Predicting Student Self-Censorship: Behavioral Insights from the Campus Expression Survey

Cali Dutta

1. Problem Statement

College students increasingly report discomfort when discussing controversial topics in class, ranging from race and gender identity to political and religious views. This self-censorship threatens open discourse on campus and presents challenges for institutional leadership.

This project develops an applied machine learning tool to predict which students are most likely to self-censor using survey data from the 2024 Campus Expression Survey. The goal is to surface latent discomfort patterns and generate actionable predictions for administrators, policy think tanks, and civic engagement organizations working in higher education.

QUESTION: Can we use student-level survey data to predict who is most likely to self-censor in classroom discussions of controversial topics?

By segmenting and classifying students based on personal and attitudinal traits, this model provides decision-making value through decreased uncertainty, allowing targeted outreach, tailored dialogue programming, and risk monitoring on college campuses.

---

2. Key Result

The **final downsampled Random Forest model** (with independent variables only) achieved:

- **AUC = 0.659**
- **Sensitivity** = 0.528
- **Specificity** = 0.713
- **Balanced Accuracy** = 0.62

This modest but meaningful improvement avoids the failure of earlier models that overfit to class imbalance (e.g., AUC = 0.66 but Specificity = 0.00, indicating no ability to identify Not Reluctant students).

3. Example Survey Question

Students were asked:

> *"How comfortable or reluctant would you feel discussing YOUR HONEST THOUGHTS, IDEAS, AND QUESTIONS on Race or Ethnicity in a relevant classroom setting?"*

Responses range from **1 = Very Reluctant** to **4 = Very Comfortable**.
This format was repeated for **10 controversial topics**, including gender identity, abortion, COVID-19, and more.

---

4. Data Summary

This project uses the **2024 Campus Expression Survey (CES)**, administered by Heterodox Academy, containing:

- **4,733 student responses**
- **10 core outcome questions** on topic-based discomfort
- **40+ features**, including Big Five personality traits, political identity, peer consequence perceptions, institutional control (public vs. private), and year in school

Note: No state or institutional identifiers are provided, which limits contextual analysis.

---

5. Future Contextual Features: Campus-Level Variables

Although the CES dataset lacks school identifiers, future work could integrate contextual data to model campus climate effects. Here are possible merges:

Category	Example Variable	Potential Source
Geographic	State or Region	IPEDS, College Scorecard
Political	County-level voting behavior	MIT Election Lab
Religious	Religious affiliation	IPEDS
Demographic	Zip-level race/ethnicity	ACS Census
Institutional	Public vs Private, Size	IPEDS
Civic Climate	Cancel culture events	FIRE Tracker
Socioeconomic	Pell % or Tuition	College Scorecard

These enrichments would help disentangle personal reluctance from systemic climates.

---

6. Feature Engineering

- `reluctance_binary`: Engineered binary outcome. Coded **1** if a student felt "very" or "somewhat reluctant" on any controversial topic.
- `controversial_discuss_score`: Mean discomfort score across all 10 topics (continuous).
- `cluster` and `profile`: Latent structures created via PCA/KMeans and LPA respectively, but excluded from final modeling due to outcome entanglement (see Section 8).

---

7. Exploratory Analysis

- **Completion check**: Used `mice` to impute missingness; retained 4,730 full rows.
- **Distribution**: 91% of students were coded as "Reluctant", confirming **class imbalance**.
- **PCA** revealed a dominant "general reluctance" factor (PC1 = 42.1% of variance).
- **LPA** identified 4 profiles showing thematic discomfort patterns (e.g., identity-specific vs. universal avoidance).
- **Cluster analysis** (K-means on PC1/PC2) provided early segmentation, but explained mostly overall intensity — not distinct types.

---

8. Modeling Approaches Considered

8.1 Logistic Regression

- AUC = **0.93**, but...
- Predicted nearly all students as "Reluctant"
- **Sensitivity** = 1.00; **Specificity** = 0.00

- *Red flag: model learned only class prevalence*

8.2 PCA / Cluster-Based Models

- Cluster predicted Reluctance perfectly (Accuracy = 99.6%)
- But profiles and clusters were structurally entangled with outcome:
  - E.g., Profile 2 = 432 students, 100% "Not Reluctant"
  - Profile 3 = 1,363 students, 100% "Reluctant"
- Overperformance due to leakage, not generalization
- Not suitable for prediction, though useful for qualitative insight

---

8.3 Independent Random Forest (No Leakage)

- Trained only on clean features:
  - Personality traits (Neuroticism, Extraversion)
  - Political identity (`politics_overall_num`)
  - Social risk (`stdnt_social_media`, `stdnt_file_complaint`)
  - Year in school, institutional control
- AUC = **0.6575**
- **Sensitivity** = 1.00
- **Specificity** = 0.00
- Still overwhelmed by class imbalance

8.4 Final Model: Downsampled Random Forest

To address class imbalance, I implement downsampling using `caret::train()` with the `sampling = "down"` option and evaluated performance via 5-fold cross-validation.

Performance Metrics:

- **AUC**: 0.659
- **Sensitivity**: 0.528
- **Specificity**: 0.713
- **Balanced Accuracy**: 0.62

This was the best-performing and most realistic model, avoiding structural leakage and overfitting. It achieved a good balance between identifying Reluctant students and correctly recognizing those who are not.

Top Predictors (by variable importance):

- **Neuroticism** (100.0)
- **Extraversion** (95.7)
- **Political Identity** (`politics_overall_num`) (65.8)
- **Year in School** (`yearinschool_num`) (62.7)
- **Formal Sanctions + Peer Risk Factors** (`stdnt_file_complaint`, `stdnt_social_media`) ~18–19

These predictors reflect psychological predispositions, ideological lean, and situational fear — all of which are theoretically plausible drivers of self-censorship.

---

9. Model Comparison Summary

Model	AUC	Sensitivity	Specificity	Notes
Dummy (All Reluctant)	—	1.00	0.00	Null baseline; reflects class imbalance
Logit (All Predict Reluctant)	0.93	1.00	0.00	Overfit; mirrors dummy
Cluster/Profile	0.99+	1.00	1.00	Structural leakage
Random Forest (Unbalanced)	0.657	1.00	0.00	Learned class only
RF Downsampled (Final)	0.659	0.528	0.713	Best generalization without leakage

---

10. Why I Chose This Approach

- Handles non-linearities: Random Forest captures complex interactions (e.g., between neuroticism and political identity) that linear models miss.
- Robust to mixed variable types: Handles both continuous and categorical predictors without transformation.
- Downsampling: Effectively corrected for severe class imbalance, which overwhelmed prior models.
- Cross-validation (5-fold): Provided reliable performance estimates and guarded against overfitting.

Alternative methods like **SMOTE** were considered but not used due to concerns over synthetic data in a sensitive binary labeling context. Downsampling offered a simpler, interpretable, and empirically balanced solution for identifying students likely to self-censor.

---

11. Risks & Limitations

- No contextual variables: The CES dataset lacks school, state, or region identifiers, preventing us from modeling broader ideological or institutional climates.
- Strict outcome labeling: A student is labeled as "Reluctant" if they report discomfort on even one topic, which may inflate the binary target and obscure partial openness.
- Structural overlap in early models: Predictors like `cluster` (PCA-based) and `profile` (LPA-based) were later shown to encode the outcome directly, undermining model validity when used for supervised learning.
- Still only moderate AUC (~0.66): This suggests that while psychological and demographic factors provide signal, much of expressive discomfort may arise from unobserved peer dynamics, current events, or local culture.
- Interpretability tradeoff: Random Forest outperformed logistic models in identifying both classes but sacrificed interpretability in favor of classification performance.
- Imputation assumptions: Multiple imputation (mice) was used to address missing values, but assumes missingness is at random. If discomfort patterns are systematically missing (e.g., students skipping uncomfortable topics), estimates may be biased.
- Class imbalance and thresholding: Even with downsampling, the underlying distribution remains skewed, and threshold decisions (e.g., 0.5 in classification) may not reflect meaningful behavioral differences between marginal and strongly reluctant students.

- Survey response bias: As with all self-reported data, social desirability and interpretation of questions may skew responses — especially on politically or emotionally charged topics — introducing unmeasured noise into the outcome.
- Limited generalizability: While the dataset is large, it may not fully represent students outside of the CES sample (e.g., community colleges, religious institutions), limiting broader application without additional validation.

---

12. Data Justification

The CES dataset is high-dimensional and behaviorally rich, making it well-suited for machine learning applications:

- Contains 10 interrelated outcome variables (topic-specific discomfort), which justified the use of dimensionality reduction techniques like PCA.
- Feature engineering expanded the dataset to include ~15 predictive variables, such as `controversial_discuss_score`, `any_sanction`, and personality measures.
- The data includes behavioral, ideological, and social risk indicators, many of which exhibit non-linear interactions — supporting the use of flexible algorithms like Random Forest.
- The dataset is managed by Heterodox Academy, the original source of data collection and curation, ensuring reliable documentation, interpretability, and data governance — in contrast to third-party aggregators.

12.1 Robustness & Reliability Check

I evaluated whether the model's results reflect real signal or spurious patterns through a combination of cross-validation, baseline comparisons, feature leakage testing, and external theoretical alignment.

Cross-Validation (CV)

All supervised models were trained and evaluated using **5-fold cross-validation** with `caret::train()`, using `summaryFunction = twoClassSummary` and appropriate resampling (e.g., downsampling for imbalance). This provided out-of-sample performance metrics for AUC, sensitivity, and specificity.

The final downsampled Random Forest model, trained on independent variables only, achieved the following average performance:

- **AUC**: 0.659
- **Sensitivity**: 0.528
- **Specificity**: 0.713
- **Balanced Accuracy**: 0.620

These results remained stable across folds, suggesting the model is not overfitting and generalizes to unseen data within this sample.

Baseline Comparison

I implemented a dummy baseline model that predicted all students as "Reluctant" (the majority class). While it achieved high accuracy (~91%), its balanced accuracy was just 0.50, confirming it could not distinguish between classes.

This reinforces that the model offers added predictive value, especially in correctly identifying the minority class (Not Reluctant).

Feature Leakage & Subsetting

I tested models that included structurally entangled features (`cluster`, `profile`) and found perfect or near-perfect performance, but only because these features encode the outcome variable. Once removed, performance dropped appropriately, confirming the final model is not driven by leakage or spurious correlations.

External Validity & Theoretical Alignment

The most important predictors — neuroticism, extraversion, political orientation, and year in school — align with established behavioral research:

- Neuroticism is linked to higher social threat sensitivity
- Extraversion is associated with verbal openness and assertiveness
- Political ideology interacts with perceived campus norms to influence expression

This theoretical alignment adds external credibility to the patterns discovered, suggesting the model captures real psychological and social dynamics.

---

13. Conclusion

This project used machine learning to tackle a complex, socially relevant problem: predicting which students are most likely to self-censor during classroom discussions of controversial topics. By applying Random Forest classifiers to a rich dataset of personality traits, political views, and peer risk perceptions, I was able to identify moderate but consistent patterns of expressive discomfort.

Machine learning was especially useful in: - Handling high-dimensional, non-linear relationships between predictors - Segmenting students based on latent reluctance profiles (LPA) - Providing predictive structure even when traditional models failed under class imbalance

However, I am cautious about overclaiming. While the model achieves a balanced accuracy of **62%**, it still misclassifies many students, and its performance is highly dependent on the limited survey data available. Importantly, I deliberately excluded features that encoded the outcome directly (e.g., `profile`, `cluster`) to maintain integrity — reducing performance but ensuring generalizability.

The findings should be viewed as a directional diagnostic tool, not a definitive classifier. They highlight that traits like neuroticism, extraversion, and political ideology carry meaningful signal — but cannot fully explain the contextual and emotional complexity of student speech behavior.

In sum, machine learning helped us identify, quantify, and compare patterns of reluctance in a way that supports future targeted interventions — while also revealing the limits of prediction in socially sensitive domains. The model is not intended as a classifier for individuals, but rather as a tool for institutions seeking to monitor trends, surface risk, and develop programming that better supports open expression.