# **APCOMP209A Project Proposal**

Title: Predicting Price Variation of Stocks with Various Financial Indicators

Author(s): Sabrina Hu (sabrinahu@college.harvard.edu)

## Background and Motivation:

Predicting the movements of stocks over time based on various financial and economic indicators is a common task in the field of finance and investment. Investors and traders rely on these predictions to assess when to buy, sell, and hold stocks, and also to gain a competitive edge in the markets and maximize returns while managing risk. However, predicting the movement of stocks is not an easy task, as prices are constantly fluctuating based on many complex factors like economic indicators, corporate earnings reports, geopolitical events, and market sentiment. In this project, our goal is to use statistical learning and machine learning techniques to (1) determine the most important financial indicators in affecting stock price and (2) develop, test, and train an accurate predictive model using these indicators for stock price variation.

## Data:

The data we will use is this Kaggle dataset: '200+ Financial Indicators of U.S. stocks (2014-2018)': https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018?select=2018_Financial_Data.csv

The dataset contains data from 5 different years, each year containing data for around 4000-5000 stocks, and including over 200 financial features for each stock such as revenue, operating income, net profit margin, liabilities, ROA, and more, and also as a response variable, containing the % increase or decrease in the stock's price over the course of the year. The goal is to identify the most important features out of these 200+ features, and then develop an effective model to predict the stock price variation (or classify the stock as increasing or decreasing in value). *Data cleanup:* After conducting a preliminary analysis of the dataset, we have determined that the main data cleanup we must conduct is selecting a smaller subset of features for our model out of the initial 200+ features. Besides that, the data is clean and organized and we did not notice any missing values.

## Scope:

Our project will focus on testing various statistical learning models like multiple linear regression k-nn regression, etc., as well as techniques like LASSO, ridge regularization, etc. to conduct feature selection and fine-tune our model. The project will be an amazing opportunity to directly apply the skills we've learned in class to a real-life, complex problem.