# Final Project: Predicting Net Income Direction using News Articles

Karthick Ramasubramanian, Rohitashwa Chakraborty, Harsh Mehta

## Problem Statement

The goal of this project is to forecast the directional change in Net Income of publicly traded companies in the following fiscal year. Accurate forecasts will allow us to create zero-cost long short sorted portfolio with higher alphas. For this project, we would be focusing on the text data obtained from news articles. Along with text data, we would also be using fundamental data which can be found in income statements, balance sheets and market & economic data including market prices, inflation, treasury bond yields, volatility index to name a few.

## Data Sources

For the scope of this project, we need the following data sources:

- Fundamentals data
    - Exclusively relying on Compustat, WRDS
- Market Indicators and Economic data
    - Publicly available sources like FRED Economic data
- Fama French 3 Factor Data
    - Data extracted from Kenneth R French Data library
- News Data
    - Fetching information from the Capital-IQ Key events data library, available with our WRDS subscription. [ Appendix - 1]

The Captial IQ Key events database provides a chronological series of corporate events. The information is highly structured and includes the type of event, event

title, short description, date and time the news was made publicly available etc. This will allow us to measure the success of a company across several categories like Acquisition, Product Launch, Sales etc., which is not available in the financial reports. With these headline topics we can also categorize the positive and negative headlines.

Since, to build a machine learning model, we would need all the data to be in tabular format. All data except for text data can be obtained in tabular, excel files. So, the text analysis or preprocessing must be done, and these values need to be converted into tabular form. These tasks are very computationally expensive. This processing will be done on local machines (laptops).

The text preprocessing we plan to do are, Sentiment Analysis, topic categorization. We also plan on using past news articles and creating an aggregated score.

Finally, our machine learning model will be trained to forecast the probability of a company earning a positive change in Net income. The companies will be sorted on the basis of the aforementioned metric and be binned in 5 categories. Should either equal or value-weighted returns of these bins exhibit a monotonically increasing or decreasing trend, we can conclude that the signal is tradable. Thereafter, we create long short portfolios using these categories and compare our portfolio's performance against the market and Fama French Models

## Model Building

Once we have collected all our tabular and text data, we will preprocess the data. The preprocessing steps include, but is not limited to:

- Null Value handling (Dropping rows, columns or imputing values depending on the percentage of data that is missing).
- Dropping duplicate data to guard against redundancy
- Normalising the data ( should we proceed with a regression based model)
- Checking for co-relation between features and eliminating unimportant features.

We plan to implement classification algorithms like Logistic Regression, Random Forest, Support Vector Machine Classifiers. Probability of positive change in Net

Income can be calculated from the '.predict_proba()' function associated with these models

## Creating Sorted Portfolio

Since, we are predicting the direction of Net Income Change, we will be using the predicted probabilities as model outputs. And then identify firms which are likely to perform the best and the worst.

Every year we will identify companies which would perform the best and worst in the next fiscal. We will then arrange these companies from best to worst. Then we pick the top 10% of the companies and the bottom 10%. The top will be put in a long position and the bottom shorted to the same amount creating a zero-cost portfolio. This portfolio will be adjusted every year and its returns will be tracked against the market.

If this Portfolio has a significant and high positive Alpha, we will conclude that this is a valid trading strategy, and we would invest more in improving the performance of the model.

## Limitations

The main issue with this assignment is to analyze text data. The size of 1 year of the News articles data files is 1.5 GB. To download and analyze 20 years of data might be a challenging task and hence, we might limit the analysis to a few sectors. We are considering using the Python API provided by WRDS that allows us to use PostgreSQL Queries to extract data

Also, text analysis is extremely slow. Using Collab is not possible as the free version kicks you out in 30 mins and therefore has to be run on our local machines.

## Past Research

We found a research paper titled – ‘Improving Earnings Prediction with Machine Learning’ authored by Joshua O.S Hunt from Mississippi University. This paper builds on another research on Earnings Prediction, Ou and Penman (1989). The original research used stepwise logit regression to predict the sign of future earnings changes. Using these predictions, they constructed a profitable hedge portfolio long in firms predicted to have an increase in earnings and short in firms predicted to have a decrease in earnings. Following the original research, the paper Hunt (2019) uses 60 financial ratios for over total of 75,489 company-year observations. Sticking only to financial ratios is smart as it takes care of standardization and makes values comparable across firms. However, we do wonder if none of the ratios were highly correlated with another. Another interesting technique they use is ranking probabilities after prediction. Instead of directly using estimated probability, they follow a technique in Holthausen and Larcker (1992) and rank the probabilities in order to have more balanced cutoffs, (i.e., we rank probabilities for each model and split the sample based on quantiles). Using this methodology not only balances the number of observations in the top and bottom groups but holds the number of observations consistent across models. We would like to dive deeper into this paper to understand data cleaning, modeling and implementation of probability ranking. Replicating their findings should allow us to construct a portfolio with over 10% abnormal returns. Adding text analysis and tweaking their input features should result in a higher alpha. Findings from this paper can not only guide us but also serve as good benchmark for our analysis.

## References


- Hunt, Joshua. “Improving Earnings Predictions with Machine Learning Hunt ...” *Improving Earnings Prediction*, 10/2019/12, https://zicklin.baruch.cuny.edu/wp-content/uploads/sites/10/2019/12/Improving-Earnings-Predictions-with-Machine- Learning-Hunt-Myers-Myers.pdf.

# APPENDIX

1) Key events dataset from Compustat - WRDS

| enterdate | entertime | gvkey | headline | situation |
|---|---|---|---|---|
| 24/04/2003 | 12:00:00 AM | 29751 | Albemarle Corporation is considering acquisitions. | Albemarle Corporation announced that it is interested in making acquisitions in the fine chemical area. |
| 27/04/2002 | 12:00:00 AM | 1581 | AT&T Corp. is considering acquisitions. | AT&T Corp. has announced that it is considering acquisitions of network assets from cash-strapped companies in Europe and Asia to enhance its |
| 1/2/03 | 12:00:00 AM | 1581 | AT&T Corp. is considering acquisitions. | AT&T Corp. announced that it could use up to $5 billion to acquire assets, both domestically and abroad. |
| 14/02/2003 | 12:00:00 AM | 1976 | Baker Hughes Incorporated is considering acquisitions. | Baker Hughes Inc. is evaluating various strategic options including external investment opportunities. |
| 26/04/2001 | 12:00:00 AM | 7647 | Bank of America Corporation is considering acquisitions. | Bank of America is considering a merger or acquisition to enhance its business. |