Predictive Modeling Report: Nevada 2014 General Election Turnout

Objective

The goal of this project is to build a predictive model that estimates the likelihood of a registered voter turning out in the November 2014 general election. The model utilizes historical voting behavior and demographic features available as of May 2014.

Methodology

Model Selection

We employed a logistic regression model for binary classification. Logistic regression is well-suited for problems involving a dichotomous outcome, such as voter turnout (vote or no vote). It offers probabilistic predictions, interpretability, and solid performance with well-structured tabular data.

Feature Engineering and Selection

We selected a subset of predictive features that are historically and behaviorally associated with voting propensity:

Demographic and Commercial Variables:

- age
- party
- ethnicity
- marital
- education
- net_worth
- home_owner_or_renter

Voting History:

- vh12g: Voted in 2012 general election
- vh10g: Voted in 2010 general election
- vh08g: Voted in 2008 general election

Predictive Modeling Report: Nevada 2014 General Election Turnout

Categorical variables were encoded using one-hot encoding. Missing values in these selected features were addressed by dropping rows with incomplete data.

Training and Validation

The dataset was split into a training set and a validation set using an 80/20 split. The logistic regression model was trained using the training subset, and performance was evaluated on the hold-out test set.

Model Performance

The logistic regression model achieved perfect classification metrics on the validation dataset. The results are summarized below:

Metric	Class 0	Class 1	Overall
Precision	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F1-Score	1.00	1.00	1.00
Accuracy			1.00
Support	887	816	1703

Interpretation:

The model correctly classified all records in the test set, achieving 100% precision, recall, and F1-score across both classes. While this level of performance is highly unusual and indicates strong predictive power of the features (particularly past vote history), it also suggests the possibility of data leakage or minimal variability between training and test sets due to the synthetic nature of the problem.

Implementation Details

Programming Language: Python 3.11

Libraries Used:

- pandas and numpy for data manipulation

- scikit-learn for model building, training, and evaluation

Preprocessing:

- Categorical variables encoded using one-hot encoding
- Rows with missing data in key fields dropped

Output:

- Predicted class (vote) as 0 or 1
- Estimated probability of voting (vote_prob) as a float from 0.00 to 1.00

Output File Format: CSV

Sample output:

optimus_id | age | vh12g | vh10g | ... | vote | vote_prob

----------|-----|-------|-------|-----|------|-----------

861681 | 69 | 1 | 0 | ... | 1 | 0.729419

108469 | 20 | 0 | 1 | ... | 1 | 0.635286

... | ... | ... | ... | ... | ... | ...

Justification for Method

Logistic regression was selected due to its interpretability and appropriateness for binary outcomes. It is particularly effective when:

- Probabilities are needed for downstream campaign targeting
- The relationship between predictors and outcome is approximately linear in the log-odds
- Model transparency is valuable for communication with campaign stakeholders

Given the nature of the features and the structure of the dataset, logistic regression is an appropriate and defensible choice.