Predicting American Engagement in Climate Activism: Features and Social Implications

Clare Milligan and Jocelyn Jiang

School of Data Science, University of Virginia

DS 3001: Foundations of Machine Learning

Professor Navya Annapareddy

May 7, 2025

Introduction

This project used Pew Research Center's American Trends Panel Wave 89 dataset, which contained response data from a 2021 survey asking respondents about their social media usage, climate change knowledge, engagement with climate change prevention, and demographic details. We wanted to study how Americans' engagement with climate change prevention could be predicted based on personal choices (e.g., following activists on social media) or if any patterns existed between identity (i.e., demographics) and climate involvement. Understanding these patterns seemed to promise insights in how to raise interest in helping the environment.

Data Pipeline

The Exploratory Data Analysis (EDA) of this dataset began with examination of the dataset's shape, feature data types, and null value counts. These initial checks uncovered concerns such as high dimensionality and high proportions of null values in some columns.

Next, we took value counts of several select features and learned why nearly every column's data type was a float: each column corresponded to a survey question, and its values represented one of the question's answer choices. This meant we would not need to do further encoding; the values were unlikely to have inconsistent format, scaling, or outlier issues (features typically needing scaling like income were encoded as integers representing ranges of income rather than raw income amounts).

We then wrote a helper function *get_feat_val_labels()*. Our data frame was retrieved from a SAV file, and pyreadstat's *read_sav()* function provided us with metadata containing explanations of the data frame's encoded values. Our helper function printed the meanings of an input column's encoded values, and using *get_feat_val_labels()*, we discovered several columns had values of 99.0, meaning the answer had been 'Refused' (Figure 1). This informed us of apparently non-null values that we should treat as missing in addition to explicitly encoded NaNs. We next aimed to reduce dimensionality and clean our data before creating visualizations.

First, columns we considered irrelevant were dropped, including the datetime and device type of survey completion. Device type, for example, was irrelevant because each of its values corresponded to an electronic device (e.g., laptop or tablet), providing no significant insight about respondents. Demographic data were kept along with two columns containing information about samples' social media usage for following climate activism. With these two columns kept, most of the remaining columns became irrelevant because their information was redundant.

Next, we addressed null values by creating a data frame with each column's proportion of missing values to total rows. Columns were dropped if they had exceptionally high null value proportions (e.g., 'F_BORN' with over 60 percent NaNs) or relatively high proportions while another existing column encapsulated similar information without so many null values (e.g., keeping 'F_PARTYLN_FINAL' was unjustified when party affiliation was also covered by 'F_PARTYSUM_FINAL,' which had no null values). We imputed the remaining columns' null or refused values using that column's mode.

Next, we checked for class imbalance with a value_counts of the target variable, Engagement. Our target experienced dramatic imbalance, with over 87 percent of the samples belonging to the majority class of low engagement (which we redefined as performing fewer than two climate engagement actions). To address the imbalance, we performed our train and test set split and called SMOTE on our train set.

After resampling, we performed a Principal Component Analysis as a final dimensionality reduction technique and found eleven components were needed to explain 95 percent of the data variance (Figure 2). Using the explained_variance_ratio_ attribute of the fitted PCA instance, we calculated the cumulative variance that each column explained across the eleven principal components. For columns with low cumulative variance, we applied our previous logic of dropping them if their data was described by a similar column with greater relevance (cumulative weight in the PCA).
We visualized potential relationships between engagement and demographic and behavior-based features. For instance, we plotted each generation's engagement in the context of party affiliation (Figure 3.1), which surprisingly indicated that across generations, the most highly climate-engaged respondents were those who did not indicate party affiliation, and the second most engaged party affiliation was the Republican Party. We also investigated engagement in the context of social media usage by generation (Figure 3.2). This plot indicated that non-social media users have greater climate engagement. Both plots proposed that we should expect no singular behavior or demographic to have dominating influence on the target even when a relationship might seem intuitive.

Model Overviews: Supervised

To classify samples' climate engagement, we chose sklearn's RandomForestClassifier model for its strong performance potential (due to pooling information from numerous decision trees). Before the RandomForestClassifier, however, we tuned a DecisionTreeClassifier, interested in the benefits of this model's interpretability when we would analyze groupings and features in our later clustering model.

DecisionTreeClassifier

To tune our Decision Tree, we used GridSearchCV and custom values for the hyperparameters *max_depth*, *min_samples_split*, and *min_samples_leaf*. We tuned these parameters hoping to prevent overfitting and make more informed decision splits. The values chosen for the Decision Tree parameter grid were the hyperparameters' default values and small multiples of the default value. These values were chosen to evaluate the performance at varying levels of intensifying the parameters.
Selecting the best estimator from GridSearchCV instance, we found that a *max_depth* of none, *min_samples_split* of 10, and *min_samples_leaf* of 4 were the hyperparameters values that produced the greatest f1_macro score (Figure 4). We wanted GridSearchCV to tune based on this parameter as another measure against our dataset's initial class imbalance. Evaluating our Decision Tree's performance, we calculated accuracy, F1 score (macro), and F1 score (weighted) with a focus on the F1 score (macro) because neither false positives nor false negatives are more costly than the other in our classification problem, and this metric considers class imbalance better, treating the engaged and unengaged respondent classes as equally important. The F1 score (macro) was about 0.58, with predictions of the majority class having an F1 score of 0.91 and minority class F1 score of 0.26 (Figure 5). The better predictions of the low climate engagement respondents was an expected result considering its much larger support.

RandomForestClassifier

Our Random Forest model was similarly tuned using GridSearchCV, and the same hyperparameters were used because of their continued advantages of preventing overfitting and decision split validity. A parameter added to the Random Forest's parameter grid was the number

of estimators, a hyperparameter significant to this model because it can capitalize on the benefits of Random Forest's ensemble learning.

The best Random Forest estimator retrieved from the GridSearchCV had *max_depth* of none, *min_samples_split* of 5, *min_samples_leaf* of 1, and *n_estimators* of 100 decision trees (Figure 4). As a result of our Decision Tree model's poor classification of the high climate-engagement class, we lowered the threshold from the default 0.5 to a custom threshold of 0.35, effectively making predictions of the minority class less strict, attempting to improve the true positive rate (TPR) and recall as a result. This custom threshold was determined after testing thresholds with our helper function *plotROC()*, took an input custom threshold and plotted a vertical line corresponding to the passed-in argument threshold on the ROC graph. Observing the tradeoff between TPR and FPR at various thresholds, at the threshold 0.35, the ROC curve had the steepest growth in TPR with the lowest growth in FPR (Figure 6). However, keeping TPR high while minimizing FPR for a stronger F1 score means the TPR at our custom threshold is not high; at this threshold, though the FPR is maintained below 0.1, the TPR approaches only 0.4. The Random Forest Classifier model's overall F1 score (macro) was 0.64. The majority class achieved an F1 score of 0.91, and the minority class achieved an F1 score of 0.37 (Figure 7), demonstrating, like the Decision Tree model, a more accurate classification of the minority than the majority class.

Model Overviews: Unsupervised

We utilized K-Means clustering as our unsupervised model to explore potential natural groupings in the survey respondents based on demographic and behavioral data. After preprocessing, we applied K-Means clustering to segment respondents into two distinct clusters based on their demographic and social behavior characteristics. We used the elbow method to determine the optimal number of clusters. Based on the plot, we identified k=2 and k=4 as potential cluster numbers (Figure 8), but after running both we decided that k=4 did not add accuracy to our analysis.

The two clusters are clearly separated, with Cluster 0 generally positioned on one side and Cluster 1 on the other (Figure 9). This separation suggests that the K-Means algorithm successfully distinguished between two groups based on the features provided. Both clusters are elongated along their axes, indicating that there is a variation in feature values across the clusters. As you approach the middle of the plot, the clusters become thinner, suggesting that the data points within each cluster are more concentrated around certain values, particularly at the extremes. The density of points near the center of the clusters is lower, which suggests that most respondents fall toward the extremes of the cluster distribution in terms of their features. This reinforces the idea that the features driving the clustering are likely concentrated at the ends of their respective scales, leading to distinct groups with less overlap. Based on our two clusters we were able to identify some general characteristics of each cluster.

Respondents in Cluster 0 have slightly higher educational attainment and political involvement but less religious devotion, observed by the higher values for features like "F_EDUCCAT5" (education level) and "F_PARTYSUM_FINAL" (political engagement) in Figure 10. Climate-related behaviors are less pronounced in Cluster 0, as seen in features like "FOLLOWS CLIMATE ACCOUNTS" and "Seeks Climate Info Online." Respondents in Cluster 1 tend to have lower educational attainment and are more likely to be religious, as indicated by higher values in the "Religious Affiliation" and "Religious Service Attendance" columns (Figure 10). This group also shows higher political engagement and climate-related behavior. They have higher average values for "F_PARTYSUMIDEO_FINAL" (political

ideology) and social media engagement ("FOLLOWS CLIMATE ACCOUNTS"). Cluster 1's respondents also have slightly higher values for "F_INTFREQ," the frequency of interaction with climate content.

When analyzing the correlation between each cluster and engagement to see if either cluster lended itself to more engagement, we found that those in Cluster 1 were slightly more likely to be engaged. However, this correlation was not large enough to be very significant. Therefore, while the clustering did uncover meaningful groups with varying behaviors and demographic profiles, the correlation indicates that clustering alone may not perfectly predict the level of engagement. However, the differences between the clusters provide actionable insights into how engagement levels and behaviors can vary across different types of individuals.

Discussion

The Decision Tree classifier demonstrated an ability to model the data but showed limitations in predicting the minority class (high engagement), with an F1 score of 0.26 for this class. The Random Forest classifier, on the other hand, provided an improved performance with an F1 score of 0.37 for the minority class, showing that ensemble learning through Random Forest had a better grasp of the underlying structure in the data, though performance was still biased towards the majority class (low engagement). The custom threshold adjustment (0.35) was effective at improving recall for the minority class, though its F1 score was still lower than for the majority class.

In terms of unsupervised learning, the K-Means clustering method successfully identified two distinct clusters in the dataset, which revealed valuable behavioral and demographic differences. One cluster (Cluster 0) was characterized by higher educational attainment, lower religious involvement, and less engagement with climate-related content. The other cluster exhibited more pronounced engagement with climate content and higher political activity, coupled with stronger religious affiliation. The clustering results confirmed that demographic and behavioral factors could segregate respondents into meaningful groups, though clustering alone did not perfectly predict engagement levels. The correlation between the clusters and engagement was weak, highlighting the complexity of engagement prediction beyond the obvious features.

In addition to our analysis of our models we must also address the privacy implications of working with individual-level survey data, particularly when it includes detailed demographic variables such as age, race, education level, political affiliation, and religious practices. Although the data used in this project is anonymized and does not contain direct identifiers like names or contact information, there remains a risk of re-identification, especially when multiple demographic features are combined. For instance, an unusual combination of traits, such as being a young, highly educated individual who identifies with a minor political party and lives in a small geographic region, could make a respondent more easily identifiable, even in a de-identified dataset. Wilkins 2021 states that with a de-identified data set, such as the one we used, “the dataset does not contain any identifiable information, but there is a way to link the information back to identifiable information.” Therefore, it could be important for further steps to be taken to not only de-identify this data but to fully anonymize it so as to better protect the privacy of the people surveyed.

This concern raises broader ethical questions about how survey participants' data is collected, shared, and used. In particular, we must consider whether individuals gave informed consent not just to participate in the survey, but also to have their data used for purposes beyond the original scope of the survey, such as in machine learning or academic research. In many public datasets, this boundary is blurred: respondents may not fully understand or anticipate the extent to which their anonymized responses might be distributed, analyzed, and potentially published in downstream

applications. Furthermore, the inclusion of sensitive demographic features introduces the potential for misuse. If not carefully framed, models built on these features could reinforce stereotypes or be misinterpreted as evidence of causal relationships—e.g., implying that certain racial or religious groups are less "engaged" in climate action. This is especially dangerous if results are taken out of context or used in policymaking without a nuanced understanding of the data. To address these issues, researchers must take extra steps to minimize harm, such as: avoiding unnecessary exposure of sensitive subgroups in reporting, ensuring transparency about how data was obtained and how it will be used, critically evaluating whether certain demographic features are ethically justifiable to include in the modeling process, committing to ethical data stewardship that considers the dignity and autonomy of survey respondents. There is also a growing concern about how data like this might be used for lobbying efforts, as political groups could use predictive insights to influence voters' decisions on climate policy. The potential for gerrymandering could also be relevant if such models were employed to draw district boundaries that disproportionately favor certain groups over others, further entrenching divisions in political engagement. This could have disastrous political consequences, especially in states like Ohio where gerrymandering has already been used to create "a supermajority legislature in Ohio [that] pass[es] bill after bill that short-circuits Ohio's ability to fight climate change" (Wessler 2021). In future work, incorporating principles from data ethics frameworks, such as differential privacy, data minimization, and participatory consent, could help ensure that research respects the individuals behind the data and avoids exacerbating existing social inequalities..

Conclusion

Based on the evaluation of both supervised and unsupervised models, we recommend the Random Forest classifier as the most suitable model for predicting climate engagement from this dataset. The Random Forest model's improved F1 score for the minority class makes it the better choice for this task. While there is room for improvement in exploring other techniques to address class imbalance, this model demonstrates a stronger overall ability to predict engagement accurately.

In terms of future applications, the predictive results from this analysis could be used to better understand which groups are more likely to engage with climate action, potentially informing public awareness campaigns or policy interventions. The unsupervised results from the clustering analysis, though not directly predictive of engagement, offer valuable insights into the profiles of individuals who are either more or less likely to engage in climate-related behaviors. These clusters could be used for targeted interventions, tailoring messages or outreach efforts to different demographic and behavioral groups.

From an ethical perspective, the potential for misuse of these models should not be underestimated. Privacy concerns, the possibility of voter profiling, and the risks of lobbying and selling data are all crucial issues to address in the deployment of such models. Clear guidelines and safeguards must be established to ensure that these findings are used in a responsible and ethical manner, with appropriate respect for individuals' rights and privacy.

References

Pew Research Center. (2021). *American Trends Panel Wave 89* [Data set]. https://www.pewresearch.org/science/dataset/american-trends-panel-wave-89/

Shankle, Derek. "Do You Know Me? The Subtle Distinction Between Anonymous and De-Identified Data in Clinical Research." *WCG Clinical*, 28 Mar. 2023, https://www.wcgclinical.com/insights/do-you-know-me-the-subtle-distinction-between-anonymous-and-de-identified-data-in-clinical-research/.

Tyson, A., Kennedy, B., Funk, C. (2021). *Gen Z, Millennials Stand Out for Climate Change Activism, Social Media Engagement With Issue*. Pew Research Center. https://www.pewresearch.org/science/2021/05/26/gen-z-millennials-stand-out-for-climate-change-activism-social-media-engagement-with-issue/

Wesseler, Sarah. "Gerrymandering Is a Climate Problem." Yale Climate Connections, 22 Nov. 2021, https://yaleclimateconnections.org/2021/11/gerrymandering-is-a-climate-problem/ .

Appendix

Figure 1: Helper Function for Analysis of Encoded Features

```
F_BORN
2.0     5082
1.0     3516
99.0     113
Name: count, dtype: int64
{1.0: 'Yes, born-again or evangelical Christian', 2.0: 'No, not born-again or evangelical Christian', 99.0: 'Refused'}
```

Figure 2

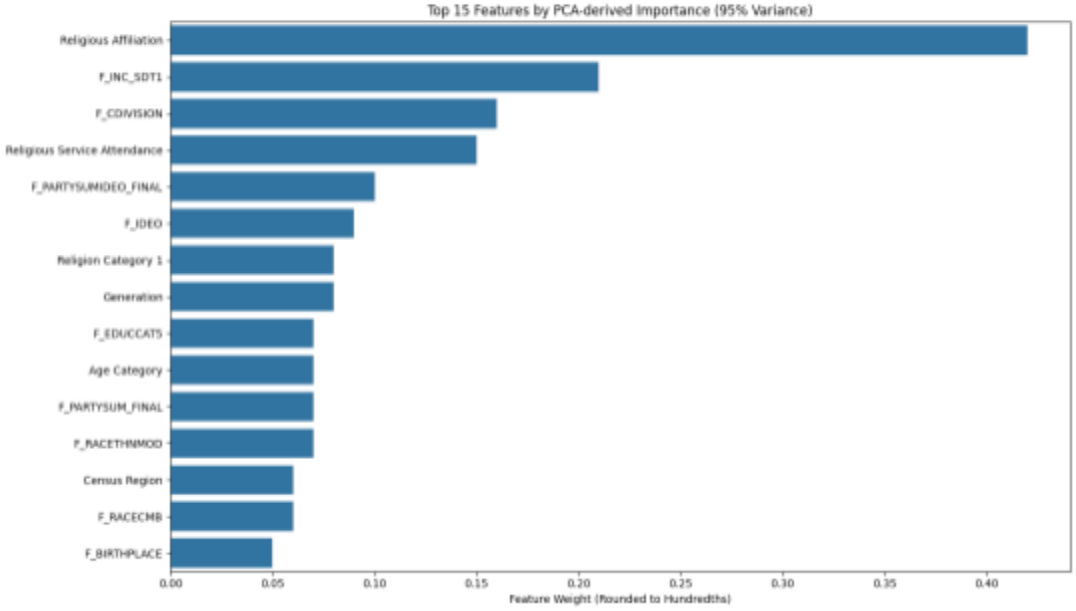

Figure 3.1

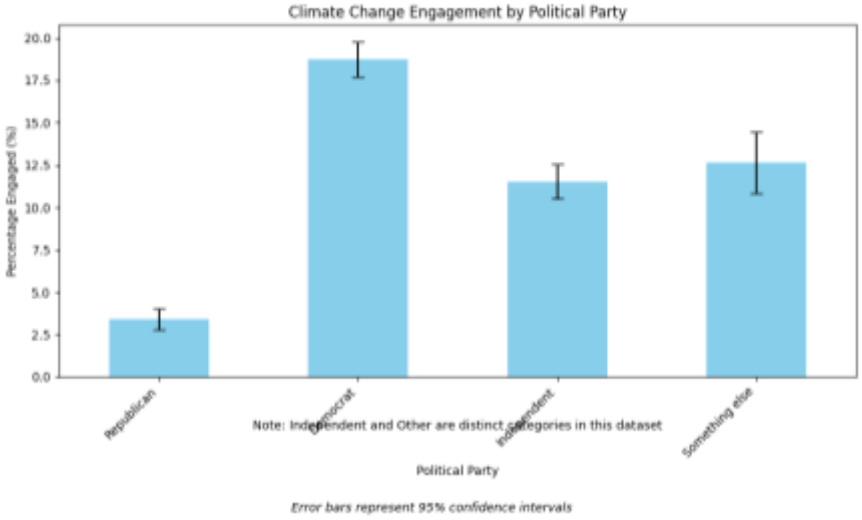

Figure 3.2

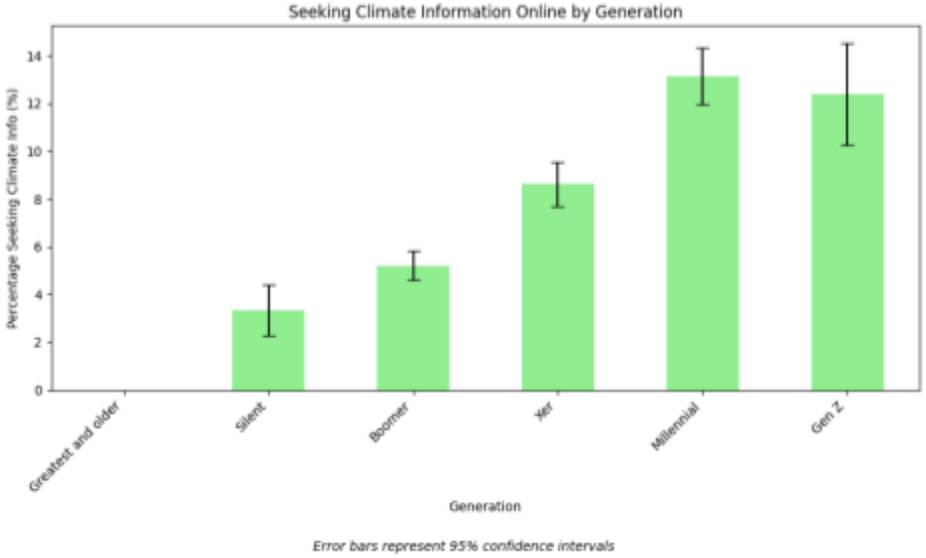

Figure 4

```
Decision Tree Best Estimator Parameter Values (GridSearch)
   max_depth  min_samples_split  min_samples_leaf
0       None                 10                 4
Random Forest Best Estimator Parameter Values (GridSearch)
    n_estimators max_depth  min_samples_split  min_samples_leaf
0            100      None                  5                 1
```

Figure 5

```
Decision Tree
Accuracy: 0.8363636363636363
F1 Score (macro): 0.5827382084346563
F1 Score (weighted): 0.8266634101088899
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91      3609
           1       0.30      0.23      0.26       516

    accuracy                           0.84      4125
   macro avg       0.60      0.58      0.58      4125
weighted avg       0.82      0.84      0.83      4125
```

Figure 6

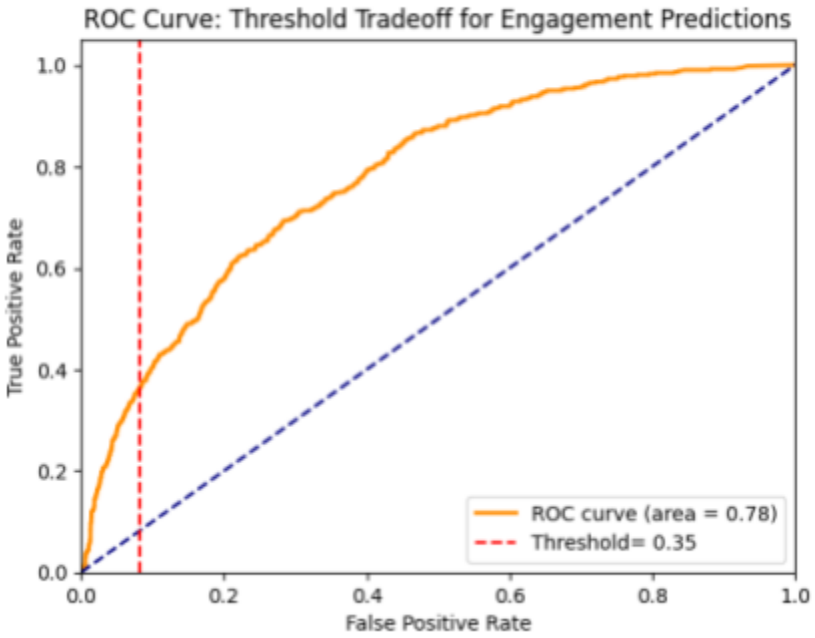

Figure 7

```
Random Forest (custom threshold)
Accuracy: 0.8487272727272728
F1 Score (macro): 0.6443521350478121
F1 Score (weighted): 0.8465050934240742
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.92      0.91      3609
           1       0.39      0.36      0.37       516

    accuracy                           0.85      4125
   macro avg       0.65      0.64      0.64      4125
weighted avg       0.84      0.85      0.85      4125
```

Figure 8

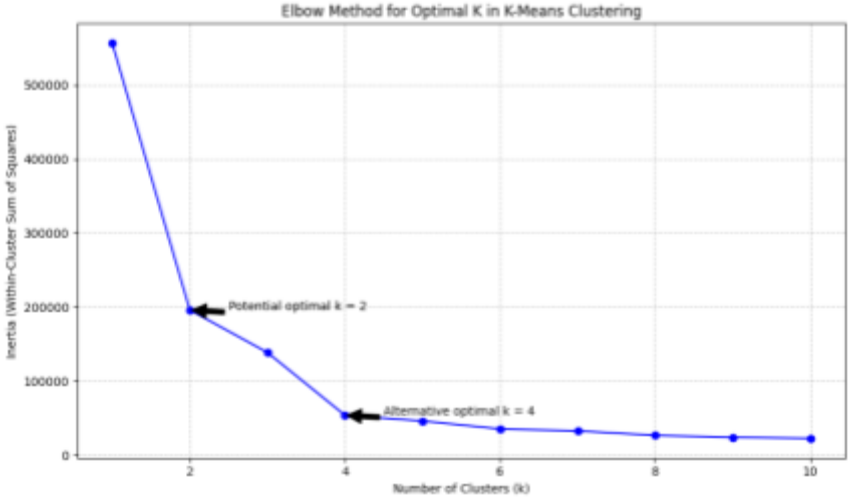

Figure 9

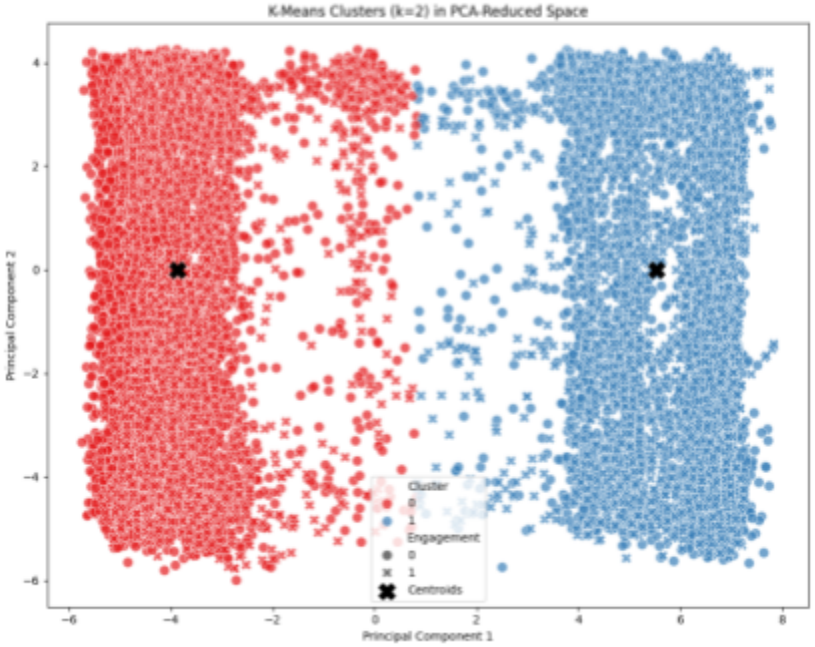

Figure 10

Cluster Characteristics (Mean Values):

Cluster	0	1
F_CDIVISION	5.09	5.46
Age Category	2.72	2.47
F_GENDER	1.60	1.56
Education Level	1.57	1.46
F_EDUCCAT5	3.67	3.83
F_HISP	1.82	1.87
F_YEARSINUS	1.30	1.27
F_RACECMB	1.40	1.42
F_RACETHNMOD	1.61	1.61
F_BIRTHPLACE	1.37	1.33
Religious Affiliation	1.62	10.70
Religious Service Attendance	3.42	5.42
Political Party	2.16	2.47
F_PARTYSUM_FINAL	1.77	2.01
F_PARTYSUMIDEO_FINAL	2.80	3.61
F_INC_SDT1	5.51	5.54
F_IDEO	2.95	3.84
F_INTFREQ	1.67	1.53
F_VOLSUM	1.46	1.50
Generation	3.78	4.07
Seeks Climate Info Online	1.88	1.85