Capstone Analysis Plan

Yimeng Xiao(yx6df), Xinru Cheng (xc9mb), Zhenzhen Zhu (zz8vy)

1 Introduction

Investment in financial securities, especially stocks, has become an important aspect of the American economy since last century. Each day, over 1.5 billion shares of stock are traded on the New York Stock Exchange with millions of people relying on stock trading as their main source of income. At the same time, professional investors and researchers have committed extensive effort to analyze the market in order to gain a competitive advantage over other investors but have not yet been able to find a way to consistently superior returns.

Almost without exception, researchers before the 1980s concluded that the efficient-market hypothesis offers a remarkably good capture of the reality of stock markets. The hypothesis states that all information available to a market is already reflected in stock price, making it impossible to predict changes in stock price without insider information. However, since the 1980s, cracks in this hypothesis started to appear that examples of anomalies seems to indicate that investors could potentially outsmart the market competition by identifying mispriced stocks following certain rules. The PE effect, for example, suggests that stocks with a low price to earnings ratio tend to later exhibit higher average risk-adjusted returns than those stocks with high price to earnings ratio. The discovery of such anomalies seems to unearthed profit opportunities that investors failed to exploit before and thus motivated investors and researches to identify more of such anomalies using statistical methods.

The use of financial ratios as a comparative tool in evaluating companies' current and future performance appeared in the mid 19th century, and still kept their fundamental power as an important supportive analysis for investment decisions till now. However, this traditional method doesn't systematically relate financial ratios to stock return, but instead only provides an abstract measurement of companies' performance such as profitability, operating efficiency and riskiness. To compensate for the lack of empirical study, this study aims to explore the relationship between financial ratios and changes in stock prices using statistical modeling and to identify ratios that possess predictive power of future stock return. Eventually, we wish to develop a robust general strategy for creating stock investment portfolios based on the relationship we find and help investors to gain abnormal gains in stock markets.

2 Questions of Interest and Study Endpoints

2.1 Questions of Interest

Stock prices reflect investors' perceptions of companies' present value and future potentials. Financial ratios which were widely accepted by investors as a comparative tool to measure companies operating efficiency, profitability, leverage and liquidity may therefore possess predicting power over stock return. Recognizing that this hypothetical relationship between financial ratios and stock prices can vary drastically from industry to industry based on business intuition, this study will only focus on the technology sector, a new but growing sector with the most publicly available financial data compared to other sectors. Therefore, we determined the questions of interest for our study as follow:

1. *Can financial ratios be used to predict stock return of firms in the technology sector?*
2. *If yes, which has the strongest predicting power? Or in other words, which ratios may contribute to existing understanding of stock market anomalies?*
3. *Is the result of our model aligned with theoretical expectations?*

Section 4 will talk about how we plan to answer these questions through statistical methods.

2.2 Study Endpoints

Data Preparation(done):

The data used for this study is provided by Simfin's free database which can be downloaded using simfin's python package. The data contains 7 datasets for annual balance sheet, annual income statement, annual statement of cash flow, daily stock price, industry information, and company information, sector information respectively. We first selected the columns we needed for computing financial ratios and stock return(listed in section 3.2), and either dropped or imputed missing values using appropriate methods. We then computed annual stock return for each firm using the formula $(P_1-P_0)/P_0$, where P_1 is the year-end stock closing price and P_0 is the year-start stock closing price. Finally, we merged all the cleaned data together on year and stock ticker, a unique identification key for each firm.

Preliminary Analysis:

The preliminary analysis includes both statistical exploratory analysis and research for relevant accounting and financial literatures which provide theoretical foundation for later analysis. For the exploratory analysis, we have and will continue to look into the sample distribution of the financial ratios and stock return so as to search for the appropriate way of data transformation. We have found out that many financial ratios have exponential distributions that most companies centered at the left end of the x axis. A log transformation will therefore rescale the data and render a less concentrated and skewed distribution which is desired for modeling. For literature research, we found out in order for the statistical result to

be interpretable, many of previous work used correlations coefficients instead of any black-box machine learning methods for their study. However, correlation only measures the strength of linear relationship between the explanatory and response variable without considering the possibilities for higher-term relationship and how the strength of relationship would be changed with the presence of other ratios. We therefore decided to use linear regression and regression tree(if time allows) for our model for its interpretability as well as flexibility.

Modle building:

We will run linear regression on annual stock return of companies in the technology sector over all financial ratios listed in section 3.2. For now, we decided to log transform the financial ratios for rescaling but we may also try using percentage change of financial ratios and other transformation methods to see which one best satisfies the modeling assumptions. We will also adjust our model by incorporating higher-order terms to account for possible quadratic and other higher-order and to see if it can improve the overall model performance. If time allows, we will also try regression tree as an alternative to see if the model performance is consistent with the result of linear regression. We will also validate the performance of our models with test data and other methods to check for stability. We will talk about this with more details in section 4.

Final Analysis and reflection:

After the modeling is complete, we will use the results of the final model to answer the question of interest. We will compare the significance of different financial ratios and determine the set of ratios that best explains the changes in stock return and hopefully summarize a general strategy that can be implemented for creating investment portfolios with high returns.

3 Data Summary

3.1 Data Description

Simfin is an open source database of financial statements and stock data of publicly listed US and Germany companies. The time range of data varies from quarterly to annually. Simfin includes 2,740 companies with 305,261 original and standardized financial statements including more than 70 company ratios. Overall, there are 7 datasets that are available for downloading: income statement, balance sheet, cash flow, companies, markets, sector/industry, share prices.

The three financial statements include 1915 companies, with annual data ranging from 2007 to 2019. The stock price data includes 2050 companies with daily stock price data from 2007 to 2019. A brief description of the financial statements can be found below:

	Income Statement	**Balance sheet**	**Cash Flow**
Description	Profit & Loss	Net worth	Cash flow received from operations and investment
Sample Columns	Operating Expense; Net Income; Gross Profit	Inventories; Total Current Assets; Short Term Debt	Dividends Paid; Cash from Debt; Net Change in Cash

We decide to focus our study on the US market and thus use the US annual data. We merged the 7 datasets based on the SimFin ID and stock ticker because each company is uniquely identified by its stock ticker and SimFin ID. In our final dataset, there exists data from 1762 companies and 71 industries ranging from 2007 to 2019. The data set has 14043 rows and 38 columns. The criteria of row and column selection will be introduced in the following section.

3.2 Inclusion and Exclusion Criteria

A. Companies
   The number of companies reduces from 1915 to 1762 since we delete all the companies with a null value in the industry column, meaning that the company does not belong to any existing industry. The reason is that we only do research on companies that belong to a certain industry sector (technology) in order to control the variance and count for the fact that business can vary tremendously from industry to industry.
B. Financial metrics

After extensive research, we determine there are 26 key financial ratios that can potentially predict a company's financial health. Thus, we only include financial metrics needed to calculate the financial ratios. A detailed description of financial ratios computed will be included in the next session.

3.3 Variables Used in the Main Analysis

The variables can be summarized into 5 aspects: liquidity ratios, leverage ratios, profitability ratios, efficiency ratios and performance ratios. We believe the selected variables can potentially predict whether a company is financially healthy or not. Overall, 26 ratios derived from 35 columns are included in the model (corresponding formula can be found in the Appendix)

	Liquidity Ratios	**Profitability Ratios**	**Leverage Ratios**
Description	Determine a company's ability to pay off current debt obligations without raising external capitals.	Measure a company's ability to generate income over a specific period of time.	Determine the relative level of debt load that a business has incurred.
Example	Current Ratio = Total Current Asset/ Total Current Liabilities	Return on Equity = After Tax Operating Income/Total Equity	Total Debt Ratio = Total Liabilities/Total Assets
Number of ratios included	4	5	6

Efficiency Ratios	**Performance Ratios**
Measure a company's ability to use its assets and manage its liabilities effectively in the current period or in the short-term.	Measure how efficiently and effectively a company is using its resources to generate sales
Asset Turnover = Revenue / Total Assets (year-start)	Sales per Share = Revenue / Shares (Basic)
5	6

3.4 Data Cleaning

A. Stock Data

We only use open price, close price, and dividends in the stock data, among which only dividends have missing values. The minimum value in the dividend column equals to 0.001, thus we can reasonably conclude that NA in the dividend column means the company at that time does not issue any dividend. We substitute all those null values with 0.

B. Financial Statement Data

Among the selected columns, 19 columns are complete and the rest have missing values. Percentage of missing value range from 0.01% to 29.84%. We have two ways of filling in the missing values: (1) substitute NA with 0. (2) fill with KNN / linear regression.

(1) Substitute NA with 0:
For certain columns, NA simply means that a certain company does not have that data available. Such columns include 'Total Equity' and 'Cash, Cash Equivalents & Short Term Investments'.

(2) Fill with KNN / linear regression
For certain columns, there exist missing values in rows where it should not be missing. For example, every publicly listed company should have either positive or negative revenue. However, 1.44% of revenue data is missing. We need to impute such missing values with either KNN or linear regression.

By finding highly correlated columns with a correlation larger than 0.8, we can fill missing values in a column by predicting it with linear regression model that regress on another highly correlated column. It is important to keep in mind that some highly correlated columns have rows that are empty at the same time. It is also possible that some columns don't have any highly correlated columns. In such cases, we would use KNN instead of linear regression.

'KNN is an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space. It can be used for data that are continuous, discrete, ordinary and categorical which makes it particularly useful for dealing with all kinds of missing data.' [1]We used KNN to fill in columns that are inappropriate to fill with linear regression. Such columns include 'Revenue' and 'Long term debt'.

[1] https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637

4 Analysis Methods

4.1 Relevance of Chosen Method

We plan to build two types of models: a linear regression model and a regression tree model. A linear regression model can show the relationship between our response variable, the annual return rate of stock, and the predictors, different financial ratios. Model selection based on p-values of individual betas and will filter out group of financial ratios that are most relevant to stock return. By interpreting the estimated parameters in the final model, we can evaluate the predictive power of each financial ratio, and hence determine which financial ratios should be looked at when making investment decisions.

Tree-based methods also help answer our question of interest because a series of splits starting at the top of the tree will inform us the most influential predictors in determining the stock return rate. The length of a branch represents the amount of improvement in RSS each split of the predictors can bring.

4.2 Method Assumptions

To check if the linear pattern is appropriate for our variables of interest, we will plot scatterplot of stock returns and various financial ratios. If the scatterplots show a linear trend, this assumption about the form of the model is satisfied. If not, we will adjust by performing some transformation methods such as log transformation.

The error term of the linear regression model has four assumptions: residuals have a zero mean, residuals are normally distributed, residuals have constant variance, and residuals are independent. We will check these assumptions by plotting the

a. residual plot (should be horizontal with no apparent curvature and has an average value of zero => a mean of zero)
b. QQ-plot (falls along the 45 degree line => normal distribution)
c. $\sqrt{standardized\ residuals}$ (constant vertical spread, not fanning in/out)
d. Cook's distance contour lines (identify influential outliers by large Cook's distance)

If the mean is nonzero, we need to transform the response variable, stock returns. If the mean is zero but the variance is not constant, then we will transform the financial ratios or include higher-order terms to adjust our data.

An alternative way to determine if stock returns need to be transformed is to plot the profile log-likelihoods for the parameter lambda of the Box-Cox power transformation. If lambda corresponding to the 95% log-likelihood threshold is 1, we do not need to transform returns. If lambda is 0, we take log transformation, otherwise, we rise returns to the power of lambda.

There seems to be no implicit assumption about decision trees.

4.3 Statistical Software

In model building, analysis, and data visualization sections, we will use R as our main statistical tool with the assistance of some Python packages and Tableau. With cleaned data

in hand, we are ready to check model assumptions and build models after we return from spring break.

We will start with the Technology sector first, then extend our analysis to other sectors if time allows. Given the time series nature of our data, we fit all of our models on training data (2007 to 2015 data) and predict the stock return rates on test set (2016 to 2019 data).

To build the linear regression model, we will (i) fit linear regression model on training data using the *lm()* function, (ii) predict the annual return of stocks on the test set and calculate the test MSE, (iii) apply shrinkage methods in the *"glmnet"* package. To compare the Ordinary Least Square model with Ridge and Lasso, we will use cross validation to find the best lambda, which will then be used to fit the models with all observations. We will evaluate the predictive power of the models by comparing their R^2.

To build a regression tree, we will fit a recursive binary splitting tree using the *tree()* function (in the *"tree"* library) on training data, predict on test set and calculate the test MSE. We will then improve the tree by methods of

a. pruning (*cv.tree()* to select an appropriate value of the tuning parameter K and size of a tree; *prune.tree()* to build the pruned tree of the best number of terminal nodes),
b. bagging (*randomForest()* with *mtry* = p, the number of covariants),
c. random forests (*randomForest()* with *mtry* = p/3), and
d. boosting (*gbm()*).

We will compare the test MSEs from all five trees and select our best regression tree model based on minimal test MSE and ease of interpretation.

4.4 Benefits and drawbacks of Chosen Method

The linear regression model is easy to interpret and can inform us which financial ratios are the most important in terms of estimating and predicting stock returns. One drawback is that it does not account for possible nonlinear relationship between the ratios and the returns. For another, since financial data can be noisy when the market volatility is high, the OLS model with high variance is not ideal. Then Ridge and Lasso can significantly reduce variance by introducing a little bias. We expect Lasso to perform better than Ridge because its parameter selection function satisfies our purpose of selecting a subset of important ratios.

Tree-based methods are easy to explain, more closely resemble human decision-making, and can be displayed graphically and hence be easily interpreted (if the tree is not too big). But they tend to not perform as well as their regression counterparts, especially if the response variable is quantitative (this is our case!). They are sensitive to individual observations; an unusual fiscal year of a company can cause a large change in the estimated tree. In particular, the recursive binary splitting tree tends to overfit the data and has many splits (high variance but low bias). Ways of improving the recursive binary splitting tree will help reduce the variance. Pruning does so by increasing the bias a little. Bagging does so by averaging the predicted value from each tree while maintaining low bias.

Random Forests do so by forming less correlated trees (this will be supreme if there is one very strong financial ratio, with a number of fairly strong ratios). And boosting builds trees sequentially and uses tuning parameters such as lambda, the number of splits of each tree.

4.5 Validation Methods

Before model building, we will split our dataset into training and test dataset and use the testing data to validate the model. Since we have data from 2007 to 2019, we will use the data before 2016 as training set and use the most recent 4 years as test set. Notice that the process is not to be done randomly like what people usually do with machine learning models. This is because the goal of our study is to develop an investing strategy that analyzes historical account information and makes predictions about future stock returns. We are not just trying to make a model that best captures the relationship between financial ratios and stock returns within a limited time horizon. Suppose we view the most recent 4 year as "future" years while using some of the data from this period to train our model, we are actually using future information to predict the future, which is obviously unrealistic and untenable. Therefore, our validation method is to keep the most recent 4 years' data out, and to see if the ratios we selected based on analyzing historical accounting data can successfully identify the firms with higher stock returns.

5 Appendix

(1) financial ratios included

Liquidity Ratios	Formula
Current ratio	Total Current Assets/Total Current Liabilities
Quick ratio	(Cash, Cash Equivalents & Short Term Investments+ Accounts & Notes Receivable)/Total Current Liabilities
Net working capital to assets ratio	(Total Current Assets-Total Current Liabilities)/Total Current Assets
Cash ratio	Cash, Cash Equivalents & Short Term Investments+ Accounts/Total Current Liabilities

Leverage Ratios	Formula
Long-term debt ratio	Long Term Debt/(Long Term Debt+ Total Equity)
Liabilities to Equity Ratio	Total Liabilities/ Total Equity)
Total Debt ratio	Total Liabilities/Total Assets
Debt to Assets Ratio	(Long Term Debt+Short Term Debt)/Total Assets
Interest coverage ratio	Operating Income (Loss)/Interest Expense, Net
Cash coverage ratio	(Operating Income (Loss)+Depreciation & Amortization)/Interest Expense, Net

Profitability Ratios	Formula
After Tax Operating Income	Revenue-Operating Expenses-Depreciation & Amortization-Income Tax (Expense) Benefit, Net
Return on Asset	After Tax Operating Income/Total Assets
Return on Capital	After Tax Operating Income/(Long Term Debt+ Total Equity)
Return on Equity	After Tax Operating Income/Total Equity
Economic Value Added	After Tax Operating Income-Cost of capital（need outside data）* (Long Term Debt+ Total Equity)

Efficiency Ratios	Formula
Operating profit margin	Operating Income (Loss)/Revenue
Net Profit Margin	Net Income (Common)/Revenue
Asset Turnover	Revenue/Total Assets(year-start)
Receivable turover	Revenue/Accounts & Notes Receivable(year-start)
Inventory turnover	Cost of Revenue/Inventories(year-start)

Performance Ratios	Formula
Market Value added	Stock Price * Shares (Basic) -Total Equity
Market to Book ratio	Stock Price * Shares (Basic)/Total Equity
Earning per Share	Net Income (Common)/Shares (Basic)
Sales per Share	Revenue/Shares (Basic)
Price to Earning Ratio	Stock Price/Earning Per Share
Price to Book Ratio	Stock Price/(Total Equity/Shares (Basic))