**Project Title:** Machine Learning for Fundamental Analysis

**Supervisor:** Lakshay Kumar Chauhan

**MDST Students:** Qiran Li

**Other Team Members:** Nick Batarilo; Vandita Singhi; Brian (Juhyuk) Lee; Enoch Lee; Chun Yin Chang

**Background:**

Machine learning can be very valuable in the investment decision-making process. Understanding business operations and intrinsic valuations are critical to evaluating a stock for long-term investment. In the professional investing community, there are principally two main-stream of thoughts – *fundamental analysis and technical analysis*. Both techniques may be used separately or in some combination depending on the strategy of the investor.

Technical analysis is a tool to evaluate securities by using statistical models on the data generated from trading activity, price movements, and volume. Fundamental analysis, on the other hand, deals with evaluating the intrinsic value of a business based on its operating characteristics and cash flows. In this project, we are focusing on the fundamental analysis which is the backbone of investment when the investing horizon is several years and not days or weeks. Every ticker symbol on New York Stock Exchange has an underlying business associated with it. To understand the business, we need to be able to speak its language of accounting. This language comprises of the balance sheet, income statement and cash flows or commonly known as the fundamentals.

We believe machine learning can play a crucial role in the process of finding good investments. It can help us to understand the business better by using the historical fundamental data. We can discover insights from historical fillings of great institutional investors and place few highly confident bets. We call this process valuation. It is a way to find the intrinsic value of the business which may or may not be close to the quoted price (market capitalization). If the intrinsic value is below the market price, say by 15%, you would be getting a $1 for 85 cents, which is great. That will become a motivation for investment.

**Objectives:**

1. To conduct a valuation of stocks with historical data.
2. To build a model that learns from the history of successful investors and use it to predict a successful investment

**Team Divisions:**

|  |  |  |
| --- | --- | --- |
| Data accessing and cleaning | Model building and testing | Data visualization |
| Chun Yin Chang | Qiran Li/ Enoch Lee | Brian (Juhyuk) Lee |
| Enoch Lee/ Qiran Li | Vandita Singhi | Nick Batarilo |

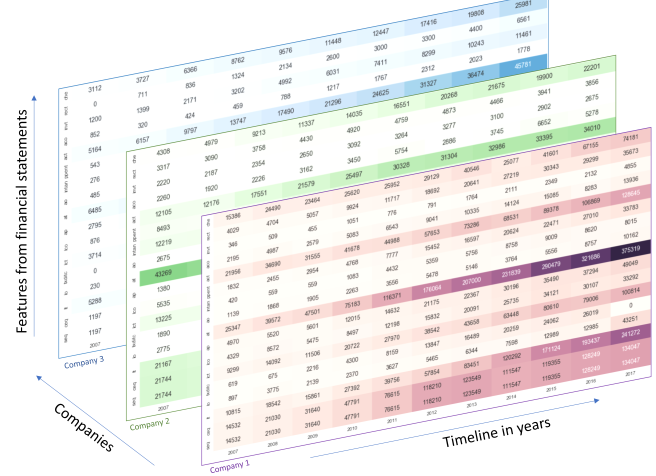
**Method:**

*Overview*

From a high-level perspective, we use Discounted Cash Flow method (DCF) which estimates the future free cash flows and discounts them back to present value. Free cash flow is the cash generated after spending the money to maintain and grow the business. It is a tangible measure of company’s financial performance. A critical element of DCF method is to project cash flows into the future which depends on various qualitative and quantitative measures. In this project, we only focus on quantitative factors. However, qualitative factors cannot be ignored for long-term investment.

|  |  |
| --- | --- |
| **Qualitative** | **Quantitative** |
| Management | Revenue, income, margins |
| Competitive advantage | Balance sheet metrics   * Assets, cast, liabilities etc |
| Customer or product related judgment | Income statement metrics |
| Competition | Cashflow statement metrics   * Operation cash flow, depreciation, amortization etc |
| Current and future risk | Interest rates, inflation etc |

We use machine learning to forecast future free cash flow based on the historical data of a company, which are those quantitative metrics listed above. In the figure below, we can see each company as a card, where the x-axis is the timeline in years and y-axis has different quantitative features coming from financial statements and economic conditions. This serves as the training data where the goal is to predict future free cash flow.



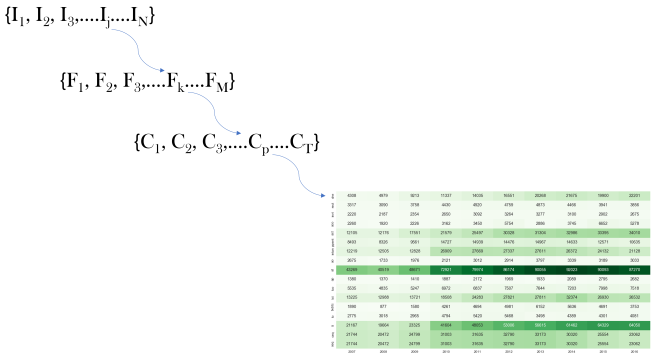
The next and major component of the project uses historical filings of successful long-term investors to build a machine learning model that learns investment decision making. We use the data collected from the cards in the figure above to formulate a classification problem where 1 signifies positive investment decision and 0 signifies neutral or negative investment decision. In this project, machine learning modeling techniques and time series modeling neural networks such as Recurrent Neural Networks (RNN) ARIMA are potentially being used.

*Data accessing and cleaning (Specific section)*

*What data?*

The SEC Form 13F is a filing with the Securities and Exchange Commission (SEC) also known as the Information Required of Institutional Investment Managers Form. It is a quarterly filing required of institutional investment managers with over $100 million in qualifying assets. For example, investor I has historical 13-f reports detailing which companies were added to their portfolio and which were reduced (or removed). Let’s say we build a list of companies that were added to the portfolio for investor I in 2016. We then build an investor stack with company cards just as shown in the figure above and do the same for other years. This serves as our training data for investor I and the target value for each card is 1 (indicating investment worthy stock). We assign the target value of 0 for company card where the position was reduced or companies which were not added to the portfolio.

Therefore, we have a group of good investors I where investor It has a list of filings F. Filing Fk consists of companies C where each company has a company card of time series historical fundamental data (figure above).



*How to collect?*

We applied for the access to Wharton Research Data Services (WRDS), it is a research platform that provides the user with one location to access over 200 terabytes of data across multiple disciplines including Accounting, Banking, and Finance etc. The data is compiled from independent sources that specialize in specific historical data, for example, S&P Global Market Intelligence and NYSE.

After getting the access, we started to build the Python script to download the fundamental dataset for top 2000 market cap equities from WRDS. It included an SQL query to get the basic equity data, price data and stock split data from the balance sheet and income sheet of every company. The database we used was Computstat which is a database of financial, statistical and market information on active and inactive global companies throughout the world with record starting from 1962. The sub-categories were the Fundamental Annual and Fundamental Quarterly which provide different data features.

Since we are creating time-series models, annual historical data is necessary. Within the 81 variables we collected, despite those basic qualitative company information, the rest of those quantitative financial metrics were classified into either trailing twelve months (TTM) or most recent quarter (MRQ). The data we got are in monthly starting from 1981 to the most recent quarter. To clean the data into the desired format for training and testing machine learning models, and for visualization, we had to recode, so the company data is in an annual manner. To do so, for TTM data, we simply took the previous month data of the last four quarters and summed them up, it would then become the annual data. For example, if I download the data on April 11th, 2018, the previous four quarters will be April-June 17, July-Sep 17, Oct-Dec 17 and Jan-March 18. We then sum up March 17, June 17, September 17 and December 17 data. For MRQ field, we simply use the last month data without further calculation.

**Results**

For my sub-team with Chun Yin Chang and Vandita Singhi, our deliverables are the python script which can update the data from time to time and the csv output files of top 2000 company data. For the python script, it will be used to update the time series data after every new fiscal quarter in order to add value in term of the longevity to our project. For the data output file, it will be used by the modeling training sub-team for creating the ARIMA or potentially the Neural Network models to predict the future cash flow (valuation) and growth rate. The output file will also be used by the visualization sub-team to create different correlation charts in order to look for the most sensitive and relevant variables towards our desired predictive output.

**Lessons Learnt**

Throughout this project from the Michigan Data Science Team, I have learned skills and knowledge that I find definitely helpful towards my career, specifically, they include the followings:

1. New programming language, Python, and SQL
2. Financial knowledge especially valuation
3. Teamwork, presentation and project management