Data 102 Final Project Report

Brian Lin, Chen Feng Tsai, Cindy Lin, Wallace Lim

I. Data Overview

This project primarily utilized election datasets from FiveThirtyEight and the Federal Election Commission.

a) FiveThirtyEight

The FiveThirtyEight dataset contains information regarding candidates' demographics, personal traits, political affiliations, endorsements, and primary results for the 2018 Democratic and Republican primary elections for US Senate, US House and State governor office positions. Data was collected for every single Democratic and Republican candidate who appeared on the ballot in 2018 and thus, can be considered a census data. There were certainly candidates who participated in the primaries but were systematically excluded from the data, particularly, these are the candidates that do not belong to either of the two major parties. Although it is unclear whether candidates were aware of the collection and use of this data, analyses utilizing this data have been published in various political news outlets. Thus, there is a high possibility that the candidates were aware that data would be collected about them and their campaign.

In this dataset, each row represents a Democratic or Republican candidate and each column represents a particular trait or endorsement. The endorsements data for each candidate were rather inconsistent and for each column, most of the entries were empty (no data). This inconsistency made it difficult for us to include endorsements as features for our analyses. As a result, for the purposes of our GLM and nonparametric model analysis we focused primarily on variables (characteristics) that not only contain more complete data, but also showed promising association with positive election results during the exploratory data analysis phase. In addition, we only considered the Democratic candidates, as the data for Republican candidates was not complete enough for any meaningful data analysis to be done.

b) Federal Election Commission (*2004-2020 Independent Expenditures Data)*

The Federal Election Commissions (FEC) collects and publishes data related to U.S. elections - in particular, FEC includes data regarding campaign finances, donations, and spending. FEC data is curated through financial reports submitted quarterly by candidate campaigns, political organizations and political committees. Any candidates, political party, political action groups or political committees that spend money to explicitly support or oppose a candidate must report any of those monetary contributions to the FEC. This data is a census data, as financial reports are required and submitted regularly from all candidates, political groups/committees.

For this project, we focused on the 2004-2020 independent expenditures data to better understand how expenditures in swing states affect election outcomes. We examined election outcomes for the 2004, 2008, 2012, 2016, and 2020 elections to determine the potential causal effect of swing state expenditure on election victory. The FEC data is self-reported and thus, the candidates, political parties and committees whose information are included in this dataset were aware of the collection and use of this data. However, some groups are exempt from FEC filings, these groups include nonprofit and other tax exempt organizations. This has potential implications for our analyses, as the expenditures data from FEC is not a complete representation of spending dedicated to supporting or opposing each candidate. Thus, this dataset poses a limitation to the conclusiveness of our analyses as depending solely on the FEC data to understand the causal effect between swing state expenditures and election victory overlooks critical funding/expenditures that were not included in the FEC data.

In the datasets for independent expenditures, each row represents an expenditure associated with a candidate and a state. In addition to information on the recipient candidate and their corresponding state, the columns in the expenditures dataset represent additional information about the expenditure such as the date, the type of payment, filing information, committee information. For the purposes of our analysis, we focused primarily on the 2020 Presidential Elections and the expenditures associated with the relevant swing states in that election.

In addition to the FEC expenditures data, we also used election data from the FEC to determine the swing states for each of the Presidential election years above. Since swing states differ for each election, it was important that we included different swing states for the different election years to best determine the effects of swing state spending and election victory.

II. Research Questions

For this project, we were particularly interested in understanding the characteristics that impact election outcomes.

a) What characteristics or political stances would most benefit a candidate in the primary election?

For this question, we were interested in understanding whether specific characteristics (i.e. candidate's demographic characteristics or political views) benefit a candidate in primary elections. In particular, we believe it would be fitting to fit both a logistic regression model (GLM) and an AdaBoost Classifier (nonparametric) and compare the performance of the two models in predicting election outcomes based on a specific set of characteristics. This method was suitable for answering our research question, because we were able to train models for predicting election outcomes and using those models, identifying the features that most significantly impact election outcomes. Understanding which characteristics benefit election outcomes the most could inform candidates which platforms or political views to invest in for positive election outcomes and potentially better inform election outcome predictions.

b) Does increased spending in swing states cause a positive impact in election outcomes (cause a win)?

Given the FEC data, we were especially interested in understanding how expenditures in swing states impacted election outcomes. The 2016 and 2020 elections both exemplified the significance of swing states in determining not only the outcome of Presidential elections, but also the majority party in the House and the Senate. To best understand whether spending extra on campaigning in swing states will **cause** a candidate to win the election, we determined that causal inference would be the best method to answer this particular research question. Understanding whether there is a causal relationship between swing state spending and winning elections can help candidates understand whether additional investments in swing state campaigns are worth it.

III. EDA

a) Data Cleaning

Prior to using the data, we need to clean it first. The main issue we saw with the individual expenditure dataset is that many of the reported expenditures had various naming schemas for each candidate. For example, Joseph R. Biden Jr. has "Biden", "Joseph", "Joseph Biden", and "Joe Biden" among the many names for 1 candidate. As a result, we cleaned the data by standardizing the naming schema and replacing the columns such that we can perform EDA on it. We performed this action on all the candidates. Also several rows containing NaN values prevented us performing useful analysis on it, so we decided to drop those rows and exclude them in all our future calculations.

To utilize the FiveThirtyEight dataset, since we discover that supporting gun safety will affect the result of democratic candidates, we drop the NA with no Gun Sense Candidate information. Due to the lack of information with Gun support materials for Republican data, we only adopt democratic dataset for modeling. To analyze whether specific characteristics affect how many votes they receive when they win, we categorize the original democratic dataset with their election results and visualize them for further exploration.

b) GLMs and Nonparametric Methods

We wanted to explore some traits that could have a beneficial impact on predicting the outcomes of primary elections based on their beliefs as well as their prior experience. In **Figure 1**, we observe that the candidate ratio between the two categories (winning/losing) are different. For the candidates who

won the election, we can see that the proportion of previously elected candidates is greater than that of the losing category. Thus, we assume that a candidate is more likely to win the election if they have been elected before compared to if they had not been elected before.

In **Figure 2**, we observe that the candidate ratio between the two categories (winning/losing) are not the same. For the candidates who won the election, the proportion of candidates that have a STEM background is relatively small compared to that of the losing category. Therefore, we assume that a candidate is more likely to lose the election if they have a STEM background compared to those who do not have a STEM background. In **Figure 3**, the proportion of candidates who support Gun sense in the winning category is greater than that of the losing category, so we assume that a candidate that supports Gun Sense is more likely to win. We decided to proceed further with these traits as they seem that they can help us distinguish and predict future primary candidate winners.

c) Causal Inference

First, we were curious how much is spent during the 2020 election in its entirety prior to evaluating the spending on swing states. From **Figure 4**, we noted the rate of spending increase dramatically as the main Nov. election approaches. Seeing this led to exploring how much Trump and Biden being the two primary candidates spent in their election process as well as in the popular swing states depicted in **Figure 5, 6** respectively. It's important to note the units for each **Figure 4 - 6** as an order of magnitude decreases each time indicating the majority of the money spent was not on swing states. We see that both candidates placed a large emphasis on trying to win Florida probably because of its high number of electoral college votes. Biden also placed a good amount in securing Rust Belt region (Ohio, Minnesota area) which seemed to benefit him in winning the state.

From our EDA, we have differing perspectives on whether spending in primary states benefits the outcome. On a grand scale, the order of magnitude of spending is dramatically less than the overall expenditure which can lead us to believe that there isn't that much of an effect; however, when comparing who spends more money, that has shown some promising results on the victor. Nonetheless, this would process to analyze whether there exists a causal relationship.

IV. GLMs and Nonparametric Methods

a) Methods

i) GLM Model - Logistic Regression

$$p = P(Y = Win)$$

$$1 - p = P(Y = Lose)$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 * Gun + \beta_2 * Stem + \beta_3 * Elected + \beta_4 * Self\_Fund + \epsilon$$

We choose logistic regression as our GLM model (**Figure 7, 8**) as it can be used to predict binary outcomes, which is winning or losing an election in this scenario. The assumptions we made are that the sample size is large enough to fit the model and that the observations are all independent of each other.

ii) Nonparametric Model - AdaBoost Classifier

We explored different nonparametric models such as decision tree, random forest, gradient boosting, Adaboost, and GaussianNB. We then compare each of the results to determine the model with the best performance (**Figure 9**). The assumption we made here is that the observations are all independent of each other.

After comparing each model, we determined that AdaBoost had the highest accuracy score and F1 score, so we selected it as our nonparametric model (**Figure 10**).

b) Results

The logistic regression model has a 72.1% accuracy score and 66.7% F-score, with two statistically significant variables in 0.1 level, which are 'Gun Senses Candidate?' and 'STEM?'. We obtain a positive coefficient for the gun sense support variable and negative coefficient for the STEM variable.

The Adaboost model has a 72.3% accuracy score and 67.3% F-score, which is almost the same as the logistic regression model.

To estimate the uncertainty of our logistic regression model, we considered the confidence interval of the coefficient of 'Gun Senses Candidate?' and 'STEM?', which are 1.74 and -0.65. This means that there is a 174% increase in the odds of winning the election if the candidate supports gun safety and 65% decreases in odds of winning the election if the candidate is a STEM major. For the interval of the odds ratio of 'Gun Senses Candidate?', there is a 95% chance between 2.6 to 12.24. For the interval of the odds ratio of 'STEM?', there is a 95% chance between 0.23 and 1.17.

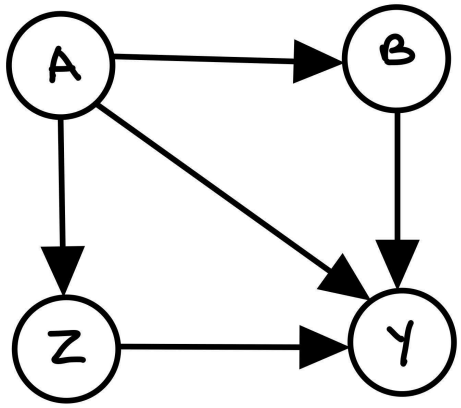
c) Discussions

The Adaboost model performed slightly better since it adjusted the weight of incorrect classifications the second time it fits the model. We are 70% confident in applying this model to the future datasets.

However, there are limitations of the models. For example, it does not perform extremely well and might need more variables to improve the models. Additionally, the Adaboost model might perform well if its guesses are always “not elected”, since the number of candidates *not elected* is far greater than those who are elected in most cases. The models are subject to error as there are only four variables to train for now. If the backgrounds of the voters or the location of the election were different, the predictions might be incorrect. Additional data such as the candidate's education level, number of attendants at their rallies, and age might be useful for improving my models.

V. Causal Inference

a) Methods



To answer our second research question, we chose to use inverse propensity weighting to estimate the effect of increased spending in swing states on the election outcome while adjusting for confounders. Our treatment variable (Z) was which party (democrat or republican) spent more during their campaign on the swing state (1 for democrat, 0 for republican), and our outcome variable (Y) was which party ended up winning that state (1 for democrat, 0 for republican). There were two confounding variables in our model. First, the difference in dem vs rep spending in the previous election (A) is a confounding variable since it affects the treatment as candidates tend to adjust their campaign spending based on the results seen from previous years, and also affects the outcome as parties gain long term favor of states when shown to provide continual financial support. Second, the election result of the previous cycle (B) is also a confounding variable, since it affects the treatment as candidates may put more money into states that were on the border the previous cycle or decide a state is not in their party's favor and put less

money into campaigning there depending on their strategy, and also affects the outcome as previous election results will psychologically influence the voters' mindset regarding the election. For example, if a state won by a landslide the previous year, voters often feel like the result is already cemented and choose not to vote, while if the previous year's result was close, voters in the state will feel like every vote counts and be more likely to place a vote. We can assume unconfoundedness which means given our confounders (A, B), the potential outcome of winning the swing state is independent of who spent more money on the state.

b) Results

Since we use the inverse propensity weighting method, we have to make a few assumptions. (1) Assume unconfoundedness. (2) A linear model would correctly describe this relationship. If these assumptions hold, we can conclude that for the party that spends the most money in a given swing state, that party has a 28.5% higher chance of winning the state.

There is always a chance that our estimate doesn't match the true percentage effect and it is related to the assumptions we made. The two confounding variables we stated might not be the only confounding variables present in the dataset, more can exist that we didn't consider. Also, we applied a logistic regression to predict the propensity function using a linear relationship of the confounding variables. This means we are assuming a linear relationship for the confounding variables, but the relationship could potentially be more complex than that.

Although we stated a 28.5% chance, it's not clear whether money is always the right answer. Other factors such as how the money is being spent, the migration of the population, and political tendencies of the population can also be additional factors that would play a role. An example is that Georgia, a swing state for the 2020 election, had Trump spending more money but Biden ended up marginally winning the state. Mike Bloomberg, a democratic presidential candidate, had a bountiful amount of money to spend on his campaign and didn't end up winning. These results should be cautiously taken into consideration due to the multitude of extraneous factors that could play an outside role on the outcome.

c) Discussions

Our method definitely has some limitations. One main limitation is the fullness of our dataset. The FEC doesn't require all organizations to report their expenditures, nonprofits for example, so our dataset is incomplete. Throughout EDA, we noticed that the overall spending was two orders of magnitude higher than the spending in a swing state. This might be because multiple state spending wasn't taken into consideration, for example, a political advertisement might have been used on Youtube which advertised it to all states but only reported as California due to the location of Google's headquarters. Our model also didn't have many entries which could have played a factor in our IPW model as well. Taking a step back, as a group, we were lacking more domain knowledge and expertise to help accurately consider all the confounding variables.

This idea can be expanded upon using additional data. As mentioned above, a key limitation of our data is the lack of nonprofit organizations reporting their support for a candidate. Having more information that encompasses the total spending would greatly benefit the confidence in our methods. Additionally, more specific details on how the money was spent potentially categorized into groups would also allow a more fine detail analysis on specific spending benefits.

We are not that confident that there is a causal relationship between spending and the outcome of a swing state. This is because there are several external factors that could impact the role of the election irregardless of money. For example, currently many democrats are populating large cities in Texas slowly increasing the state blue which definitely would play a role in future elections. After our analysis, we recognize that money isn't the sole factor and many external factors we consider, we failed to find a sufficient dataset to account for it.

VI. Conclusions

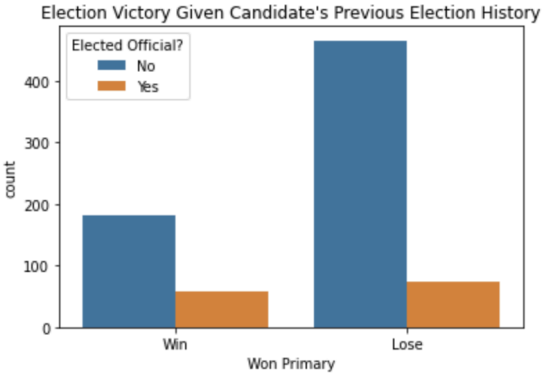
To answer our first research question, we utilized GLMs and Nonparametric methods to understand the relationship between particular characteristics of candidates and their election outcome. From the results, we observe a positive coefficient for the gun sense support variable and negative coefficient for the STEM variable in our logistic regression model, indicating that there is a correlation between Democratic candidates who support gun control (have the support of gun sense) and whose background is not in STEM and winning the Democratic primary elections. Our results are not the most generalizable, as the FiveThirtyEight dataset focuses primarily on the 2018 primaries and voter behavior can vary considerably in different voting years depending on the key issues of concerns in each of those years. However, our results are interesting for better understanding the behavior of voters in 2018 and the impact particular political views can have on election outcomes. Based on our results, we observe that those who vote for Democratic candidates are particularly interested in addressing gun safety and potentially lack interest in candidates with a background in STEM. Candidates interested in winning the Democratic primaries could potentially invest more effort in promoting and improving gun control policies and could perhaps consider not focusing on highlighting their background in STEM, if they have any. One limitation of this dataset is that there is a lot of missing information that makes the data particularly difficult for us to use. For example, data for Republican candidates was largely missing so we weren't able to investigate the characteristics that impact Republican primary election results. Even for our analysis on the Democratic candidates, we were unable to look into more variables and characteristics because many of our columns (characteristics) simply did not have a data entry.

For our second research question, we conclude from our causal inference that if a Democratic candidate spends more than a Republican candidate on swing state campaigning (Treatment = 1), then there is roughly a 28% greater chance that they will win the electoral votes in that particular swing state. Our results considered elections from the five most recent Presidential elections (2004 - 2020) and accounted for confounders such as spending in previous elections and the results from previous elections. Based on the conclusions above, Democratic Presidential candidates can consider investing in campaigning in swing states to improve their chances of victory in each swing state as well as in the overall election. However, our results are not generalizable because there are a lot of limitations from the dataset and other potential confounders that we were not able to account for. The FEC data excludes expenditures from organizations such as nonprofits, so our analysis fails to account for additional expenses in support or opposition of each campaign. In addition, the P-scores we observe were mostly around 0.4 - 0.6, suggesting that our model does not perform well predicting the treatment variable. This could be because the variables and confounders we considered are not well differentiated and fail to fit our data with the model. We also did not have a lot of entries in our dataset to train our model, which may be another reason for poor fit. Finally, due to limitations in domain knowledge and breadth of our data, we may have overlooked confounders that can have significant impacts on the result.

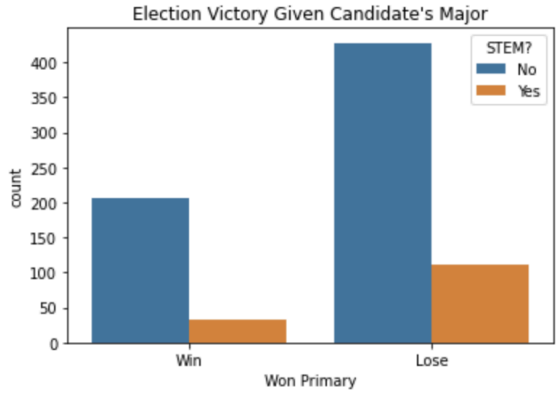
Building on our work, future studies could further investigate the granularities in expenditures and examine what types of campaign spending will cause or be associated with positive election outcomes. For example, future studies could analyze whether additional spending on broadcasted campaign ads or billboard ads is more beneficial to election victory. This could better inform election spending and encourage more efficient use of money. Additionally, building on top of our models for predicting election outcomes with expenditures, future studies could develop models for more accurate election forecasting.

VII. Figures

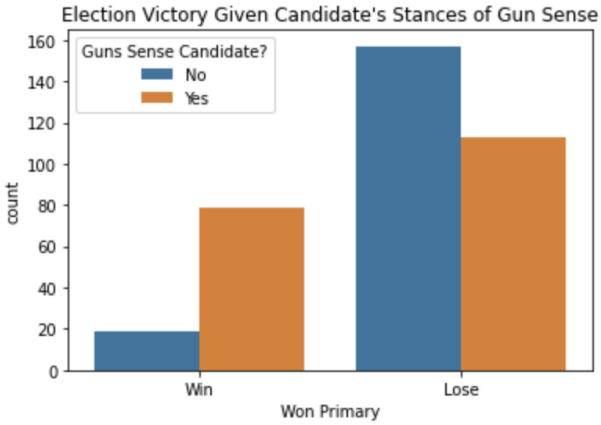
1. **Figure 1:** Relationship between whether a candidate wins the current election and whether they have been elected before



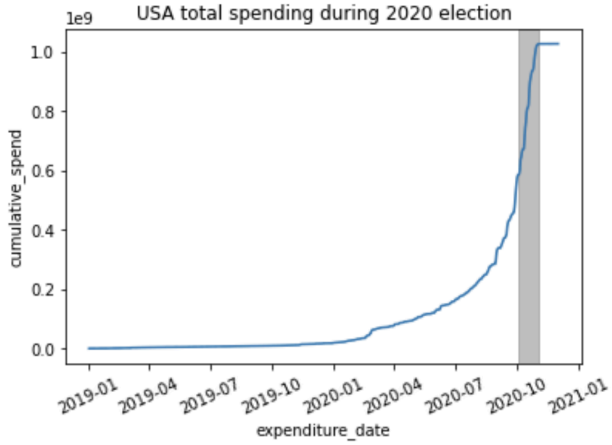
2. **Figure 2:** Relationship between whether the candidate win the election this time and whether he is STEM



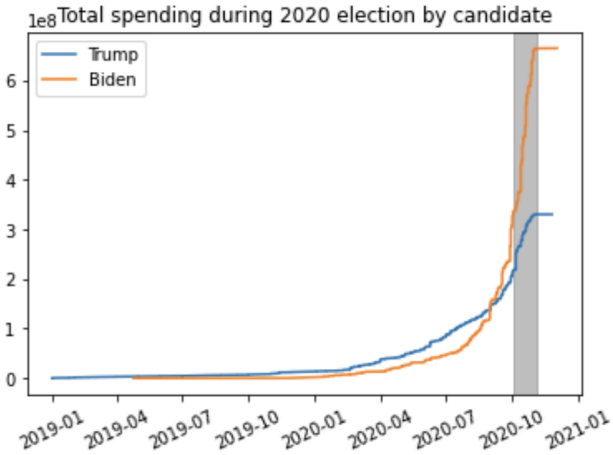
3. **Figure 3**: Relationship between whether the candidate win the election this time and whether he supports Gun Sense



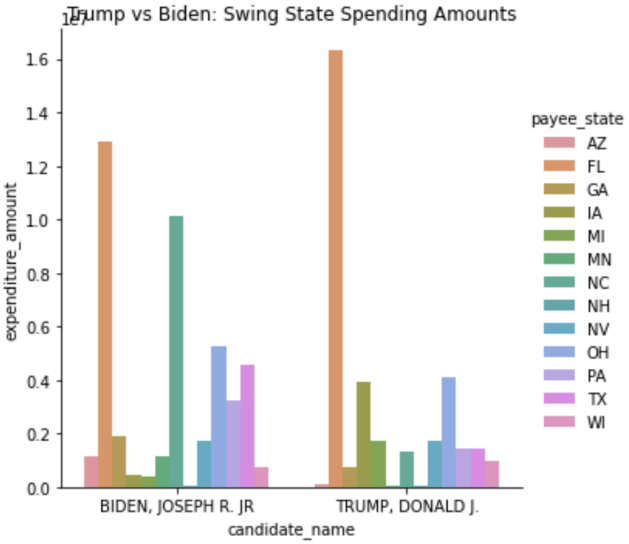
4. **Figure 4:** Total spending overall in the election by all candidates



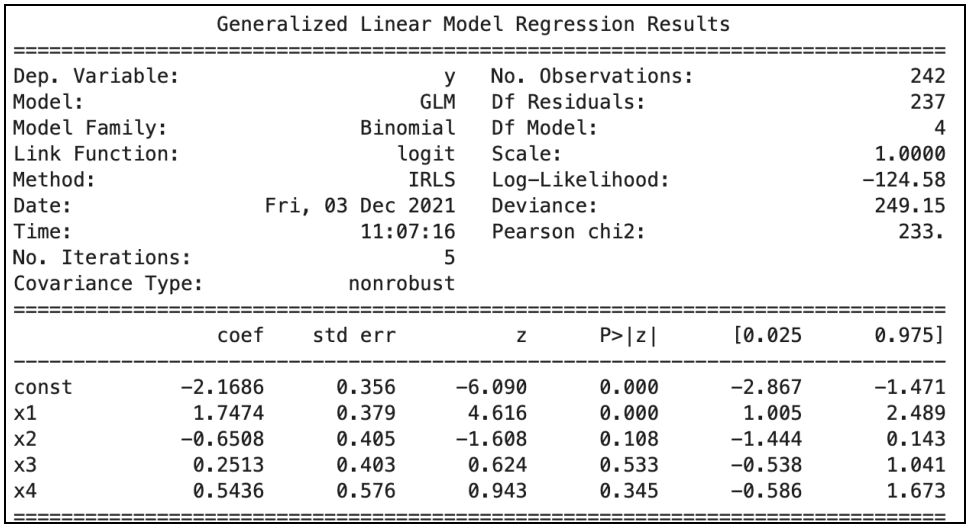
5. **Figure 5:** Total spending from 2019 - 2020 by Trump and Biden



6. **Figure 6:** Trump and Biden Swing State Spending Comparison



7. **Figure 7:** Generalized Linear Model Regression results



Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	242
Model:	GLM	Df Residuals:	237
Model Family:	Binomial	Df Model:	4
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-124.58
Date:	Fri, 03 Dec 2021	Deviance:	249.15
Time:	11:07:16	Pearson chi2:	233.
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1686	0.356	-6.090	0.000	-2.867	-1.471
x1	1.7474	0.379	4.616	0.000	1.005	2.489
x2	-0.6508	0.405	-1.608	0.108	-1.444	0.143
x3	0.2513	0.403	0.624	0.533	-0.538	1.041
x4	0.5436	0.576	0.943	0.345	-0.586	1.673

8. **Figure 8:** Classification report for GLM

Classification report :

	precision	recall	f1-score	support
0	0.72	0.98	0.83	86
1	0.33	0.03	0.05	34
accuracy			0.71	120
macro avg	0.53	0.50	0.44	120
weighted avg	0.61	0.71	0.61	120

9. **Figure 9:** Comparison of Performance for Different Nonparametric Methods

Model with CV=8	Mean accuracy score	Mean F-score
Logistic Regression	0.721	0.667
Decision Tree	0.715	0.614
Random Forest	0.698	0.629
Gradient Boosting	0.701	0.664
Adaboost	0.723	0.673
GuassianNB	0.715	0.614

10. **Figure 10:** Classification report for AdaBoost

```
Classification report :
              precision    recall  f1-score   support

           0       0.72      0.98      0.83        86
           1       0.33      0.03      0.05        34

    accuracy                           0.71       120
   macro avg       0.53      0.50      0.44       120
weighted avg       0.61      0.71      0.61       120
```