

# STAT 425 Graduate Project Proposal

**Name:** Sheetal Tewary

**NetID:** stewary2

---

## Summary and Motivation (1–2 paragraphs, 4 pts)

**Project Title:** *Tech Layoffs: Probability of Occurrence, Roles Impacted, and Event Magnitude*

The technology sector has experienced repeated waves of layoffs across multiple industries and regions. To understand this pattern in depth, this project studies **three complementary questions**:

- 1) *Probability*: Which company characteristics predict **whether** a firm lays off in a given period?
- 2) *Roles*: Which **job roles** (e.g., Engineering, Product, Sales/Marketing, HR/People Ops, Support/Operations) are disproportionately impacted during layoff events?
- 3) *Magnitude*: Conditional on a layoff occurring, what factors drive the **number of employees laid off**?

The study demonstrates STAT 425 competencies in **linear modeling, logistic regression, regularization, ANOVA/ANCOVA, splines, diagnostics, and mixed-effects**. Results can inform workforce-planning discussions and provide a rigorous, reproducible analysis using public data.

---

## Methods (1–3 paragraphs, 4 pts)

### A) Probability of Layoff (Binary Classification)

- **Unit of analysis:** Company–month panel
- **Response:** `layoff_event` {0,1} indicates any layoff by the firm in month  $t$ .
- **Predictors:** `employee_count`, `funding_total_usd`, `stage` (seed/early/late/public), `industry` (SaaS, FinTech, Hardware, etc.), `country/region`, and **time splines** for month index.
- **Models:** Logistic regression (baseline) and **penalized logistic lasso** (`glmnet`) for feature selection and multicollinearity control.
- **Evaluation:** 5–10 fold CV; **ROC/AUC**, precision/recall; calibration plot; confusion matrix at selected thresholds.
- **Diagnostics:** Influence diagnostics, separation checks, multicollinearity review, and comparison of with/without splines.

### B) Roles Impacted (ANOVA/ANCOVA and Multinomial Modeling)

- **Objective:** Test whether certain **job roles** are disproportionately affected, accounting the industry.
- **Role construction:** Derive role labels from event notes/press releases using text rules/regex (e.g., “engineer/dev/data/QA” → *Engineering*; “PM/product” → *Product*; “sales/marketing/advertising” → *Sales/Marketing*; “HR/recruiting/people/talent” → *HR/People*; “support/ops” → *Support/Ops*). Where unavailable, treat as *Unspecified*.
- **Analyses:**
  - **Two-way ANOVA/ANCOVA:** Role share (e.g., percent of laid-off employees in a role) ~ `industry` + `stage` + covariates (size/funding).
  - **Multinomial logistic (optional):** Predict the **most-impacted role** category using firm/industry features.
- **Diagnostics:** Normality/variance checks for ANOVA residuals; multiple-comparison (Tukey) for role contrasts.

### C) Magnitude of Layoff (Continuous Regression)

- **Response:** `employees_laid_off` + 1 for each event.
  - **Predictors:** `employee_count`, `funding_total_usd`, `stage`, `industry`, `country`, **time splines**, and proportions of roles affected (if available) to test compositional influence.
  - **Models:** Multiple linear regression (baseline), **ridge/lasso** regularization; **GLS** with AR(1) if temporal correlation persists; **mixed-effects** with random intercepts for firm or industry if repeated events per firm.
  - **Evaluation & Diagnostics:** CV-RMSE/MAE, Adjusted ( $R^2$ ), AIC/BIC; residual vs fitted, QQ-plot, Cook's distance; consider WLS if heteroscedasticity detected.
- 

### Data (4 pts)

- **Primary layoff events:** Public tech layoff trackers (Layoffs.fyi mirrors/Kaggle), including: `company`, `date`, `employees_laid_off`, `percentage_laid_off`, `industry`, `country`, and any textual **notes** that may mention **roles**.
  - **Firm features:** Where available, `employee_count` (headcount), `funding_total_usd`, and `stage`. For public firms, supplement headcount from annual filings when possible.
  - **Role labels:** Create using keyword-based tagging from event **notes/press**. Maintain a lookup table for regex patterns → role category. Keep an *Unspecified* bucket to avoid bias.
  - **Panel build (Probability model):** Create company-month grid, mark `layoff_event` if any event occurs that month; left-join firm features; add month index and spline basis.
  - **Cleaning:** Harmonize company names; deduplicate events; winsorize top 1% of numeric predictors; standardize numeric features for penalized models.
- 

### Results (4 pts)

#### Anticipated findings:

- **Probability:** Larger firms and later-stage companies are more likely to announce a layoff in a given month; industry and time splines capture sectoral cycles. Penalized logistic models

expected to improve AUC and yield a concise feature set.

- **Roles:** HR/People Ops and Support/Ops may show higher proportional impact early in cost-cut cycles; Engineering/Product share may vary by industry and period. ANOVA/ANCOVA should identify statistically significant role contrasts after adjusting for firm size and stage.
- **Magnitude:** Conditional on occurrence, layoff counts increase with firm size and funding; splines reveal surge periods. Regularization expected to reduce prediction error vs OLS.

---

## Conclusions (2 pts)

**Research questions to answer:**

- 1) Which firm features best predict **whether** a layoff happens?
- 2) Which **roles** are disproportionately affected, controlling for size, stage, and industry?
- 3) Among layoff events, what drives the **severity** of cuts?
- 4) Do time trends and sector differences materially shift these relationships?

---

## References (2 pts)

- Faraway, J. J. (2002). *Practical Regression and ANOVA using R*. CRAN.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *JRSS B*, 58(1), 267–288.
- James, G., Witten, T., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with R (2nd ed.)*. Springer.
- Public tech layoff trackers and cleaned CSV mirrors (2020–2025) for event- and firm-level features.
- Layoffs.fyi (2025). Tech Layoff Tracker Dataset. <https://layoffs.fyi>
- Kaggle (2025). Tech Layoffs 2020–2025 Dataset (cleaned mirror).