Progress Report I - Financial Forecasting using Annual/Quarterly Filings

Kechengjie Zhu
kz2407@columbia.edu

Xuchen Wang
xw2747@columbia.edu

Yao Xiao
yx2696@columbia.edu

Zhenyu Yuan
zy2492@columbia.edu

Zixiang Yin
zy2444@columbia.edu

1 Problem Definition and Progress Overview

The Financial Forecasting using Annual/Quarterly Filings project is proposed and supervised by Simran Lamba from JPMorgan. Our research goal is to forecast companies' financial performance using: 1) annually/quarterly released 10k/10Q filings including Balance Sheet, Income Statement, and Cash Flow Statement; 2) market sentiment; and 3) macroeconomic indicators. A high-quality performance forecast aids market participants such as investors to make better trading decisions and manage their portfolios more suitably while outperforming the market.

Mathematically, our project aims at building a multivariate model f that takes a constant $k \in N^*$ and time series data $\mathbf{X}$ as input and outputs Y_{t+1} where $X_i \in R^m$, $Y_i \in R^n$, and

$$\mathbf{X} = (X_{t-k+1}, X_{t-k+2}, \ldots, X_t, Y_{t-k+1}, Y_{t-k+2}, \ldots, Y_t)$$

In the first stage of the project, we have completed data wrangling and prepossessing that mainly handles the missing values. After that, we built a baseline uni-variate ARIMA model that predicts Y_{t+1} only based on $Y_{t-k+1}, Y_{t-k+2}, \ldots, Y_t$.

2 Exploratory Data Analysis

2.1 Data Collection

At the current stage, the baseline mode only requires fundamental data sourced from a public stock market data API called EODHIstoricalData (EODHD). For a given list of stock symbols, it provides fundamentals in various temporal dimensions covering detailed terms in three financial statements (Balance Sheet, Income Statement, and Cash Flow Statement). We have therefore pulled the last 5 years of quarterly fundamental data for all components of the NASDAQ Composite. After collecting the data, we found that a few companies had less than 20 entries and eventually decided to exclude them. We ended up with 46640, 46920, and 46580 entries for Income Statement, Balance Sheet, and Cash Flow Statement respectively.

2.2 Handling Missing Values

During the data collection, we notice that many features contain missing values so we plot the percentage of missing values in each feature per statement as shown below.

Preprint. Under review.

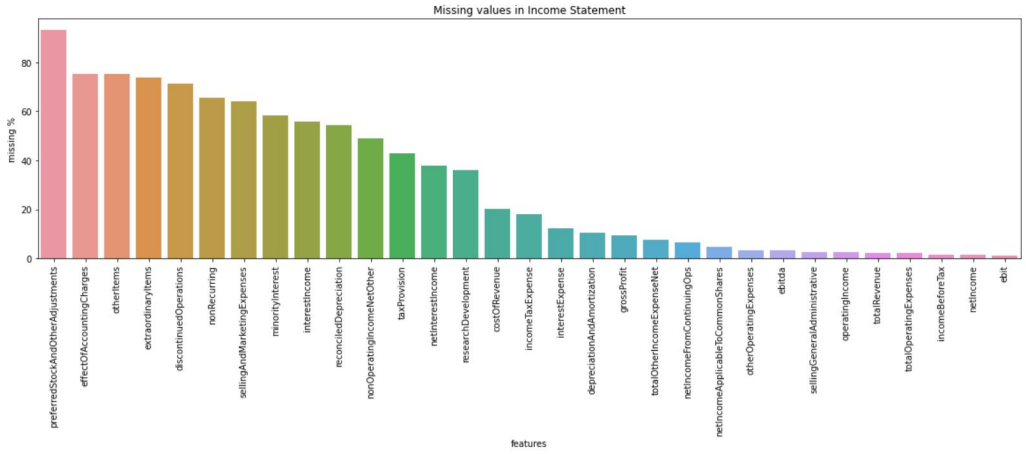

Figure 1: Percentage of missing values for Income Statement

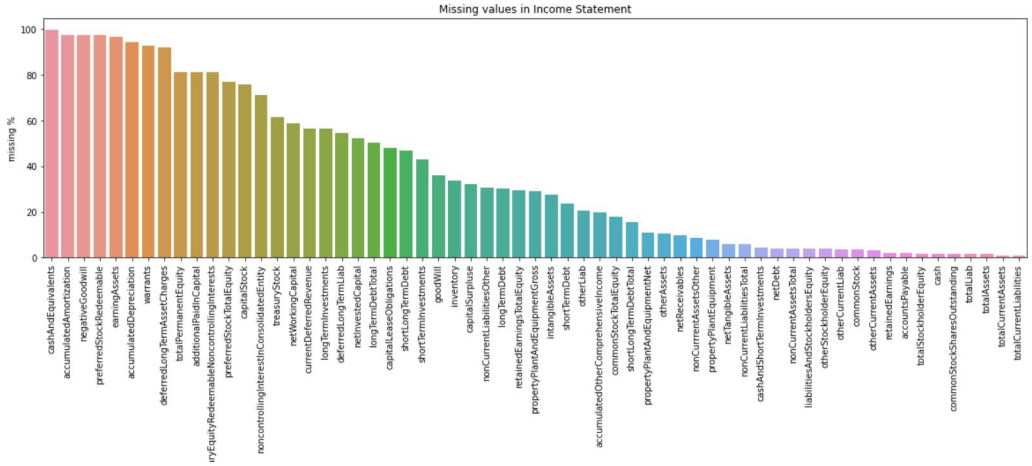

Figure 2: Percentage of missing values for Balance Sheet

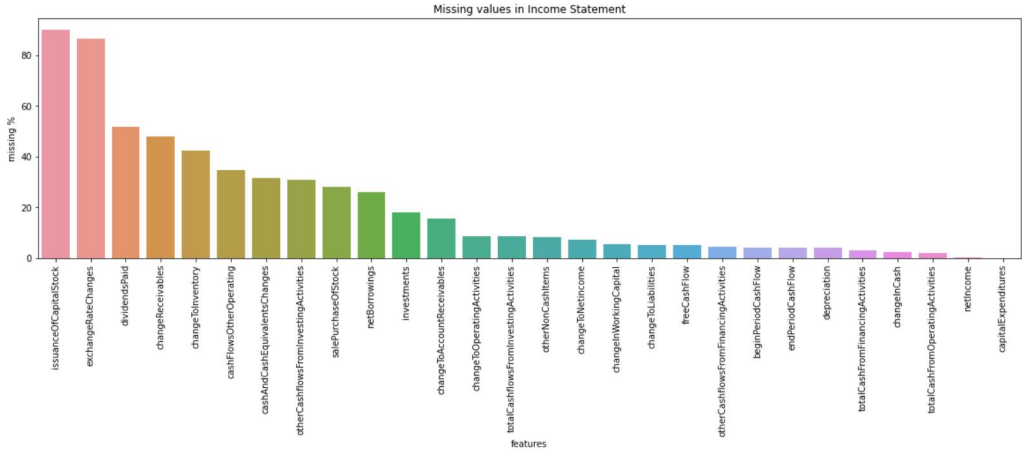

Figure 3: Percentage of missing values for Cash Flow Statement

The cutoff point is set at 10% since features with too many missing values cannot be used for time series analysis. That is to say, features that contain more than 10% missing values will be dropped. Eventually, we got 16 features from Income statements and 22 features each from Balance Sheet and Cash Flow Statement.

3 Methodology

3.1 Related Literature

Financial forecasting is a framework for assessing a company's future prospects based on its industry, strategy, and financial performance. Even highly skilled financial analysts are incapable of perfect prediction. Scholars have proposed several quantitative methods to get the forecast more accurate on a case-by-case basis.

The use of uni-variate historical time series by statisticians has been proven successful in numerous tasks including stock price prediction. Given the assumption of stationarity, Auto-Regressive (AR) and Moving Average (MA) models as well as their combination ARIMA model were introduced and used in early trials.

In recent decades, there is a great advancement in the field of deep learning, opening a new track for solving time series questions. For instance, a seq2seq structure that combines conventional AR models with RNNs was introduced to do predictions. In addition, a multi-faceted modeling approach was developed to leverage univariate and multivariate models and identify the best-performing model setting.

3.2 Our Approach

Our baseline model is nothing but seasonal ARIMA model which predicts Y_{t+1} only based on $Y_{t-k+1}, Y_{t-k+2}, \ldots, Y_t$. We did model selection by a grid search with hyperparameters (max_p, max_d, max_q) $= (5, 2, 5)$ and an evaluation metric of AIC. Using feature *ebit* from Income Statement as an example, the optimal model for the company with symbol 'A' is ARIMA(0, 0, 2) which gives the following prediction on the last four periods.

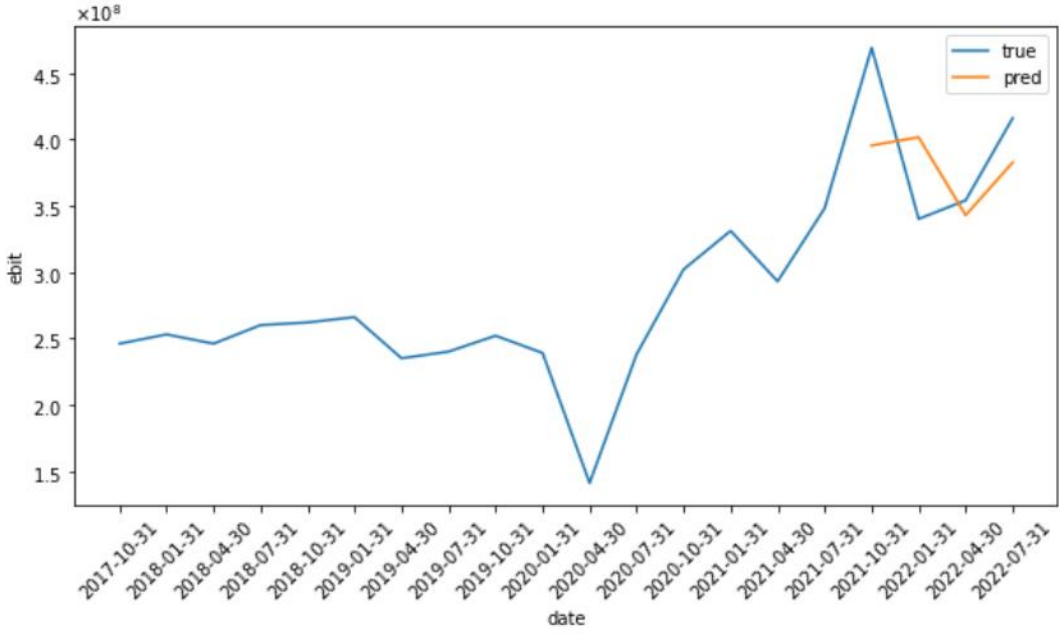

Figure 4: ARIMA prediction on ebit for the last four periods

Then, we repeat the approach described above and calculate the prediction accuracy using Mean Absolute Percentage Error (MAPE) where A_i and P_i are actual and predicted values respectively:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}|\frac{A_i - P_i}{A_i}|$$

The distribution of log MAPE is plotted below in Figure 5. Based on the plot, we can find that the model works fine for most companies as the majority is on the left.

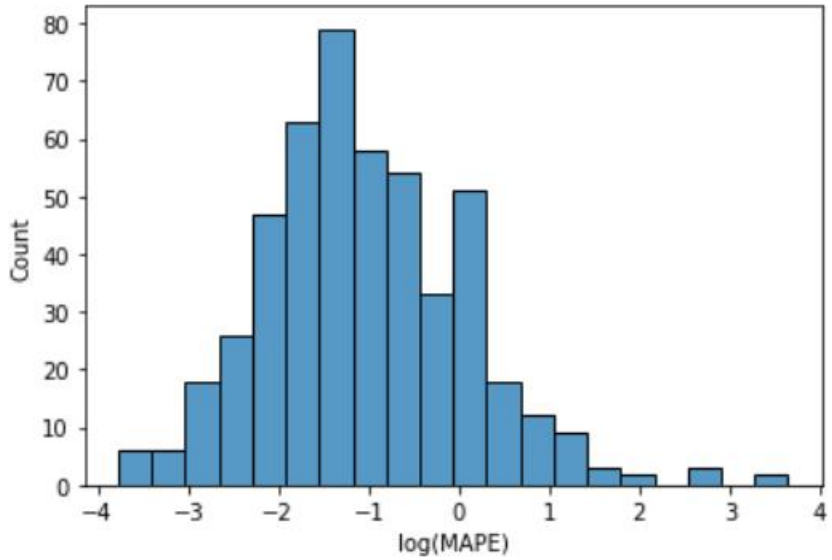

Figure 5: Distribution of MAPE in log scale

4 Next Steps

In terms of data, we plan to collect more extended time series fundamentals data and use social media sentiments as well as macro indicators to improve our forecast accuracy. We should also consider using a multivariate instead of a uni-variate model next, which makes feature selection necessary. This is because some features in financial statements are highly correlated or even exactly equivalent, possibly resulting in overestimated model accuracy. The last but most important thing is using deep learning, especially transformers to make predictions, which we believe will give us a huge performance lift. Considering the model evaluation, we can alternatively compare our forecast again experts' to get a sense of how our model works in the real world.

References

Antony Papadimitriou, Urjitkumar Patel, Lisa Kim, Grace Bang, Azadeh Nematzadeh, and Xiaomo Liu. 2020. A multi-faceted approach to large scale financial forecasting. In Proceedings of the First ACM International Conference on AI in Finance (ICAIF '20). Association for Computing Machinery, New York, NY, USA, Article 5, 1–8.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2019. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting (2019).

Jeffrey E Jarrett and Eric Kyper. (2011) ARIMA modeling with intervention to forecast and analyze chinese stock prices. International Journal of Engineering Business Management 3, 3 (2011), 53–58.

Peter Whittle. 1951. Hypothesis testing in time series analysis. Vol. 4. Almqvist & Wiksells boktr.

Xiaodong Sun, Dinesh K Gauri, and Scott Webster. 2011. Forecasting for cruise line revenue management. Journal of Revenue and Pricing management 10, 4 (2011), 306–324.