Report:
Analyzing Voter Persuasion Using Uplift Modeling

Introduction and Business Understanding:

When it comes to voting in elections, there are many factors that motivate some to vote one way or another. It can be helpful to utilize voter data to see what persuades voters into making their decisions. The uplift model can be used to share insights on voter persuasion and identify subsets of individuals who could be persuaded into voting a certain way.

Data Understanding:

There are two types of data that go into uplift modeling, demographic data and design of experiment data, which were given. The two data sets ("VoterPersuasion Demographics and DOE Results) were then combined into "VoterPersuasionmergeddata" which will be used for the analysis. The combined data set contains 10,000 rows of data and was checked for outliers and missing values, resulting in 26 rows being excluded. There are 24 covariates for each voter.

Analysis:

Firstly, a cluster analysis was performed on the data to shed more insights on voters. In JMP, Analyze then Clustering then K-Means Clustering with 5 clusters was selected. The variables that were used were: variables: AGE, NH_WHITE, GENDER_F, PARTY_D, PARTY_R, PARTY_I, MOVED_AD. MOVED_AD was recoded to continuous for this part. The clusters can be seen in Appendix A and vary in size. Cluster 3 is the largest with cluster a count of 2655 and cluster 1 was the smallest with a count of 1373. Cluster 2 had the lowest mean age of 47.6398 and cluster 4 had the highest mean age, 53.9229. Cluster 1 had the lowest "Gender_F" value,0, and cluster 4 had the highest value with 1. Cluster 2 had the highest "Party_D" value of 0.9875 and clusters 1, 4, and 5 had a value of 0. For "Party_R", cluster 4 had the highest value with 0.9888, and clusters 2,3, and 5 had a value of 0. All clusters had a value of 0 for "Party_I" except for cluster 5 with a value of 1. Cluster 3 had a value of 1, the highest value, for "Moved_AD" and cluster 2 had the lowest with a value of 0.

A randomized experimental design allows for randomly assigning experimental conditions which removes bias from the process. This design is also ideal for data sets with many variables because there should be few interactions between them. The procedure involves the random assignment of treatments, verifying each variable has the same probability of receiving treatment, and the randomization itself. All the previous forms of information collection such as mailouts could be used to gain general insights, but this design is ideal for this situation. The design that was selected was a simple randomized experiment with the data divided into two randomly chosen samples and the results were provided in "DOE Results (1).jmp". The "Flyer" column is used as the treatment column and the column Moved_AD served as a binary variable, showing favorability to candidate A of the Democratic party. The information was recorded in the "DOE results" file. The two tables were merged by voter ID, resulting in"VoterPersuasionmergeddata", which was given.

Next, the development of the uplift model began. After obtaining data on variables of interest like demographics, the randomized design experiment was performed. The experiment was used to evaluate the effect of the treatment on the response, using Yes/No for both treatment and response. Then, the results of that experiment were added to the original data set. From there, the data was split into training and validation sets, and the model was built to estimate outcomes based on covariates and treatment level, using the validation set to score the best model. The model was used in reserved treatment levels to score responses again to compute the probabilities. The probabilities are used to compute the differences between them of success for different treatments thus making the uplift model.

In JMP, the uplift function was used by going to Analyze then Consumer Research then Uplift. This function allowed for the last three steps mentioned above in developing the uplift model to be combined into one step. MOVED_AD was recoded back to nominal for this part. As seen in Appendix B, the uplift graph showed that for the first 15% of the sorted dataset, the uplift was 0.15. This means there was a 15% probability of an increase in the likelihood to respond when given a flyer then compared to not getting a flyer. For most of the right half of the population, the uplift was at a 5% probability of an increase in likelihood to respond, until the end of the graph where there is no chance of response.

Finally, the difference in probabilities was saved to the merged data to be used for another cluster analysis. The K means clustering for 5 clusters was used where Moved_AD was replaced with the difference in probabilities. The cluster analysis is shown in Appendix C. Cluster 2 is now the largest with a count of 4467 and cluster 5 is the smallest with 575. Cluster 1 now has the highest median age of 64.2502 and cluster four has the lowest median age of 41.6346. Cluster 4 had the lowest "NH_White" value and cluster 5 had the largest with 82.1130. For "Gender_F", cluster 5 had the lowest value of 0.2678, and cluster 2 had the highest value of 0.6447. "Party_D" has the highest value in cluster 2 but the lowest in clusters 1, 3, and 4 with 0. For "Party_R", cluster 4 was the highest in value, and clusters 2,3, and 5 were lowest with 0. "Party_I" had the highest values in cluster 3 and the lowest in clusters 1,2, and 4. The difference in probability had the highest value in cluster 5 and the lowest value in cluster 1 1. Comparing the two cluster analyses, cluster 5 should be

contacted in the future since that cluster has a high probability of responding.

Conclusion

Understanding what persuades voters to vote one way versus another is crucial to examining past elections and informing strategies for current and future elections. The first cluster analysis gave an initial idea of clusters within the population and voters who were most likely to move to Democrat was cluster 3. Using the randomized experimental design and eventually preparing for the uplift model allowed for more analysis of the population to determine that the first 15% of the dataset in the uplift graph had a 15% increased likelihood of responding to a flyer. The final cluster analysis with the added difference in probabilities showed that although clusters shifted that cluster 5 should be contacted in the future as they have a higher probability of responding from the flyer. Utilizing uplift modeling can shed invaluable insights into voter persuasion and can be used to contact the right group of voters for an election.

Appendix A: Cluster Analysis for "VoterPersuasion merged data.jmp"

K Means NCluster=5

Columns Scaled Individually

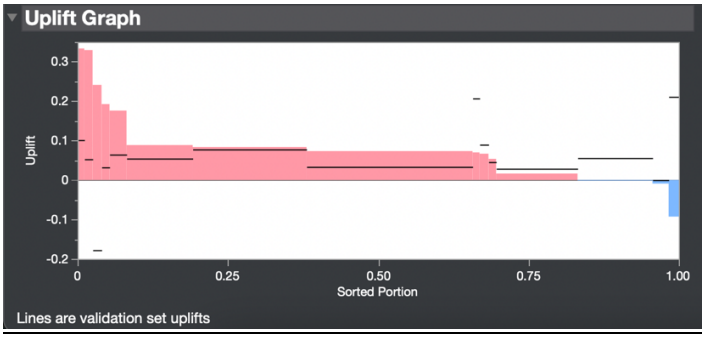
Cluster Summary

Cluster	Count	Step	Criterion
1	1373	17	0
2	2160		
3	2655		
4	1517		
5	2269		

Cluster Means

Cluster	AGE	NH_WHITE	GENDER_F	PARTY_D	PARTY_R	PARTY_I	Moved_AD
1	51.6744355	70.0473416	0	0	0.9817917	0	0.03277495
2	47.6398148	63.9106481	0.53564815	0.9875	0	0	0
3	53.1258004	62.653484	0.68474576	0.97853107	0	0	1
4	53.9228741	69.7567568	1	0	0.98879367	0	0.07448912
5	48.7148524	68.7020714	0.52842662	0	0	1	0.39973557

Appendix B: Uplift Graph



Appendix C: Cluster Analysis for "VoterPersuasion merged data.jmp" with Difference in Probabilities

K Means NCluster=5

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	1427	32	0
2	4467		
3	2030		
4	1475		
5	575		

Cluster Means

Cluster	AGE	NH_WHITE	GENDER_F	PARTY_D	PARTY_R	PARTY_I	Difference Prob(Moved_AD==1)
1	64.2501752	78.5711282	0.59074982	0	0.97897687	0	0.01576884
2	50.4235505	61.8791135	0.64472801	0.98791135	0	0	0.07615744
3	48.4014778	67.0339901	0.56108374	0	0	0.99704433	0.02749547
4	41.6345763	61.3179661	0.45762712	0	0.98372881	0	0.07115218
5	53.2904348	82.1130435	0.26782609	0.55304348	0	0.42608696	0.26166765