# 

# **Applied Practical Data Science**

# **Report**

**Team Members:**

**Apoorva Palle Z248H929 (Code) ||**

**Areeba Mohiuddin D655Y483 (Summary) ||**

**Lakshmi Vasanthi Godi V567R326 (Report) ||**

**Sri Manikanta Reddy Tangirala G332B644 (PPT)**

**1. Executive Summary**

**a) Introductory paragraph that summarized your introduction section**

In Western democracies, the convergence of personality traits and voting behavior has emerged as a pivotal area of study, particularly in the era of personalized politics [11]. This systematic examination analyzes 288 studies, culminating in 12 relevant works (Mondak & Halperin, 2008) that unveil four pivotal [12]. These findings illuminate the profound influence of both voters' inherent traits and their perceptions of candidates' characteristics on electoral outcomes [13]. By elucidating the complex interplay between personality traits and political preferences, this analysis offers invaluable insights into the customization of politics within democratic frameworks [11]. The accuracy is based on previous election criteria as well [12].

**b) Results – what you found in one paragraph**

Analysis pinpointed influential predictors of voting behavior in the 2016 presidential election, such as political affiliation, past voting trends, and socio-economic indicators. These factors played a crucial role in shaping the election outcome, highlighting the intricate nature of voting behavior. Consideration of demographic, economic, and personal belief factors is imperative for accurately predicting election results.

**c) Why this is important – how it can be used, future research implications**

Understanding the significant predictors of voting behavior, such as political affiliation and socio-economic factors, is vital for accurately interpreting election outcomes. This knowledge can inform political campaigns, policy-making, and strategic decision-making for businesses and organizations, helping them tailor their approaches to better resonate with voters. Moreover, by delving deeper into emerging trends in voting behavior and exploring the impact of evolving media landscapes, researchers can provide actionable insights for improving voter engagement and participation in democratic processes.

**2. a. Why is this important? (1 paragraph)**

Understanding voting behavior is crucial as it determines leadership in cities and nations, shaping the future. Voters seek leaders who enhance their and their children's prospects. Through voting, people express opinions and ensure their voices count. Non-voters forfeit influence over their lives and society. It's essential to grasp voter motivations for effective governance [7]. Voting behavior dictates election results and government formation [5].

**b. Background research: what’s happening now, what are the big influencers, and what have others done (~3 paragraphs) MUST INCLUDE CITATIONS**

Research into voting behavior is a multifaceted exploration of socio-cultural, political, and economic influences that shape individual decisions at the polls. Utilizing diverse data sources such as demographic profiles, political surveys, historical election data, and economic indicators [9], scholars aim to comprehend the complexities of voter behavior. This study considers a range of independent variables (IVs) including education, income, gender, age, ethnicity, race, and religion [9][11]. For instance, higher levels of education have been correlated with increased political awareness and participation [12], while income disparities can influence access to political resources and engagement [13]. By investigating these relationships, researchers seek to elucidate how socio-economic factors influence voting behavior in diverse populations and contexts [3].

In navigating the realm of voting behavior, businesses and organizations encounter significant challenges in understanding voter preferences and addressing barriers to voter participation [2][6]. Demographic targeting strategies can assist businesses in identifying segments likely to support specific parties or causes, but they must also ensure equitable access to voting options for all eligible voters [6]. Scholars have delved into various aspects of voting behavior, exploring the impact of registration strategies on youth voter participation [1][3][4], the role of personality traits in candidate perception and the personalization of politics [2][5][6], and the influence of factors such as polarization and electability in primary elections [7][8]. These studies contribute to a nuanced understanding of voting behavior and its implications for democratic processes, offering insights into the intricate mechanisms that drive electoral outcomes.

In the study of voting behavior, scholars employ diverse methodologies and data sources to unravel the complex interplay of socio-economic, cultural, and political influences on electoral decisions. By analyzing demographic profiles, historical data, and economic indicators, researchers can gain deeper insights into the drivers of voter behavior. Key independent variables (IVs) such as education, income, gender, age, ethnicity, race, and religion are scrutinized to discern their impact on political engagement and participation. Through this comprehensive approach, researchers aim to illuminate the underlying factors that shape voting behavior across diverse populations and electoral contexts.

**c. What are you doing?**

Using Support Vector Machine (SVM) to identify voting patterns based on a few factors that influence voting behavior. Source of the dataset is taken from The General Social Survey (GSS).

**3. Research Questions and Hypothesis**

**Research Question:**

What factors significantly influence voting behavior in democratic processes? How do businesses, organizations, and industries address challenges related to understanding and responding to voting behavior?[1][2][5]

**Hypothesis:**

Hypothesis posits previous election outcomes (PRES12), significantly predict voting behavior and can be used to accurately forecast the 2016 presidential election outcome (PRES16).

**4. Methodology**

**a. Data Used: where did you get it, what format, how many observations (1 paragraph+)**

Source of the dataset is taken from **The General Social Survey (GSS)** for the **2022**. The reference link for the dataset is <https://gss.norc.org/get-the-data> . The format of the data having is **.CSV**. And the size of the dataset dealing with is having **2349** rows which include the details of voting behavior.

**b. Cleaning: What was wrong with the data? How did you clean the data and calculate DV? Reference visuals (1 paragraph +)**

Cleaning of dataset involves following steps –

1. **Handling Missing Values:** Identify and visualize missing values in the dataset using a heatmap. This can be done by using “**data.isna().sum()”** to count the number of missing values in each column.
2. **Removing Duplicates:** Identify and remove duplicate rows from the dataset. This can be done by using **“data.duplicated().sum()”** to count total number of duplicate values in data.
3. **Handling Outliers:** Visualize outliers using box plots and filter out extreme values based on the Interquartile Range (IQR) method. This can be done by **“sns.boxplot”** to identify outliers in numerical columns.

Further, have chosen the Dependent variable (DV) based on the aspects of the voting behavior. Have chosen the DV as **PRES16,** based on what factors the voters are voting which will have the list of voters who are involved in elections and based on the value of the DV people choose to vote for.

**c. Variables: what fields did you test and why? How did you transform these variables? Reference visuals (1 paragraph +)**

Fields tested in independent variables are: **PRES12, PARTYID, OTHER, JEW, AGE, ROWGUN, RELIG, OTH16, CLASS, SEI10EDUC**.

Fields tested in dependent variables is: **PRES16**

Have chosen these variables to transform based on coefficient values with the dependent variable and based on background research. Below visuals are some of the univariate graphs.

The below graph demonstrates the count of the voting for particular candidates who are involved in the voting. PRES16 is the column in dataset which describes the candidate details. As from the manual observation the count for IAP is more which means people who are In- Applivable for voting range is more- category of under or over aged people and visuals show Clintion is on top.

A graph of red bars

Description automatically generated with medium confidence

A graph of different religious groups

Description automatically generated with medium confidence

A graph of a bar graph

Description automatically generated

The above graph demonstrates the count of people who have voted in the year 2012 for the candidates. Clearly from the visual that the count of IAP is more and highest voted candidate is Obama as per 2012 voting election results. This would be probably have more impact and rely on current voting.

A graph of age and histogram

Description automatically generated

The above graph demonstrates the frequency of the voters in elections based on the age. The histogram visual clearly says that age is main factor for the elibity of voting and the number of people age who have voted in election is ranging from 20-90. Tha maximum people age who have voted is around 30-40.

**d) Software used and analytics performed: why did you choose that software and algorithms?**

Used Python for algorithm. Based on the analysis of calculating the better accuracy, when compared between the models like Decision Tree, SVM, Logistic Regression, Random Forest which use categorical data, algorithm Random Forest has given better accuracy. So, using the algorithm **“Support Vector Machine Model”** is suitable for modeling categorical outcomes. It is used for classification and regression tasks. SVM is widely used due to its ability to handle complex datasets and its flexibility in choosing different kernel functions to capture intricate patterns in the data, making it suitable for a variety of real-world applications. Calculated the R-squared score and accuracy for the Decision Tree Regression and obtained the results as R-squared score for Decision Tree is -0.2793 and the accuracy is 61%. Calculated the R-squared score and accuracy for the SVM model and obtained the results as R-squared score for SVM is 0.264 and the accuracy is 72%. By calculating the accuracy and R- squared score clearly SVM is the best fit.

**5. Results**

**a. Was your hypothesis supported? Research question answered? Reference visuals?**

Yes, Hypothesis is supported. By analyzing the coefficients and statistical significance of predictor variables in the model, identify which factors have the most significant influence on voting behavior. For instance, if the coefficient for PRES12 is highly significant and positive, it suggests that party votes are a strong predictor of voting behavior, holding other factors constant. As per the visuals PRES12 is highly dependent for 2016 election votes. Research questions are also answered.

A graph with different colored bars

Description automatically generated

**b. What was significant? Reference visuals.**

Support Vector Machine Model is more significant because when accuracy performance is tested, SVM gave an accuracy of 72% while Decision tree gave an accuracy of 61%.

|  |  |  |
| --- | --- | --- |
| **Model** | **Visual** | **Result** |
| SVM | Accuracy\_score(y\_test, S\_pred) | 72% |
| Decision tree | Accuracy\_score(y\_test,D\_pred) | 61% |

**c. Interpretation: Why are these variables significant? Reference visuals and prior research**

Based on Background research and calculation of correlation values have chosen the variables below as significant.

|  |  |
| --- | --- |
| **Variable** | **Correlation Value** |
| PRES12 | 0.555662 |
| PARTYID | 0.239217 |
| OTHER | 0.185116 |
| Jew | 0.166840 |
| AGE | 0.162831 |
| ROWNGUN | 0.157964 |
| RELIG | 0.152613 |
| OTH16 | 0.150806 |
| CLASS | 0.141861 |
| SEI10EDUC | 0.140594 |

The visual shows the influence of the independent variables on the dependent variables in decreasing order. PRES12 is the most influential and DIPGED is the least influential. The correlation between PRES12 and PRES16 is more compared to any other independent variable with PRES16. So that variable will be the most influential variable.

**6.Discussion**

**a. What were the limitations of the project? (data, resources, access, time, knowledge, etc.)**

**Limitations:**

**1.Data Limitations**

* Data quality and Completeness – The dataset may have certain limitations in terms of data quality, such as missing values/ null values. It has more inapplicable values.
* It has limited variables related, which made us challenging to explore all aspects of voting behavior.

**2.Resource Constraints:**

* Impact of depth and complexity of analysis.
* Additional datasets or extensive data cleaning and preprocessing.

**3. Access:**

* Conducting in-depth analyses of voting behavior requires expertise in political science, sociology, or related fields, which may not have been fully available within the project team.
* Access to additional Data Sources could have enhanced the analysis but might have been challenging due to access constraints.

**4. Time Constraints:**

* They may have limited depth of analysis or the exploration of more nuanced aspects of voting patterns.
* Time limitations may have impacted the ability to iterate on the analysis.

**b. What are the recommendations for future research?**

**Recommendations:**

To enhance the efficiency and effectiveness of the project focused on voting behavior classification using machine learning models, especially by considering a broader range of machine learning techniques.

**1. Exploration of Alternative ML Models:** Evaluate and select the most suitable machine learning models based on performance metrics such as accuracy, precision, recall, computational efficiency, and interpretability.

**2. Fine – Tuning:** Fine – tune preexisting models on domain specific data to adapt them for voting behavior classification tasks.

**3. Feature Engineering and Domain Specific Knowledge:** Incorporate domain knowledge and expertise the insights to enhance the new features and capture nuanced aspects of voting behavior such as political affiliations and historical voting patterns. Integrate temporal features into modeling process to capture dynamic changes in voting behavior.

**4. Model Deployment and Real Time Prediction:** Implement mechanisms for continuous model monitoring, performance tracking, and periodic updates to adapt to evolving voter preferences and trends.

By adopting these recommendations, future research endeavors can leverage a broader repertoire of machine learning techniques, foster interdisciplinary collaboration, and enhance the applicability and impact of predictive models for understanding and predicting voting behavior.

**c. What are the use cases and implications for this research? How could people/companies/organizations use this? Why are the findings helpful and important?**

**Use Cases and Implications:**

**1. Political Campaigns and Strategists:**

Political campaigns can benefit greatly from leveraging predictive models to understand voter behavior trends and optimize campaign strategies. By analyzing historical voting patterns and demographic data, campaigns can identify key demographic groups that are more likely to support certain candidates or parties. For example, predictive modeling techniques can be used to analyze voter behavior in past presidential elections, such as those in 2012 and 2016, to identify trends and patterns [1][2]. Campaign strategists can then tailor their messaging, advertising, and outreach efforts to resonate with these specific voter groups, maximizing their impact and efficiency.

**2. Election Administration and Governance:**

Election officials and administrators can leverage predictive modeling to enhance election administration and governance practices. One valuable application is voter turnout prediction, where predictive models are used to forecast voter turnout in upcoming elections [3][4]. By accurately predicting voter turnout, election officials can allocate resources more effectively, such as deploying voter registration drives or establishing polling stations in areas with higher predicted turnout rates. This proactive approach helps optimize resource allocation and enhances overall election management. These applications of predictive modeling in political campaigns and election administration highlight the potential for data-driven insights to improve decision-making processes and optimize resource allocation in the electoral context. By leveraging historical data and advanced analytical techniques, political stakeholders can make informed strategic decisions that enhance campaign effectiveness and contribute to more efficient and inclusive electoral processes.

**Importance and Significance of Findings:**

**1. Data-Driven Decision Making:** Your findings enable data-driven decision-making in various domains, from politics to business, by providing actionable insights into voter behavior and preferences.

**2. Resource Optimization:** By understanding the factors influencing voting behavior, organizations can optimize resource allocation, focusing efforts and resources on segments of the population that are more likely to be receptive to specific messages or campaigns.

**3. Enhanced Civic Engagement:** Knowledge of voter behavior promotes civic engagement and informed participation in democratic processes by shedding light on the underlying drivers of electoral outcomes.

**4. Transparency and Accountability:** The use of empirical models like Random Forests enhances transparency and accountability in electoral analysis, offering a systematic approach to understanding complex voter dynamics.

Overall, research has practical applications for political campaigns, governance, market research, and academic inquiry, highlighting the value of data-driven methodologies in uncovering insights that inform decision-making and contribute to a deeper understanding of voting behavior patterns over time.

**8. Visuals**

**i. Descriptive Statistics (Outliers, distributions, etc.)**

For performing descriptive statistics, Identified Outliers, Assess Variable Distributions, Summary Statistics, Visualization and Interpretation and Reporting.

**Example Visuals:**

A graph of a distribution of age

Description automatically generated

A graph with a box plot of income

Description automatically generated

**ii. Bi-variate statistics (relationship between DV and key IVs)**

To analyze relationships between (DV) PRES 16 (voting for CLINTON or TRUMP), we can perform bivariate statistical analysis.

a) Used chi-square test of independence to assess the relationship between categorical variables such as PRES 16 (DV) and PERTYID, RELIG, POLVIEWS, SEXNOW, ETC.

b) Examined the frequency distribution and conducted chi- square tests to determine if there are significant associations between the voting behavior and categorical IVs.

A graph with different colored bars

Description automatically generated

**iii. Model output (significance, accuracy, key features)**

**Significance:**

Based on coefficients have used Support Vector Machine (SVM) to get better results. Performed chi-square testing to show Bi-Variate statistics between IV and key IVs.

**Accuracy:**

Accuracy after performing the SVM model to the dataset has given 72% and have calculated the R-square score and got as 0.2647734482683177.

A red rectangular object with white text

Description automatically generated

**Key features:**

**PRES12 has more influence on PRES16 which is clearly observed and some of the IVs which has influenced DV is listed down, and the Model Output is calculated accordingly.**

|  |  |
| --- | --- |
| **PRES 16**  **(Dependent Variable)** | **Vote CLINTON or TRUMP**  **Whom the responder has voted for 16 elections** |
| PRES 12 | Vote OBAMA or ROMNEY |
| AGE | AGE of the that particular person |
| RELIG | RELIGION of that person |
| CLASS | Subjective class identification (Working, middle, lower class etc |
| PARTYID | Republican, Independent or Democrat |
| SEXNOW | Gender of that person |
| CONRINC | Respondent income in dollars |
| POLVIEWS | Whether the person is liberal or conservative |
| SATFAM7 | Family satisfaction in general |
| DIPGED | Education qualification of the responder |

**9. References:**

**a. At least 10 references from reputable sources**

[1] Hall, Heather, "Voting Behavior Among Young Adults: An Analysis of Youth Voters and how Behavioral Economic Concepts can be Applied to Increase Young Voter Turnout" (2019). Honors College Theses. 256.

[2] Braha, D., & Aguiar, M. A. (2017). Voting contagion: Modeling and analysis of a century of U.S. presidential elections. Plos One,​ 12​(5). doi: 10.1371/journal.pone.0177970

[3] Gohmann, T. (2017) How Donald Trump Won the Election: A Behavioral Economics Explanation. Retrieved from http://www.behavioraleconomics.com/how-donald-trump-won-the-election-a-behav ioral-economics-explanation/

[4] CIRCLE Staff at Jonathan M. Tisch College of Civic Life at Tufts University. (2016) Young Voters in the 2016 General Election. Online. Retrieved from: https://civicyouth.org/wp-content/uploads/2016/11/CIRCLE-Full-Exit-Poll-Analysis\_ Final.pdf

[5] Falcão, F., Sousa, B., Pereira, D.S.M. et al. We vote for the person, not the policies: a systematic review on how personality traits influence voting behaviour. Discov Psychol 3, 1 (2023). https://doi.org/10.1007/s44202-022-00057-z

[6] Anderson, Amber (2023) "Voting Behavior and Primary Elections," Res Publica - Journal of Undergraduate Research: Vol. 28

[7] Kulachai,Waiphot, and Unisa Lerdtomornsakul, and Patipol Homyamyen. 2023. Factors Influencing Voting Decision: A Comprehensive Literature Review. Social Sciences 12: 469. https:// doi.org/10.3390/socsci12090469

[8] Bakker, Bert N., and Yphtach Lelkes. 2018. Selling ourselves short? How abbreviated measures of personality change the way we thinkabout personality and politics. The Journal of Politics 80: 1311–25.

[9] Basil, Michael D., William J. Brown, and Mihai C. Bocarnea. 2018. Candidate image building during the 2016 U.S. presidential election:Visual and verbal strategies in Donald Trump’s Instagram posts. Visual Communication Quarterly 25: 88–103

[10] Campbell, Amy L. 2008. The young and the realigning: A longitudinal analysis of youth political behavior. American Journal of PoliticalScience 52: 685–96.

[11] Caprara, G. V., Barbaranelli, C., & Zimbardo, P. G. (2006). Personality profiles and political parties. Political Psychology, 27(1), 55-68.

[12] Feldman, S. (1988). Structure and consistency in public opinion: The role of core beliefs and values. American Journal of Political Science, 32(2), 416-440.

[13] Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). Personality traits and the consumption of political information. American Politics Research, 39(1), 32-84.

[14] Mondak, J. J., & Halperin, K. D. (2008). A framework for the study of personality and political behavior. British Journal of Political Science, 38(2), 335-362.

[15] Sibley, C. G., & Duckitt, J. (2010). Personality and prejudice: A meta-analysis and theoretical review. Personality and Social Psychology Review, 14(2), 126-145.

**b. References in APA or MLA format**

[1] Brunnerová O. It’s Getting Personal: Personalisation of political campaigns in four Prague districts during the 2018 Czech Senate elections. Polit Central Eur. 2019; 15:279–308. <https://doi.org/10.2478/pce-2019-0018>.

[2] Anderson, Amber (2023) "Voting Behavior and Primary Elections," Res Publica - Journal of Undergraduate Research: Vol. 28

[3]Abramowitz,A.I.(1989).Viability,electability,andcandidatechoiceinapresidentialprimary election:Atestofcompetingmodels.TheJournalofPolitics,51(4),977–992. <https://doi.org/10.2307/2131544>

[4] Kulachai,Waiphot, and Unisa Lerdtomornsakul, and Patipol Homyamyen. 2023. Factors Influencing Voting Decision: A Comprehensive Literature Review. Social Sciences 12: 469. https:// doi.org/10.3390/socsci12090469

[5] Abramowitz, Alan I. 2010. The Disappearing Center: Engaged Citizens, Polarization, and American Democracy. New Haven: Yale University Press.

[6] Ansolabehere, S., & Hersh, E. (2012). Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. Political Analysis, 20(3), 437-459.

[7] Hopkins, D. J. (2018). The Hidden American Voter: Women, Racial Change, and the Election of Barack Obama. Political Behavior, 40(1), 73-99.

[8] Lauderdale, B. E. (2016). Using state laws to forecast voter turnout in US presidential elections. State Politics & Policy Quarterly, 16(4), 453-473.

[9] Krimmer, R., Volkamer, M., & Duenas-Cid, D. (2020). Strategic Decisions in Election Administration. Springer International Publishing.

**c. References are cited appropriately in the paper.**

[1] Hall, Heather, "Voting Behavior Among Young Adults: An Analysis of Youth Voters and how Behavioral Economic Concepts can be Applied to Increase Young Voter Turnout" (2019). Honors College Theses. 256.

[2] Atkin, C. K., & Gantz, W. (1978). Television News and Political Socialization. Public Opinion Quarterly,​ 42​(2), 183. doi:10.1086/268442

[3] Bandura, A. (2001, 02). Social Cognitive Theory: An Agentic Perspective. Annual Review of Psychology,​ 52​(1), 1-26. doi: 10.1146/annurev.psych.52.1.1

[4] Bettinger, E., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2009, 09). The Role of Simplification and Information in College Decisions: Results from the H&R Block FAFSA Experiment. doi:10.3386/w15361

[5] Falcão, F., Sousa, B., Pereira, D.S.M. et al. We vote for the person, not the policies: a systematic review on how personality traits influence voting behaviour. Discov Psychol 3, 1 (2023). <https://doi.org/10.1007/s44202-022-00057-z>

[6] <https://www.icpsr.umich.edu/web/pages/instructors/setups2008/exercises/analysis.html>