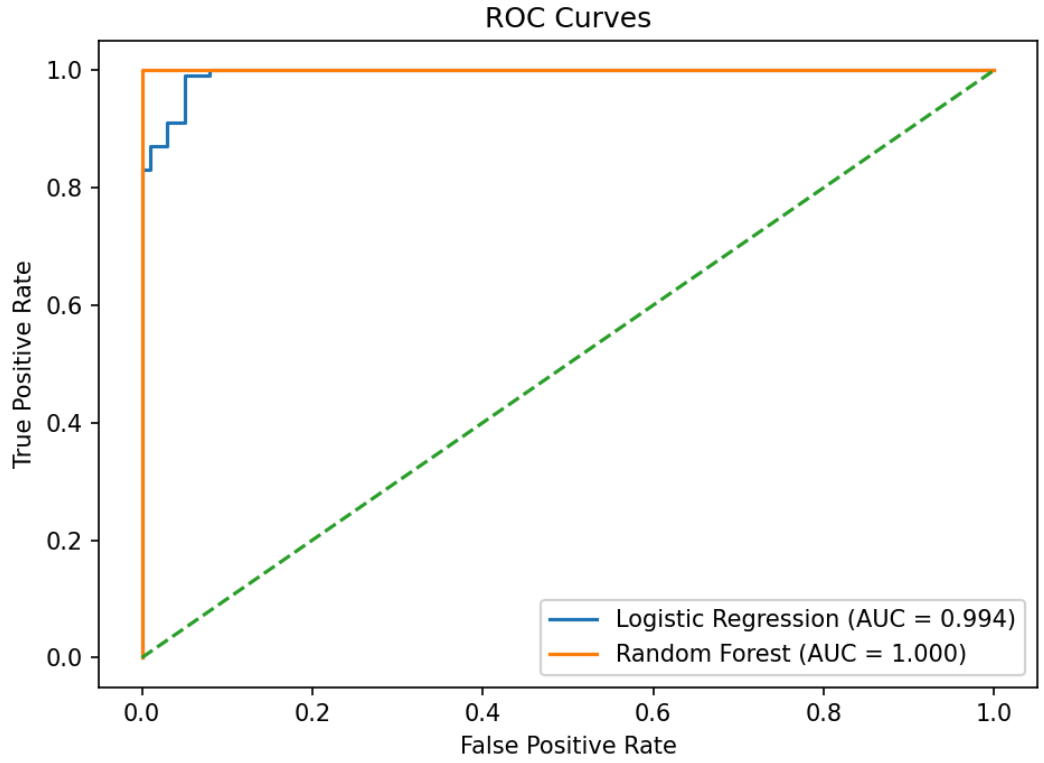

Figure 6: ROC curves before removal of data leakage

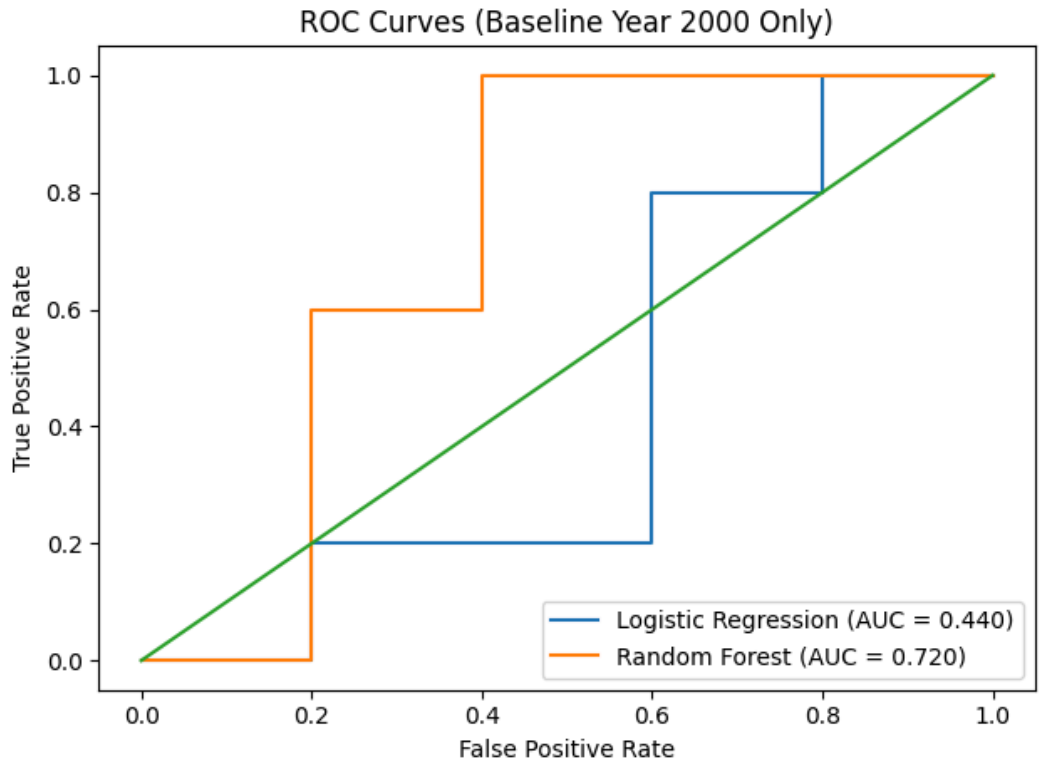

Figure 7: ROC Curves after removal of data leakage

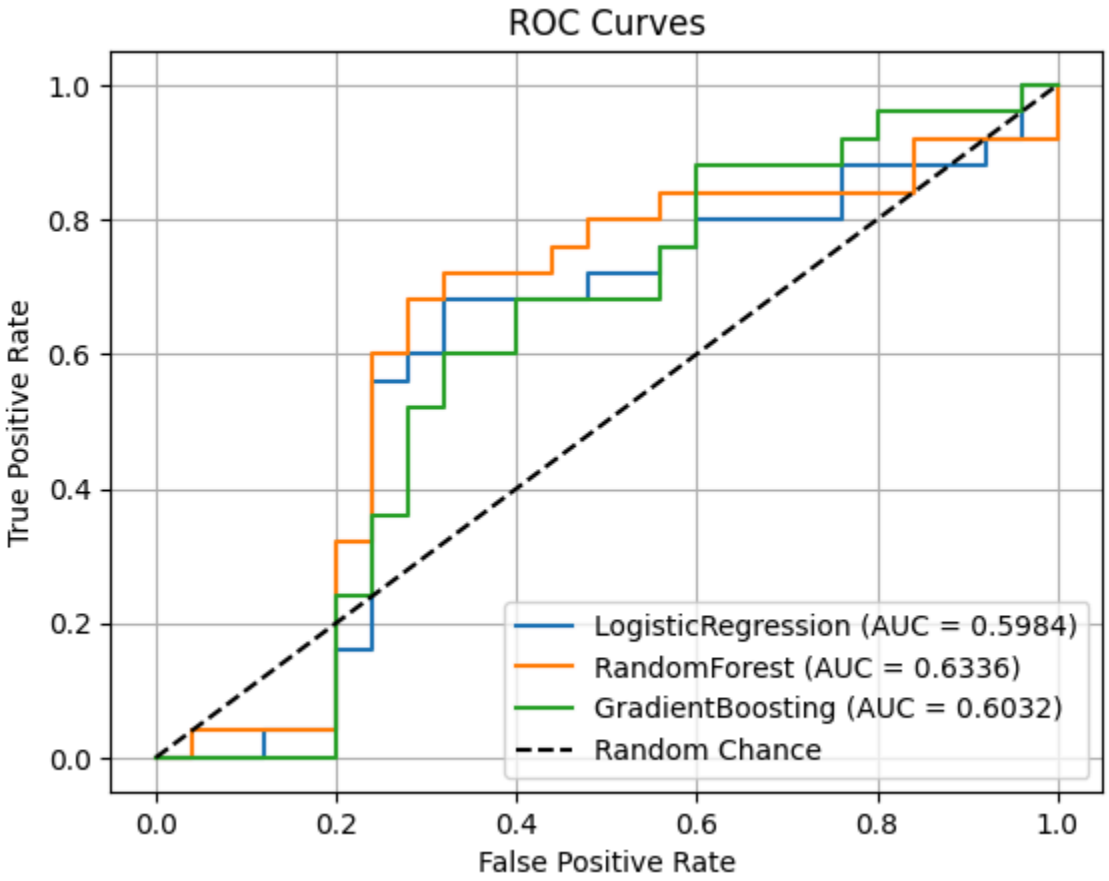

Figure 8: ROC Curves for our final three models.

Code File	Model Type	Model Name	Parameters	Accuracy	Precision	Recall	F1-Score	AUC	Notes
baseline_models	Logistic Regression	log_reg	max_iter=2000	0.6	0.571	0.8	0.667	0.44	
baseline_models	Random Forest	rf	n_estimators=300, random_state=42	0.6	0.667	0.4	0.5	0.72	
SupervisedModels	Logistic Regression	log_reg	Default	0.66	0.7412	0.578	0.6495	0.6584	
SupervisedModels	Random Forest	rf_clf	Default	1	1	1	1	1	Overfit
SupervisedModels	Gradient Boosting	gb_clf	Default	0.995	0.9909	1	0.9954	1	Overfit
SupervisedModels	LightGBM	lgbm_clf	Default	0.985	0.9818	0.9908	0.9863	0.9998	Overfit
SupervisedModels	XGBoost	xgb_clf	use_label_encoder=False, eval_metric='logloss'	1	1	1	1	1	Overfit
Eval_Interpretation_Pipeline	Gradient Boosting		"model__subsample": 1.0, "model__n_estimators": 150, "model__max_depth": 2, "model__learning_rate": 0.1	0.58	0.6032		0.6316		
Eval_Interpretation_Pipeline	Logistic Regression		"model__C": 0.1, "model__penalty": "l2"	0.64	0.5984		0.625		
Eval_Interpretation_Pipeline	Random Forest		"model__subsample": 1.0, "model__n_estimators": 150, "model__max_depth": 2, "model__learning_rate": 0.1	0.66	0.6336		0.6909		

Table 3: Model Performance Summary

Summary of Final Model and Tuning

- Leakage removed: Only baseline (2000) predictors used.
- Cross-validation: 5-fold stratified CV, optimizing ROC-AUC.
- Best model: Random Forest (ROC-AUC = 0.63, Accuracy ≈ 0.66, F1 ≈ 0.69).
- Top predictors:
    1. Democratic presidential vote share
    2. Renewable energy consumption (BTU)
    3. Republican presidential vote share
    4. Median income
    5. Renewable generation (MWh)

Interpretation: Political lean, economic capacity, and baseline energy footprint matter more than raw counts of early policies.

Conclusion

This project set out to answer the question of what combination of state characteristics and renewable energy policy design best predict a successful reduction in fossil fuel shares per state in the U.S.? By combining various datasets containing energy generation and consumption data, renewable energy policy implementation, median income, and political polling, we developed a machine learning model to predict whether a state is above or below the national median in decreasing their usage of fossil fuels. Our final Random Forest model achieved a modest 66% accuracy, with room to grow with more fine tuning. It is clear that the highest feature predictors for our model are renewable energy generation and consumption, presidential voting, and median income, which was also what was predicted.

These findings align with other researchers as well. In an article by Carley, Engle, and Konisky in 2021, they found that unless equity considerations are involved in a given policy, there are many groups that may not be able to reap the benefits of this renewable energy usage growth. This aligns with our thoughts that not only the presence of a policy is important, but also the content and the enforcement of these policies. (Carley, Sanya & Engle, Caroline & Konisky, David M., 2021)

Discussion & Next Steps

Key Takeaways

Our question simply asked what combination of state characteristics and renewable energy policy design best predict reductions in fossil fuel usage. The results of our modeling suggest that baseline structural conditions, such as political leaning, median income, and initial energy dependence, are stronger predictors of long-term fossil fuel reduction than the base number of policies put in place. As seen in table 2, variables such as Democratic presidential vote share (0.25) and median income (0.07) were more influential to the model than total policy counts (0.02). This reinforces our finding that it is not the amount of policies put in place, but the content and way that it is implemented.

It is also important to discuss limitations. The final model resulted in an AUC value of 0.63, which leaves room for improvement. Some variables, such as state area and imputed income, were static or estimated, which may understate their dynamic effects. Another dynamic environment that is not well accounted for is the election vote share. We use this information to assign the political leaning for states, but it does not incorporate a dynamic way of changing this, as political climates do so often. Some additional caveats consist of data gaps in state-level policy tracking limited coverage, multicollinearity in energy metrics required pruning, reducing feature diversity, and generalizability limited to U.S. states; external validity may differ globally.

In the future, there are many areas for improvement. Our future work should incorporate policy design and enforcement data, since our results suggest that quantity of policies alone is insufficient. Including measures of grid infrastructure investment, federal incentive timing, and renewable resource availability would likely improve model accuracy. We would also consider temporal/sequential models (e.g., recurrent neural nets) to capture policy dynamics.

Code Availability

GitHub: https://github.com/TaxMuttle/DSE6311

Appendix

Data Dictionary

Variable	Type	Unit	Description	Derivation / Source	Missing (n)	Missing (%)	Unique	Min	Max	Example
Area (SqMi)	float	square miles	State land area in square miles.	Original or merged from sources	1000	100.0	0			
Dem_Pres_Percent	float	percent (0–100)	Percent of votes for the Democratic presidential candidate that year.	Original or merged from sources	0	0.0	250	21.628900876393036	71.45291011787819	27.666339823504693
Dem_Sen_Percent	float	percent (0–100)	Percent of votes for Democratic candidates in Senate elections	Original or merged from sources	0	0.0	346	0.0	131.7568385145137	0.0

			that year (where applicable).							
Fossil Energy Consumption (BTU)	float	BTU	Total energy consumption from fossil sources.	Original or merged from sources	0	0.0	1000	419233.4	53302215.0	3497341.0
Median Income ($)	int	USD	State median household income (USD).	Original or merged from sources	0	0.0	986	35548.0	95572.0	78675
Non-Renewable Generation (MWh)	float	MWh	Electricity generated from non-renewable sources within the state.	Original or merged from sources	0	0.0	1000	1913.0	402730736.4	5154706.0
Renewable Energy Consumption (BTU)	float	BTU	Total energy consumption from	Original or merged from sources	0	0.0	999	1692.0	1140436.8	10891.0

			renewable sources.							
Renewable Generation (MWh)	float	MWh	Electricity generated from renewable sources within the state.	Original or merged from sources	0	0.0	996	0.0	99832043.94	1001819.0
Rep_Pres_Percent	float	percent (0–100)	Percent of votes for the Republican presidential candidate that year.	Original or merged from sources	0	0.0	250	26.4362019365703 1	72.7905331026891	58.620955315870575
Rep_Sen_Percent	float	percent (0–100)	Percent of votes for Republican candidates in Senate elections that year (where applicable).	Original or merged from sources	0	0.0	347	0.0	169.6995499935216 6	0.0

State	string	string	U.S. state name (50 states only).	Original or merged from sources	0	0.0	50			AK
Total Energy Consumption (BTU)	float	BTU	Total energy consumption; sum of fossil and renewable consumption.	Renewable Energy Consumption + Fossil Energy Consumption	0	0.0	1000	456711.2	54125065.8	3508232.0
Total Generation (MWh)	float	MWh	Total electricity generation; sum of renewable and non-renewable generation.	Renewable Generation + Non-Renewable Generation	0	0.0	1000	1911207.0	483201031.0	6156525.0
Total_Policies	float	count	Count of active renewable/clean-energy policies in	Original or merged from sources	0	0.0	24	0.0	48.0	2.0

			effect in the given year.						
Y_OUTCOME	int	binary {0,1}	Label: 1 if the state's fossil share reduction (2000–2019) exceeded the national median, else 0.	1 if (fossil_share_2019 − fossil_share_2000) < national median change; else 0	0	0.0	2	0.0	1.0	0
Year	int	year	Calendar year of observation.	Original or merged from sources	0	0.0	20	2000.0	2019.0	2000
fossil_share	float	fraction (0–1)	Share of total energy consumption from fossil sources for the given row/year.	Fossil Energy Consumption ÷ Total Energy Consumption	0	0.0	1000	0.816671470985143	0.997921157244056	0.996895587292972 8

fossil_share_2000	float	fraction (0–1)	Fossil energy share in baseline year 2000.	Original or merged from sources	0	0.0	50	0.8423892719344019	0.99689558729729728	0.9968955872929728
fossil_share_2019	float	fraction (0–1)	Fossil energy share in 2019.	Original or merged from sources	0	0.0	50	0.8255218389793596	0.99360774552717663	0.9904568775213516

Works Cited

Carley, Sanya & Engle, Caroline & Konisky, David M., 2021. "An analysis of energy justice programs across the United States," Energy Policy, Elsevier, vol. 152(C).

MIT Election Data and Science Lab. *U.S. President 1976–2020*, 2017, https://doi.org/10.7910/DVN/42MVDX. Accessed 27 8 2025.

MIT Election Data and Science Lab. “U.S. Senate statewide 1976–2020.” *Harvard Dataverse*, 2017, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PEJ5QU. Accessed 27 August 2025.

NC Clean Energy Technology Center. “Programs.” *DSIRE*, https://programs.dsireusa.org/system/program. Accessed 27 August 2025.

Sakellaris, A. “How politics shapes the future of renewable energy: Policies, priorities, and potential.” *Telkes*, 12 12 2024, https://www.telkes.org/energynews/how-politics-shapes-the-future-of-renewable-energy-policies-priorities-and-potential. Accessed 27 8 2025.

U.S. Census Bureau. “Income in the Past 12 Months (in 2023 Inflation-Adjusted Dollars) American Community Survey, ACS 1-Year Estimates Subject Tables, Table S1901.” *U.S. Census Bureau*, U.S. Census Bureau, U.S. Department of Commerce., https://data.census.gov/table/ACSST1Y2023.S1901?q=median+household+income+by+state. Accessed 27 8 2025.

U.S. Census Bureau. “State Area Measurements and Internal Point Coordinates.” *U.S. Census Bureau*, 16 December 2021, https://www.census.gov/geographies/reference-files/2010/geo/state-area.html. Accessed 27 August 2025.

U.S. EIA. “Historical State Data - U.S. Energy Information Administration.” *U.S. EIA*, 4 10 2024, https://www.eia.gov/electricity/data/state/. Accessed 27 August 2025.

U.S. EIA. “U.S. Energy Information Administration.” *U.S. Energy Information Administration - EIA - Independent Statistics and Analysis*, 27 6 2025, https://www.eia.gov/state/seds/seds-data-complete.php. Accessed 27 August 2025.