The Rise of AI and the Fall of Juniors? Analyzing Tech Industry Layoffs from 2020 to 2024

PROJECT PROPOSAL

Group:	22
Contributors:	Ahmet ATAR — 22290230, Yiğit GÜLBEYAZ — 22290275
Course:	COM2502 Introduction to Data Science
Date:	02.04.2025

Introduction

Background and Motivation

The emergence and rapid proliferation of Generative AI technologies—such as ChatGPT, GitHub Copilot, and similar tools—have brought transformative changes to the field of software development. These tools are capable of assisting with or even automating various programming tasks, thereby reshaping the dynamics of developer productivity and the broader software engineering landscape. While they offer significant gains in efficiency, their disruptive potential raises essential questions about workforce implications, particularly for entry-level roles such as junior software developers.

This project aims to critically examine the hypothesis that the adoption of Generative AI has contributed to a decline in demand for junior developers in the tech sector. Rather than approaching this issue purely through qualitative observation, the study will use quantitative data analysis techniques to determine whether observable patterns support this concern.

Research Objectives

The primary goals of this study are as follows:

- To analyze layoff trends in the global tech industry from 2020 through 2024.
- To explore whether the period following the widespread introduction of Generative AI (beginning in late 2022) correlates with increased layoffs, particularly in junior-level roles.
- To investigate whether factors such as company size, industry type, or investment stage influence layoff decisions.
- To construct predictive models to estimate layoff likelihood and magnitude using historical company and industry data.
- To present findings in a manner that is both visually intuitive and analytically rigorous.

Dataset Description

This study will leverage a publicly available dataset titled **"Tech Layoffs 2020–2024"**, sourced from **layoffs.fyi** and hosted on Kaggle. The dataset consists of 1,672 records and 16 features that document real-world tech layoffs across multiple countries. Key attributes include:

- Company name and location
- Country and continent of headquarters
- Number of employees laid off
- Total company size before and after layoffs
- Industry classification and investment stage
- Date of layoffs and year of event
- Total capital raised

The dataset satisfies all requirements set by the course: it is publicly accessible, contains well over 1,000 records, and includes multiple numerical and categorical features suitable for a variety of machine learning techniques.

Methodology

Data Preprocessing

The dataset will first undergo comprehensive preprocessing to ensure consistency and analytical readiness. This will include:

- Parsing and formatting dates to enable time-series analysis
- Imputing or removing missing values in fields such as funding and company size
- Normalizing and encoding categorical variables (e.g., industry, country, company stage)
- Creating new binary and ratio features (e.g., "layoff_occurred", "layoff_ratio")

Exploratory Data Analysis (EDA)

Following preprocessing, extensive EDA will be conducted to uncover patterns and insights:

- Temporal analysis to track layoff trends before and after the emergence of Generative AI
- Comparative analysis across industries, continents, and company stages
- Geospatial visualizations to highlight regions most affected by layoffs
- Correlation heatmaps and scatterplots to examine interdependencies between variables

Feature Engineering

Custom features will be engineered to enrich the model's predictive capabilities, including:

- Ratio features such as layoffs as a percentage of company size
- Categorical time segments (e.g., Pre-AI Era vs. Post-AI Era)
- Funding tiers based on total capital raised

Predictive Modeling

Two primary modeling objectives will be pursued:

- **Classification** – Predicting whether a company experienced a layoff event using logistic regression, decision trees, and ensemble methods such as random forests
- **Regression** – Estimating the number of laid-off employees using models such as linear regression, ridge regression, and tree-based regressors

Model Evaluation

The following metrics will be used to evaluate and compare model performance:

- For classification: Accuracy, Precision, Recall, and F1-Score
- For regression: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared

Tools and Technologies

The entire analysis will be conducted using Python, primarily in Jupyter Notebook. The following libraries and frameworks will be utilized:

- Pandas and NumPy for data manipulation
- Matplotlib and Seaborn for data visualization
- Scikit-learn for machine learning and model evaluation

References

1. Herold, U. (2024). Tech Layoffs 2020–2024 Dataset, Kaggle. https://www.kaggle.com/datasets/ulrikeherold/tech-layoffs-2020-2024
2. Layoffs.fyi (2024). Real-time Layoffs Tracker. https://layoffs.fyi
3. OpenAI. (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4
4. GitHub Copilot Documentation. (2023). https://docs.github.com/en/copilot