ANLY-502 Final Report
Lujia Deng, Tianxing Jiang, Shuyang Yu, Yunfei Zhang
05/03/2020

Big Data Financial Statement Interpreter

I. Summary & Code Files

1. Data Collection

Acquired datasets from S3 (s3://dataset.secdatabase.com/), saved into our S3 bucket through AWS Athena

2. Preliminary Data Exploration

Explored the structures and variables in different datasets in the data source.

See "01-Dataset Exploration.ipynb"

3. Dataset Manipulation

Merged the datasets together using AWS Athena and select features based on our need using SparkSQL

See "02-Merge Data" folder

Reformat and cleaned the dataset

See "03-Reformatting Dataset.ipynb"

4. The Merged Dataset Exploration and Visualization

See "04-Visuals of the Final Dataset.ipynb"

5. Data Preparation for Machine Learning Modeling

See "05-Preprocessing before ML"

6. Predictive Modeling and Interpretation

Task 1: Predicting profit from selected variables ## See "06-ML task1.ipynb"

Task 2: Predicting industries from companies' performance ## See "07-ML task2.ipynb"

Failed Trial: Predicting profit increase / decrease ## See "08-ML trial task.ipynb"

7. Conclusion and Future Work

Details please see the report.

ANLY-502 Final Report
Lujia Deng, Tianxing Jiang, Shuyang Yu, Yunfei Zhang
05/03/2020

II. Introduction

Industries rise and fall and many current prosperous businesses did not exist decades ago. To understand the changing landscape of business over time, the SEC financial statements could shine the light for us by presenting the most critical financial information of all the public-traded companies. Our focus is on analyzing numerical data in all annual reports from January 2009 to December 2019, by wisely selecting valuable information and building predictive models to better understand the operations and profitabilities of those companies.

III. Data Overview and Exploration

1. Data Source and Collection

We collected the dataset from a github repository (https://tinyurl.com/ycdhs2ow) maintained by SECDatabase.com. The dataset was extracted from the online public database, EDGAR, of the U.S. Securities and Exchange Commission (SEC). It contains text and numerical information of all corporate annual and quarterly reports filed with the SEC since January 2009 by public companies selling securities in the U.S. or to U.S. investors.

The dataset consists of 7 tables stored on an S3 bucket (s3://dataset.secdatabase.com/), owned by SECDatabase.com. They are compressed in the Parquet format and together with a size exceeding 20GB.

SECDatabase suggested users of this dataset to leverage the AWS Athena, an interactive query service used to analyze data in S3 using SQL. Detailed instructions are provided on how to set up the database on Athena. We decided to follow the instructions and use the Athena platform to conduct preliminary data exploration and data preprocessing.

2. Initial Data Exploration

In the data source, there were 7 data tables in total and we mainly used 4 of them. We first saved the dataset into our S3 through Athena in AWS and conducted initial data exploration as follows.

Data Table 1: company_submission

This data table contains the submission information of companies, types of documents submitted, and the fiscal year. Each report has a unique identifier **accession_number_int**. The **document_type** indicates the type of the report, such as

annual report, quarterly report, to name a few. Through aggregating the dataset, we found that

1) There were 15,807 unique companies submitted documents during the last decade (2010-2019).
2) Table 1 shows the summary of types of documents submitted. Based on the distribution of the type of document, we decided to only include 10-K, as it contained a wide range of annual financial data while most of the companies submitted the 10-K form.

document_type	cnt
10-Q	9318
10-K	8423
null	5330
10-Q/A	3103
10-K/A	1558
20-F	847
20-F/A	338
40-F	149
10-KT	124
40-F/A	33
10-QT	27
10-KT/A	11
424B3	6
10-12G/A	3
10-QT/A	3
DEF 14A	1
10-12G	1
10-D	1

Table 1

3) Figure 2 shows the distribution of companies. It can be seen that a large proportion of companies are located in CA, NY, and TX.

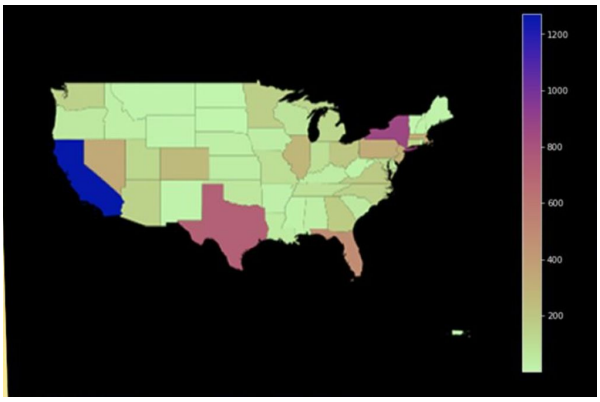


Figure 2

Data Table 2: report_presentation_section

This data table contains reported section information within each document. Two important data columns are **statement_type** and **section_sequence_id**. There are 10 unique statement types in **statement_type** (Table 3) and the section corresponding to each statement can be located through **section_sequence_id**.

As widely known, Statement of Income (I), Balance Sheet (B), and Statement of Cash Flows (C) would provide a comprehensive description of the companies. So, we summarized the sections to make sure there are enough submissions to those three.

Based on the result, we only include those 3 sectors in our dataset during the data cleaning procedure. Table 4 shows the count for each statement type in the dataset.

Statement Types
I - Statement of Income
IP - Statement of Income (Parenthetical)
CI - Statement of Comprehensive Income
CIP - Statement of Comprehensive Income (Parenthetical)
B - Balance Sheet
BP - Balance Sheet (Parenthetical)
C - Statement of Cash Flows
CP - Statement of Cash Flows (Parenthetical)
SE - Statement of Equity
SEP - Statement of Equity (Parenthetical)

Table 3

statement_type	cnt
\N	10857756
B	242149
C	240038
BP	226214
I	199841
SE	151089
CI	134102
SEP	47356
CIP	27601
IP	15550
CP	15103

Table 4

Data Table 3: report_presentation_line_item

This data table contains the line item sequence for each report section. It is used to link the **report_presentation_section** table and **data_point** table, using **section_sequence_id** and **datapoint_id**, respectively.

Data Table 4: data_point

Each data point represents a numerical value in the submitted documents. This table contains information about data point name, value, company, and represented period of time (quarterly or annually). There are more than 100,000 types of data points, i.e. attributes, across all the selected documents. Example data point types are Revenue, Net Income Loss, Inventories, etc. Therefore, we further manipulated our datasets by (Details explained in the next section):

1) Selecting 48 variables from 100,000 data points, by referring to related research and picking variables from top 150 frequently filled data points based on our domain knowledge in finance and accounting (detailed procedure please see "02-Merge Data" folder and explained in the Merge datasets section).
2) Identifying data duplication and missing value issues, and the need of reformatting the dataset.

IV. Data Preprocessing

1. Merge Data Tables

The goal is to extract all numerical data points related to the three financial statements (cash flow statement, balance sheet and income statement) in annual reports (10-K). The data values are stored in the datapoint table, but information we can leverage on to filter the data values that we want are stored in other tables. The challenge is to understand the relationships among the tables and join and filter them properly to extract the information we need. We provide the following diagrams to illustrate our data merge procedures. Detailed information is attached in the appendix.

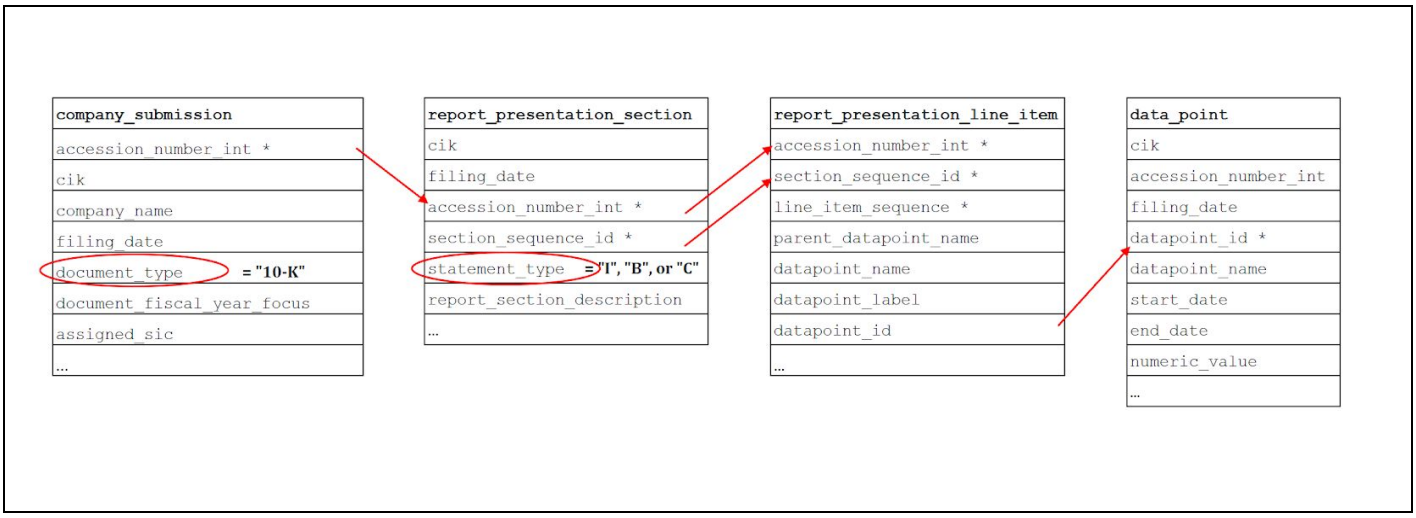


Diagram 5

We used **document_type** in **company_submission** table to select all annual reports (10-K) and **statement_type** in the **report_presentation_section** to select Income statement ("I"), Balance Sheet ("B"), and Statement of Cash Flow ("C"). Arrows show the procedures of merging tables using matching keys. The resulting table contains all the attributes in the **data_point** table with company and document information from the **company_submission** table.

2. Pivot the Dataset

For the dataset resulting from the previous step, each row is a unique data point associated with an annual report filed by a company. For example, the first row might be Amazon's net income loss in its 2019 filing and the second row might be its revenue in its 2019 filing.

Our analysis requires the dataset in the format that each row represents a company's annual report, containing all the data points within this report as attributes. Therefore, we pivoted our dataset to get this format.

(Code in **03-Reformat Dataset.ipynb**)

3. Feature Selection

The dataset contains 10,000+ unique types of data points (attributes). Attributes such as Revenue, NetIncomeLoss, and StockholdersEquity are more generic and applicable for almost all companies, while attributes such as "GasAndOilAreaReserve" are industry-specific and applicable for only a few ones. To make the most of our

dataset, we wanted to include more companies into our analysis with fewer missing values. Therefore, we selected 48 generic features from top 150 features with the most occurrences, based on our domain knowledge in finance as well as features used in other related literature.

A complete list of attributes selected is attached in the appendix.

4. **More Data Cleaning**

By looking at the selected dataset, we identified 3 main data cleanliness issues:

1) Duplication

Companies may fill the same data point in different documents, so there were a great proportion of duplications. To solve this, we dropped all the duplicates if two or more records have the same "data point name", "data point value", "company", and "document_fiscal_year_focus".

2) Different Time Periods

The data points may represent the quarter, semi-annual, or annual performance. To simplify the tasks for predictive models, we only kept the annual data.

3) Missing Values

As not all the data points are required to file, the final dataset is relatively sparse, where around 12% of the features have over 50% missing values. Yet, we kept all the features to feed the models.

5. **Explore Our Clean Dataset**

The clean dataset contains 41,741 rows and 36 columns. Each row represents an annual report associated with a company and a specific filing date. Each type of the data points is an attribute, such as revenue, net income and inventory.

In our analysis, net income (also referred as profit) is our target variable. Other attributes are treated as predictors. We explored the relationships between the response and predictors. Two examples are shown in Figure 6 and 7. While Revenue shows a stronger and positive relationship with Net Income, the relationship between R&D Expenses and Net Income are obscure. We used our predictive models to further make sense of which features are more important in predicting the target variable.

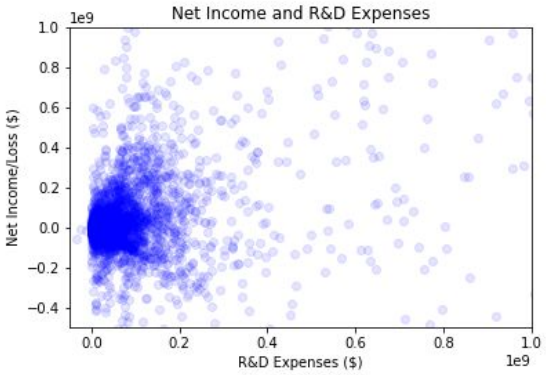


Figure 6

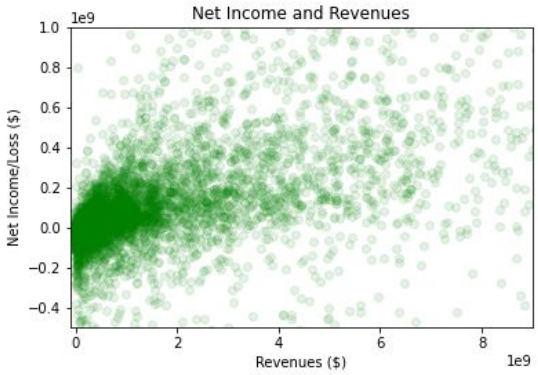


Figure 7

V. Methods and Results

Task 1: Predicting next-year profits from selected variables

We aimed to examine whether it is possible to use a company's first-year public financial to predict the second-year profit. We divide this task into classification and regression subtasks.

Methods we used are Logistic Regression and Light Gradient Boosting Machines (LightGBM). Logistic Regression uses a logistic function to statistically model a binary dependent variable, commonly applied in classification tasks. Light GBM is a fast, high-performance gradient boosting framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks.

Classification Task:

We created a label column according to each company's profit in each fiscal year, assigning value 1 if profit is greater than or equal to zero and -1 otherwise.

Logistic Regression Results

Accuracy: 0.75

Test Area Under ROC: 0.90

Confusion Matrix:

	Class 0 predicted	Class 1 predicted
Class 0 actual	1094	80
Class 1 actual	572	876

Table 8

LightGBM Results

Accuracy: 0.89

Test Area Under ROC: 0.94

Confusion Matrix:

	Class 0 predicted	Class 1 predicted
Class 0 actual	1021	153
Class 1 actual	156	1292

Table 9

From the results above, we found that

1) The LightGBM model outperforms the Logistic Regression model. Specifically it does a better job at identifying those profit-making companies.
2) The overall accuracy is high, which means the company's next-year gain/loss is predictable using last year's financial data.

Regression Task:

We further applied the LightGBM regressor model to predict the exact value of profit. We randomly selected 100 companies and compared the true and predicted values of their profits (Figure 10).

The blue line and the red line mostly overlap with each other, suggesting that our model performs well in predicting the profits for average companies. The prediction is off for those exceptional companies, indicated by the red circle.

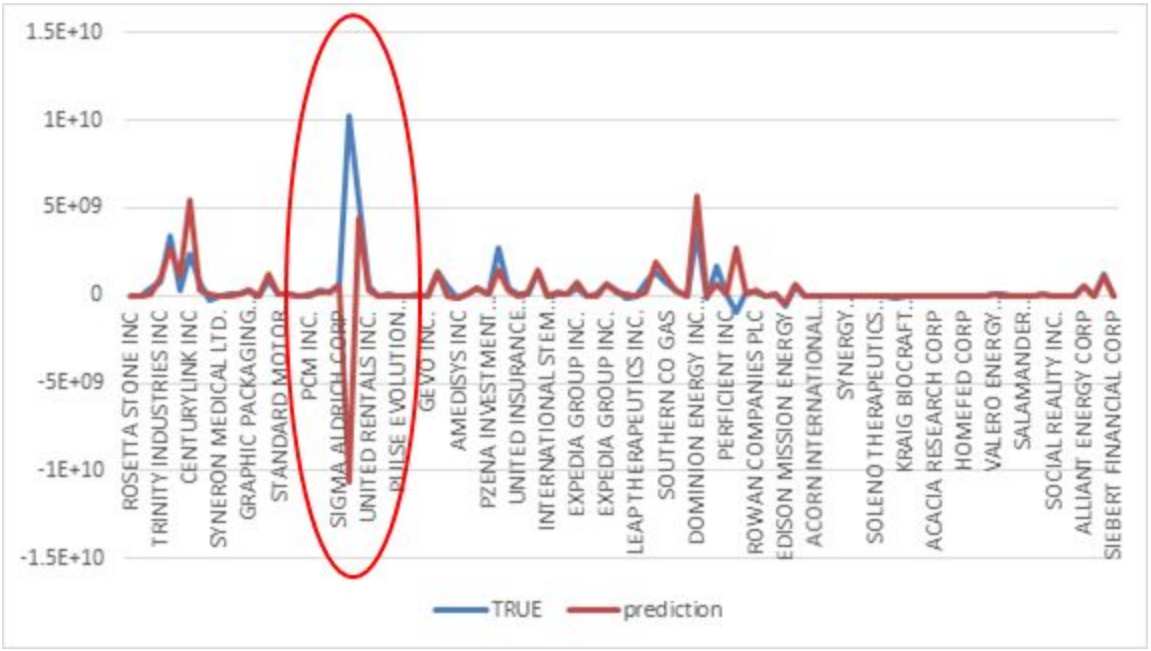


Figure 10

The result indicates the values of profits are predictable as well. The R square is 0.88, suggesting that most of the variance in the data is explained by the model. The feature importance chart shows that the most important features can be categorized into earnings (e.g. EarningsPerShareBasic, Revenue, OperatingIncomeLoss), and cash flow (e.g. Net Cash Provided by Used in Financing Activities), which shed light on the indicators of next-year profitability.

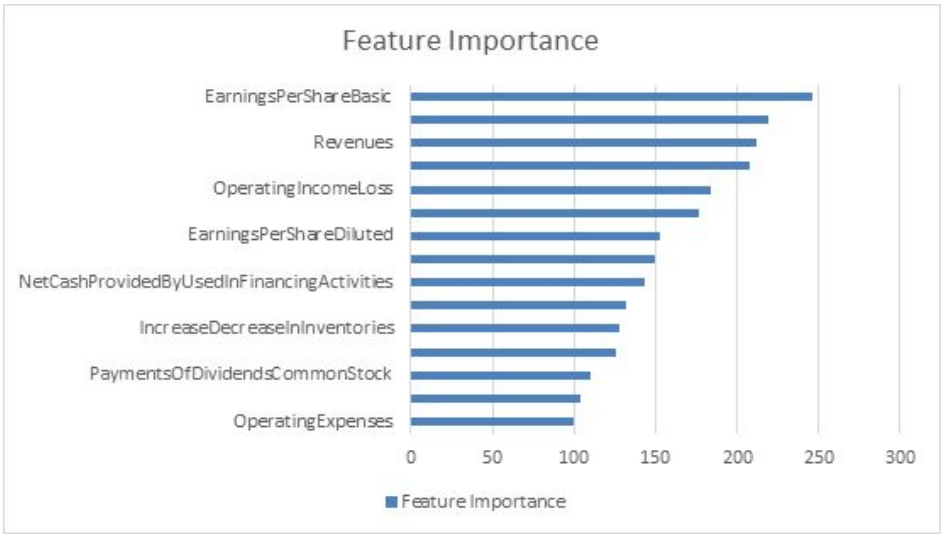


Figure 11

ANLY-502 Final Report
Lujia Deng, Tianxing Jiang, Shuyang Yu, Yunfei Zhang
05/03/2020

Task 2: Predicting industries from companies' performance

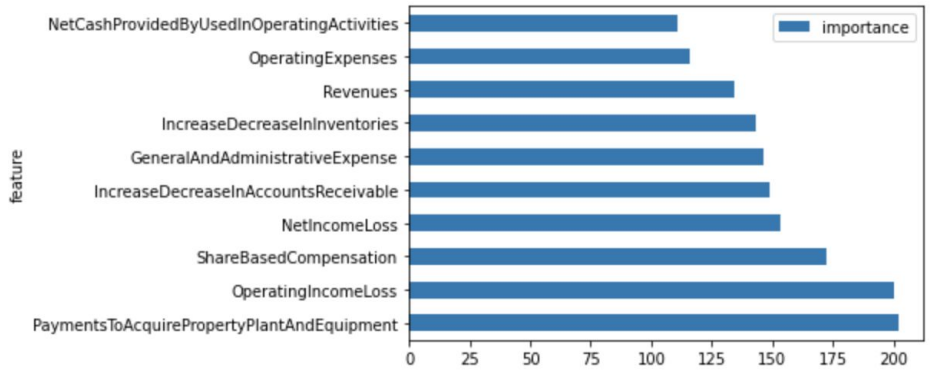
Each company was assigned with an SIC code (Standard Industry Classification). There are 10 sectors of industry. For simplification, we categorized them into 2 groups:

1) Agriculture and manufacturing, labeled as "1", including: Agriculture, Forestry and Fishing; Mining; Construction; Manufacturing; Transportation, Communications, Electric, Gas and Sanitary service.

2) Sales and services, labeled as "2", including Wholesale Trade; Retail Trade; Finance, Insurance and Real Estate; Services; Public Administration.

The aim of this task is to use the financial performance of the companies (excluding company name), to predict which industry a company belongs to. Again, we applied LightGBMClassifier to build the classification model.

Shown by the following diagram, we got an accuracy of 0.83 and the area under roc was 0.92, which was good. By looking at the feature importance and comparing the mean in those features between the two categories, we found that:



PaymentsToAcquirePropertyPlantAndEquipment	6.601431e+08
OperatingIncomeLoss	1.068636e+09
ShareBasedCompensation	1.335568e+07
NetIncomeLoss	8.773825e+08
IncreaseDecreaseInAccountsReceivable	3.391371e+07
GeneralAndAdministrativeExpense	2.407248e+07
IncreaseDecreaseInInventories	8.529190e+07
Revenues	8.762541e+09
OperatingExpenses	1.069280e+09
NetCashProvidedByUsedInOperatingActivities	1.816322e+09

Mean of the Features, Agriculture and Manufacturing

PaymentsToAcquirePropertyPlantAndEquipment	1.747440e+08
OperatingIncomeLoss	1.961909e+08
ShareBasedCompensation	1.891467e+07
NetIncomeLoss	8.128655e+08
IncreaseDecreaseInAccountsReceivable	8.036808e+06
GeneralAndAdministrativeExpense	1.617463e+08
IncreaseDecreaseInInventories	-1.922402e+06
Revenues	1.966043e+09
OperatingExpenses	1.738115e+08
NetCashProvidedByUsedInOperatingActivities	9.838758e+08

Mean of the Features, Sales and Services

Diagram 12

1) The most important feature that separates the two industry groups is "Payment to Acquire Property, Plant, and Equipment", where agriculture and manufacturing companies spend more in this category than sales and services companies do. This indicates that the industry group 1 generally needs more expenditure for long-lived, physical assets.

2) The two industries have similar profitability in terms of net income loss; however, Agriculture and Manufacturing has higher operating income, which means it spends more on tax and interests than the Sales and Services industry.

3) Agriculture and Manufacturing has lower cost in G&A expenses (including rent, utilities, insurance, legal fees, and certain salaries).

4) While the two industry groups have the same level of profits, Agriculture and Manufacturing has higher average values of revenue, operating expenses, and account receivable. This indicates that Agriculture and Manufacturing needs more expenditure and large volumes in sales to keep up with the Sales and Services industry's profit level.

This task provides insights for us to better understand the differences of industries.

VI. Conclusion, Takeaways, and Future Work

In this study, we mined the SEC financial statement dataset to understand 1) which predictors are more important in predicting a company's profit, and 2) financial differences between Agriculture & Manufacturing and Sales & Services industry groups.

After using logistic regression and Light GBM models to predict profits and industry groups, we found that 1) features related to **current year's earnings** (e.g. Earnings Per Share, Revenue, Operating Income Loss etc.) and **cash flow** (e.g. Net Cash Provided by Used in Financing Activities, etc.) have the greatest predictive power in forecasting next-year profits. The second finding is aligned with our expectation, since a company with a good cash flow has more resources to leverage on in its investment and operations activities; therefore, it is more likely to generate more profits, compared with those with limited cash flow. The first finding implies a probably obvious but often neglected fact that, in forecasting a company's future profits, the most informative data is its most recent financials; data from even longer before will provide less value added, given that a company's next-year performance is strongly correlated with the current year's performance.

Our good model performance in predicting industry groups suggests that their financial structures are significantly different. For example, the most important feature that distinguishes the two groups is "Payment to Acquire Property, Plant, and Equipment", aligned with the fact that Agriculture and Manufacturing requires more investment in infrastructure.

Overall, our models performed well in analyzing financials. In the future, we think it will be meaningful to further engineer our features. For example, to calculate some financial ratios as predictors, and make a model comparison with our current one.

that often used by financial analysts. In this way, we can better mimic the workflow of a financial analyst and build a better financial statement interpreter.

VII. Reference

Gepp, A., & Kumar, K. (2015). Predicting financial distress: a comparison of survival analysis and decision tree techniques. Procedia Computer Science, 54, 396-404.

Al-Khatib, H. B., & Al-Horani, A. (2012). Predicting financial distress of public companies listed in Amman Stock Exchange. European Scientific Journal, 8(15).

Wikipedia: Standard Industry Classification

Investopedia: to help us understand most of the financial terms

IX. Appendix

1. Selected Features Contained in the Final Dataset

Selected Features
CashAndCashEquivalentsAtCarryingValue
NetIncomeLoss
OperatingIncomeLoss
Revenues
SalesRevenueNet
CostOfRevenue
EarningsPerShareBasic
EarningsPerShareDiluted
NetCashProvidedByUsedInOperatingActivities
NetCashProvidedByUsedInFinancingActivities
NetCashProvidedByUsedInInvestingActivities
NetCashProvidedByUsedInOperatingActivitiesContinuingOperations
NetCashProvidedByUsedInFinancingActivitiesContinuingOperations
NetCashProvidedByUsedInInvestingActivitiesContinuingOperations
ShareBasedCompensation
PaymentsToAcquirePropertyPlantAndEquipment
OperatingExpenses
GeneralAndAdministrativeExpense
SellingGeneralAndAdministrativeExpense
SellingAndMarketingExpense
IncomeTaxesPaid
ResearchAndDevelopmentExpense
PaymentsForRepurchaseOfCommonStock
CostOfGoodsSold
CostOfGoodsAndServicesSold
CostOfServices
RepaymentsOfLongTermDebt
PaymentsToAcquireBusinessesNetOfCashAcquired
PaymentsOfDividendsCommonStock
PaymentsOfDividends
LaborAndRelatedExpense
PaymentsOfFinancingCosts
IncreaseDecreaseInAccountsReceivable
AccountsReceivableNetCurrent
IncreaseDecreaseInInventories
IncreaseDecreaseInAccruedLiabilities
IncreaseDecreaseInAccountsPayable
LiabilitiesCurrent
Liabilities
AccountsPayableCurrent
StockholdersEquity
Assets
AssetsCurrent
GainLossOnDispositionOfAssets
CommonStockValue
PreferredStockValue
Goodwill
PropertyPlantAndEquipmentNet

2. Failed Task: Predicting profit increase/decrease

We aimed to predict the relative increase/decrease of profit compared to last year. However, the models failed to have satisfying results, which we will discuss later.

We divided this task into classification and regression subtasks as task 1. The label for classification is defined as whether the difference between this year's profit and last profit is greater than zero or not. The regression task aimed to fit this difference. Before applying machine learning models to fit our data, we firstly did some feature engineering.

Feature Engineering

We selected the top 15 most important features based on the feature importance provided by a LightGBM regression model.

Classification Task

We created label columns with 1 representing increase in profit and 0 representing decrease.

LightGBM

Accuracy: 0.57

Test Area under ROC: 0.55

Confusion Matrix:

	Class 0 predicted	Class 1 predicted
Class 0 actual	524	565
Class 1 actual	521	922

From the above result, we can see that the model failed to give a precise prediction.

Regression Task

We now use LightGBM Regressor to fit change in profit. As a result, we got a R square of -0.04. The result was not satisfying as the R square is negative.

Discussion

When looking back at our data preprocessing procedure before fitting the model, we found one mistake we made might lead to the failure in models' bad performance. In this case, we filled the missing value for profit as "0", so that if a year's profit is null, it will influence the label data for this year, its last year. For example, if the actual profit for 3 consecutive years is 100, 200, and 300, the actual difference for the later 2 years should be 100 and 100. However, if the data of year 2 is missing (200 filled as 0), the labels for the later 2 years under our rule would become -100 and 300, which have

dramatically more bias than the actual situation. So, we believe this mistake influenced our result.

However, in task 1 our model was already able to predict the exact value of the profit next year, which seems similar and even more difficult to predict than this task. So, we did not refine this model which aimed to predict the performance difference in profit compared to last year.

X. Division of Labor

Frame the data science problem, find the dataset, think of paths to collect and process data -- All of us

Data Collection --Yunfei

Data Exploration, Manipulation, and Cleaning --Lujia & Yunfei

Understand Financial Statement --Tianxing

Feature Selection --Tianxing, Lujia, & Yunfei

Predictive Models --Shuyang & Yunfei

Writeup --Lujia, Yunfei & Shuyang

Presentation --All of us