Inter IIT Tech Meet 10.0 - Digital Alpha

March 2022

1 Problem Statement

SaaS companies are customer-driven and are heavily dependent on their customer base. There are a set of metrics that can showcase the health of the SaaS companies and their aspects such as **cash and cash equivalents, total current liabilities, total current assets, etc**. Since these matrices are not readily available on publicly reported SEC Fillings, we will chalk out these values from the available 10-K, 10-Q 8-K forms. And we'll use these values to determine the health of SaaS companies and present it on our interactive dashboard so that users can get it easily.

2 Brief Description of codes

2.1 ML Model

- Input data files were generated from the API calls with the help of tickers integrated by us in different python files. For each data file of 8-K, we obtained the dates of filing of that particular form.
- Our target labels are implemented as the following
  - If the percentage change of close market value for two consecutive filings (interval of three months) is less than 2% assigned the output values as -1.
  - If the percentage change of close market value for two consecutive filings is more than 2% assigned the output values as 1.
  - If the percentage change of close market value for two consecutive filings is from -2% *to* 2% assigned the output values as 0.
- **Pre-processing**: After obtaining the scrapped *.htm* files from the url, we tokenise the file conents, tag each words from tokenized list, chunk the sentence and extract entity name for each tree in chunked sentences. This will reverse the input features in the form of numpy matrices

- **Model:** In the MLPClassifiers we pass in the required parameters i.e. adam optimizer and relu activation, with max iteration 750. Then we fit the data into the model, dump the model to another file.
- For predictions we load the model and pass in the required parameters.
- For analysing our model performance, we use measures like accuracy, precision, recall, F1-score.

The following subsections describe the functions performed by each code file present in GitHub.

2.2 get_8k.py

This program gives us two lists which contain 8-K forms filed within specified date-range and the corresponding date of files.

2.3 10kplot.py

This gives the json data which can be used to plot various graphs to display on the dashboard.

2.4 10qplot.py

This gives the json data which can be used to plot various graphs to display on the dashboard.

2.5 htm_str.py

This gives us the html content of the file with the specified name.

2.6 nasdaq.py

This will return the stock market values and date for given ticker, start date and end date.

3 Technologies used

- **Web Scrapping:** Used python libraries like Beautiful Soup to get the training data for running our MLPClassifier machine learning model.
- **Machine learning libraries:** Sklearn, vectorizers, nltk, pandas, pickle, numpy and other customized python modules written by us for easy access to various forms and date ranges of various companies.
- **Website Frontend:** We are using React js to power front-end.
- **Website Backend:** Node js is used to render the backend.
- **UI/UX:** Styled using Material UI.

4 Resources

- Finnhub stock API
- Stock Market API from `www.alphavantage.co`
- Edgar database `https://www.sec.gov/`

5 How to use the Webpage?

- On opening the webpage, you will be asked to enter the company name and the range of years.
- On clicking search, you will be taken to a dashboard where plots for different metrics will be shown for the given range of time.
- There are 3 tabs including 10-K form analysis, 10-Q form analysis and market predictions.
- The market prediction is based on supervised machine learning models which predicts rise or fall of stock values based on the 8-K trained model for various companies.
- The accuracy of the trained ML model comes out to be nearly **80%**.
- Precision comes out to be **0.833**, Recall as **0.9375** and F1-Score as **0.8823**.