Milestone 3

Final White Paper

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

Bobga-Herman Gwanvoma

Bellevue University

DSC 680 – Applied Data Science

Professor Amirfarrokh Iranitalab

June 01, 2025

**Abstract**  
VotePulse is a data science initiative designed to evaluate whether political campaign financing can predict electoral outcomes. Utilizing R for data analysis, the project begins with structured datasets from the Federal Election Commission (FEC) and the U.S. Census Bureau to model candidate performance. It employs three predictive modeling techniques—logistic regression, random forest, and XGBoost—highlighting financial indicators like disbursements and cash-on-hand as key predictors. The broader goal is to incorporate sentiment analysis from social media in future iterations. This paper outlines our initial results, ethical considerations, limitations, and implementation timeline.

**Business Problem**  
Polling is traditionally used to understand public political sentiment, but it is time-consuming and often subject to bias. Social media platforms offer real-time insights but present challenges in data structure and reliability. VotePulse aims to bridge this gap by starting with a clean, structured dataset to model electoral trends, setting the stage for future integration of unstructured sentiment data. By aligning robust financial indicators with advanced machine learning, we seek to build a foundation for more timely and scalable political forecasting tools.

**Background / History**  
The project draws on foundational research such as Tumasjan et al. (2010), which highlighted Twitter's predictive potential in elections. However, Bessi and Ferrara (2016) cautioned that bots and misinformation distort online political discussions. These insights inform VotePulse’s decision to begin with validated financial data before expanding to social media sources. Additionally, the rapid digitization of campaign operations has made campaign finance data both more accessible and more essential for political analysis.

**Data Explanation**  
The FEC dataset includes candidate-level data such as total receipts, disbursements, and cash-on-hand. These metrics provide quantitative indicators of a campaign’s financial health and potential voter reach. Supplementary demographic context is drawn from the U.S. Census Bureau, providing insights into population distributions that can be used to understand how campaign finance effectiveness varies by region. The Pew Research dataset serves as a model for how future sentiment data might be structured and analyzed, especially in multilingual or cross-national contexts.

**Data Cleaning & Preparation**  
The data was imported into R using readr and readxl. Numeric columns were cleaned and converted to ensure accuracy in downstream analysis. Missing values were removed with na.omit() to prevent data leakage and bias in model evaluation. Categorical variables such as party affiliation were formatted using factor(). A binary classification target was created by labeling candidates above the median in total receipts as “high\_funding,” allowing for effective classification modeling.

**Methods**  
Three modeling approaches were used. Logistic regression offered interpretability and a reliable statistical baseline. Random forest helped identify variable importance and offered robustness to overfitting. XGBoost provided enhanced predictive performance and allowed for hyperparameter tuning and feature interaction analysis. Each model used the same cleaned dataset and was evaluated using accuracy, feature importance, and interpretability. Code was implemented entirely in R, leveraging packages like randomForest, xgboost, and caret.

**Analysis**  
Logistic regression demonstrated that cash-on-hand and party affiliation significantly influence candidate funding success. The odds of being classified as high-funding increased substantially with higher ending balances. Random forest modeling emphasized total disbursements as a crucial variable, suggesting that spending strategy may be as critical as fundraising itself. XGBoost showed the highest classification accuracy, reinforcing its utility as a primary model for future work. Visuals of each model’s output clearly demonstrate differing variable influence.

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

AI-generated content may be incorrect.

**Illustration 2: Random Forest Feature Importance**

A graph of a number of objects

AI-generated content may be incorrect.

**Illustration 3: XGBoost Feature Importance  
A graph of a graph with text

AI-generated content may be incorrect.**

**Assumptions**  
We assume the FEC dataset is accurate and reflective of actual financial performance. Using the median value for defining high funding is treated as a valid threshold. We also assume that campaign finance has a consistent relationship with candidate success across election cycles, though this will require further validation.

**Limitations**  
This phase excludes sentiment data and does not factor in voter outreach, incumbency, or media influence. Sentiment integration will introduce additional complexity. The generalizability of the models depends on future validation across cycles and jurisdictions. Financial metrics alone cannot fully predict voter sentiment or turnout, but they provide a solid analytical foundation.

**Challenges**  
Key technical hurdles included handling missing data, formatting variables for XGBoost, and preparing the dataset for classification. Selecting meaningful thresholds and balancing class distributions posed additional complexity. Adding unstructured text analysis in the next phase will require preprocessing pipelines, language model selection, and platform-specific considerations.

**Future Uses / Additional Applications**  
VotePulse can incorporate sentiment scores from Twitter or Reddit using APIs and tools like VADER and BERT. Beyond election prediction, the framework can be adapted to forecast campaign effectiveness, policy reception, and voter turnout. Over time, the model can support real-time dashboards for campaigns and voter insight tools, potentially integrating geolocation and time-sensitive factors.

**Recommendations**  
Adopt XGBoost as the core model given its superior accuracy. Expand dataset breadth by incorporating unstructured sentiment in Phase 3. Test on additional election cycles and control for demographic variables to improve robustness. Develop interactive dashboards for public and organizational use.

**Implementation Plan**  
The project is structured in three phases over four weeks. In Phase 1, data from the FEC and Census is acquired and cleaned using R. Phase 2 includes model development using logistic regression, random forest, and XGBoost. Phase 3 finalizes reporting and integrates sentiment analysis techniques. Visualizations and documentation will be completed in RMarkdown and ggplot2, with potential for dashboarding tools like Shiny.

**Ethical Assessment**  
All social media data will follow platform policies and avoid capturing any personally identifiable information. No personal data will be stored or processed. Bias in models will be tested and addressed using fairness metrics, and results will be interpreted through an ethical lens to avoid reinforcing structural inequalities. All code and findings will be transparently documented and made available for public review.

**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/