Financial Forecasting using Quarterly Filings

Data Science Institute
COLUMBIA UNIVERSITY

Zixiang Yin, Xuchen Wang, Yao Xiao, Zhenyu Yuan, Kechengjie Zhu
Mentor: Simran Lamba & Instructor: Sining Chen

Data Science Capstone Project with JPMorgan

Background and Methods

- In recent years, artificial intelligence has been widely used to predict stock prices, but limited attention was paid to companies' financial status which is one of the most influencing factors against stock price. Creating a high-quality performance forecast would definitely aid market participants such as investors to make better trading decisions and manage their portfolios more suitably while outperforming the market.
- For this reason, our research focused on forecasting companies' financial performance using quarterly released 10-K/10-Q filings including Balance Sheet, Income Statement and Cash Flow.
- Mathematically, we aimed at building a multivariate multi-target model f_θ that takes historical time series data **X** and k as input and output **Y** where k is the number of time steps looked back:

$$\mathbf{X} = (X_{t-k+1}, X_{t-k+2}, \dots, X_t)^T, \quad \mathbf{Y} = X_{t+1}, \quad X_i \in R^d$$

- We leveraged the EODHistoricalData API to collect financial fundamentals (Balance Sheet, Income Statement and Cash Flow) of 2,321 Nasdaq Composite components for at least 5 years and ended up with a dataset of 171,560 entries as well as 128 features.
- Based on the exploratory analysis, missing values were prevalent in most features so we set a cut-off point of 5% to filter out "bad" ones. We also dropped companies that contain NAs in the remaining features from the dataset.

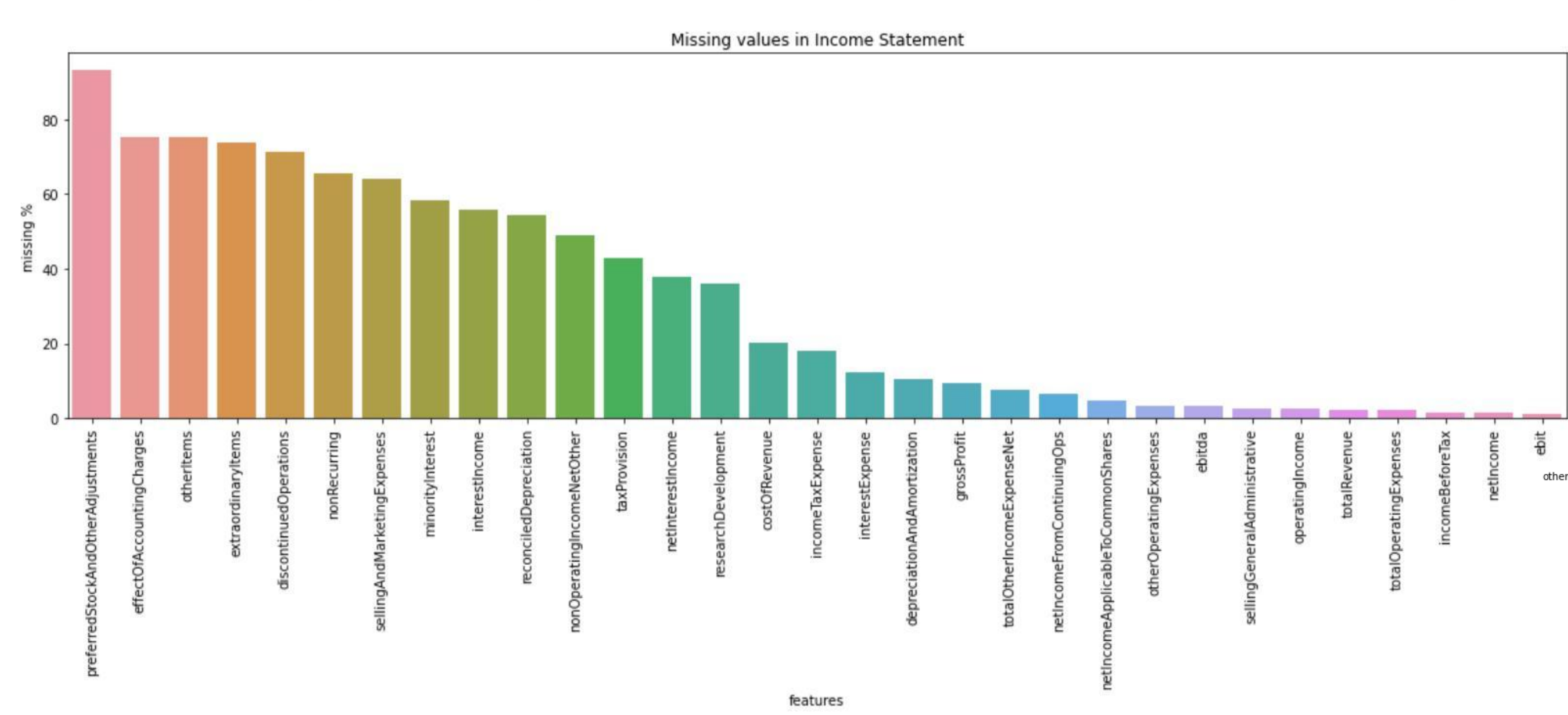

Figure 1. Percentage of NAs in Income Statement.

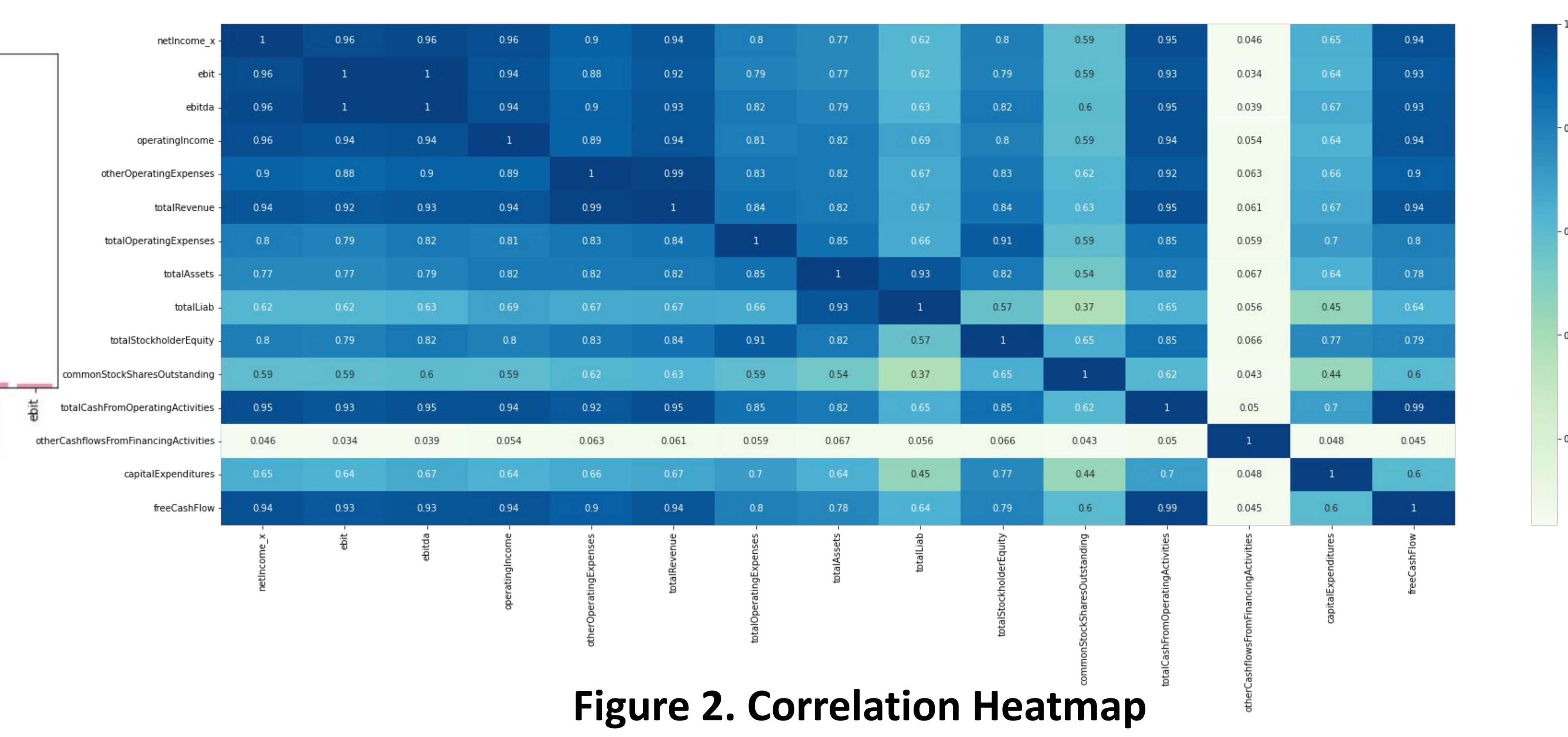

Figure 2. Correlation Heatmap

- Since financial fundamentals are highly correlated, we selectively deleted features with high correlation by drawing a heatmap and finally got a "good" dataset of 16,368 entries as well as 7 features among which 3 are from Balance Sheet, 2 are from Income Statement, and 2 are from Cash Flow.
- We used univariate ARIMA models as the baseline and built LSTM as well as Transformers for performance improvement. During model selection, AIC and MSE were used as loss functions for ARIMA and DL models respectively.

Results

- We built company-wise ARIMA models for each of the seven targets by automatically searching best hyperparameters (p, d, q).
- By doing a grid search on LSTM units and the number of steps looked back, the optimal model is achieved with a look_back of 14 quarters and 76 LSTM units.
- We tuned the transformer by look_back only with six attention layers and a slight dropout of 0.1. The smallest loss is obtained when look_back equals 4.
- Since companies' size vary from one to another, we used Symmetric Mean Absolute Percentage Error (SMAPE) as the evaluation metric for model comparison where:

$$\mathbf{SMAPE} = \frac{200\%}{N} \sum_{i=1}^{N} \frac{|F_i - A_i|}{|F_i| + |A_i|}$$

- The final result is shown in the right table, it's obvious that company-wise ARIMA models dominate the most in terms of SMAPE. But, the gap between ARIMA and LSTM is quite small.

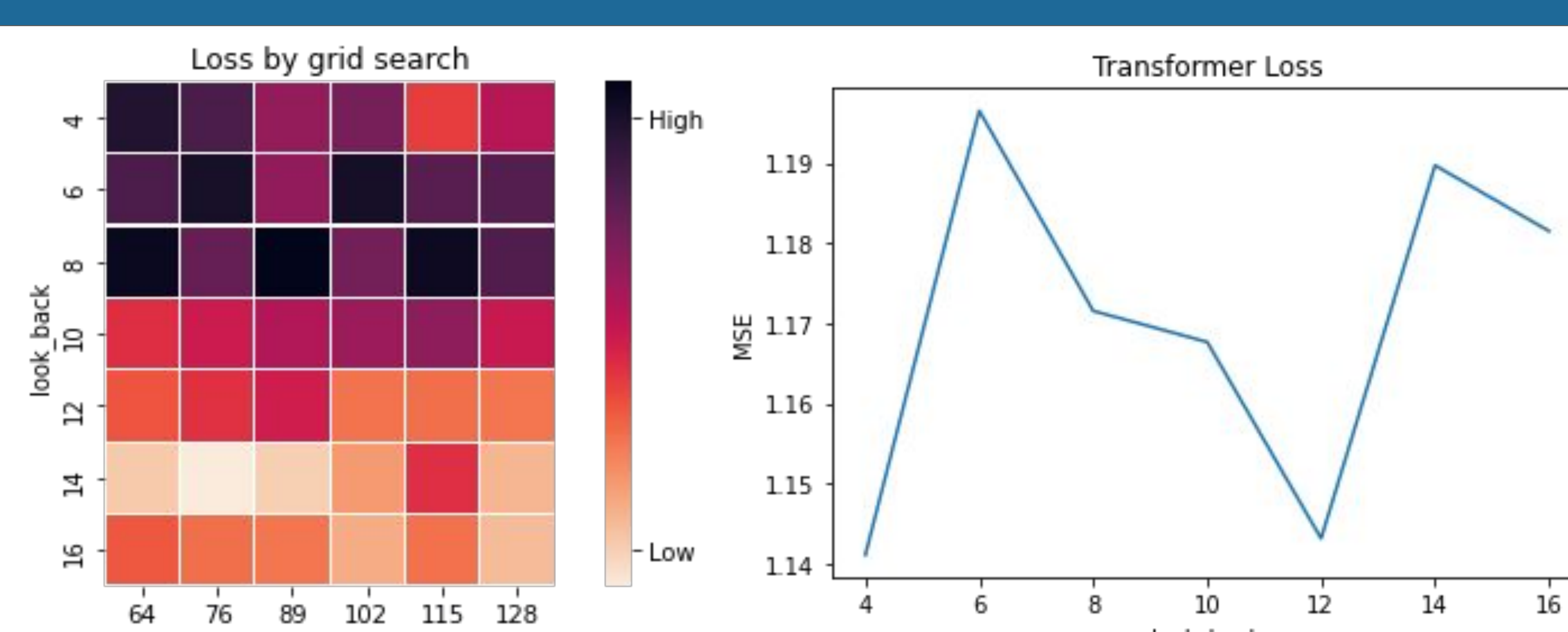

Figure 3. Grid search on LSTM and Transformer

	EBIT	OE	Liab.	Equity	CSS	CFFA	CE
ARIMA	23.57	19.90	8.83	9.16	20.18	34.36	44.83
LSTM	23.42	19.99	8.68	11.53	21.20	37.91	48.72
Transformer	38.81	42.83	26.47	28.22	35.78	58.95	82.21

Table 1. Model Performance

Acknowledge:

We are grateful to our mentor Simran Lamba for the guidance and support during the project as well as Sining Chen for facilitating the project and offering feedback.

References:

1. Ariyo et al. (2014). Stock Price Prediction Using the ARIMA Model, 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014, pp. 106-112.
2. Papadimitriou et al. (2015). 2021. A multi-faceted approach to large scale financial forecasting. International Conference on AI in Finance, Article 5, 1–8.
3. Vaswani, Ashish, et al. (2017). Attention Is All You Need. arXiv, https://doi.org/10.48550/arXiv.1706.03762.