Braden Bourgeois **ISDS 7303, Final Assignment** **3/2/24**

Uplift Models on Voter Data

Introduction and Business Understanding: There are many strategies that go into determining which voters should be contacted in order to influence their vote. The objective is to use data in order to determine which voters are persuadable and worth contacting. The goal is to use an Uplift Model on the data collected in order to make well informed decisions.

Data Understanding: The data being used comes from the 'VoterPersuasion Demographics.jmp' file including twenty-four variables of voter information. Moved_AD was a binary response variable created by running a randomized experiment and indicates if the voter moved toward favoring the candidate of the Democratic Party.. A Validation column was created for the DOE data set in order to evaluate the treatment on the output, Moved_AD. A value order was established to identify specific voter groups for targeted analysis. By assigning numerical values (1 and 0) to certain categories of variables such as Moved_AD and Flyer, the prioritization of these groups was clarified. The selection of the particular voter segment for treatment during an election is determined through additional analysis of datasets. The final table, VoterPersuasion merged data.jmp, includes the original twenty-four variables as well as Treatment, Validation, Flyer, and Moved_AD. This table was screened for missing values and did not have any. The assumption was made that for this analysis, it is assumed outliers were already excluded given the format of this data set.

Randomized Experiment: Initially, a randomized experiment was undertaken using the dataset. It is imperative to ensure complete randomization in the experiment to minimize the introduction of bias. Randomly assigned groups of subjects receive either the treatment or control conditions. Covariates play a crucial role in understanding potential sources of bias, identifying variables that may act as blockers, and recognizing confounders. Regression analyses are conducted on both the response variables and treatment factors to address these considerations. If a variable shows significance for both the treatment and response, it is considered a confounder and should be incorporated into the experiment. In randomized experiments, random assignment reduces bias and ensures comparability between treatment groups. Furthermore, causal inference is facilitated, enabling researchers to establish cause-and-effect relationships. Controlling for confounding variables enhances the internal validity of the study. However, conducting randomized experiments can be resource-intensive and time-consuming. Ethical considerations may limit the feasibility of randomization in certain scenarios, and while randomized experiments offer strong internal validity, generalizability to real-world settings may be limited.

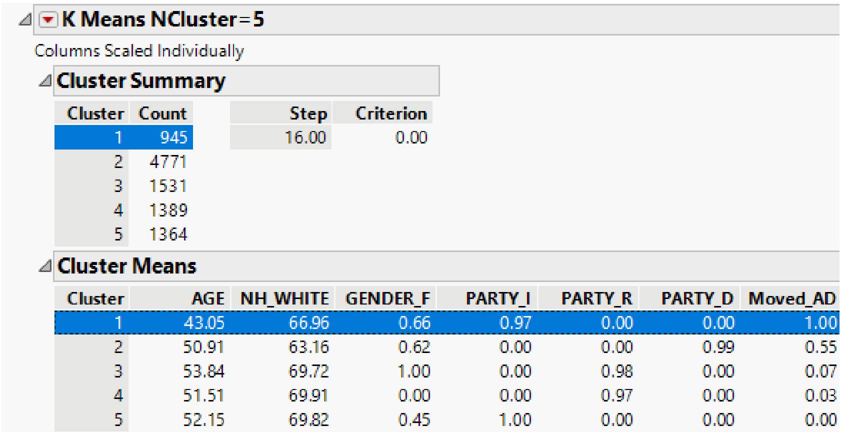
In contrast, if data are collected from past mail-outs and surveys instead of from a randomized design, observational methods are utilized to assess associations between treatment and outcome variables. While observational data offer insights into real-world scenarios, they may be vulnerable to bias from unobserved confounding variables. Additionally, causal inference in observational studies is more challenging due to the absence of randomization, making it difficult to establish causality.

Cluster Analysis: A K-means cluster with 5 clusters was run on Age, NH_White, Gender_F, Party_I, Party_R, Party_D and Moved_AD. The results are in Figure 1. Cluster 1 with the lowest count (945) represents people who had a party of independent. All of them had a Moved_AD of 1 representing they were moved to be more supportive of the democratic candidate. Cluster 2 is the biggest cluster (4771) representing democrats where half of them moved to be supportive of the democratic candidate. Clusters 3 and 4 represent the female and male republicans, who have low Moved_AD (.07 and .03). Cluster 5 shows the independents who were not moved. Another cluster analysis was run after creating the Uplift Model. The difference in probabilities was used in place of Moved_AD. The results are in Figure 2. Figure 2 incorporates information about treatment effects into the clustering process. These clusters are based on the predicted uplift scores, which indicate the expected change in behavior or outcome as a result of the treatment. These clusters show a much lower average Moved_AD than the clusters in Figure 1 with only one cluster surpassing 20% which is cluster 1 with a count of only 439.

Uplift Model: An uplift model was created using all the variables from the original dataset. Flyer was used as the treatment, Validation was used for Validation, and the Y-variable was Moved_AD. The objective is to identify persuadable voters which are indicated by groups in the top right. The results are shown in Figure 3. The model has 12 splits and shows that if PARTY_R>=1, H_F1<1, and AGE<69 would most likely have negative effects putting them in the lost causes and do not disturb categories. The Uplift Graph (Figure 4) sorts the population and shows that about 15% of the dataset has an uplift score of .15 and for about 55% of the population the uplift score is about .05 making it worthwhile to send a flyer.

Conclusion: In conclusion, uplift analysis provides valuable insights into voter behavior and the effectiveness of targeted interventions like sending flyers. Targeting segments of the population with higher uplift scores, such as the top 15%, can enhance campaign impact and effectiveness. Moreover, understanding the characteristics associated with treatment effects informs strategic decision-making for future outreach efforts, allowing for more informed targeting and resource allocation, ultimately leading to more successful persuasion campaigns.

Figure 1: K-Means Clustering Means



K Means NCluster=5

Columns Scaled Individually

Cluster Summary

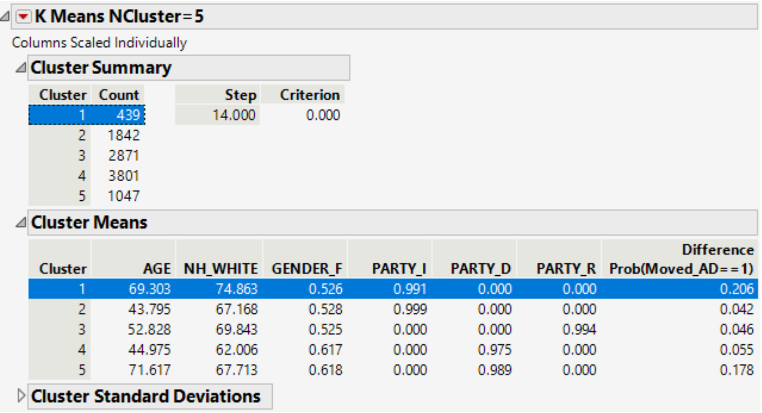
Cluster	Count
1	945
2	4771
3	1531
4	1389
5	1364

Step	Criterion
16.00	0.00

Cluster Means

Cluster	AGE	NH_WHITE	GENDER_F	PARTY_I	PARTY_R	PARTY_D	Moved_AD
1	43.05	66.96	0.66	0.97	0.00	0.00	1.00
2	50.91	63.16	0.62	0.00	0.00	0.99	0.55
3	53.84	69.72	1.00	0.00	0.98	0.00	0.07
4	51.51	69.91	0.00	0.00	0.97	0.00	0.03
5	52.15	69.82	0.45	1.00	0.00	0.00	0.00

Figure 2: K-Means Clustering Means (Lift)



K Means NCluster=5

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	439
2	1842
3	2871
4	3801
5	1047

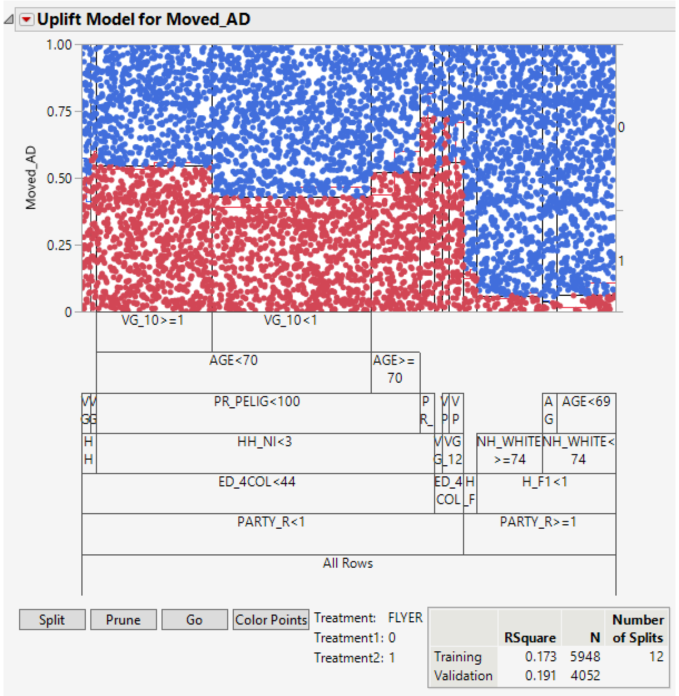
Step	Criterion
14.000	0.000

Cluster Means

Cluster	AGE	NH_WHITE	GENDER_F	PARTY_I	PARTY_D	PARTY_R	Difference Prob(Moved_AD==1)
1	69.303	74.863	0.526	0.991	0.000	0.000	0.206
2	43.795	67.168	0.528	0.999	0.000	0.000	0.042
3	52.828	69.843	0.525	0.000	0.000	0.994	0.046
4	44.975	62.006	0.617	0.000	0.975	0.000	0.055
5	71.617	67.713	0.618	0.000	0.989	0.000	0.178

Cluster Standard Deviations

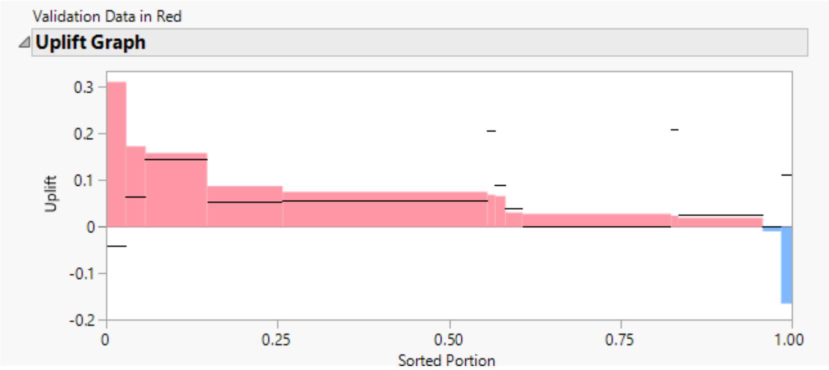
Figure 3: Uplift Model



Treatment: FLYER
Treatment1: 0
Treatment2: 1

	RSquare	N	Number of Splits
Training	0.173	5948	12
Validation	0.191	4052	

Figure 4: Uplift Graph


Validation Data in Red
Uplift Graph
Uplift
Sorted Portion