COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Capstone Report - I

Business Health Simulator with Capital One

Made by:

Ankita Agrawal, UNI: aa4229

Akanksha Rajput, UNI: ar3879

Gurkanwar Singh, UNI: gs3006

Sheetal Reddy, UNI: kr2793

Contents

1 Introduction & Problem Statement

This Capstone project is a collaboration between the Data Science Institute at Columbia University and Capital One. The aim of this project is to understand the business health of small businesses and develop a model to predict it 6 months into the future. Looking at the possibility of a recession in the coming year, Capital One aims to evaluate which small businesses are safe to invest in, in the near future.

The health of businesses can be quantified in various ways like Revenue, EBITDA (Earnings Before Interest, Taxes, Depreciation, and Amortization), Net profit margin, etc. To make a baseline model, we select EBITDA as our target variable as it is a good measure of the profitability of a company. The idea is to first develop a satisfactory model to predict EBITDA at company level and industry level, and later explore other target variables and make a mixture model to predict multiple indicators that constitute our definition of business health. One of the first goals of our project is to explore relevant features that can contribute to our model towards this prediction.

2 Literature Review

Among the many other objectives of this project, one of the objectives was to perform a thorough literature review and create our own datasets. We decided to focus on the 1425 small business companies that are part of the Vanguard Small-Cap Business Index Fund, for which the information is available on the Vanguard website. We learned and researched about various financial terms like EBITDA, Dividend Yield, price-earnings ratio, price-cash flow ratio, Current Ratio, Debt-Equity Ratio, Value at Risk (VaR), etc. to get a better understanding of which variable to select as our target variable and other features for our dataset. We also studied some macroeconomic variables like the US unemployment rate, Real GDP, CPI (inflation), Yield curve, etc.

We went through several articles on different factors that contribute to the business health of a small business, and various ways business health has been defined in the financial world, including literature on the Small Business Optimism Index which is published monthly by the National Federation of Independent Business. We read articles like Road to Recession published by Moody's on the liklihood of a recession in the coming year, which helped us to identify financial features for our problem.

For building a predictive model, we went through several techniques like ARIMA, ARIMAX, Multivariate Time Series Modeling with Vector AutoRegression (VAR), deep learning techniques like Multi-Step LSTM (Long Short Term Memory), etc. In addition, We educated ourselves on time series concepts like stationarity, autocorrelation, moving averages, etc.

3 Data Collection

- We extracted part of the data from the Bloomberg Terminal. We collected additional information like historical data of some of the macroeconomic indicators such as the US Unemployment Rate, Consumer Price Index (inflation), GDP etc, from it. We also extracted the historical values of

NFIB Small Business Optimism Index separately from the bloomberg terminal.

- To identify small businesses, we explored the Vanguard Small-Cap Index Fund which has an investment portfolio of 1425 small business companies. We pulled the historical data of this fund from the Bloomberg terminal in form of financial indicators like Price-earnings ratio, Value at Risk (VaR), Current Ratio, etc. for time period Jan 2016 - Sep 2019.
- To extract the target variable EBITDA, we performed web scraping of quarterly EBITDA values of all the small business companies from here using selenium in python.

Our final dataset has individual quarterly data for all 1425 small business companies for the time period Jan 2016 - Sep 2019 with financial features such as P/E, P/CF, Dividend Yield, VaR etc., macroeconomic features such as Unemployment rate, inflation, GDP etc., NFIB Index and the target variable as EBITDA.

4 Exploratory Data Analysis

1. We first explored the composition of different industries represented in the Vanguard Small-Cap Index Fund *(Figure 1)*. We see that Communication services take up the maximum share in the fund portfolio.

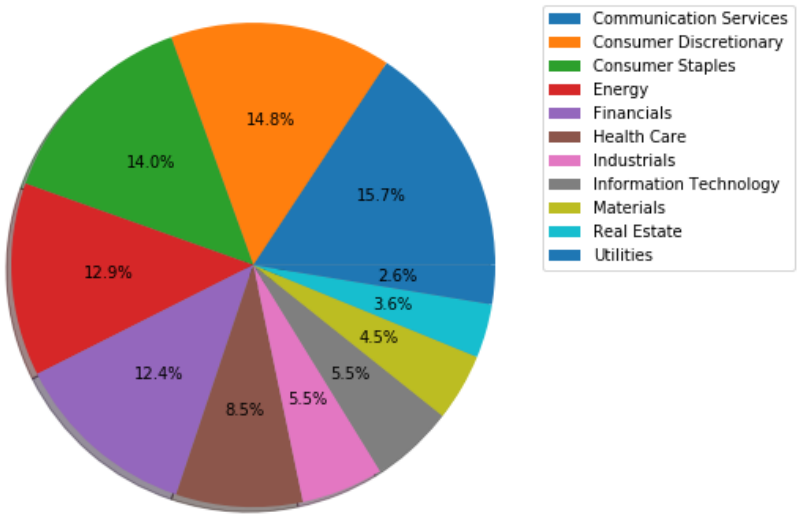

Figure 1: Industry wise composition of Vanguard Small-Cap Index Fund

2. Next, we made a correlogram plot *(Figure 2)* to investigate features which are correlated.

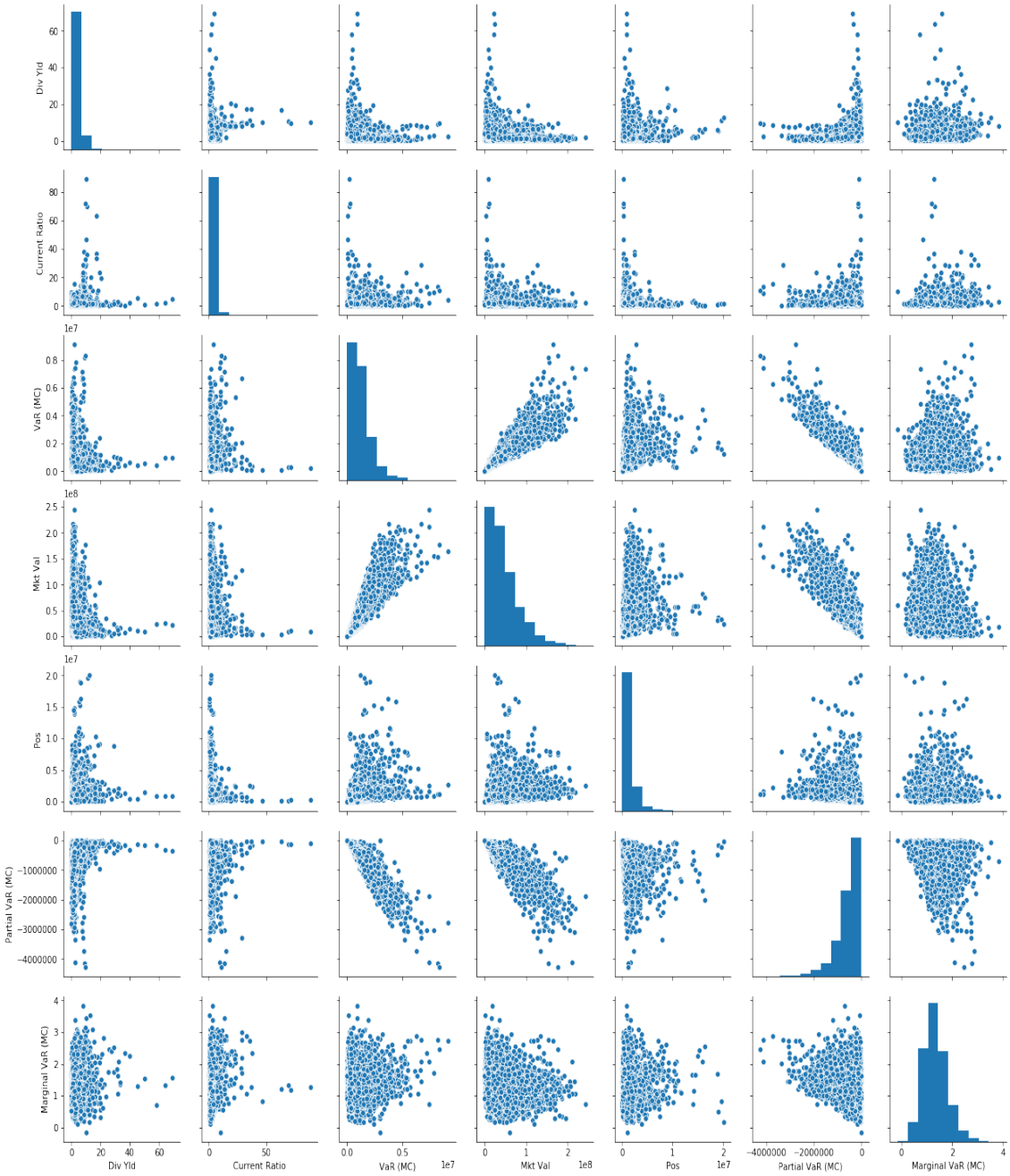

Figure 2: Correlogram of dataset features

We observed that the Conditional Value at Risk (CVaR) is positively correlated with VaR, which is to be expected as CVaR values are derived from the calculations of VaR itself. Partial VaR is negatively correlated with VaR, CVaR and Market Value.

3. We plotted a heatmap to investigate the missing values in our data *(Figure 3)*.

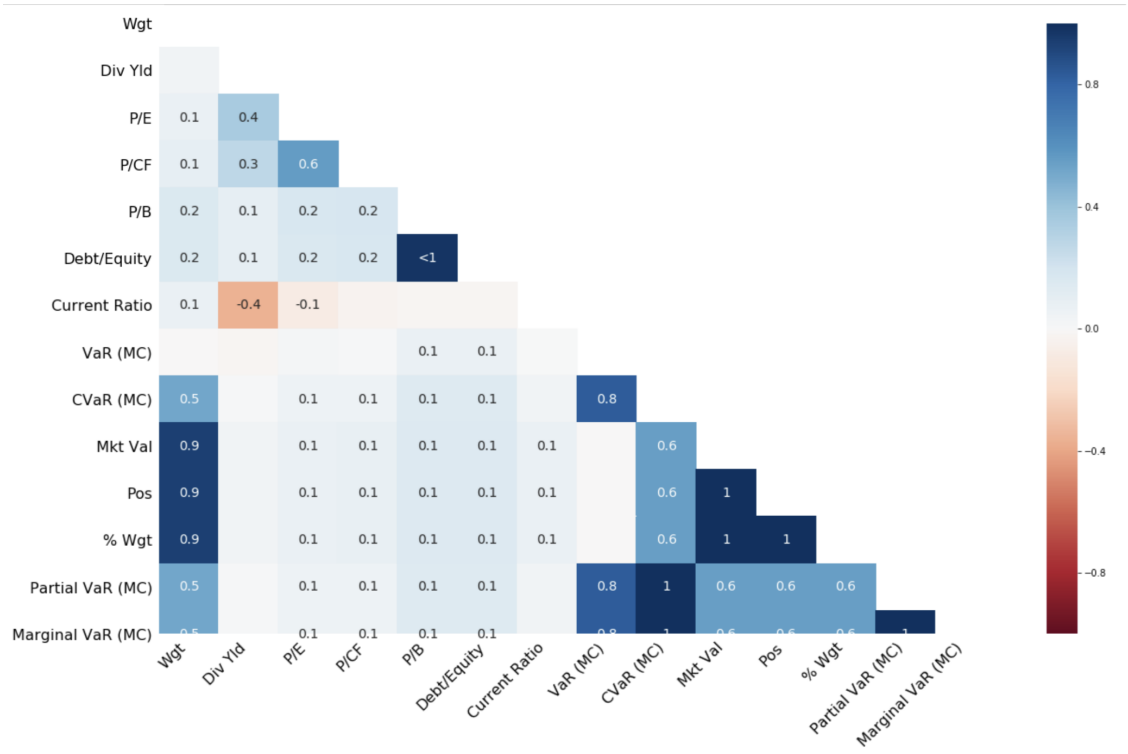

Figure 3: Heatmap of missing values

We observed that most of the values missing are in features related to variance. When P/E is missing, Div Yld is missing 40% of the time. When P/CF is missing, P/E is missing 60% of the time.

4. We then plotted the macroeconomic indicators like unemployment rate and inflation with the NFIB Small Business Optimism Index *(Figure 4)*.

We observed that though inflation does not show a monotonic trend, as the unemployment rate has decreased over the last decade, the NFIB Index has mostly increased, indicating that more employment leads to more optimism in the small business health.

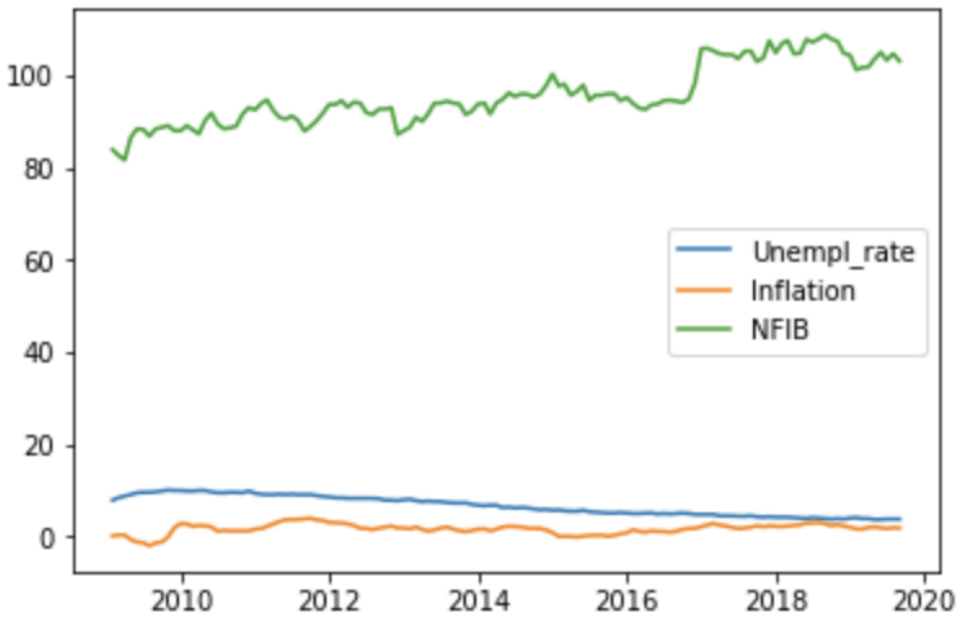

Figure 4: Comparison of the trend of macroeconomic indicators with NFIB

5 Baseline Model

To make an initial model, we shifted the values of our target variable EBITDA ahead by one quarter so that the features of Q1 can be used to predict the EBITDA of Q2 and so on. Shifting of EBITDA is performed at the company level.

1. Data Preparation
    - We dropped the rows with missing EBITDA values.
    - We also dropped the rows for the last quarter (current quarter) as the EBIDTA value for the future is missing and hence it can not help the model learn.
    - To fill the missing values for other features, we used random forest missing data algorithm.
2. Model Evaluation
    - We are using a simple linear regression model as a baseline to fit on our dataset.
    - With this model we achieved Mean Absolute Error (MAE) of $\tilde{4}3$, and Root Mean Square Error (RMSE) of $\tilde{1}03$.
    - The RMSE is relatively on the higher scale, that may indicate that there are some outliers in our data.

- Currently, we are in the process of analyzing EBITDA and other features of the test data to better understand these outliers and discrepancy in the results of evaluation metrics.

6 Goals and Next Steps

Since our model is trained on limited data , it is not fully comprehensible and reliable as of now. We further aim to enhance our model by delving more into the following sections.

- Data Collection:
  1. Extract more historical data from the years 2019-2015 to expand our data-set
- Missing Value Treatment:
  1. Try imputation using EM algorithm
- Feature Engineering and Feature Selection:
  1. We plan to explore more relevant financial features and target variables and will implement feature engineering and feature selection after doing research and leveraging the domain knowledge of our mentors at Capital One, to improve our models.
  2. We plan to find out the feature importance and other latent information from the features by considering lagging, concurrent and leading variables. These patterns can be found with more data.
- Modeling Approaches:
  1. Try ARIMAX
  2. Multivariate Time Series modeling with Vector Auto Regression (VAR)
  3. Multi-step LSTM

  Once we have a satisfactory model for predicting EBITDA, we will develop similar models for other business health indicators and combine them, in the end, to predict a single value for business health.

7 Report Contribution

- Akanksha Rajput - Data extraction from Bloomberg Terminal, web scraping, baseline model
- Ankita Agrawal - Data munging, exploration and visualization
- Gurkanwar Singh - Data extraction from Bloomberg Terminal, visualization, report compilation
- Sheetal Reddy - Literature review, data exploration, report compilation

8 Acknowledgement

We would like to thank our faculty advisor Adam Kelleher for his valuable feed-backs and guidance through the first phase of this work. We would also like to extend our gratitude to the team at Capital One (Peter Deng and Chaitanya Dara) for giving us this opportunity to work with them and for their continuous support and mentorship.

9 References

1. https://pypi.org/project/missingpy/
2. https://www.sec.gov/edgar.shtml
3. https://www.macrotrends.net