Quantitative Social Science in the Age of Big Data and AI

Lecture 11: Classification

Zhaopeng Qu
Hopkins-Nanjing Center
May 27 2025

Outline

1. Review of the last lecture
2. Classification
3. Logistic Regression
4. Predictions Assessment

Classification

Introduction

- So far our dependent variable or outcome is always continuous, however, in many cases, we are also interested in the discrete or categorical outcomes as well.

- Examples in traditional data:
  - Y= person smokes, or not; X = cigarette tax rate in the state, income, education, age, gender, etc.
  - Y= Civil War, or not; X = economic, social, political, etc.
  - Y= Presidential election, or not; X = economic, social, political, etc.

- Examples in Big Data:
  - sentiment analysis in text: positive or negative
  - topic classification for news or social media: news, sports, entertainment, etc.
  - image classification: cat or dog

- In the ML literature, this kind of problem is called **classification**
  - Binary classification: two categories
  - Multi-class classification: more than two categories

Binary Classification: Linear Probability Model

- The linear probability model(LPM) is a simple extension of the linear regression model to the case of a binary outcome.
- If a outcome variable Y is binary, thus

$$y_i = \begin{cases} 1 & \text{if} D = 1 \\ 0 & \text{if} D = 0 \end{cases}$$

- The unconditional expectation of Y is

$$E[Y] = 1 \times Pr(Y = 1) + 0 \times Pr(Y = 0) = Pr(Y = 1)$$

  which is the probability of $Y = 1$.

- Then we can extend it to the conditional expectation of Y on X, thus $E[Y|X_{1i}, \ldots, X_{ki}]$ equals to the the probability of $Y = 1$ conditional on $X_{1i}, \ldots, X_{ki}$,thus

$$E[Y|X_{1i}, \ldots, X_{ki}] = Pr(Y = 1|X_{1i}, \ldots, X_{ki})$$

  which is the probability of $Y = 1$ conditional on $X_{1i}, \ldots, X_{ki}$.

Multiple OLS Regression

- Suppose our regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

- Based on supervised learning assumption, thus

$$E[\epsilon_i | X_{1i}, \ldots, X_{ki}] = 0$$

- The conditional expectation equals the probability that $Y_i = 1$ conditional on $X_{1i}, \ldots, X_{ki}$

$$E[Y | X_{1i}, \ldots, X_{ki}] = Pr(Y = 1 | X_{1i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

- Now a Linear Probability Model can be defined as following

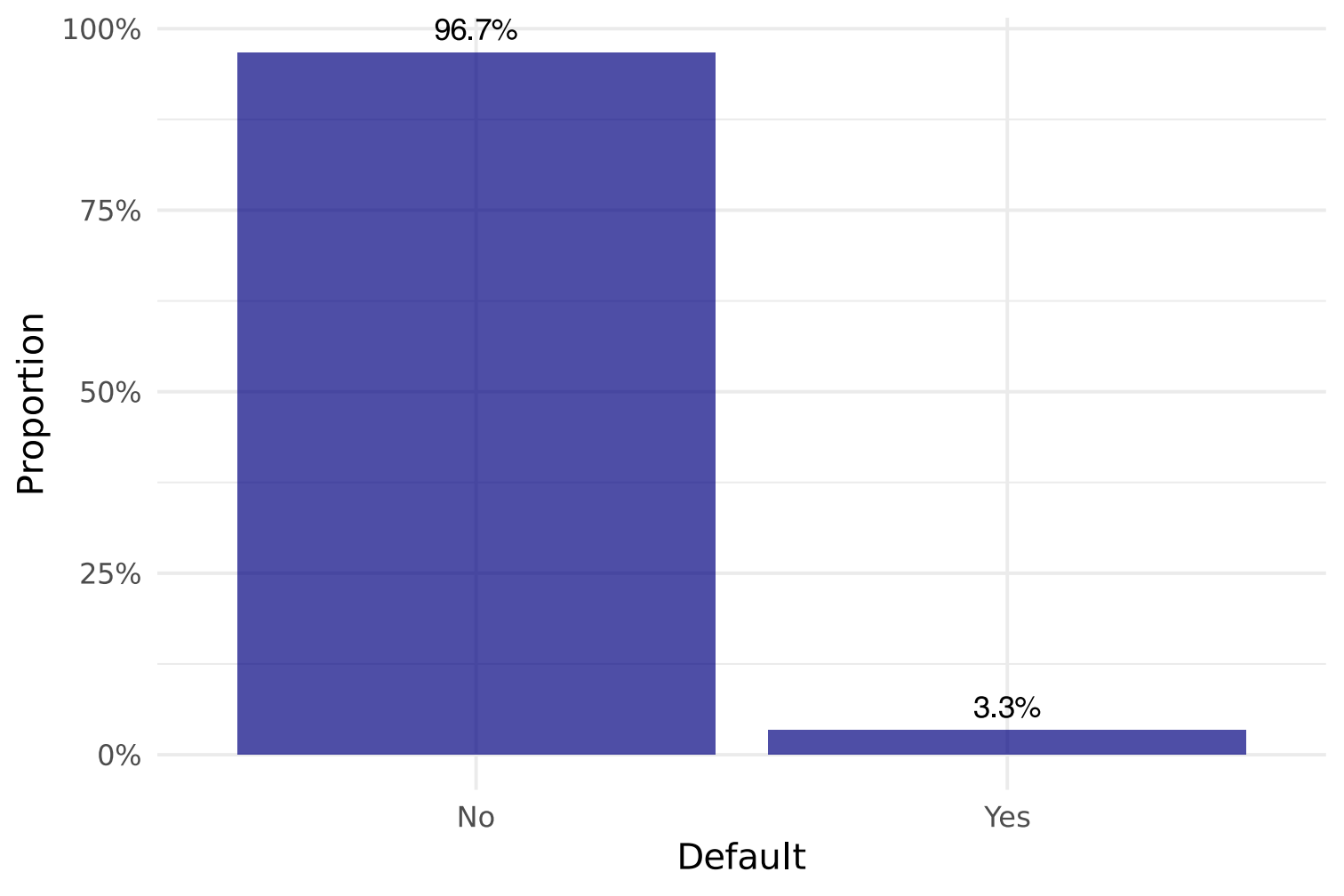
$$Pr(Y = 1 | X_{1i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

- In other words, we could use the linear regression model to predict the probability of $Y = 1$ based on $X_{1i}, \ldots, X_{ki}$

Example: Default Data

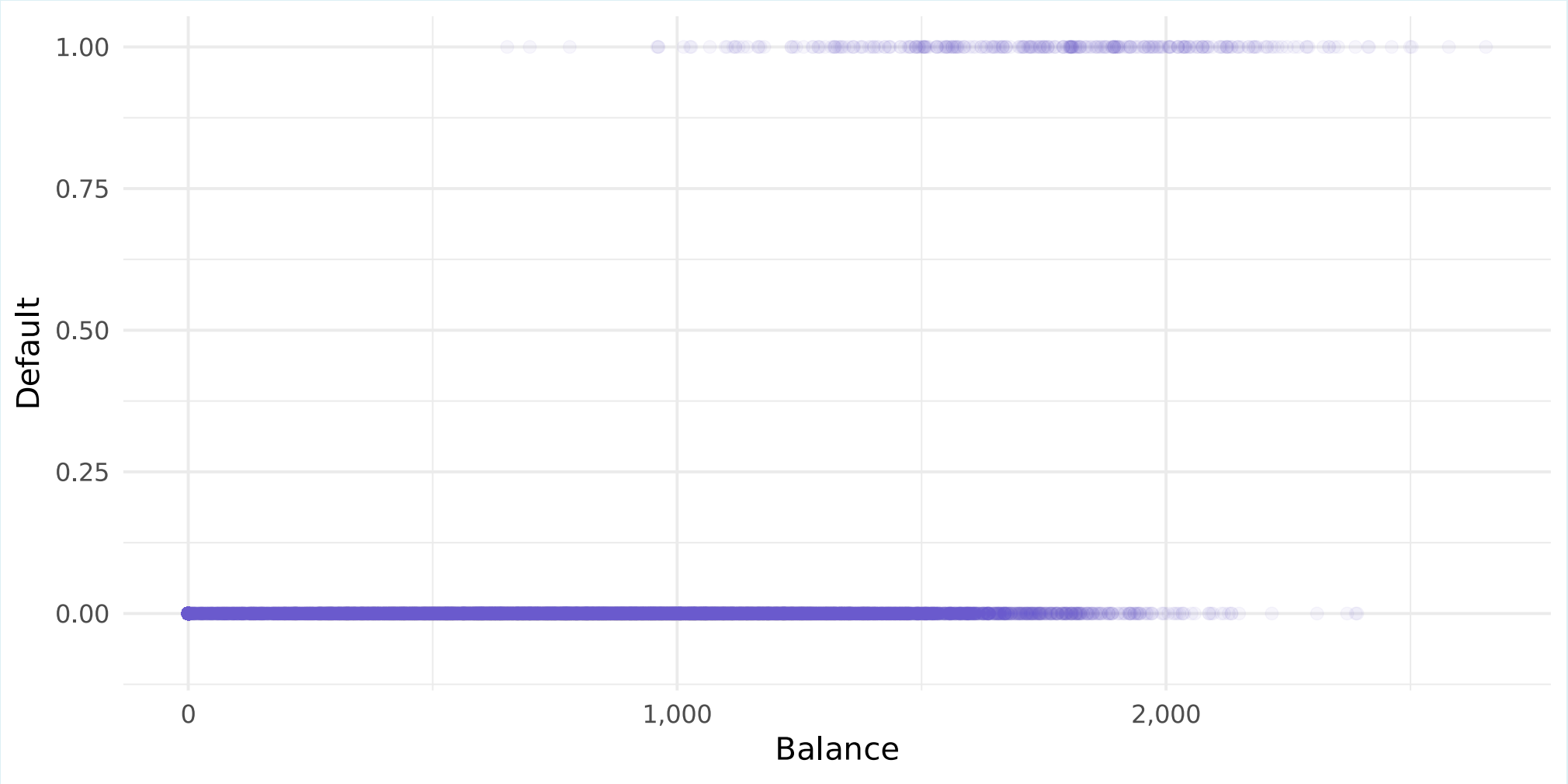
default	student	balance	income
No	No	939.10	45,519
No	Yes	397.54	22,711
Yes	No	1,511.61	53,507
No	No	301.32	51,540
No	No	878.45	29,562
Yes	No	1,673.49	49,310
No	No	310.13	37,697
No	No	1,272.05	44,896
No	No	887.20	41,641
No	No	230.87	32,799

Example: Default Data

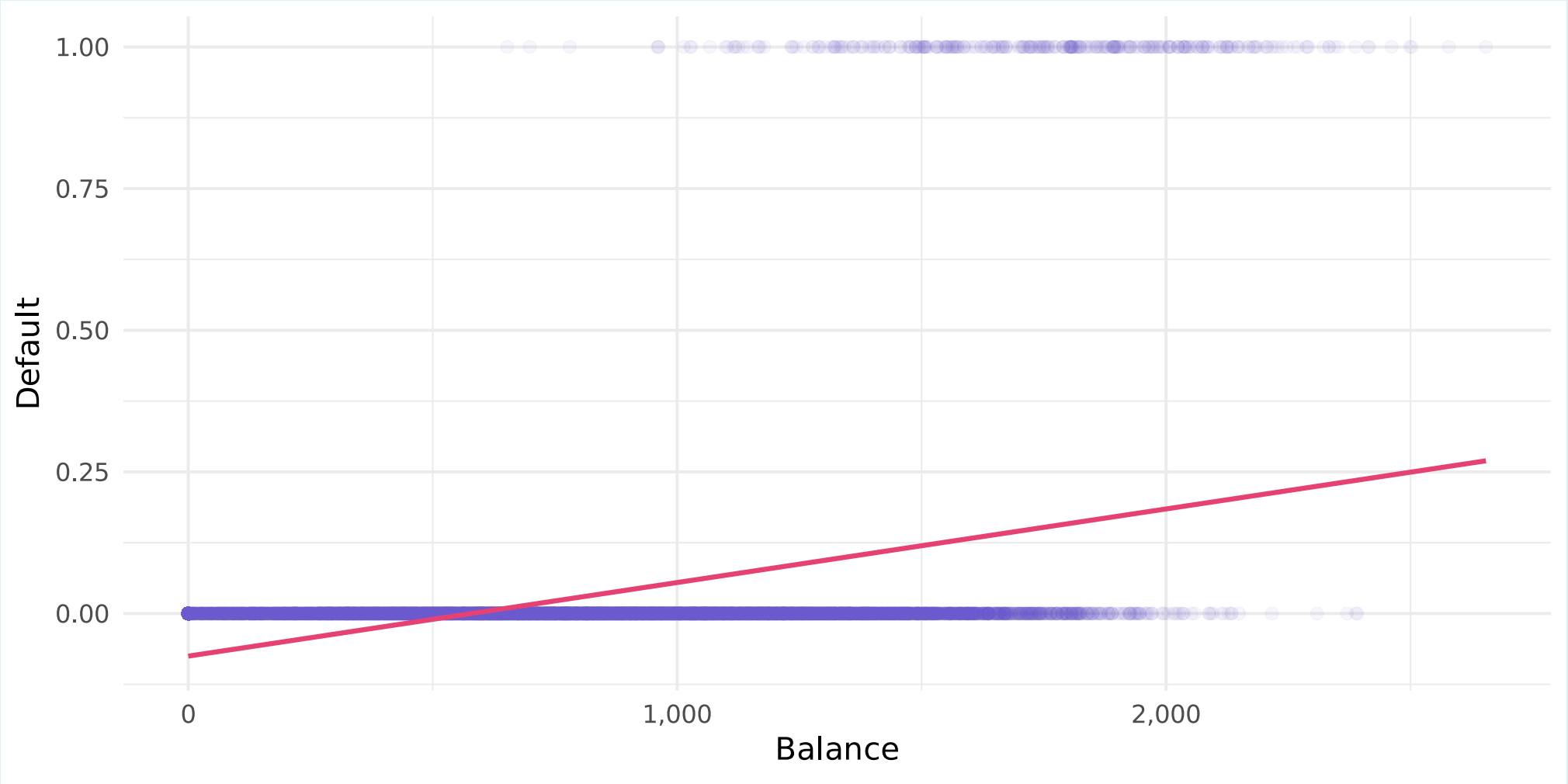


- The outcome, default, only takes two values (only 3.3% default).

Scatter Plot: Balance vs Default



Linear Probability Model



- The serious problem of the LPM is that the predicted probability can be less than 0 or greater than 1!

Case Study: Presidential Election Prediction

- Let's consider a more politically relevant example: predicting voter choice in presidential elections

Research Question:

- Can we predict whether a voter will vote for the incumbent party candidate based on economic and demographic factors?

Variables:

- Y: Vote for incumbent (1) or challenger (0)
- X_1: Personal income change (%)
- X_2: National unemployment rate (%)
- X_3: Age of voter
- X_4: Education level (years)
- X_5: Party identification (scale 1-7)

Why Classification?

- Voting is a binary choice
- Traditional polling has limitations
- Economic voting theory suggests economic conditions influence electoral outcomes
- Can help campaigns target resources

Applications:

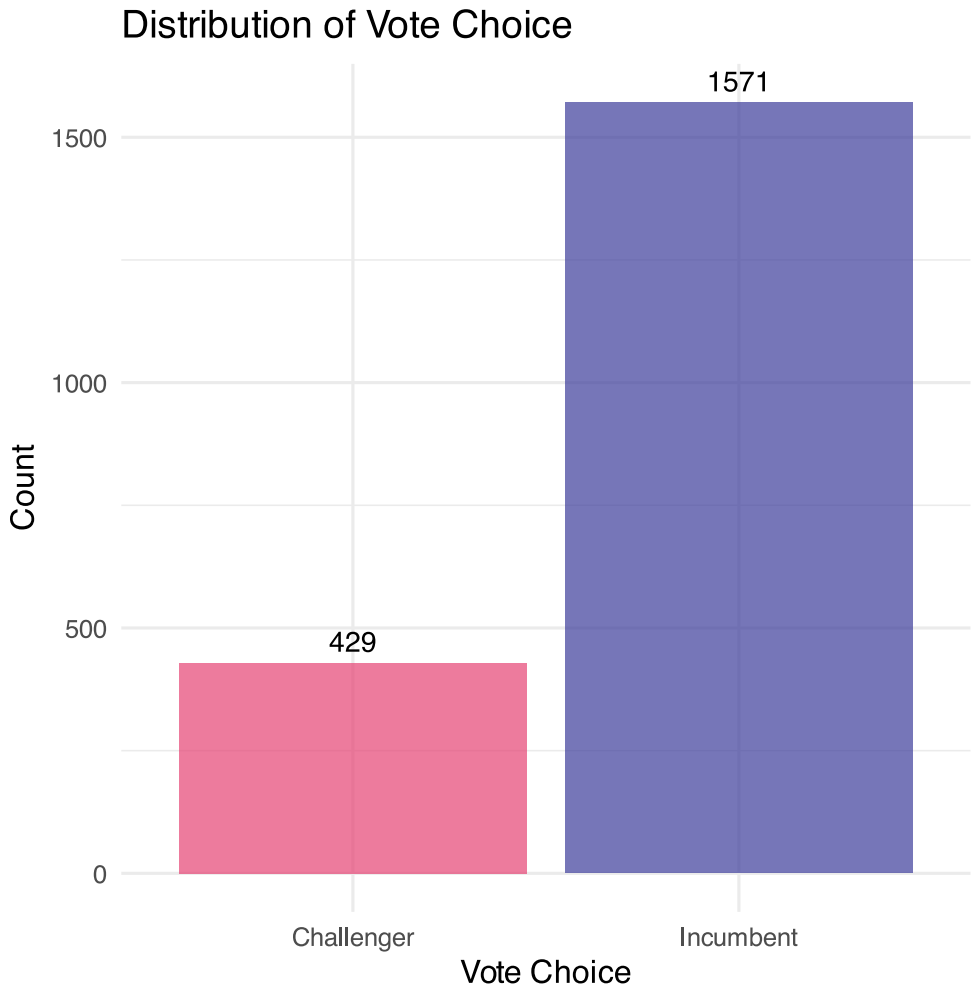
- Campaign strategy
- Electoral forecasting
- Understanding voter behavior
- Resource allocation for campaigns

Election Data: Simulated Example

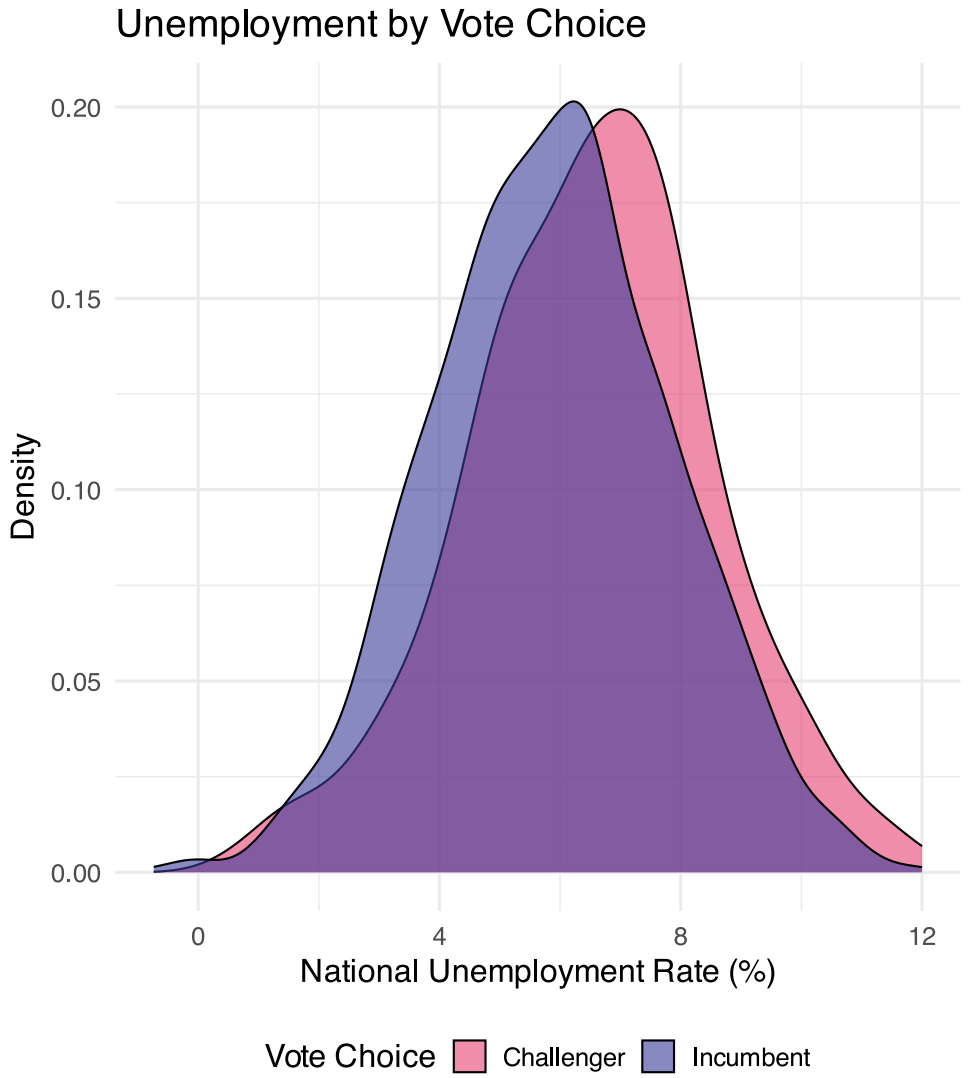
Sample of Election Data (First 10 observations)

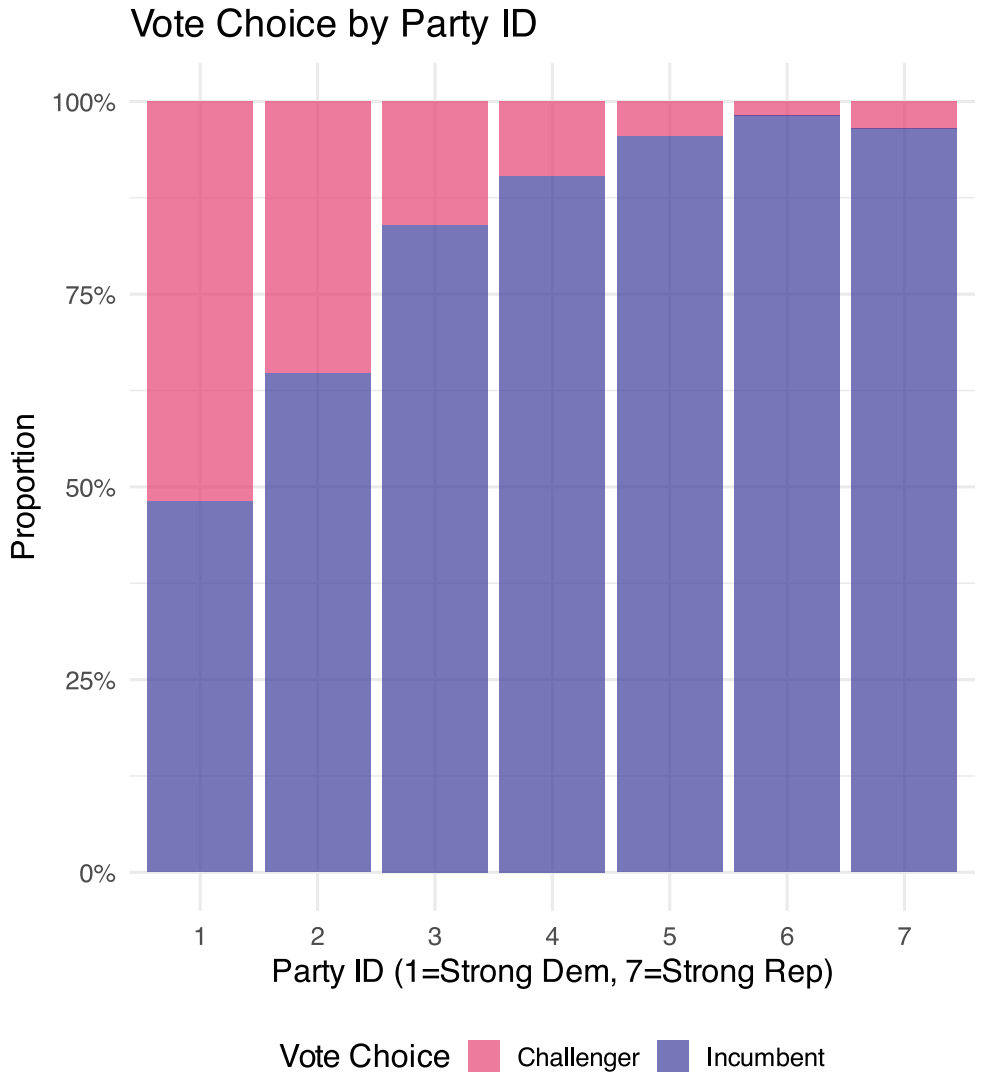
Vote Choice	Income Change (%)	Unemployment (%)	Age	Education (years)	Party ID (1-7)
Incumbent	-1.8	4.6	36	13	4
Incumbent	1.0	1.7	48	20	1
Incumbent	-3.2	4.5	33	17	6
Incumbent	1.0	5.6	55	19	5
Incumbent	-0.3	6.5	52	17	5
Incumbent	2.3	3.1	59	18	1
Challenger	1.4	5.5	36	13	2
Challenger	2.1	8.7	21	18	1
Challenger	3.6	5.0	30	18	1
Incumbent	-1.1	6.6	73	19	5

Election Data: Descriptive Statistics



Election Data: Descriptive Statistics





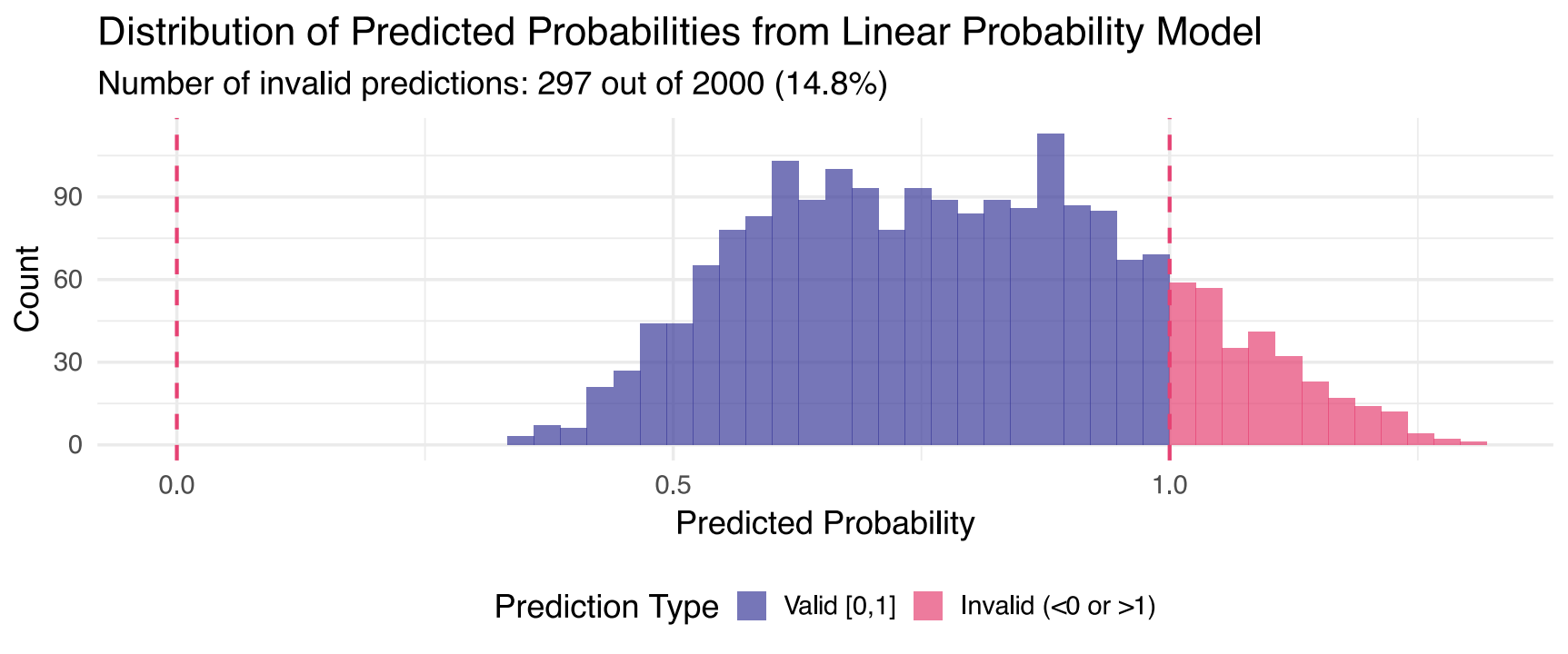
Linear Probability Model: Election Example

Linear Probability Model Results: Predicting Vote for Incumbent

Variable	Coefficient	Std. Error	t-statistic	p-value
Intercept	**0.4305**	0.0636	6.77	< 0.001
Income Change (%)	**0.0184**	0.0028	6.64	< 0.001
Unemployment (%)	**-0.0255**	0.0041	-6.29	< 0.001
Age	**0.0007**	0.0004	1.61	0.109
Education (years)	**0.0067**	0.0031	2.12	0.034
Party ID (1-7)	**0.0957**	0.0045	21.25	< 0.001

- A 1% increase in personal income change increases probability of voting incumbent by 0.018.
- A 1% increase in unemployment decreases probability by 0.026.
- Strong party identification effect: each unit increase in party ID (toward Republican) increases incumbent vote probability by 0.096.

Problems with Linear Probability Model



Key Problems:

1. 297 predictions fall outside the valid probability range [0,1]
2. Heteroskedasticity: Error variance depends on X values
3. Non-normal errors: Binary outcomes violate normality assumption

Logistic Regression

Introduction to Logistic Regression

- Probabilities must be bounded between 0 and 1.
- To address this limitation, we consider a general probability model:

$$Pr(Y_i = 1|X_1, \ldots X_k) = G(Z) = G(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i})$$

  - where $Z = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i}$
- The function $G(\cdot)$ must satisfy two essential conditions:
  - $0 \leq G(Z) \leq 1$
  - monotonicity and continuity

Introduction to Logistic Regression

- Using the standard logistic cumulative distribution function

$$Pr(Y_i = 1|X_{1,i}, \ldots, X_{k,i}) = \frac{1}{1 + e^{-Z}} = \frac{e^Z}{1 + e^Z}$$

  - Where $Z = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i}$

- Like LPM model, the logistic regression estimates the probability of $Y = 1$ conditional on $X_{1,i}, \ldots, X_{k,i}$

- And then we could use the **estimated probabilities** to **make predictions**

  - if $p(\text{Balance}) \geq 0.5$, we could predict "Yes" for Default

  - or, to be conservative, we could predict "Yes" if $p(\text{Balance}) \geq 0.1$

Logistic Regression

- Suppose we only have one feature, X, then $Z = \beta_0 + \beta_1 X$
- Then the logistic regression model is

$$p(x) = Pr(Y_i = 1 | X_{1,i}, \ldots, X_{k,i}) = \frac{1}{1 + e^{\beta_0 + \beta_1 X_{1,i}}}$$

- With a little math, you can show

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \implies \log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- This is the **log odds ratio**, which is the log of the odds ratio of $Y = 1$ and $Y = 0$
- If we could know the values of β_js, we could predict $p(x)$ for any x
- Question: how to obtain the values of β_js?

Estimation: Maximum Likelihood Estimation

- Because Logit model is a non-linear model, we cannot use the OLS or other linear regression methods to estimate the parameters.Instead, we use the **Maximum Likelihood Estimation** (MLE) method to estimate the parameters.
- The likelihood function is a *joint probability distribution* of the data, treated as a function of the unknown coefficients. It describes the probability of the data we observed or the sample from the population, given the unknown coefficients.
- The maximum likelihood estimator (MLE) are the estimate values of the unknown coefficients that maximize the likelihood function.
- MLE's logic

  > the most likely function is the function to have produce the data we observed.

- In other words, the MLE method searches for the β_js that make our data most likely given the model we've written.

Estimates and predictions

- Thus, our estimates are $\hat{\beta}_0 \approx -10.65$ and $\hat{\beta}_1 \approx 0.0055$.

Remember: These coefficients are for the **log odds**.

- If we want **to make predictions** for y_i (whether or not i defaults),
  then we first must **estimate the probability** $p(x)$, more specifically, $p(\text{Balance})$ here,

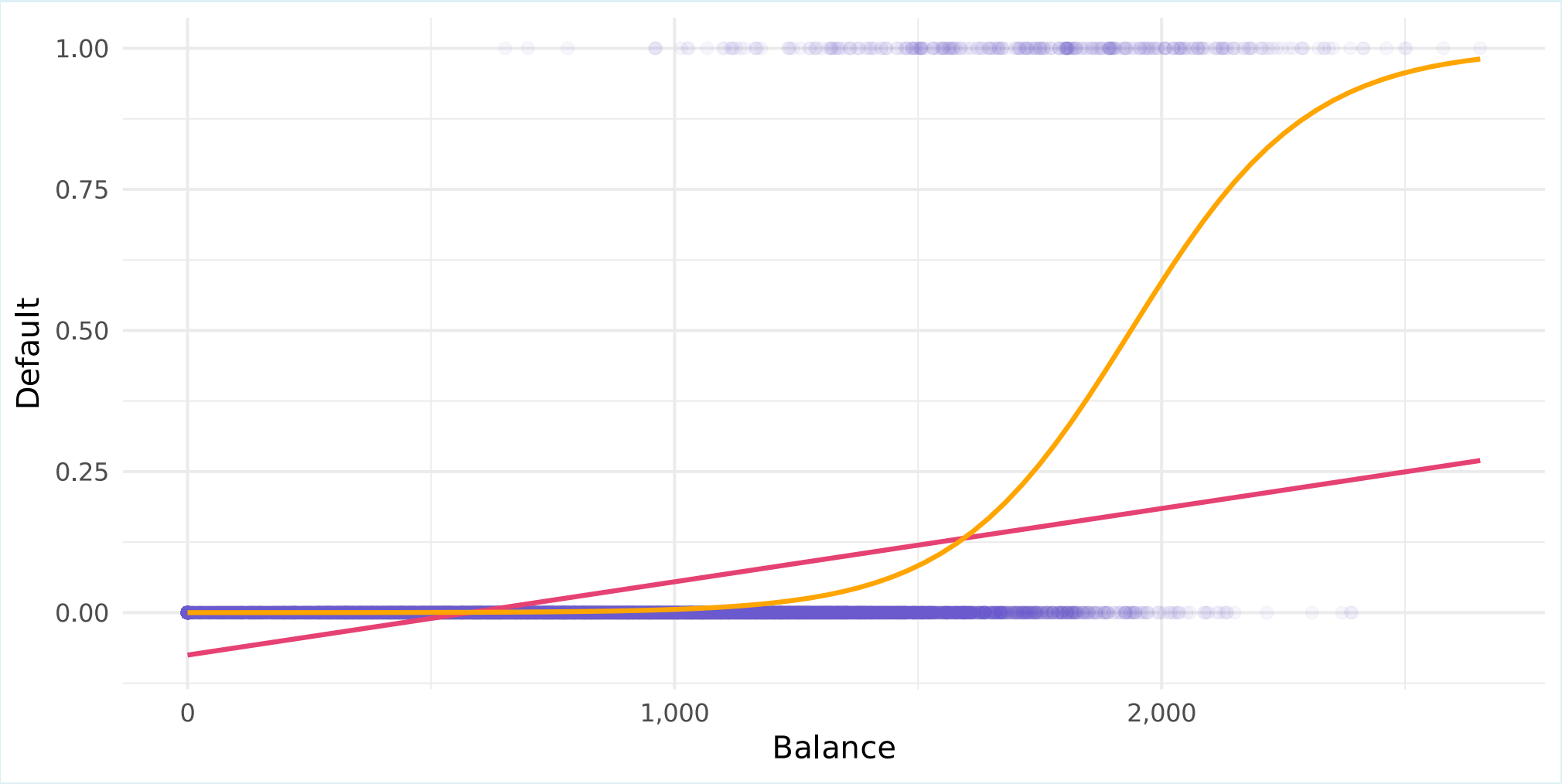
$$\hat{p}(\text{Balance}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Balance}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Balance}}} \approx \frac{e^{-10.65 + 0.0055 \cdot \text{Balance}}}{1 + e^{-10.65 + 0.0055 \cdot \text{Balance}}}$$

- If $\text{Balance} = 0$, we then estimate $\hat{p} \approx 0.000024$

- If $\text{Balance} = 2,000$, we then estimate $\hat{p} \approx 0.586$
- If $\text{Balance} = 3,000$, we then estimate $\hat{p} \approx 0.997$ †

† You get a sense of the nonlinearity of the predictors' effects.

Logistic Regression for default data



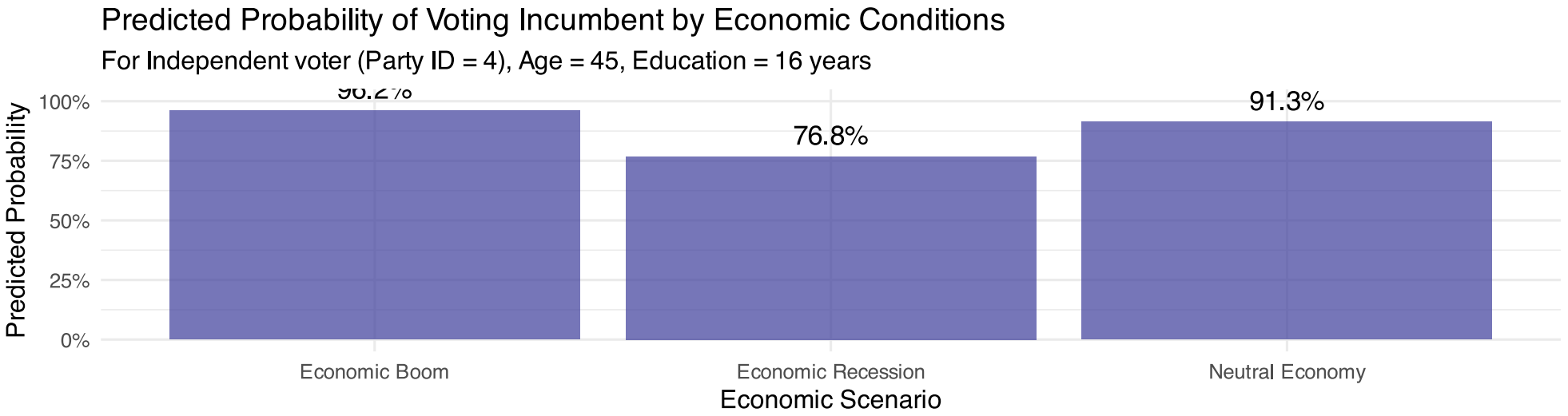
Logistic Regression: Election Example

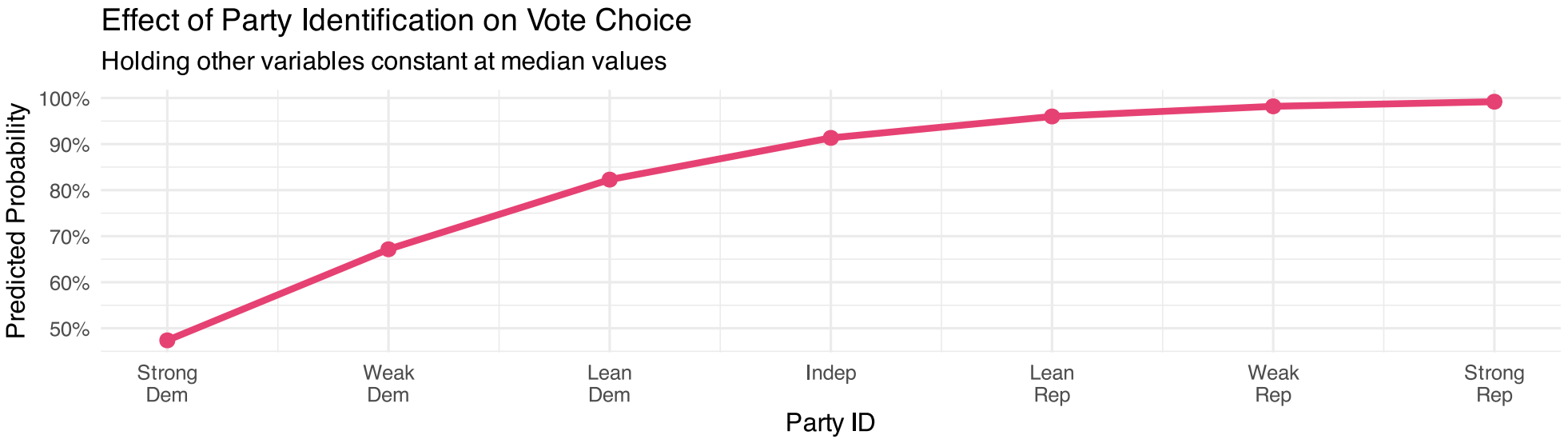
Logistic Regression Results: Predicting Vote for Incumbent

Variable	Coefficient	Std. Error	z-statistic	p-value
Intercept	**-1.1250**	0.4921	-2.29	0.022
Income Change (%)	**0.1472**	0.0226	6.50	< 0.001
Unemployment (%)	**-0.2104**	0.0323	-6.51	< 0.001
Age	**0.0060**	0.0035	1.72	0.085
Education (years)	**0.0561**	0.0244	2.30	0.022
Party ID (1-7)	**0.8203**	0.0482	17.00	< 0.001

- All variables are statistically significant predictors of vote choice
- Party ID has the strongest effect (coefficient = 0.8203)
- Economic variables matter: income growth helps incumbent, unemployment hurts
- Demographics also play a role: older and more educated voters slightly favor incumbent

Predicted Probabilities: Election Example





Predictions Assessment

Predictions Assessment: the correct rate

- First, we need to set a **threshold** for the predicted probability.
  - For example, if the predicted probability is greater than 0.5, we predict $\hat{y} = 1$ (vote for incumbent), otherwise $\hat{y} = 0$ (vote for challenger).
- Then we could calculate the accuracy of the predictions as the ratio of the number of **correct predictions** to the total number of **predictions**.
  - The correct prediction rate here is 82.6%
- Question: 82.6% is fairly good?
- However, recall that 78.6% of voters actually voted for the incumbent.
- It means that even if we guessed "Incumbent" for everyone, we would get 78.6% right.
- The formal assessment of the classification model is to use the **confusion matrix**.

Predictions Assessment

The confusion matrix

- It is used to display the **correct** and **incorrect** predictions for each class of the outcome.

Confusion Matrix: Election Prediction

	Actual Vote Choice	
Prediction	**Challenger**	**Incumbent**
Challenger	True Negative (TN)	False Negative (FN)
Incumbent	False Positive (FP)	True Positive (TP)

- The accuracy of a method is the share of correct predictions, *i.e.*,

$$\textbf{Accuracy} = (\text{TN} + \text{TP}) \; / \; (\text{TN} + \text{TP} + \text{FN} + \text{FP})$$

Predictions Assessment

The confusion matrix

Confusion Matrix: Election Prediction

	Actual Vote Choice	
Prediction	**Challenger**	**Incumbent**
Challenger	True Negative (TN)	False Negative (FN)
Incumbent	False Positive (FP)	True Positive (TP)

- Sensitivity : the share of positive outcomes $Y = 1$, where it is the incumbent voters, that we correctly predict.

$$\textbf{Sensitivity} = \text{TP} \ / \ (\text{TP} + \text{FN})$$

- Sensitivity is also called recall and the true-positive rate.

Predictions Assessment

The confusion matrix

Confusion Matrix: Election Prediction

	Actual Vote Choice	
Prediction	**Challenger**	**Incumbent**
Challenger	True Negative (TN)	False Negative (FN)
Incumbent	False Positive (FP)	True Positive (TP)

- Specificity : the share of negative outcomes $Y = 0$, where it is the challenger voters, that we correctly predict.

$$\textbf{Specificity} = \text{TN} \ / \ (\text{TN} + \text{FP})$$

Predictions Assessment

The confusion matrix

Confusion Matrix: Election Prediction

	Actual Vote Choice	
Prediction	**Challenger**	**Incumbent**
Challenger	True Negative (TN)	False Negative (FN)
Incumbent	False Positive (FP)	True Positive (TP)

- Precision : the share of predicted positives $(\hat{Y} = 1)$ that are correct.

$$\textbf{Precision} = \text{TP} \ / \ (\text{TP} + \text{FP})$$

Predictions Assessment

The confusion matrix

Confusion Matrix: Election Prediction

	Actual Vote Choice	
Prediction	**Challenger**	**Incumbent**
Challenger	True Negative (TN)	False Negative (FN)
Incumbent	False Positive (FP)	True Positive (TP)

- F1 Score : the harmonic mean of precision and sensitivity, thus it is a good measure of the overall performance of the model.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} = \frac{2 \times \text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

Which one should we use?

- Question: which criterion should we use?
- Answer: use the *right* criterion based on specific context.
- Are true positives more valuable than true negatives?
  If yes, then *Sensitivity* will be the key.
  eg: Campaign wants to identify all potential incumbent voters for mobilization (missing supporters is costly)
- Do you want to have high confidence in predicted positives?
  Precision will be the key.
  eg: Targeted advertising budget - want to be sure before spending money on "incumbent voters"
- Are all errors equal?
  Accuracy is perfect.
  eg: General election forecasting where both types of prediction errors are equally important

Election Case: Model Performance Assessment

Confusion Matrix: Election Prediction Results

	Actual Vote Choice	
Prediction	**Challenger**	**Incumbent**
Challenger	167 (TN)	262 (FP)
Incumbent	87 (FN)	1484 (TP)

Model Performance Metrics

Metric	Value	Interpretation
Accuracy	**82.6%**	Overall correct prediction rate
Sensitivity	**94.5%**	Correctly identified incumbent voters
Specificity	**38.9%**	Correctly identified challenger voters
Precision	**85.0%**	Accuracy of incumbent predictions
F1 Score	**89.5%**	Balanced precision-recall measure

- Don't forget that all the rates are based on the **threshold** we set.

Multiple thresholds

- If we change the threshold, we will get different rates.
- Suppose we set the threshold to 0.3 instead of 0.5, then we will get the following rates:

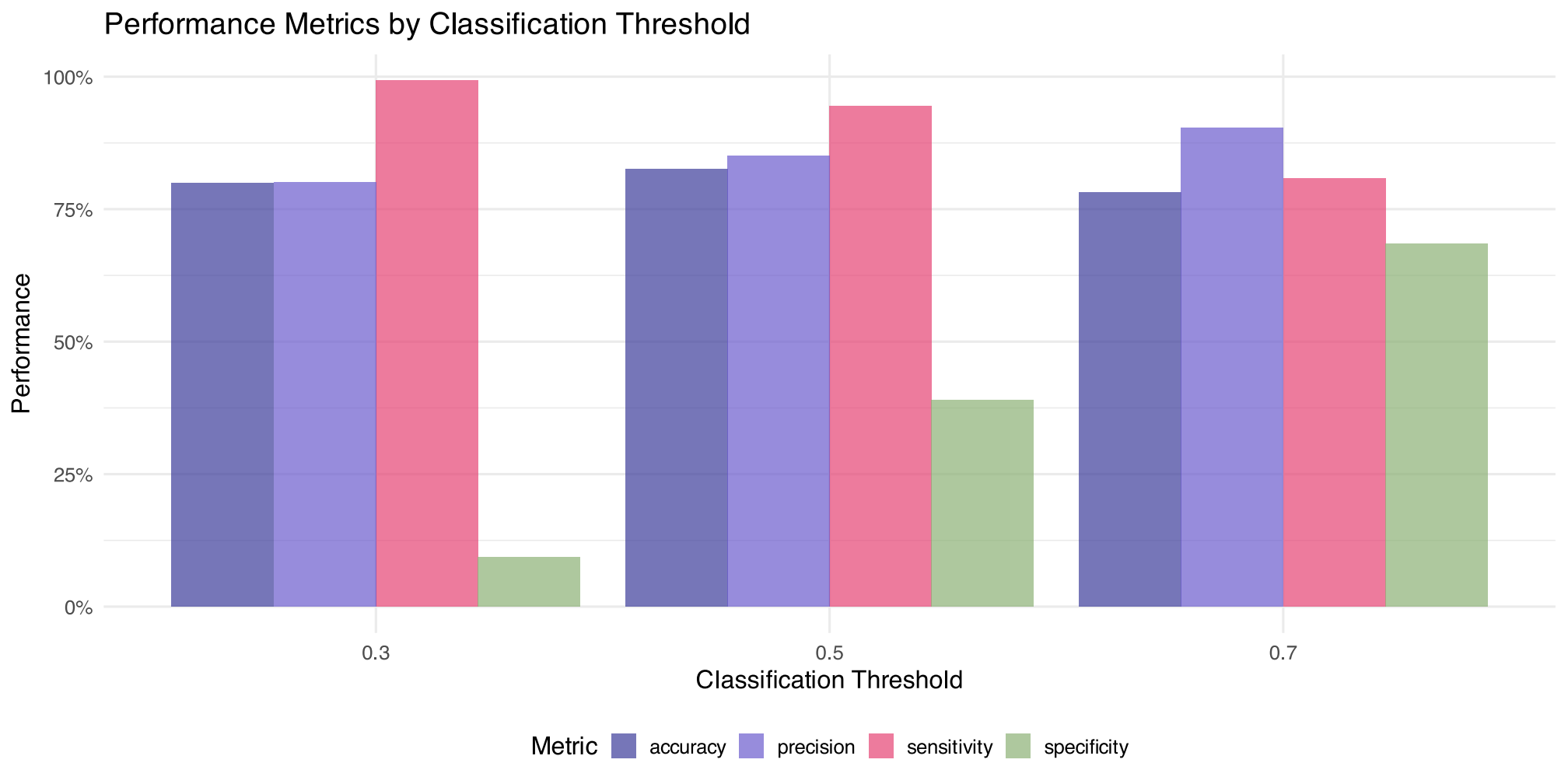
Confusion Matrix: Election Prediction Results

	Actual Vote Choice	
Prediction	**Challenger**	**Incumbent**
Challenger	40 (TN)	389 (FP)
Incumbent	12 (FN)	1559 (TP)

Model Performance Metrics

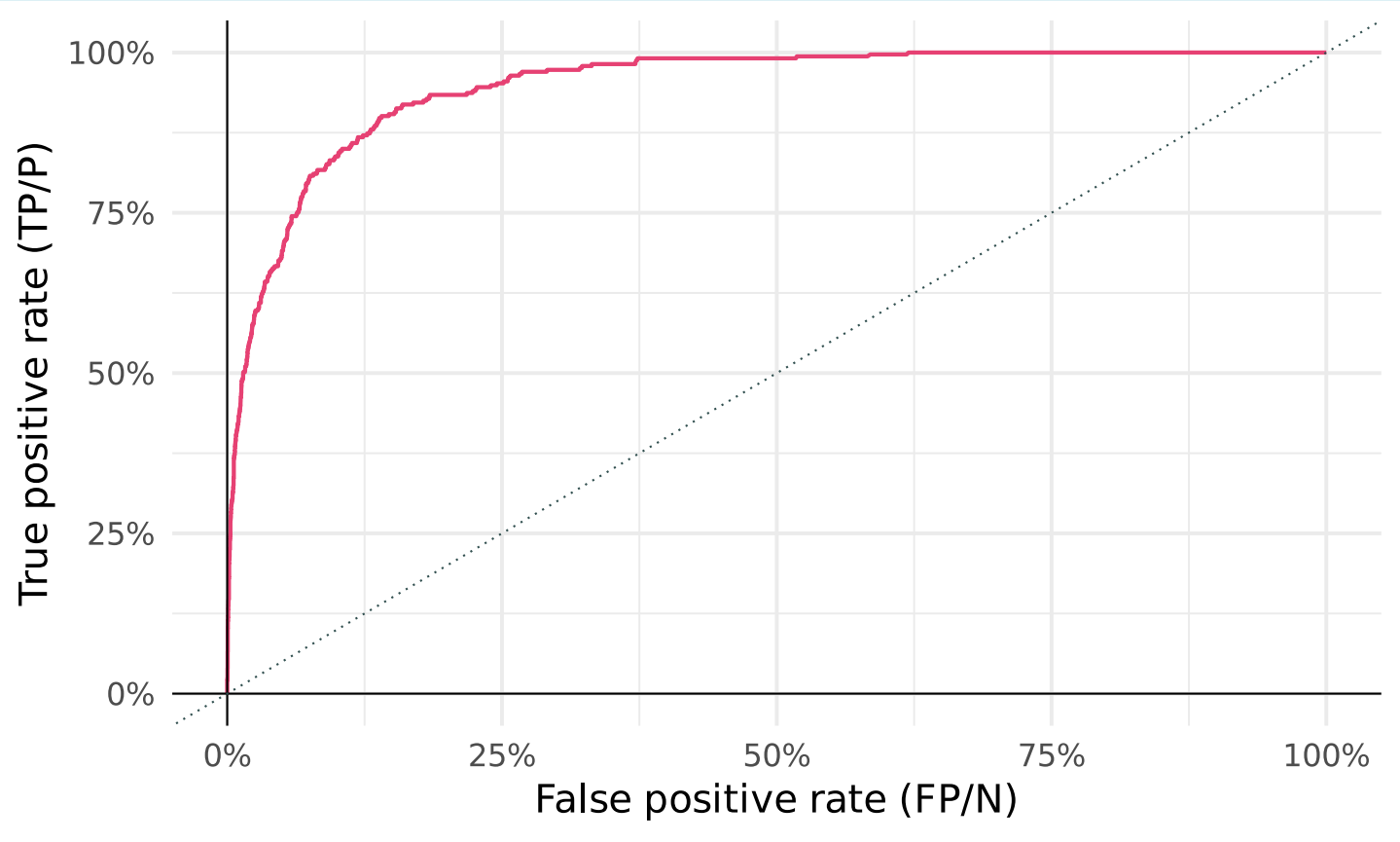
Metric	Value	Interpretation
Accuracy	**80.0%**	Overall correct prediction rate
Sensitivity	**99.2%**	Correctly identified incumbent voters
Specificity	**9.3%**	Correctly identified challenger voters
Precision	**80.0%**	Accuracy of incumbent predictions
F1 Score	**88.6%**	Balanced precision-recall measure

Multiple thresholds



Visualizing the Sensitivity-Specificity Trade-off

- Receiver Operating Characteristic (ROC) curve plots the true(TP/P) and false positive rates(FP/N).



Choosing the Optimal Threshold from ROC

- Question: How do we choose the best threshold from the ROC curve?
- Several approaches for threshold selection:

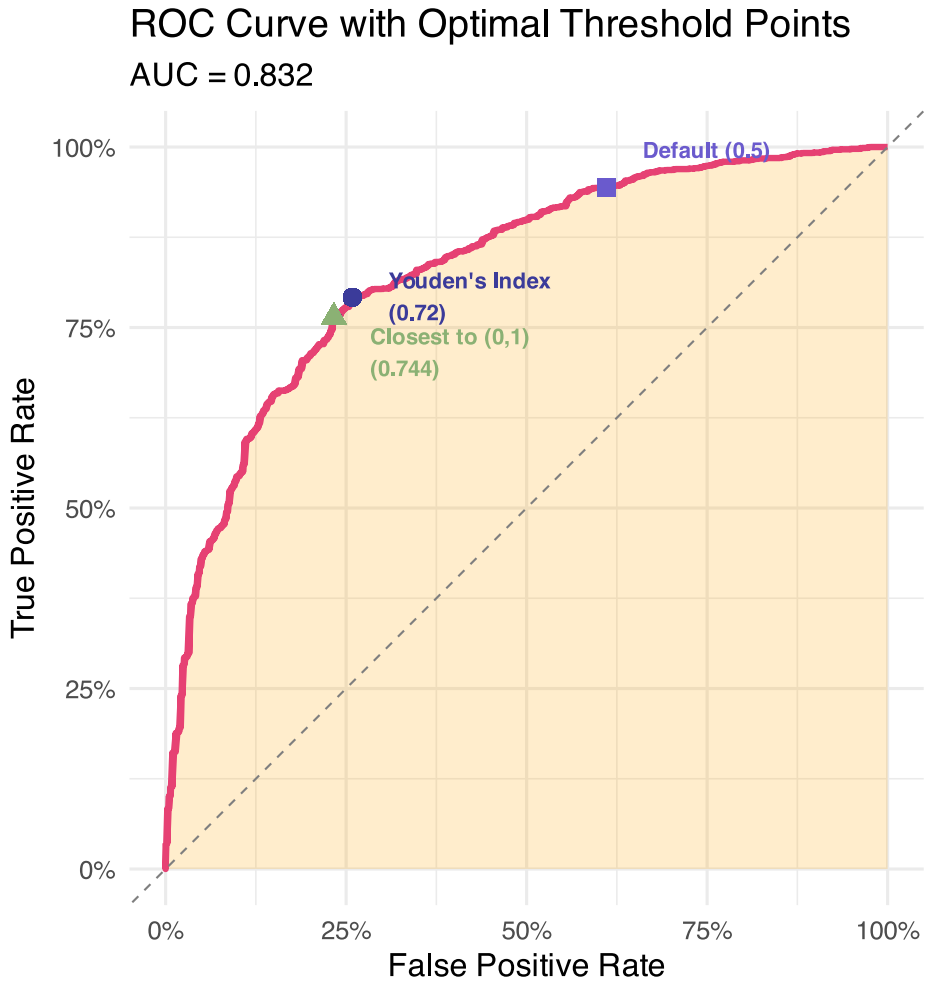
1. Youden's Index: Maximize (Sensitivity + Specificity - 1)
2. Closest to (0,1): Minimize distance to perfect classifier
3. Cost-sensitive: Based on relative costs of false positives vs false negatives
4. Domain-specific: Based on practical considerations

Comparison of Threshold Selection Methods

Method	Threshold	Sensitivity	Specificity	Youden_Index
Youden's Index	**0.720**	0.792	0.741	0.533
Closest to (0,1)	**0.744**	0.764	0.767	0.531
Default (0.5)	**0.500**	0.945	0.389	0.334

- Youden's Index often provides a good balance between sensitivity and specificity.

Choosing the Optimal Threshold from ROC



Threshold Selection: Practical Considerations

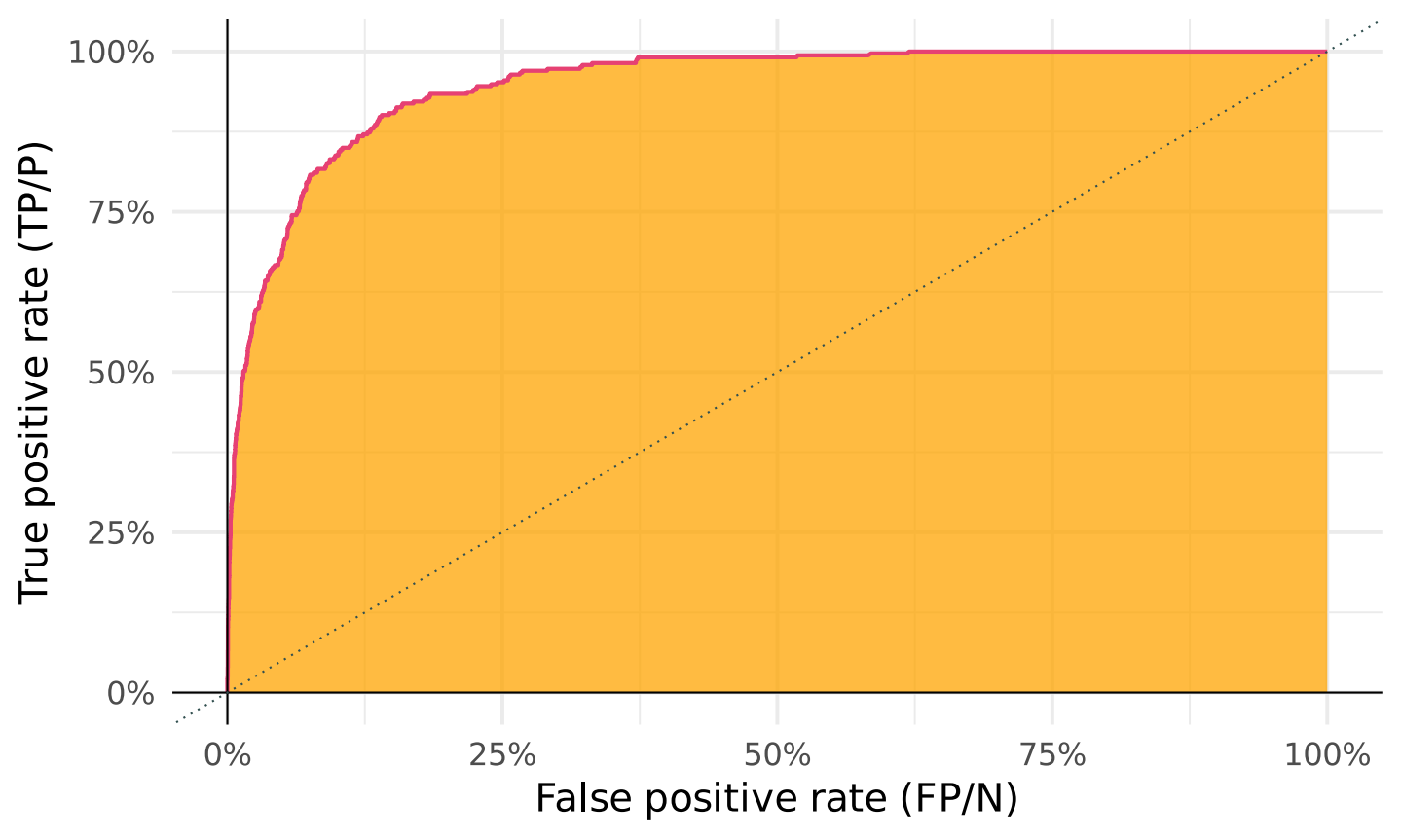
Threshold Selection by Campaign Context

Scenario	Priority	Reasoning	Suggested_Threshold
Voter Mobilization	High Sensitivity	Don't want to miss potential supporters	**Lower (0.3-0.4)**
Targeted Advertising	High Precision	Limited budget - need confident predictions	**Higher (0.6-0.7)**
General Polling	Balanced Accuracy	Equal importance of both voter types	**Optimal (0.5)**

Key Insights:

- No universal "best" threshold - depends on context and costs
- Lower thresholds → Higher sensitivity, more false positives
- Higher thresholds → Higher precision, more false negatives
- ROC curve helps visualize trade-offs but domain knowledge guides final decision

The Area Under the Curve (AUC)



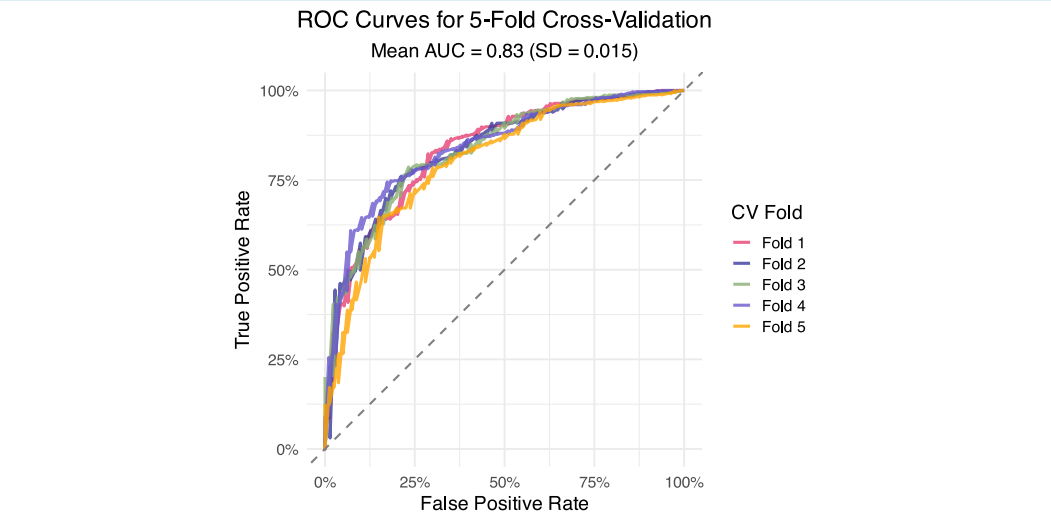
- It can be used to calculate the area under the curve (AUC) which is the probability that a random positive example is ranked higher than a random negative example.

Election Case: K-fold Cross-Validation

- Suppose that Accuracy is our criterion of interest.
- Then we will still use K-fold cross-validation to estimate the accuracy of the predictions.
- Suppose $K = 5$, then we could estimate the accuracy of the predictions as

$$\hat{CV}_{Accuracy} = \frac{1}{5} \sum_{k=1}^{5} \frac{TP_k + TN_k}{TP_k + TN_k + FP_k + FN_k}$$

Election Case: K=5 Cross-Validation ROC Curves



Cross-Validation AUC Results

Fold	AUC
1	0.834
2	0.834
3	0.834
4	0.842
5	0.804
Mean	**0.830**
SD	**0.015**