**Project Part 3**

Based on your work in this course to date, prepare a report for the campaign for whichever candidate you choose. The report should have the following parts:

1. Overall summary of the statistical approach taken, written in language that is accessible to the campaign management?

*Summary*

I use different features taken from the voter profile such as age, registered party, household income, education level, and previous voting records to predict whether the voter will vote for the Democrats or the Republicans. I also predict how I can persuade a user to vote for a different party.

1. Specific recommendations for what to do with each voter, messaging-wise

*Recommendations*

For highly supportive voters send content to reinforce support. When it is close to the election, focus on GOTV (go out to vote) efforts and ask them to help spread the candidate’s message.

For persuadable voters, provide a clear message aiming to persuade them based on policy rather than party loyalty. Use surveys or interactive tools to engage them on issues and collect data to refine further messaging targeting those voters with issue focused ads.

For undecided opposing voters, respectfully highlight the differences between the candidates and opponents on the key issues. Share testimony and stories from constitute that have similar background or concerns. If there is evidence of wide spread misconception or misinformation, provide factual correction in a non-confrontational manner.

For low information voters, focus on basic information about the candidates’ backgrounds and what they stand for. Avoid jargon or complex details.

For non-voters, or disengaged individuals, emphasize the importance of voting as a civic duty and the impact of their vote. Address common barriers to voting and provide solutions.

1. A technical section to document what was done, covering handling and data-prep, model-building, model assessment and scoring

*Technical*

I use logistic regression for partisanship prediction. The dependent variable is D2 indicating Democrat or Republican affiliation. I use the model to predict turnout in an off-year election. Use past data from a similar off year rather than more recent general election and the rationale that the behavior in a similar election would be more predictive. I also use SVM and random forest model to predict and combine the models to have a better accuracy score. It was observed that the decision tree performed better on the larger data set, indicating a certain method that can benefit more data due to increased granularity and range of scores.

The voter data set contains demographic and historical voting data of potential voters. The pre-processing step includes importing necessary python libraries for data manipulation and machine learning. Next, I read the data set and stored in a dataframe mapping categorical variables to binary indicators. For partisanship, I converted from CAND1S column with lambda function mapping SD to 1, otherwise 0. The turnout converted from VG\_14\_DV column with a lambda function mapping y to 1 else 0. Extracting relevant features such as age, number of democrats or republicans in a household, median age, non-Hispanic, Caucasian, Hispanic, median household income, percent of family with children under 18, etc. For model input and then standardizing features with standard scaler to ensure they contribute equally to the analysis.

Two separate logistic regression models were developed. Modelpartisanship predicts the likelihood of voter partisanship. Modelturnout forecasts the probability of voter turnout. The models were trained using 70-30 train test split to evaluate their performance on unseen data. The models were accessed and assessed using accuracy scores and classification reports which include precision, recall, and f1 score for both predictions. The partisanship model showed high precision for the negative class, but poor precision for the positive class, suggesting an imbalance in the data or a need for additional feature engineering. The turnout model demonstrated moderate accuracy, indicating that turnout prediction is more challenging and may require a more sophisticated model or additional features.

Predictions from both models were used to assign campaign messages. A function named assignmessage was created to assign messages based on the prediction for partisanship and turnout. Voters likely to support the campaign and likely to vote received tailored messages encouraging them to vote. Voters likely to support the campaign but less likely to vote received messages aimed at increasing turnout. Voters not predicted to support the campaign received a general campaign message or no message based on whether the campaign aimed to persuade them.

Model types include: LogisticRegression, DecisionTree, RandomForest, and SVM.

Method 1

For the accuracy turnout the SVM model is most accurate among those models with accuracyturnout:0.641

For accuracy partisanship the Logistic Regression is the most accurate with accuracypartisanship:0.6033333333333334

Since the accuracy score difference is more for the accuracy partisanship, I use logistic regression as the best model among those 4. Logistic Regression can be used as the final and best model.

Method 2(more accurate)

For the PartisanshipAccuracy Score the SVM Model has the best accuracy score of 0.627

The logistic regression, the random forest, and the decision tree all have the best turnout accuracy score.

You can choose a dependent variable such as 'MOVED\_RD', 'MOVED\_DR' or 'MOVED\_AW', etc., for persuadable:

Results: I used featured MOVED\_RD which indicates people moved from republican to democrat between wave 1 and wave 2 IDs and used a logistic regression model to find out the statistics of the persuadable voters' information.

Explain how the results will be utilized in identifying electorate segments for the campaign to address:  
1. High Support, High Turnout: These voters are predicted to support the campaign and are likely to vote. They can be targeted with messages that reinforce their support and encourage them to influence their social circles. This group is critical for word-of-mouth campaign efforts and should receive the most detailed and frequent communications.  
  
2. High Support, Low Turnout: Voters in this segment are supportive but might need extra motivation to vote. The campaign should focus on turnout efforts, such as reminders of voting dates, providing information on voting locations, and offering assistance for registration if needed.  
  
3. Low Support, High Turnout: This segment might vote but are not inclined to support the campaign. Persuasive messaging that aligns the campaign's key issues with the interests of these voters can be beneficial. They should be engaged with discussions about policy positions and candidate strengths that might sway their vote.  
  
4. Low Support, Low Turnout: Voters who are unlikely to support the campaign or vote may not be the primary target for campaign resources. However, if the campaign strategy includes outreach to all potential voters, then messages designed to inform and engage on a broad level, without heavy resource investment, can be used.  
  
5. Undecided or Unknown Preferences: For voters without clear predictions, surveys and additional data collection could be employed to better understand their stance and likelihood to vote. These voters might receive a general campaign message initially, followed by more targeted outreach as more information becomes available.

*Uplift Section Results:*

A positive uplift value suggests the treatment is likely to increase the desired outcome.

In our case, it means that a positive value indicates that the voter will vote democrat.

The larger the number the more effective the treatment is assumed to be for that individual. A negative uplift value suggests that the treatment might actually decrease the likelihood of the desired outcome and an uplift close to 0 implies that the treatment has little or no effect on the likelihood of the desired outcome.