DSC-550 Data Mining
Bellevue University
Final Project Write up
Author: Debasish Panda
Date: 11/18/2022
Topic: Congressional Resignation

Introduction:

Since March 4, 1901 — the first day of the 57th Congress — 615 members of Congress have resigned or been removed from office. The reasons members of Congress give for stepping aside can tell us a lot about the political era in which they occurred. The 115th Congress owes its historic turnover to the confluence of two events, one normal and one abnormal. First, there was the start of a new presidential administration. Five of the first six members to resign, did so to accept jobs in President Trump's administration. That's not unusual. It's similar to the seven members who resigned in 2009 to join the Obama administration2 and the five members who left in 1993 to join Bill Clinton's. But in addition, three of the four most recent members to resign from the 115th Congress did so because they were accused of unwanted sexual advances. However, a retirement from Congress at the end of one's regularly scheduled term is not the same as a mid-session resignation, which is what we're looking at here.

The extraordinary string of sexual misconduct allegations over the past years has led many people to conclude we are in the midst of an unprecedented cultural moment. In the political world, at least, the data bears that out. There has never been a concentration of sexual misconduct allegations that has caused as much public fallout before.

But in the big picture, sexual misbehavior, whether consensual or not, has not been a common reason for politicians to resign. Only 3 percent of congressional departures since 1901 have had to do with sex at all, according to media reports.

Now, for any electoral campaign management company to run a successful campaign, avoid the problem of an unsuccessful campaign, and to ensure successful office election, CampaignMe will gather and analyze prior candidate history, including resignation history in past offices held. CampaignMe (the campaign management company) is looking to take on candidates with the highest probability of winning elections, and devise strategies to ensure that they will get elected.

Hence, we are going to leverage the congressional resignation data set to find out resigning member's party and district, the date they resigned, the reason for their resignation and the source of the information about their resignation.

Justify why it is important/useful to solve this problem:

To ensure a successful candidate campaign and election into office at any US government level, an analysis of each CampaignMe candidate's profile needs to be completed. In this specific analysis, CampaignMe will look at past US Congressional resignations and associated political party. In some cases, resignation reasons are benign (e.g., being elected to a higher office), while others are more negative and illegal in nature (e.g., sexual harassment). A playbook for developing and executing a successful campaign will then be created and will be derived based off a created statistical model to help a potential candidate to get elected. This model analysis would be useful to Campaign Managers specifically so that they can have groundwork ready in the event of the Congressional seat opening up early.

Pitching the problem to a group of stakeholders:

In pitching this problem to Campaign Manager stakeholders to gain buy-in, the main focus is not repeating past mistakes, "Those that fail to learn from history are doomed to repeat it." as Winston Churchill said. Members of Congress have resigned for different reasons (good and bad) in the past, and to be prepared for future campaigns, we need to look at past data to understand how to best position our candidates to get elected. On average, a Senate campaign costs 15.7 million and House campaign costs 2 million, with Republicans spending the most on Senate campaigns (19 million). With this information in mind, Campaign Managers could allocate their time and contracts to parties with sufficient campaign budgets. This model could help identify which seats would resign early (Democrat or Republican) which would allow proper decision making for picking up a campaign.

Data source:

The dataset used is a Congressional Resignation dataset of 615 members of Congress who resigned or were removed from office from March 4, 1901 (the first day of the 57th Congress) through January 15, 2018, including the resigning member's party and district, the date they resigned, the reason for their resignation and the source of the information about their resignation

Link to Kaggle: https://www.kaggle.com/datasets/fivethirtyeight/fivethirtyeight-congress-resignations-dataset?select=congressional_resignations.csv

The dataset features are defined as follows:

- Member: The full name of the resigned Congress member o Party: The affiliated party of the resigned Congress member (D: Democrat, R: Republican).
- District: The affiliated congressional district of the resigned Congress member.
- Congress: The affiliated United States Congress of the resigned Congress member.
- Resignation Date: The resignation date of the resigned Congress member.
- Reason: The text resignation reason why the Congress member resigned.

- Source: The source in which the Congress member resignation was published and found in.
- Category: The categorical resignation reason why the Congress member resigned
    - X: Unwanted sexual contact
    - A: Consensual sex scandals
    - B Other scandals
    - C Other office
    - D Private sector
    - E Health/family
    - F Other
    - G Left early
    - H Military service
    - I Election overturned

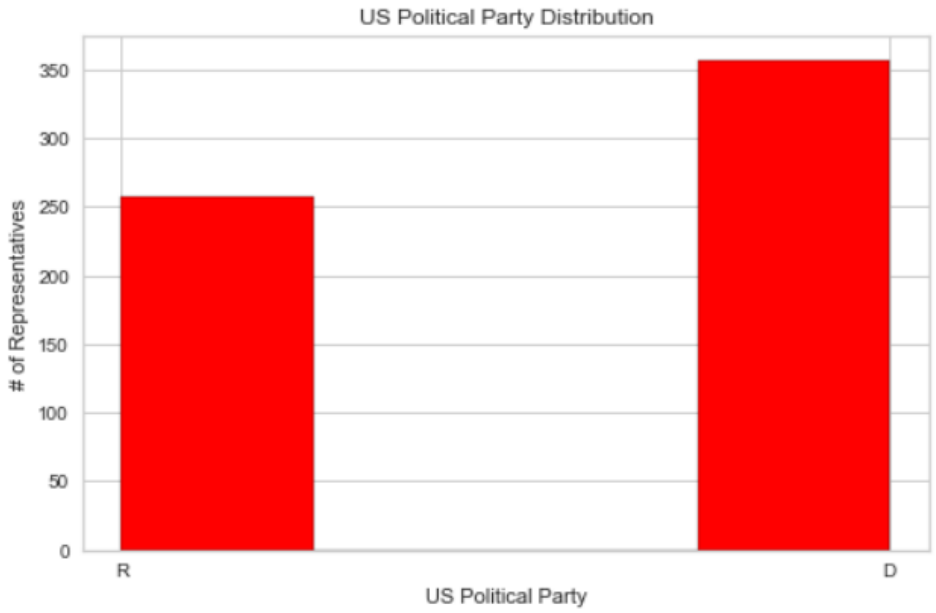
Summary of Milestones 1-3:

Exploratory Data Analysis (EDA)

- The dataset has 615 rows, 8 columns and there is no missing data.
- Being appointment to be a federal judge is the most common reason for resignation in the 'Reason' and 'Category' column (category C is "Appointed federal judge).
- The most common Party to be a part of while resigning is the Democratic Party.
- December 31, 1974 is the most common date to resign which is associated with the 93rd Congress (1973-1974).
- Some of the columns will not be useful for model building, e.g., 'Member', and 'Source'.
- All data is in object format, and Resignation Date will need to be converted to date/time format.
- The target of the model will be Party: Democrat is 0 and Republican is 1.

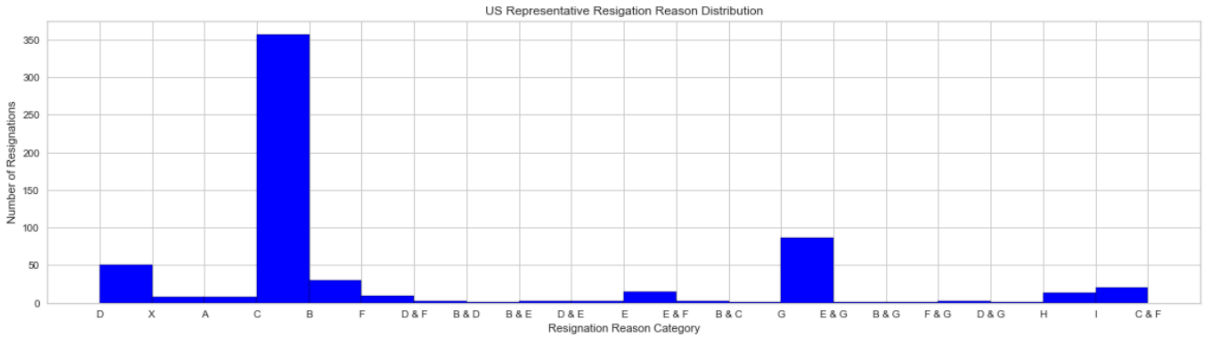
Visualization-1:

- Histogram of US Political Party Affiliation: Most of the senators in the Congressional resignation dataset are a part of the Democratic party (350 vs 250 Republicans).



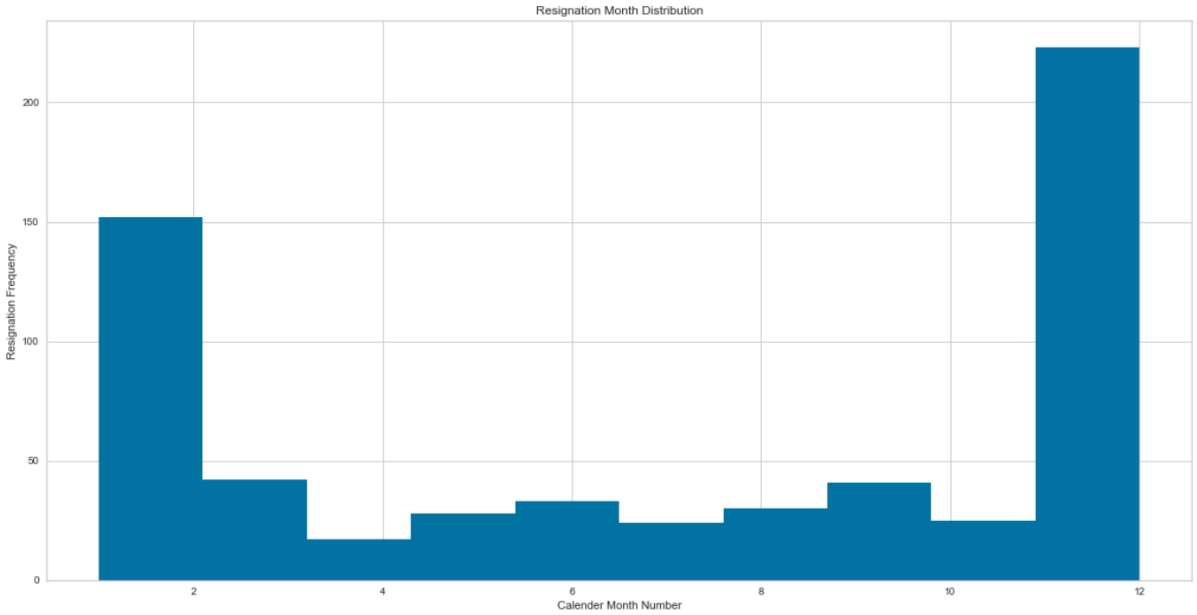
Visualization 2:

Bar graph of Resignation Reason Categories: The reason most likely to have resigned are as follows: 1. Appointed or Elected to a different role in Government (C), 2. Retired, lost re-election bid, appointment was expiring (G), 3. Took a role outside of government (D), 4. Ethics or corruption issues (B), 5. Their election was successfully contested and overturned (I).



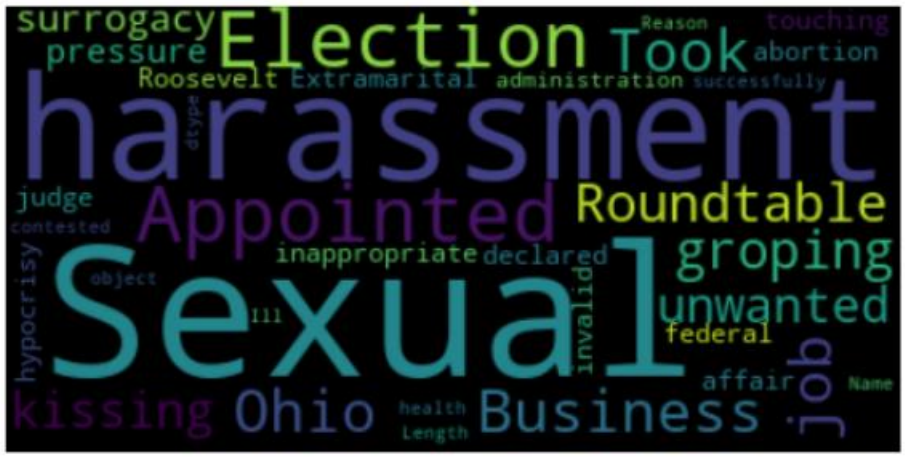
Visualization 3:

Resignation Data Histogram: Most of the Congressional resignations happen in either January or December.



Visualization-4:

Word Cloud: The bigger the words in the cloud, the more often they are seen in the dataframe column. So, in this case, Sexual, Appointed, Harassment and Election are the most common words found in the Reason column. Although this Word Cloud is very useful, this is out of alignment with the findings from Visualization #2 (Ethics Issues including sexual harassment being the 4th most common instead of most common), so in subsequent analyses we will look specifically at the Reason column as target variable to understand this better.



Data Preparation:

- Creation of New Features:
    - A new US "State" feature was added to the dataframe by splitting the 'District' feature at the "- ". This new feature ensures that the data can be analyzed at the higher state level, and at the lower district level.
    - Some of the "Category" values have two separated by an "&" (e.g., B & C). So, a new "Secondary Category" feature was created by splitting the "Category" feature values at the '&', and the "Category" feature was renamed "Primary Category".
    - The "Reason" feature contains resignation reasons in string text format. The PorterStemmer function was used to predict of the resignation reasons are positive or negative, and a new feature "Reason Score" was added to contain these values. 0 is a neutral/positive sentiment, and 1 is a negative sentiment of the "Reason" feature.

- Dropped Features:
    - "Member", "Resignation Date", "Source", and features related to the creation of the "Reason Score" feature are not useful in this model building because when creating a playbook for future elections, the Congress Member's name will not be important to analyze. Also, the source where the resignation is published will also not be important to create the playbook. Resignation Date is included in the Congress feature with a margin (2-year period). For this analysis, we are specifically looking for Party affiliations of the resigned Congress Member (e.g., State, District, Congress, Resignation Date) that could be applied to future candidates.

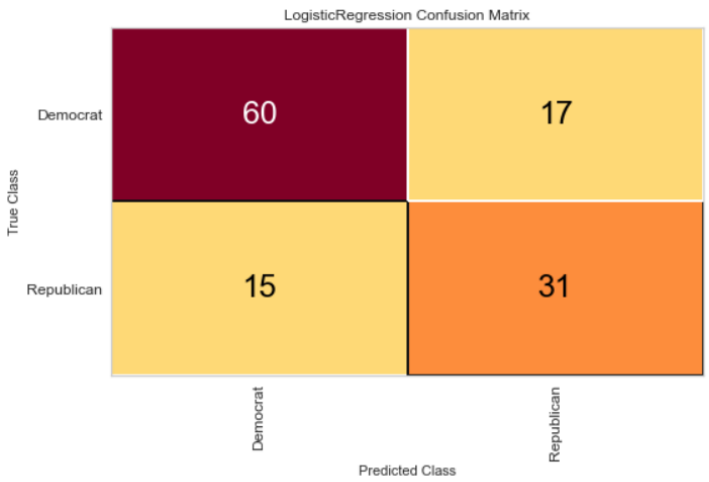
- Transformed Features:
    - As the "Party" feature is the target, the values "D" and "R" were converted into numbers (Democrat is 0 and Republican is 1) as the models can only run analysis on integers and not strings.

- Dummy Variables:
    - Dummy variables of the 'State', 'District', 'Congress', 'Category', and 'Secondary Category' features were created to enable the use of a single regression equation to represent multiple groups. Creating dummy variables allows us to convert the encoded categorical values of these features so they will have no impact on the mathematical models or the associated predictions.

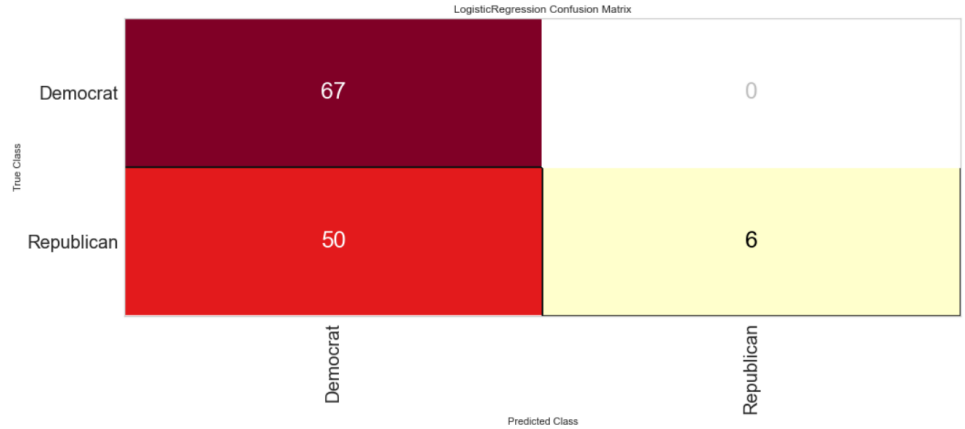
- Test and Train sets:
    - The data was split into a test (20%) and train (80%) set with the "Party" feature as the target, with the following features dropped for reasons in the above bullet point: 'Party', 'Reason', 'Reason_tokenized', 'Reason_stemmed', 'Reason_stemmed_sentence', 'polarity', 'subjectivity'. This test/train set will be used in the modeling with 171 features.
    - An additional test and train set was created using X2 to find the best 5 features in the dataset, and re-run the models with only these 5 features (as opposed to all 171 features).

Model building and evaluation

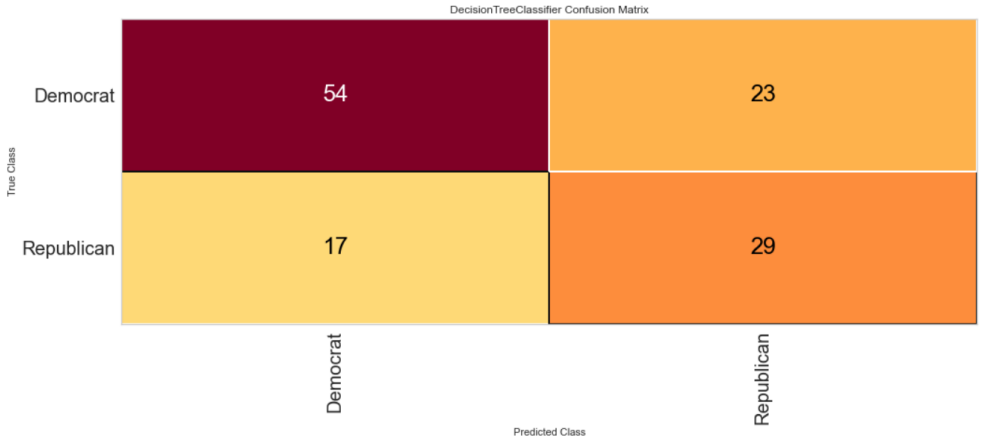
- Three types of models were created for this analysis: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier.
    - **Logistic Regression Model**: In a US Congressional Campaign, the most important thing to focus on is the campaign success. The goal is to have the campaign playbook ready and available prior to the campaign start, and with the logistic regression you have a way to predict this outcome (Party is the target).
    - **Decision Tree Classifier**: The idea of a Decision Tree is for one to see if decisions are made at each step, there will be another decision to made at the next step, until you finally work your way down the decision tree to an ultimate outcome (Party is the target).
    - **Random Forest Classifier:** "Random Forest" classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object." (Patel, 2017). In this study, it will predict based on the majority of votes from each decision tree created.
- Based on the best features ('State_IN', 'State_KS', 'State_NE', 'State_TX', 'Congress_77th') the Campaign Managers can have a better idea of what things can lead to resignations. These features provide play an important role in predicting Party resignations for the Campaign Managers, and have an overarching theme of being Midwest states (Kansas, Nebraska and Texas all sit along the 100W longitude line) with Indiana being close (90W), and the 77th Congress (world events - see below):
    - State_IN is Indiana
    - State_KS is Kansas
    - State_NE is Nebraska
    - State_TX is Texas
    - Congress_77th was from 1941-1942 (start of World War II, President FDR).
- Logistic Regression model (171 features) predicts Party (Democrat vs. Republican) with approximately 73.98% accuracy.

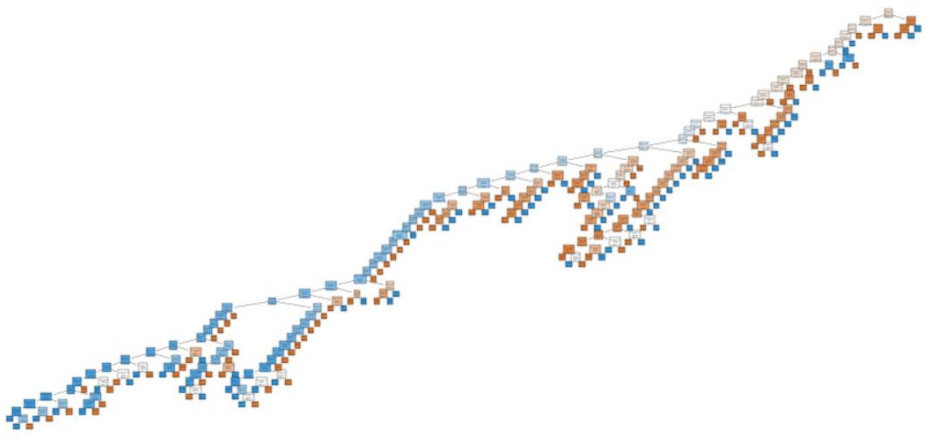


- Logistic Regression model with only the 5 best features predicts Party (Democrat vs. Republican) with approximately 59.35% accuracy (these 5 features are the 'State_IN', 'State_KS', 'State_NE', 'State_TX', 'Congress_77th').



- The Decision Tree Classifier (171 features) predicts Party (Democrat vs. Republican) with approximately 67.48% accuracy, with a visualization of the Decision Tree created.

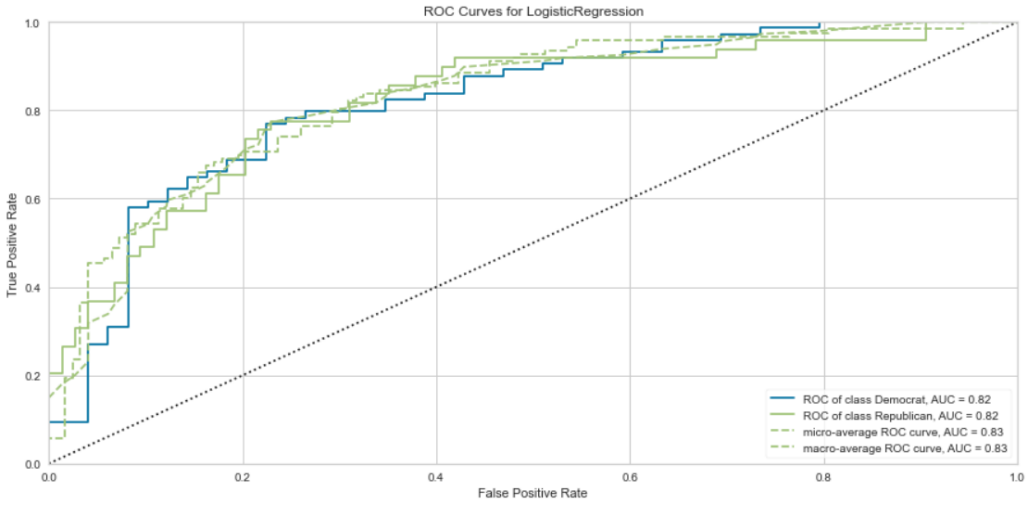




Conclusion:

What does the analysis/model building tell you?

- Classification report and ROC curve all suggest the model has a good predictive power (AUC of 0.82), "The area under the ROC curve (AUC) results were considered excellent for AUC values between 0.9-1, good for AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8, poor for AUC values between 0.6-0.7 and failed for AUC values between 0.5-0.6." (El Khouli, 2009).



- Between all three of the models (Logistic, Decision Tree and Random Forest), the Logistic Regression Model had the best accuracy with all 171 features (73.98%), while the Random Forest Classifier had the best accuracy with the 5 best features (63.41%).
- The best 5 features (Indiana, Kansas, Nebraska, Texas, 77th Congress) tell us what factors are important in predicting the political party (democrat or republican) of the seat resignation. This information will help develop the playbook as these 5 features will be looked at as a priority within the model to predict party affiliation. Depending on the US area, the campaign playbook can be provided to the campaign manager to start managing the campaign.

Is this model ready to be deployed?

- This model has several features from the dataset (Member, Party, District, Congress, Resignation Date, Reason, Source, Category), however it is not ready to be deployed yet. To ensure that the playbook is robust enough to support a successful political campaign, more data is required. Examples of different data sources to be added to the model include:
    - Voter Turnout by district.
    - District demographics (race, sex, income, education, etc.), included in US Census data, broken down by US Districts associated with the resigned Congress Person.

What are your recommendations?

- To build the most robust model, the dataframe should be augmented with voter turnout and district demographic data to ultimately support the CampaignMe playbook.

What are some of the potential challenges or additional opportunities that still need to be explored?

- This study aimed to develop a playbook for executing a successful campaign will then be created and will be derived based off a created statistical model to help a potential candidate to get elected. This model analysis would be useful to Campaign Managers specifically so that they can have the groundwork ready in the event of the Congressional seat opening early. As a subsequent opportunity, an additional playbook could be developed to support and maintain a current, elected Member of Congress' seat. For example, what policies in specific areas or events could maximize their probability of being re-elected.