Predicting Company Layoffs

ale22, iacosta1, kwang111, cpak4

Goal:
With the recent economic downturn ensuing the Covid Pandemic, many US companies have experienced mass layoffs. Tech giants like Microsoft, Amazon and Facebook have laid off thousands of employees. Our team is primarily interested in the following task: <u>given 2020-2023 layoff data from over 1000 companies, can we create a model using company attributes like stock price and revenue to predict if a company will have another layoff and how many employees will be laid off.</u>

Data:
We collected the data from Kaggle, Crunchbase and IEX API. The stock history data was obtained from IEX Cloud, while the layoffs data was collected from Kaggle's layoffs.fyi dataset, which tracks layoffs in tech companies and startups. The company_attributes data was downloaded from a Crunchbase dataset on a publically accessible Github repository. To enhance the Crunchbase dataset, additional columns were added using information from the Kaggle "Fortune_1000" dataset. While the sources are generally reputable, some limitations and biases should be noted. The samples used are random but may be biased towards tech companies and startups. The data is mostly representative of American economic activity, and some caution is advised when relying on user-uploaded data. We conducted 3 statistical tests on the data and concluded that none of our hypotheses were significant. We first hypothesized that there is a significant association between tech companies and whether their layoff rate is above the mean of the dataset. We also hypothesized that there is a significant difference in average stock prices for each company between August 2022 and February 2023. Finally, we hypothesized that there is a significant difference between layoff rates in 2021 and 2022. However, all of our P values were above 0.05 and thus, we failed to reject the null hypothesis.

Model and Evaluation Setup
We primarily focused on predicting the number of employees being laid off from a company using Gamma Regression because it is a regression technique that can prevent overfitting in a linear regression model and guarantee positive predictions. We also use RandomForest and decision trees to determine if there will be another layoff and which attributes contribute most to the prediction of future layoff. Because most of our data was biased to the US, we did not expect to generalize to companies in different countries. We only trained on data from 2020-2023 and the features contributing to a current outcome are heavily influenced by other economic factors that were unable to be captured by the datasets. Thus, the models are only reliable when predicting layoffs within this timespan within the US. To evaluate the performance of our model, we partitioned our data into two sets: a training set comprising 75% of the data, and a testing set consisting of the remaining 25%. Our assessment of accuracy was based on the Mean Squared Error (MSE) metric.

Results and Analysis:

Claim #1: Average volume, number of acquisitions, estimated revenue range, and state HQ being in Washington are the 4 important features in determining the scale of a layoff.

Support for Claim #1:
Comparing the top 10 coefficients in terms of magnitude from our Ridge Regression and from our Gamma Regression, we find that the four features that are included in both tables are average volume, number of acquisitions, estimated revenue range, and state HQ being in Washington. It is interesting to note that features with negative coefficients from Ridge are not amongst the top 10 for Gamma. This may be because Gamma ensures positive predictions and thus puts more emphasis on finding positively correlated features.

Top 10 coefficients in terms of magnitude from Ridge Regression:

avg_volume	5095.085763255015
num_acquisitions	4250.273127120718
Industry_Hardware	-3194.553382090601
total_funding	-1187.0025877540425
estimated_revenue_range	912.984467283423
founded_date	-678.2668960401885
avg_change_percent	-421.22715465410585
state_hq_Washington	374.5208594030576
state_hq_Michigan	-366.40659144138334
Industry_Retail	339.7902347177163

Top 10 coefficients in terms of magnitude from Gamma Regression:

constant	4.8961510311623995
num_employees	0.4293287864613737
stage_Post-IPO	0.3086272719098574
estimated_revenue_range	0.302439138355664
ticker	0.29292790195335283
acquisition_status	-0.2527691669674705
avg_volume	0.1645392429493997
num_acquisitions	0.15533277875549176
multiple	0.13880000544525226
state_hq_Washington	0.11457918783290658

Claim #2: Technology companies have different sets of important features for determining the scale of the layoff as compared to non-tech companies.

Support for Claim #2:
Coefficients for our Ridge and Gamma models for layoff data from companies labeled as a tech company. Our previous Chi-Square hypothesis testing concluded no significant difference between layoff rates of tech versus non-tech companies. Nonetheless, we find through our regression models that the features determining the scale of layoffs differ between tech and non-tech companies. The top 5 features for tech companies in ridge regression are: num_acquisitions, Industry_Hardware, founded_date,

estimated_revenue_range, and total_funding. This is different from the top 5 features for non-tech companies in ridge regression which are: avg_volume, num_employees, Industry_Support, avg_close, and Industry_Product. Similarly, these features are different for gamma regression results as well as the top 5 for tech companies are num_employees, ticker, estimated_revenue_range, stage_Post-IPO, and acquisition_status while the top5 for non-tech companies are multiple, stage_Post-IPO, acquisition_status, num_employees, and state_hq_New York. Thus, both regression models demonstrate how features in determining layoff count for tech versus non-tech companies different significantly.

Coefficients for our Ridge and Gamma models for layoff data from companies labeled as a tech company.

Top 10 coefficients in terms of magnitude from Ridge Regression:

num_acquisitions	6079.520859882977
Industry_Hardware	-1806.85601651443
founded_date	-883.9606266899092
estimated_revenue_range	882.8149396944001
total_funding	-672.620982356271
Industry_Consumer	404.8311229857972
Industry_Infrastructure	-367.660387916765
state_hq_Arizona	328.7678352965678
avg_volume	-320.5845896299044
Industry_Transportation	319.6863662718202

Top 10 coefficients in terms of magnitude from Gamma Regression:

constant	4.843871968984618
num_employees	0.43933625269491394
ticker	0.37754602885739774
estimated_revenue_range	0.3550372563024669
stage_Post-IPO	0.3404542537889789
acquisition_status	-0.2638765179561676
num_acquisitions	0.24900268725235172
state_hq_New York	-0.12591672310646268
stage_Series C	-0.11742526794271478
multiple	0.11151799558681508

Coefficients for our Ridge and Gamma models for layoff data from companies labeled as a non-tech company.

Top 10 coefficients in terms of magnitude from Ridge Regression:

avg_volume	1552.282835520634
num_employees	649.9003235320264
Industry_Support	-377.628974597995

Top 10 coefficients in terms of magnitude from Gamma Regression:

constant	4.866887150401001
multiple	0.2983377906614766
stage_Post-IPO	0.2318887381309567

avg_close	312.2550603359858
Industry_Product	241.46348249326692
total_funding	-233.9252800683768
trend_score_30	232.00805020463432
num_acquisitions	-224.88202420714057
estimated_revenue_range	-215.45877518872425
stage_Unknown	207.38391372989125

acquisition_status	-0.2230640642567936
num_employees	0.22027832718703896
state_hq_New York	0.20658657147236492
Industry_Real Estate	0.1669513585293378
ticker	0.16693878858012412
avg_volume	0.1237796013416466
state_hq_California	-0.10501386777473735

Claim #3: Industry being consumer, average volume, industry being healthcare, and industry being support are the top 4 features in determining if a company will have another layoff given that the company performed a layoff during 2020~2023.

Support for Claim #3:
In order to determine the most important features in determining the likelihood of future layoffs, we first determined the companies that have multiple layoff entries in our layoffs database, which we used as the label. Using the random forest classifier from a library called XGBoost, we plotted the most important features and ranked it by the F1 score, which measures a model's accuracy using precision and recall. Notice the F1 score on the graph is not the absolute value, but scaled so that it shows the relative contribution of each feature for each tree in the model. From this, we determined that companies in the consumer, healthcare, and support industries as well as average stock volume changes have the highest chance of having another wave of layoffs.