# School of Computer Science and Engineering

# COMP\_SCI-5588-0001-16348-2024SP-DS Capstone

Thambabattula Ramya Kumari

**Professor: Yugyung Lee**

**Technical Report : Assignment - 4**

**Introduction**

In today's dynamic business environment, organizations often face the challenging decision of downsizing their workforce due to various economic and operational factors. Identifying potential layoffs in advance can help companies strategize and mitigate the negative impacts on both employees and the organization. This study aims to develop a machine learning model that can predict employee layoffs based on historical data, enabling proactive measures and informed decision-making. This report presents an analysis of employee layoff data using a machine learning model to predict layoffs based on various features. The goal is to develop a model that can accurately identify potential layoffs, which could assist companies in making informed decisions and implementing preventive measures.

**DATASET**

The dataset used in this analysis is stored in the file 'layoffs.csv' and contains information about employees, including their location, industry, date, country, and whether they were laid off (indicated by the 'percentage\_laid\_off' column). The data is preprocessed, and the 'percentage\_laid\_off' column is converted to binary labels (0 for employees not laid off, and 1 for those laid off.

**Data Analysis**

The dataset is split into training and testing sets, the data used for training and testing. The categorical features (location, industry, date, and country) are one-hot encoded, and numerical features are standardized using Standard Scaler. The processed data is then used to train and evaluate the machine learning model.

**Results:**

A graph of training loss

Description automatically generated

Image 5: The image displays a line plot depicting the training and validation loss curves throughout the training of a machine learning model. The y-axis represents the loss value, while the x-axis represents the epoch (iteration) number. Initially, both the training and validation losses are high, but they gradually decrease as the training proceeds. The validation loss curve closely tracks the training loss curve, suggesting that the model is neither overfitting nor underfitting. This plot serves as a valuable tool for monitoring the model's convergence during training and identifying potential issues such as overfitting or underfitting.

A graph of a training and validation accuracy

Description automatically generated

Image 6: The graph shows the training and validation accuracy curves for a machine learning model during the training process. The training accuracy curve shows a steady increase as the model learns from the training data over epochs. However, the validation accuracy curve initially rises but then plateaus or even declines slightly, indicating overfitting - the model is starting to memorize the training data instead of generalizing well to unseen data.

A screen shot of a graph

Description automatically generated

Image 13: The ROC curve illustrates the true positive rate plotted against the false positive rate across different threshold settings. Here, the area under the ROC curve is 1.0, suggesting flawless classification performance.

**Model**

The **distilbert-base-uncased-finetuned-sst-2-english model** is a variant of the **DistilBERT** model, which itself is a distilled version of the BERT (Bidirectional Encoder Representations from Transformers) model. This specific variant has been fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset, which is a benchmark dataset for sentiment analysis tasks. Here's what each part of the model’s name signifies. The model used in this analysis is a KNeighborsClassifier wrapped in a CalibratedClassifierCV for better probability estimation. The model is trained on preprocessed data using a Pipeline that includes the column transformer for feature preprocessing and the classifier. In summary, the distilbert-base-uncased-finetuned-sst-2-english model is a smaller and faster version of BERT that has been fine-tuned on the SST-2 dataset for sentiment analysis, making it particularly suitable for classifying the sentiment (positive or negative) of English text.

1. Converts the 'percentage\_laid\_off' column to binary labels (0 or 1) using a lambda function.
2. Drops the 'percentage\_laid\_off' column from the feature set (X) and separates the target variable (y).
3. Splits the data into training and testing sets.
4. Defines categorical and numerical columns for feature preprocessing.
5. Creates a Column Transformer to apply StandardAero for numerical features and One Hot Encoder for categorical features.
6. Trains the model on the training data.
7. Evaluates the model's performance on the test data by calculating accuracy, precision, recall, and ROC AUC score (if possible). The model's performance is evaluated using various metrics, such as accuracy score (if available). That provides a visual representation of the model's predictions, allowing for further analysis and interpretation.

Overall, this report presents a comprehensive analysis of employee layoff data using a machine learning model. The model's performance can be further improved by exploring different algorithms, feature engineering techniques, or acquiring additional relevant data.