Milestone 3

Summary of the Data

To begin our analysis, we briefly summarize the features of our dataset. The shape of our concatenated dataset is (22077, 46), containing 22077 data points from companies throughout 2014-2018 with 46 predictor variables. The breakdown of companies for each year is as follows: 3,808 companies in 2014, 4,120 companies in 2015, 4,797 companies in 2016, 4,960 companies in 2017, and 4,392 companies in 2018. The data types for most predictor variables are 64-bit floats. However, two exceptions are observed: one of the predictors (`sector`) contains objects, and one of the response variables (`class`) contains boolean classifications. Please refer to our notebook for the summary statistics including count, mean, std , min, 25%, 50%, 75%, and max for each variable.

From Exhibit 5, we can see that our features have very different scales. For instance, revenue is in the billions while eps_growth is close to zero. Such disparities can indeed cause issues with many machine learning algorithms that are sensitive to the scale of the data, particularly those that use distance measures (such as logistic regression). We will fix this issue with Min-Max Scaling.

Deeper Understanding & Meaningful Insights

Through our initial EDA, we have identified several key issues:

1. Some financial indicator values are missing (nan), leading us to examine missingness by using `missingno` library to visualize the patterns of missing data.
2. Of the independent variables, multicollinearity is probable and will likely impact the results

Missingness: We use the missingno (missing "no") library to visualize the patterns of missing data (see exhibits 1, 2 & 3). Interpretation strategies for missing data patterns:

1. Randomness: If the missingness appears random in the matrix plot, it may indicate data is Missing Completely At Random (MCAR).
2. Patterns or Bands: Systematic patterns or bands in the matrix plot suggest Missing At Random (MAR), where missingness is related to observed variables.
3. Blocks: Blocks of missing data indicate Missing Not At Random (MNAR), suggesting issues with data collection.
4. Correlation in heatmap: High correlation in the heatmap implies that missingness in one variable is not independent of another, often seen in MAR or MNAR scenarios.

From Exhibit 1, we can see that there are clearly patterns in the white lines indicating the location of missing values. The sparkline at right summarizes the general shape of the data completeness and points out the rows with the maximum and minimum nullity in the dataset. We notice that the features with the most clusters of white spaces include net debt to ebitda, free cash flow yield, current ratio, net debt, and return on assets. Additionally, we see that there are three general groups of white lines that span across all of the variables, signifying companies are missing data across all features. These companies with high-missing-value columns across all years may be dropped. Given the missing data seems not to be MCAR, mean or median imputation methods may not be suitable. Mode imputation is also not applicable as most features are numerical. Therefore, we can use kNN imputation to replace missing values using the k-nearest neighbor mean for each feature.

From Exhibit 2, we see that the leftmost bar, sector, operating income, operating cash flow has the least number of missing values and the rightmost bars, net debt to ebitda, net debt, return on asset, current

ratio, has the most number of missing values. We will keep in mind that the rightmost ones will have 10-15% of values imputed

Finally, from Exhibit 3. The nullity correlations range between 0.2 and 1. Certain correlations, including EPS and dividend per share, return on equity, and net debt to EBITDA, remain close to 1.

Imbalance: From Exhibit 4, sector distribution does show some imbalance, which is common in real-world datasets where some categories (in our case, sectors like Financial Services, Healthcare, and Technology) are more prevalent than others (like Utilities and Communication Services).

Data Scaling: Our feature values have very different scales. Since we don't want to make any assumptions about the distribution of our data, we normalized our data using Min-Max Scaling which rescales the data to a fixed range between [0, 1] in order to ensure that all features contribute equally to the analysis, preventing variables with larger scales from dominating those with smaller scales.

Correlation Between Predictors: We plotted a correlation heatmap as seen in Exhibit 6, to determine confounding variables which may yield multicollinearity that we will need to drop (or do PCA) before modeling. This provides us with a visual representation of how strongly variables are correlated with each other. Positive values, colored red, indicate a positive correlation, while negative values (in blue) indicate a negative correlation. Gray values are close to zero, and suggest a weak or no correlation. Though the plot is difficult to interpret due to having many features, we observe a significant orange hue in the top-left corner, indicating a positive correlation ranging from 0.5 to 0.75. Additionally, there is a blue stripe near the top left, signifying a negative correlation in the range of -0.25 to -0.5. This tells us that financial metrics such as revenue, net income, profit margin, etc. are very correlated, which makes sense logically. The negatively correlated variables correspond with other metrics such as depreciation, investing cash flow, etc. that correlate negatively with revenue and positive metrics. In our final model, we will definitely need to further reduce financial metrics that give similar information.

Labeled Visualization

Exhibit 1

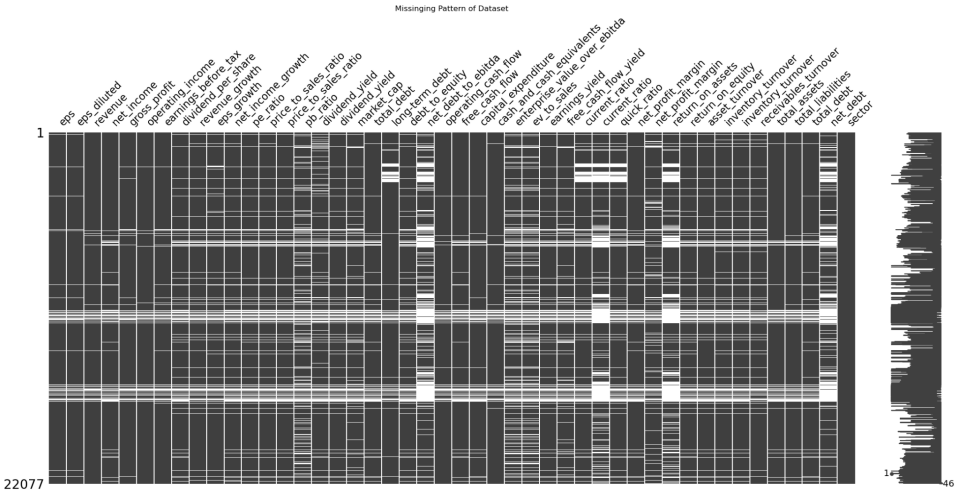

Exhibit 2

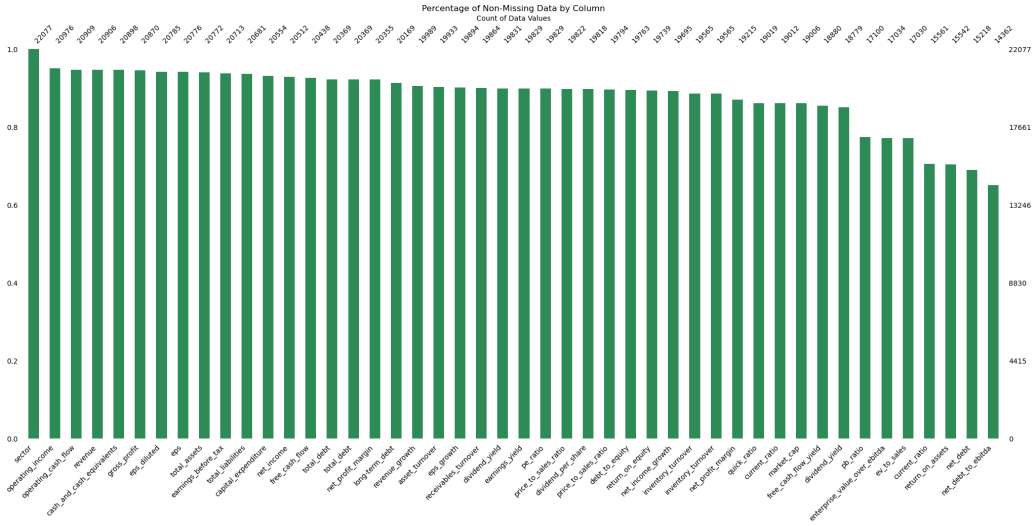

Exhibit 3

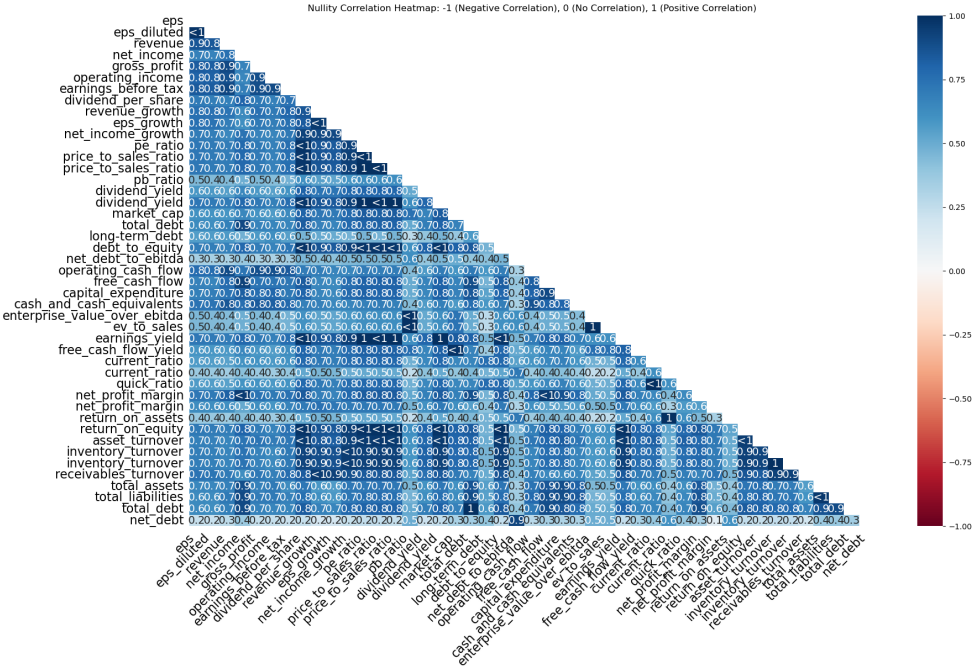

Exhibit 4

	Class	Percentage	Feature	Count
0	Financial Services	0.213797	sector	4720
1	Healthcare	0.149703	sector	3305
2	Technology	0.141595	sector	3126
3	Industrials	0.125379	sector	2768
4	Consumer Cyclical	0.111926	sector	2471
5	Basic Materials	0.060878	sector	1344
6	Real Estate	0.056212	sector	1241
7	Energy	0.055125	sector	1217
8	Consumer Defensive	0.041763	sector	922
9	Utilities	0.023463	sector	518
10	Communication Services	0.020157	sector	445

Exhibit 5

```
Means of numeric columns:
eps: -10657.48
eps_diluted: -10735.82
revenue: 5161618858.18
net_income: 388672668.87
gross_profit: 1970452466.97
operating_income: 589697890.73
earnings_before_tax: 492500290.56
dividend_per_share: 1.22
revenue_growth: 3.62
eps_growth: 0.28
net_income_growth: -2.23
pe_ratio: 37.90
price_to_sales_ratio: 88.05
price_to_sales_ratio: 87.95
pb_ratio: 20328.17
dividend_yield: 0.29
dividend_yield: 0.02
market_cap: 30560346775.25
total_debt: 4558784756.37
long-term_debt: 3294063118.38
debt_to_equity: 0.56
net_debt_to_ebitda: 2.88
operating_cash_flow: 1073238606.85
free_cash_flow: 453355548.55
...
total_assets: 20407238233.88
total_liabilities: 16089038257.03
total_debt: 4558784756.37
net_debt: 1839178489.07
```

Exhibit 6

Data Description

Our dataset consists of 4000 publicly traded companies, on average, between 2014 to 2018, and 200+ key financial performance indicators obtained from their annual 10-K filings. The original data was built by the author of Financial Modeling Prep API and pandas_datareader and can be accessed through Kaggle. There are a total 5 datasets: 2014_Financial_Data, 2015_Financial_Data, 2016_Financial_Data, 2017_Financial_Data, and 2018_Financial_Data. Key columns include:

1) `SECTOR` classifies stocks into sectors (e.g. Technology) for sector-specific analyses.
2) `PRICE VAR [%]` provides yearly price variation data for each stock
3) `CLASS` is a binary classification column that indicates stock price movement, where 1 signals increase and 0 decrease, serving as potential buy or not-buy indicators for trading strategies

These datasets are equipped for classification tasks using the 'class' column and regression tasks using the 'PRICE VAR [%]' column to predict stock values.

Initial Preprocessing and Cleaning: To enhance readability and maintain consistency for future analysis, we took the following steps before concatenating the 5 datasets from 2014-2018 into 1 dataset

1) Renamed all column heads to follow the industry standard – lower case and '_' as separator
2) Renamed the first column as 'ticker'
3) Renamed the price_var column removing the year
4) Reduced the number of predictors from 223 to 46
5) Separated the predictors (all other features) and targets (class and price variation %)

For steps 4 and 5, we researched factors that influence stock prices and determined a set of predictors based on the reasoning below from initial explorations.

Initial Exploration of Predictors and Targets: Stock prices are ultimately determined by supply and demand, which are influenced by both fundamental and technical factors.

Fundamental factors relate to a company's financial performance, especially earnings. The earnings per share (EPS) represents the owner's proportional share of earnings. The P/E ratio expresses the valuation multiple, i.e. how much an investor is willing to pay for those future earnings. Earnings can also be measured by cash flow per share or dividends per share. The valuation multiple depends on the expected earnings growth rate and the discount rate, which represents inflation and the perceived riskiness of the stock. Higher expected growth leads to a higher multiple, while higher risk or inflation leads to a lower multiple.

Technical factors are from external market conditions that drive supply and demand. Historically, low inflation boosts multiples while high inflation lowers them. Deflation hurts pricing power. Stocks tend to move with the overall market and their industry peers, and stocks compete with other asset classes like bonds and commodities. Regarding investors, middle-aged investors tend to invest more in stocks. Furthermore, stocks build momentum or revert to the mean. Large cap stocks have high liquidity while small caps often have a liquidity discount. Finally, unforeseen positive or negative events impact sentiment, and psychology often weighs more heavily than fundamentals in the short term. Behavioral finance aims to explain irrational market behavior.

Reconciliation: While technicals often dominate in the short run, fundamentals determine long-term prices. Different investors weigh the factors differently, but both play a critical role in the complicated dynamics of supply, demand, and human psychology. We decide to have the categories of predictors:
Fundamentals:

- Earnings Per Share (EPS) and Diluted EPS: Directly related to earnings power
- Revenue, Net Income, Gross Profit, Operating Income, Earnings Before Tax: Critical components of earnings base
- Revenue Growth, EPS Growth, Net Income Growth: Reflect expected earnings growth
- P/E Ratio, P/S Ratio, P/B Ratio, Dividend Yield: Valuation multiples based on earnings
- Dividends Per Share: Important for mature dividend-paying companies

Technical Factors:

- Market Capitalization: Relates to liquidity and size
- Trading Volume: Reflects liquidity and investor interest
- Total Debt, Long Term Debt, Debt-to-Equity, Net Debt to EBITDA: Influence risk profile and cost of capital
- Operating Cash Flow, Free Cash Flow, Capital Expenditures, Cash & Cash Equivalents: Reflect ability to generate cash

Market Sentiment:

- EV/EBITDA, EV/Sales, Earnings Yield, FCF Yield: May capture some market sentiment

General Financial Health:

- Current Ratio, Quick Ratio: Assess short-term liquidity
- Net Profit Margin, ROA, ROE: Show how efficiently company is run
- Asset Turnover, Inventory Turnover, Receivables Turnover: Reflect asset utilization
- Total Assets, Total Liabilities, Total Debt, Net Debt: consolidated metrics

Sector Information:

- Sector: Captures industry-specific effects

By carefully selecting the most relevant predictors based on the Investopedia framework, we have reduced the number of features in our dataset from 223 down to 46. This significant reduction will allow us to focus on the core factors driving stock prices.

Project Question

We want to train a machine learning model to classify stocks as either buy-worthy or not buy-worthy, or predict the future value of a stock using our 46 predictors. The dependent variables are percent price variation and investment recommendations/class, respectively. We can either use consolidated financials rather than components or use dimensionality reduction techniques like PCA to remove redundant or highly correlated variables. We will conduct our analysis for a single year, subsequently repeating the model on 2014-2018 to assess its robustness. Furthermore, we can perform sector-specific analyses and comparisons, considering that each sector likely exhibits distinct behavior and characteristics. To incorporate macro trends, we can delve into time series analysis.

Baseline Model or Implementation Plan

In accordance with our three primary guiding questions, we will begin with a linear regression, and incorporate predictors that produce statistically significant coefficients. If any of these predictors are better represented by a higher-degree polynomial in our regression, we can use cross-validation to determine the optimal degree of the polynomial.

Subsequently, we will apply logistic regression to the same independent variables, using the classification of investment recommendation as the predictor. While logistic regression results may bear some similarities to the linear regression on continuous data, we expect to see some distinctions. While

predicting the sign alone theoretically seems easier than magnitude, predicting numbers around the bound of approximately 0 may be challenging.

Finally, we will assess the performance of decision trees. We will test various ensemble techniques such as bagging, random forests, boosting, and Bayesian additive regression trees. We will compare our regression-based and tree-based approaches based on testing accuracy.