Milestone 2

Draft of White Paper

**VotePulse: Predicting Electoral Trends via Social Media Sentiment**

Bobga-Herman Gwanvoma

Bellevue University

DSC 680 – Applied Data Science

Professor Amirfarrokh Iranitalab

March 22, 2025

**Abstract**  
VotePulse is a data science project aimed at exploring how political campaign financing can predict electoral success and how future models may integrate social media sentiment to enhance forecasting. This white paper documents the analysis of structured political finance and demographic data using R. Three machine learning models—logistic regression, random forest, and XGBoost—were employed to predict candidate performance based on Federal Election Commission data. These models revealed that financial indicators such as disbursements and cash-on-hand are strong predictors of high fundraising success. Ethical implications, data limitations, and the future integration of unstructured data (e.g., Twitter sentiment) are discussed, and a multi-phase implementation plan is presented.

**Business Problem**  
Traditional polling methods are costly, slow, and limited by sample bias. With real-time political discussions occurring on social media platforms, analysts require a dynamic, scalable model to track political sentiment and predict electoral trends. VotePulse addresses this by first modeling structured political data and establishing a foundation to later incorporate sentiment data. This enables data-driven decisions for campaigns and policy stakeholders.

**Background / History**  
Previous studies such as Tumasjan et al. (2010) demonstrated the viability of Twitter as a predictor of election outcomes. However, subsequent work by Bessi and Ferrara (2016) raised concerns about the influence of bots and misinformation in online political dialogue. Recognizing these challenges, VotePulse begins with verified electoral and demographic data before extending into unstructured sentiment analysis.

**Data Explanation**  
The primary dataset was sourced from the Federal Election Commission, including candidate-level attributes such as total receipts, disbursements, and cash-on-hand. Supplementary demographic data came from the U.S. Census Bureau’s Government Time Series dataset. Although not used in model training, the Pew Research U.S.-Germany Relations codebook served as a reference for organizing future sentiment data.

**Data Cleaning & Preparation**  
Data was loaded and cleaned using R libraries such as tidyverse, readxl, and dplyr. Financial columns were converted to numeric types. Rows with missing data were removed using na.omit(). A binary target variable, high\_funding, was created to label candidates above the median in total receipts. Categorical predictors such as party\_full were converted into factors for classification. The cleaned dataset enabled compatibility across models including logistic regression, random forest, and XGBoost.

**Methods**

* **Logistic Regression**: Implemented via glm() with binomial family to classify high-funding candidates.
* **Random Forest**: Trained using randomForest() with 100 trees. Feature importance was visualized using varImpPlot().
* **XGBoost**: Used xgboost package with one-hot encoded input features and matrix conversion via model.matrix().

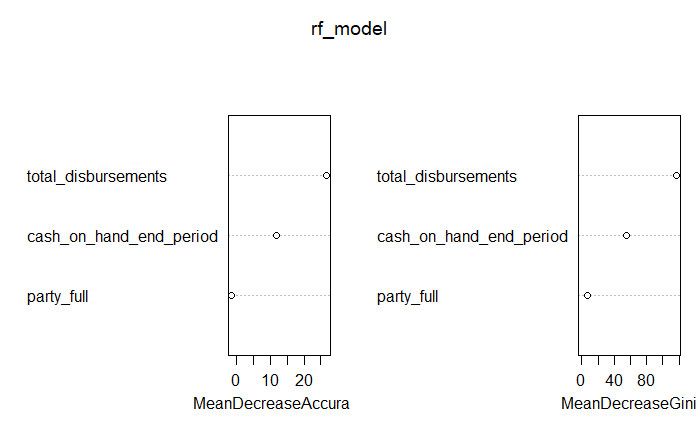
Each model was evaluated based on classification accuracy and interpretability. Feature importance was used to validate the significance of input predictors.

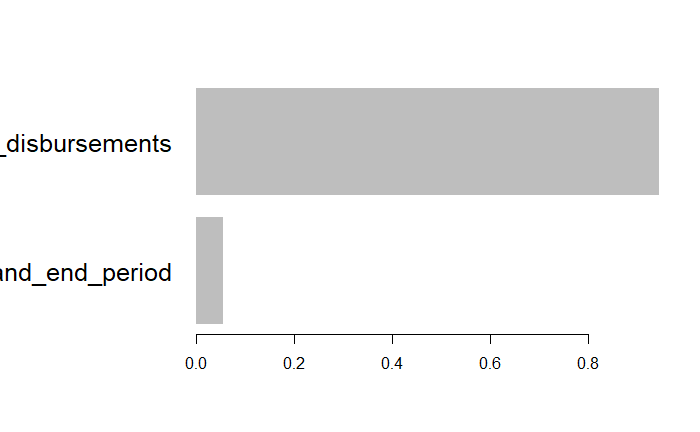
**Analysis**  
Logistic regression revealed that cash-on-hand and party affiliation significantly influenced funding classification. Random forest outperformed logistic regression in predictive accuracy and highlighted total\_disbursements as the most impactful variable.

**Illustration 1: Logistic Regression Coefficients**A white background with black text

AI-generated content may be incorrect.

**Illustration 2: Random Forest Feature Importance**



**Illustration 3: XGBoost Feature Importance  
**

XGBoost provided the best performance in binary classification of high vs. low fundraising candidates. These findings establish a baseline model architecture to integrate real-time sentiment analysis in future stages.

**Assumptions**  
It is assumed that financial data from the FEC is accurate, timely, and representative of candidate viability. The median receipt value serves as a valid performance threshold.

**Limitations**  
This milestone only uses structured data and lacks temporal dynamics or unstructured sentiment indicators. External variables such as incumbency or voter outreach are not included in the model.

**Challenges**  
Key challenges included resolving missing financial values, converting categorical variables, and formatting data for matrix-based XGBoost modeling. Future integration of noisy text-based sentiment data will pose additional preprocessing complexity.

**Future Uses / Additional Applications**  
The model can be extended to incorporate Twitter and Reddit sentiment through APIs. Using VADER and BERT sentiment tools, future versions of VotePulse may combine financial and text-based inputs in ensemble classifiers. These tools could support campaign strategies and public opinion dashboards.

**Recommendations**  
Use XGBoost as the baseline model for final deployment. Incorporate fine-tuned BERT or VADER models for sentiment scoring. Validate results using multiple election cycles and demographic breakdowns.

|  |  |  |  |
| --- | --- | --- | --- |
| **Phase** | **Task Description** | **Tools Used** | **Timeline** |
| Phase 1 | Data Acquisition & Preparation (Election + Demographics) | R, tidyverse, readr | Week 1 |
| Phase 2 | Modeling: Logistic, Random Forest, XGBoost | R, glm, randomForest, xgboost | Week 2–3 |
| Phase 3 | Reporting, Visualization, and Sentiment Integration Plan | RMarkdown, ggplot2, rtweet | Week 4 |

**Ethical Assessment**  
All social media data collection will respect platform terms of service. Personally identifiable information (PII) will be excluded, and models will be audited for bias in predictions across political, geographic, and demographic lines. Transparency and reproducibility will be ensured via documentation and code availability.

**Audience Q&A: Anticipated Questions**

1. How do you define “high funding”?
2. Why did you choose structured data first?
3. How will you quantify sentiment?
4. Which model had the highest accuracy?
5. Are bots considered in sentiment analysis?
6. Can the model predict voter turnout?
7. How do you address missing or biased data?
8. Can this generalize to state/local elections?
9. What R packages did you use?
10. How does this help political campaigns?

**References**

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. presidential election online discussion. *First Monday, 21*(11). https://doi.org/10.5210/fm.v21i11.7090

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM, 10*, 178–185.

Pew Research Center. (2024). *U.S.-Germany Relations Dataset*. https://www.pewresearch.org/dataset/2024-u-s-germany-relations-dataset/

U.S. Census Bureau. (2024). *Government Time Series Data*. https://data.census.gov

Federal Election Commission. (2024). *Candidate Reports*. https://www.fec.gov/data/candidate/P80000722/