Monthly Revenue Forecasting for Paychex

Shijing Li, Yuan Wang, Lingyu Ye, Yangxin Fan
Paychex-2 Team, DSC 483
Goergen Institute for Data Science
University of Rochester

1: Introduction

Paychex is a leading provider of human capital management solutions for payroll, benefits, human resources, and insurance services. Paychex's revenue is the total amount of income generated by the sales of its different service products. An accurate prediction of future revenue will better Paychex in risk management since future revenue trends have a profound impact on stock performances [1]. Forecasting revenue is also essential for future budget planning and resources allocation to avoid any possible downward trends in revenue. Identifying highly correlated external features or significant events like the COVID-19 pandemic will potentially bring valuable business insights regarding revenue activities.

Monthly revenue prediction forecasting is well suited for time-series prediction models. There are two main categories of time series prediction methods, one being the traditional method and the other is the more complex deep learning-based method. The traditional methods mainly focus on parametric models, including Autoregressive (AR) [2], Moving Average (MA) [3], Seasonal Autoregressive Integrated Moving-Average (SARIMA) [4], Vector Autoregression (VAR) [5], and Exponential Smoothing [6]. Some of the popular deep learning-based methods are Recurrent Neural Networks (RNNs) [7], Long Short-Term Memory (LSTM) [8], Gated Recurrent Unit (GRU) [9], Autoencoder [10], and Attention Mechanism [11]. Under the condition of sufficient training data, deep learning-based methods can perform much better than the traditional methods by better recognizing complex patterns and handling uncertainties in long-term prediction.

We construct four different time-series forecasting models: SARIMA, ThymeBoost [12], FB-Prophet [13], and LSTM. They are all trained by monthly revenue data from June 2017 to November 2020 (41 months) and evaluate with data from December 2020 to August 2021 (9 months). Our best model FB-Prophet plus peak indication achieves promising results with only 4.9% test MAPE and $13.33 million test MAE. We make the following four major contributions. To summarize, (1) identify the increase in GDP per capita, inflation rate, and job opening rate may indicate an increase in Paychex revenue; (2) propose a promising revenue forecasting model with less than 5% MAPE; (3) discover the potential revenue growth opportunities in Texas and Illinois, and construction industry; (4) provide suggestions to boost forecasting performance by incorporating more data into the LSTM model, adding future marketing peaks into the FB-Prophet model, and collect more client data like monthly employee count of each client as features for all the prediction models.

The remainder of the paper is structured as follows. In section 2, we describe our datasets and Data Preprocessing. In section 3, we discuss our works in exploratory data analysis (EDA), model constructions, tunings, and selections are presented. In section 4, results and comparisons of model performance are shown in section 5. Finally, section 6 summarizes our works and lays out a roadmap for future improvements.

2: Dataset

To identify the external activities that have significant impacts on Paychex's revenue, our main dataset is the Paychex's monthly revenue data. From the wide range of external activity sources, we build three well-structured datasets that capture COVID-19, economic and labor activities for us to discover correlations with Paychex's internal revenue activity.

2.1: Internal Datasets

We received internal monthly revenue datasets from Paychex over the fiscal year of 2017 to the fiscal year of 2021. More specifically, it covers the period from July 2017 to September 2021. There are 823,963 rows in total with 247 columns including client information (Client ID, EE count, Business Type, State code), total monthly revenue, total monthly product count, and monthly revenue of 7 service products. Overall, the dataset is clean with no missing or duplicate values.

2.2: External Datasets

We explore and test a broad range of external features. After discussing with the client, we finalize our external datasets with three aspects. The first dataset is constructed based on the data we get from the U.S. Bureau of Labor Statistics which contains features like job opening rate, hiring rate, separation rate, and unemployment rate. We also collect data from the Bureau of Economic Analysis with features like GDP per capita, personal income per capita, and inflation rate. Since the revenue data that we get includes the COVID-19 period, we also collect COVID-19 related data from the Centers for Disease Control and Prevention such as the case number, death number of vaccination numbers. All the datasets are monthly except the inflation rate which is quarterly.

2.3: Data Preprocessing

To better perform on the data, we do all-sided data preprocessing before fitting them into the models. There are three stages to our data preprocessing. The first stage is data cleaning where we check the missing and duplicated values. Then we standardize the data into the same granularity with either monthly or quarterly data. The second stage is data transformation where we transform the internal and external data into the same data type, and then merge the internal and external datasets based on Client ID and State Code. The final stage is data reduction. We select and aggregate the monthly revenue of all industries with the corresponding Client ID, State Code, and industry. We then perform feature selection and extract related features for modeling.

3: Exploratory Data Analysis

To better understand the internal and external datasets. We perform some exploratory data analysis to analyze them more in-depth and extract findings. We first decompose Paychex's monthly total revenue to discover any trend, seasonality, and residuals. Then we conduct correlation analyses to identify external features that are highly or moderately correlated with Paychex's revenue. To get a better sense of the distribution of revenue by industry and by service product, we visualize the distribution. From these visualizations, we can identify the top states and service products that take up the highest percentages of the total revenue.

3.1: Revenue Overview and Decomposition

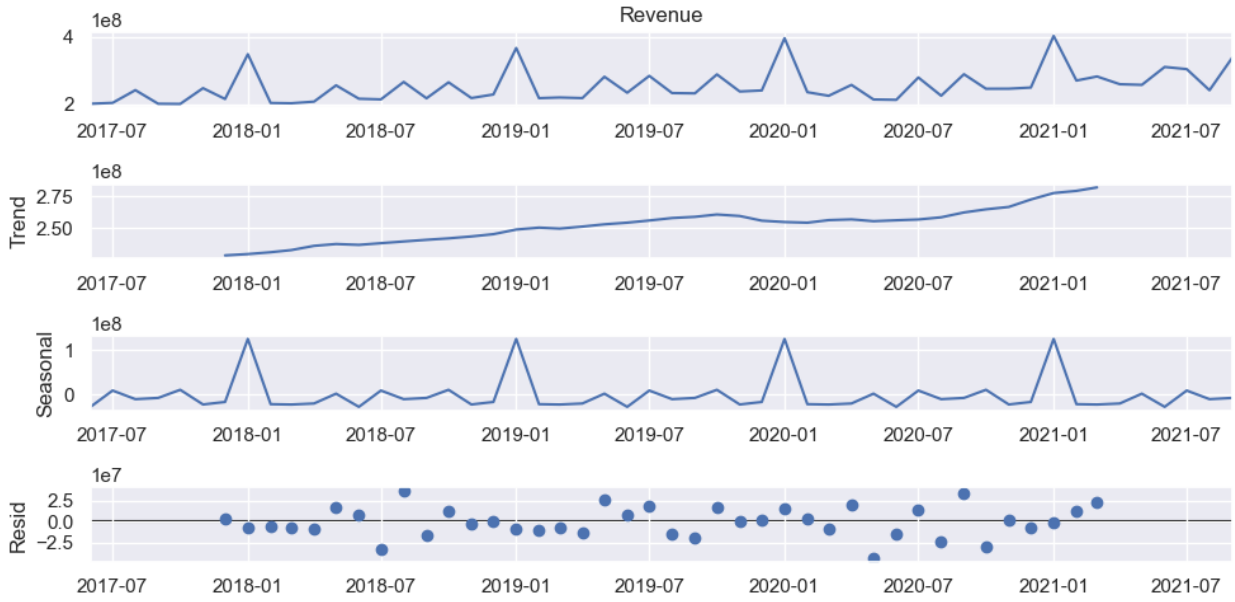

Fig. 3-1-1 Decomposition of Paychex Monthly Total Revenue

We decompose Paychex's monthly total revenue from June-2017 to August-2021 into three parts, trend, seasonality, and residual (see Fig. 3.1). The trend is calculated from a 12-month moving average. We observe an overall increasing trend. As for seasonality, January is always the peak of every fiscal year while the revenues for other months stay relatively flat. We find that there are only white noises (no obvious patterns) in residuals, which is desirable in time-series data decomposition.

3.2: Correlation Analysis

We conduct a correlation analysis of the period under COVID-19 and the period without COVID-19 to better understand the correlation between external and internal data for modeling. We use seasonality adjusted revenue, which is the seasonality adjusted revenue's residuals and trend.

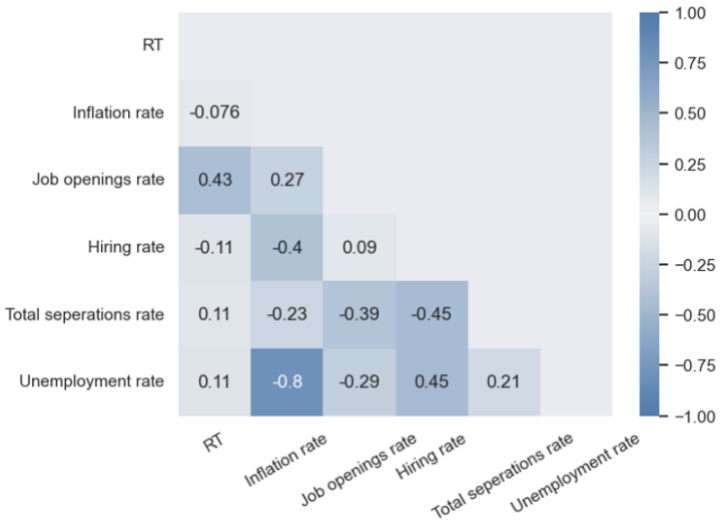

Fig. 3-2-1 SAR and external features correlation analysis

The correlation analysis of Paychex seasonality adjusted revenue (SAR) and external features from the Bureau of Labor Statistics (BLS) is shown in Figure 3-2-1. We use monthly data from June 2017 to August 2021 (about 4 years) to perform the analysis. Compared with other features that only have an absolute correlation of around 0.1, the Job Opening Rate is moderately correlated with seasonality adjusted revenue with a correlation of 0.43. Thus, the Job Opening Rate is a feature that we will emphasize in the further implementation.

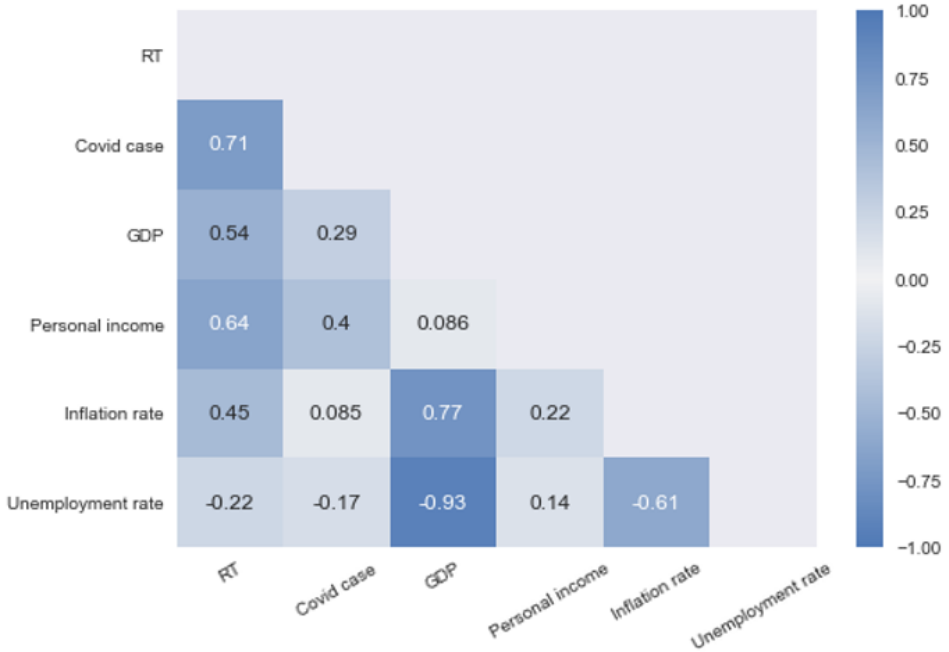

Fig. 3-2-2 SAR and external features correlation analysis (Covid period)

We then conduct the correlation analysis for SAR and external features during the COVID-19 period to figure out which features have a stronger correlation with the revenue. The data is

divided into quarters, ranging from Q1 2020 to Q3 2021. All external features have a stronger correlation with revenue under COVID-19 period, which is shown in the graph above. Overall, with a correlation of 0.71, COVID-19 case has the largest positive correlation with SAR, and Personal Income has the second-highest correlation with it. We anticipate that the Covid case and Personal Income will be important features for us to use in our future modeling.

3.3: Data Visualization

Over the past four years, Paychex's total revenue and the revenue proportion of different service products have been increasing. The percentage of Payroll service revenue falls from 57% ($1,566 million) to 49% ($1,582 million) while the ASO's increases from 18% ($0.503 million) to 24% ($0.776 million). The percentage of revenue in Retirement and PEO also have the same increasing trend. See Fig. 3-3-1 Paycheck Revenue by Service Product and notice the 2022 fiscal year is incomplete.

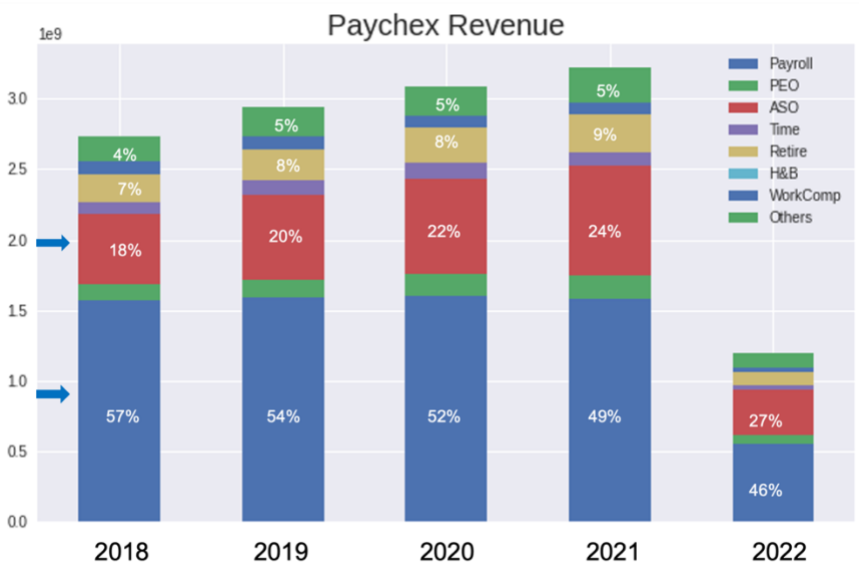

Fig. 3-3-1 Paycheck Revenue by Service Product

The revenue distribution by industry appears to have no significant change over the years. The top-3 industries with the highest revenue percentage are Professional, scientific, and technical service (14%), Health care and social assistance (14%), and Construction (9%). Since the one trillion infrastructure bill passed on November 2021, we suggest there is potential revenue growth in the construction industry. See Fig. 3-3-2 Paycheck Revenue by Industry.

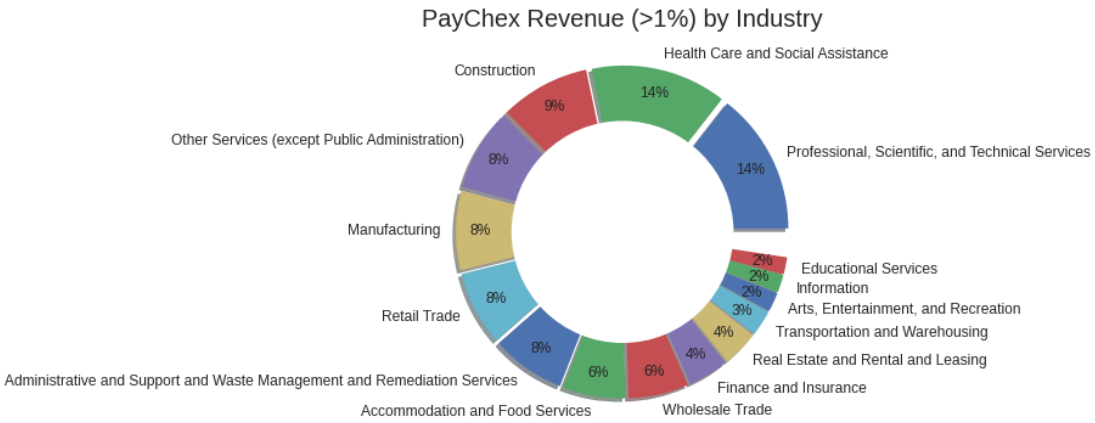

Fig. 3-3-2 Paycheck Revenue by Industry

From the visualization of revenue for different industries, we can see almost all the industries share the same seasonality and trend, except the appear on the bottom of the graph. Therefore, it is reasonable to use the total revenue instead of each individual industry for the future revenue prediction. More specifically, the industries at the bottom of the graph are Agriculture, forestry, fishing, and hunting, Management of companies and enterprises, Mining, quarrying, and oil and gas extraction, Public administration, and Utilities). See Fig. 3-3-3 Paycheck Revenue by Industry.

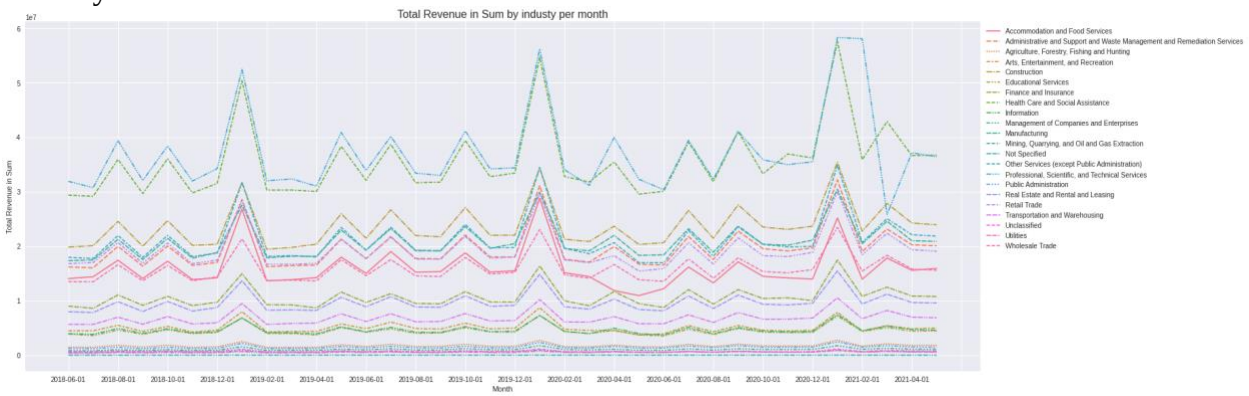

Fig. 3-3-3 Paycheck Revenue by Industry

The revenue distribution among 50 states also shows no change over the years. The top-5 states with the highest revenue are CA, NY, FL, NJ, and MA. However, this ranking is not aligned with the GDP rank among all states. We believe there are more opportunities for Paychex to capture more revenue from TX and IL as their GDP rank 2nd and 5th in the United States. See Fig. 3-3-4 Paycheck Revenue by State

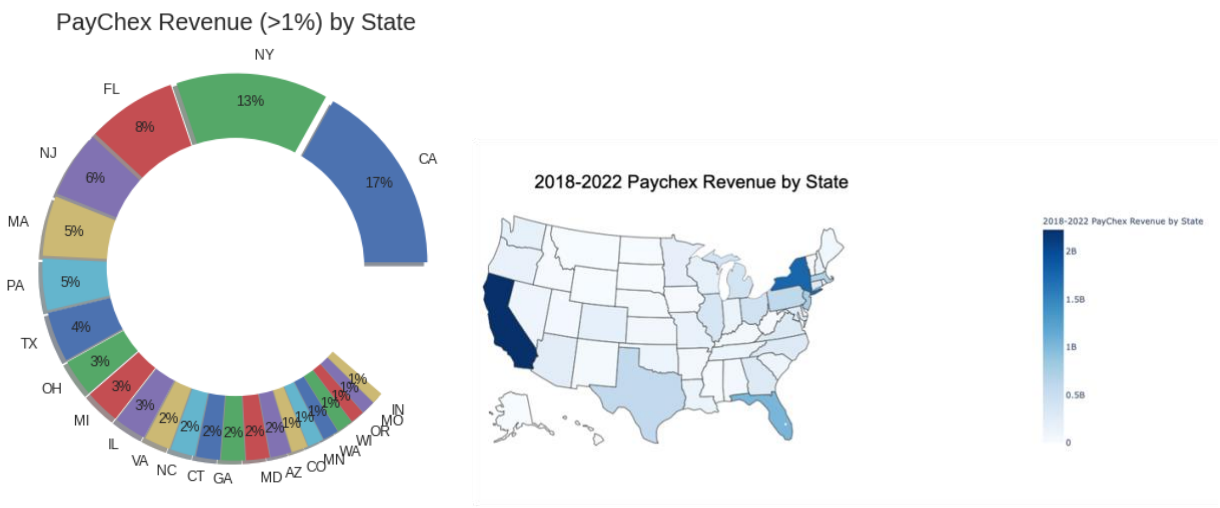

Fig. 3-3-4 Paycheck Revenue by State

4: Model Development

Since almost all the industries' revenue share the same seasonality and trend (part 3.3, except a small portion of six industries), it is reasonable to use total revenue for the overall revenue prediction, and each industry's revenue can be calculated by its proportion. Our model selection includes four types of models: we use SARIMA as our baseline model given it represents a traditional statistic approach for time-series data analysis. We then pick ThymeBoost as our hybrid model, FB-Prophet for our GAM model, and LSTM for the deep learning model. There is an iterative step for each model from feature engineering to its hyperparameter tuning. The best model is selected based on our predefined performance metrics. See Fig. 4-1 Overall methodology of Paychex revenue analysis.

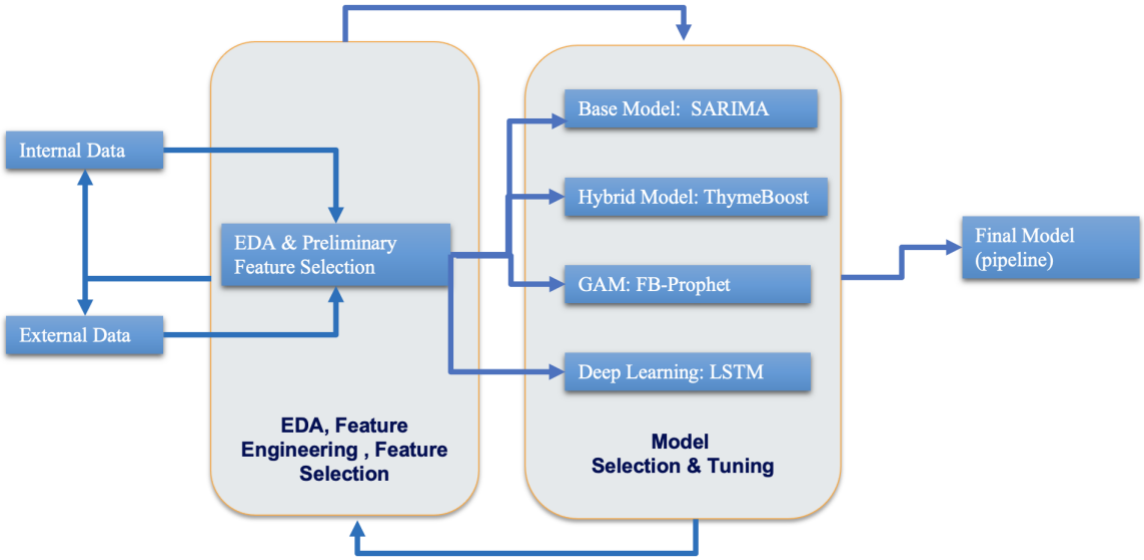

Fig. 4-1 Overall methodologies of Paychex revenue analysis

Our performance metrics include three Key Performance Index (KPI): Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE). Our models are evaluated using revenue data from the testing period (from December 2020 to August 2021). We pick MAPE as our performance measurement since it provides us with intuitive interpretations in terms of relative errors. MAPE is defined as follow:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|$$

Where:

- n is the number of fitted points,
- A_i is the actual value
- F_i is the forecast value

We strive for achieving a MAPE below 5% while constructing the revenue forecasting model.

4.1: SARIMA

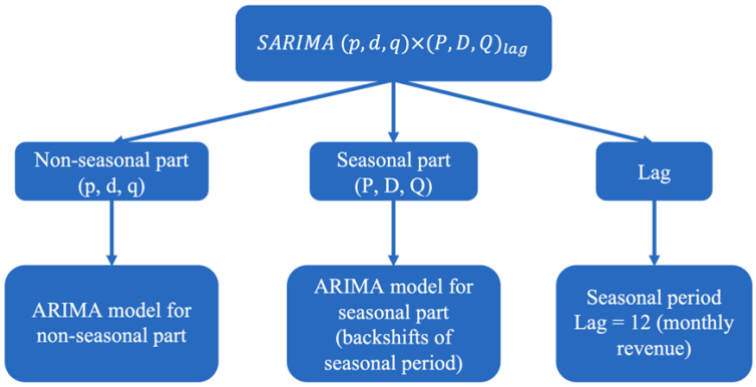

Fig. 4-1-1 SARIMA Model Architecture

SARIMA is our baseline model since it is the most prevalent traditional time-series prediction model. The SARIMA model has three components: two ARIMA models for non-seasonal and seasonal parts separately and the lag (see Fig. 4-1-1). In our analysis, the lag is 12 since revenue data is monthly. (p, d, q) and (P, D, Q) are non-negative integers referring to the order of autoregressive, integrated, and moving average parts of the model accordingly.

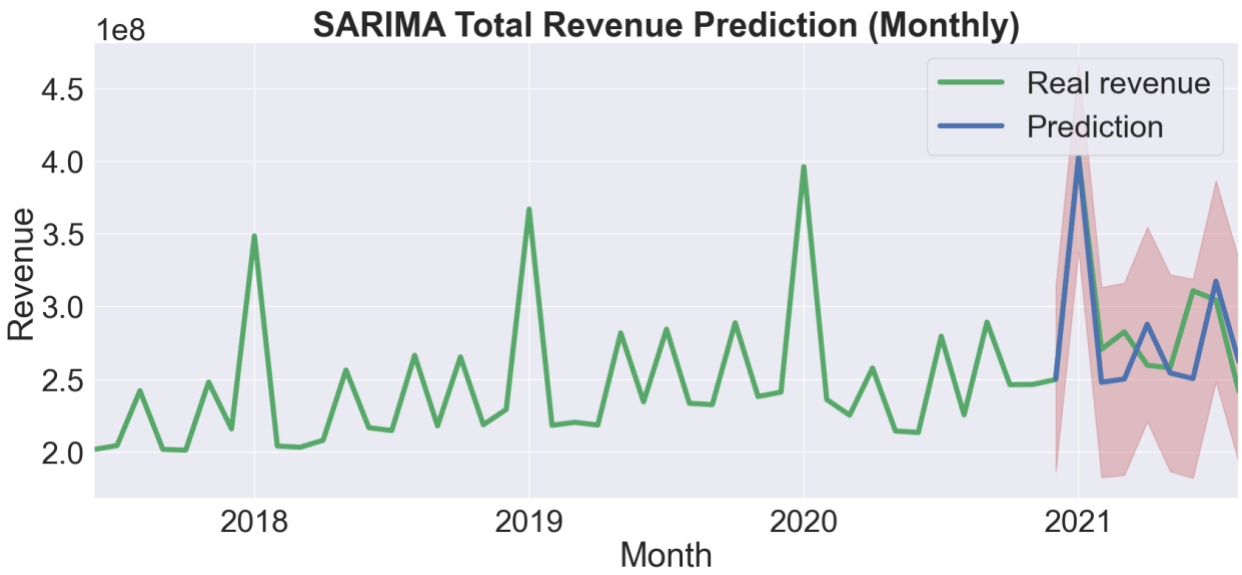

Fig. 4-1-2 SARIMA Total Monthly Revenue Prediction

Since we observe that Job Openings Rate is correlated with revenue, we include it as the external factor in our SARIMA model. The parameters of the SARIMA model were chosen by minimizing AIC (Akaike Information Criterion). Our final SARIMA model parameter after tuning is $SARIMA\ (0,\ 1,\ 1) \times (1,\ 1,\ 1)_{12}$ + Job Openings Rate, which achieves 8.0% MAPE, $22.28 million MAE, and $28.71 million RMSE in the testing period. SARIMA model performance in the testing period is illustrated in Fig. 4-1-2. The green line indicates real monthly revenue while the blue line is the predicted monthly revenues in the testing period. The red area is the 95% confidence interval for predicted revenues in the testing period.

4.2: ThymeBoost

ThymeBoost combines the traditional decomposition process with gradient boosting to provide a flexible mix-and-match time series framework for trend/seasonality/exogenous decomposition and forecasting [15]. The model has three estimators: trend estimator, seasonality estimator, and exogenous estimator. After interactively tuning hyperparameters, our model reaches the best MAPE with linear regression for trend estimator, Fourier for seasonality estimator. The gradient boost model will then calculate the residuals after the decomposition, the parameters with minimized cost (use AIC) will be used for the final model. See Fig. 4-2-1 ThymeBoost algorithm structure.

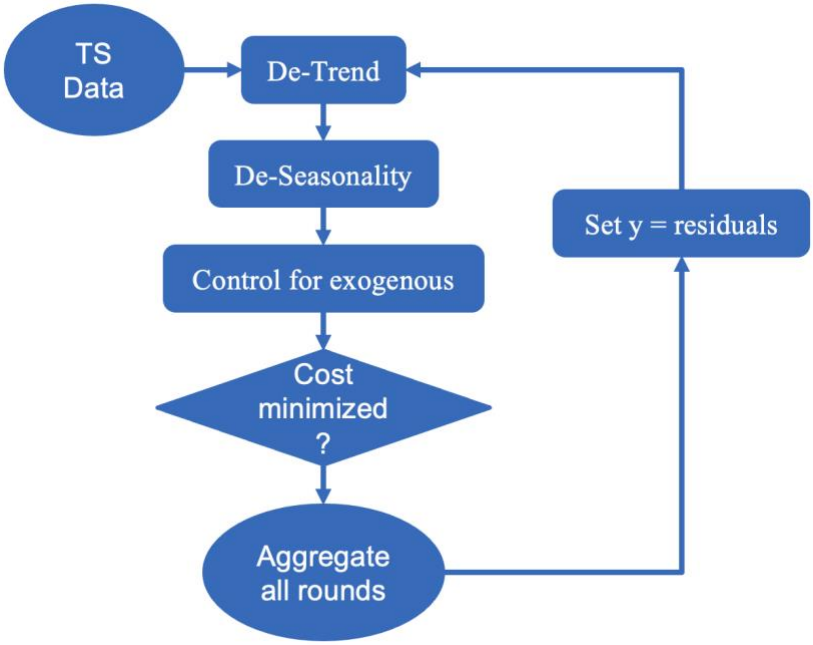

Fig. 4-2-1 ThymeBoost algorithm structure

Our best MAPE for this model is 9%. ThymeBoost successfully predicts the revenue in annual peak time (January) but fails to address the rest of the small peaks, and falsely predicted some of the trends. Given the limited time-series data, the model could perform better in the future if we have more data to feed. See Fig 4-2-2 ThymeBoost Total Revenue Prediction (Monthly)

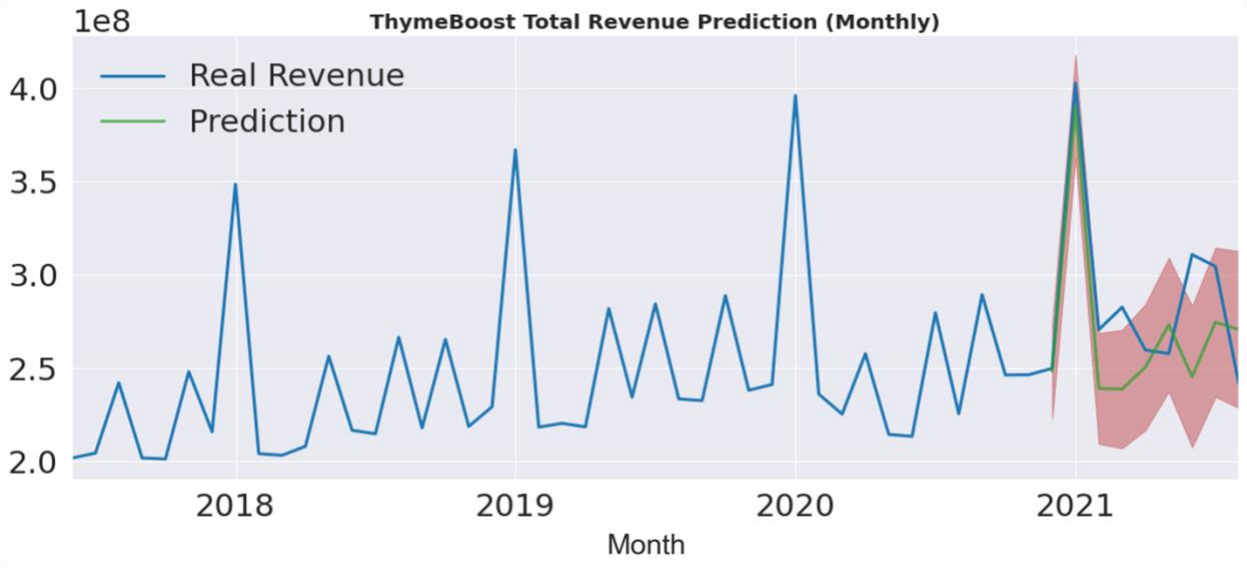

Fig 4-2-2 ThymeBoost Total Revenue Prediction (Monthly)

4.3: FB-Prophet

FB-Prophet was invented by Meta (Facebook) for time series analysis. It works best with time series that have strong seasonal effects and several seasons of historical data. The model has four parts: see Fig 4-3-1 FB-Prophet algorithm structure

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

Fig 4-3-1 FB-Prophet algorithm structure

g(t) is a trend function which models the non-periodic changes.
s(t) represents a periodic change.
h(t) is a function that represents the effect of holidays which occur on irregular schedules. (n≥1 days)
e(t) represents error changes that are not accommodated by the model.

As we are familiar with the traditional methods of times series decomposition, the meaning of g(t), s(t), and e(t) represents trend, seasonality, and residual correspondingly. The beauty of this model is the h(t) part, called the "holiday" function, which will tell the model about the potential peak period in the past and in the future to improve the accuracy of prediction. Another highlight of this model is that it also allows us to add external features to the model. In our project, we add the following external features to this model: Job Openings Rate, Hiring Rate, Total Separations Rate, Unemployment Rate, and Inflation Rate. We also input the potential peak period time stamp (holiday) as an additional regressor to the model and obtain the best MAPE 4.9% and MAE $13.33 million, RMSE $15.98 million. See Fig 4-3-2 FB-Prophet Total Revenue Prediction (Monthly).
If we don't include the peak period in the model, the MAPE degrades to 9%. Since the model can predict 95% correctly of Paychex revenue in the future, we recommend Paychex collect their marketing strategy information for the model which will improve the accuracy of the model's prediction.

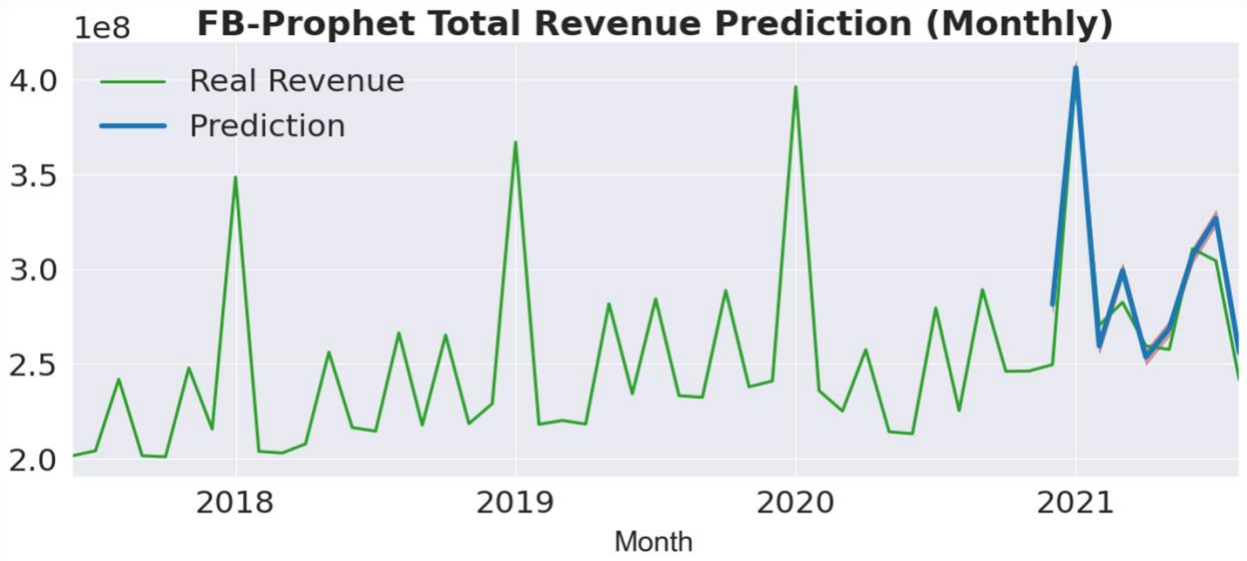

Fig 4-3-2 FB-Prophet Total Revenue Prediction (Monthly)

4.4: LSTM

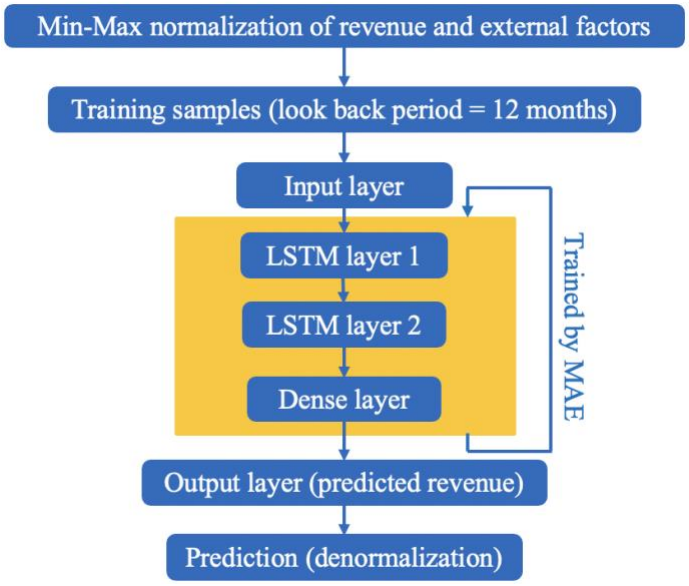

Fig. 4-4-1 LSTM Model Architecture

Our LSTM model consists of two LSTM layers and one dense layer, trained by MAE loss (see Fig. 4-4-1). Since our look back period is 12 months, we only have 30 months (from June 2018 to November 2020) revenue data for training. Compared with the other three models that we constructed, the deep learning-based model LSTM has much less training data.

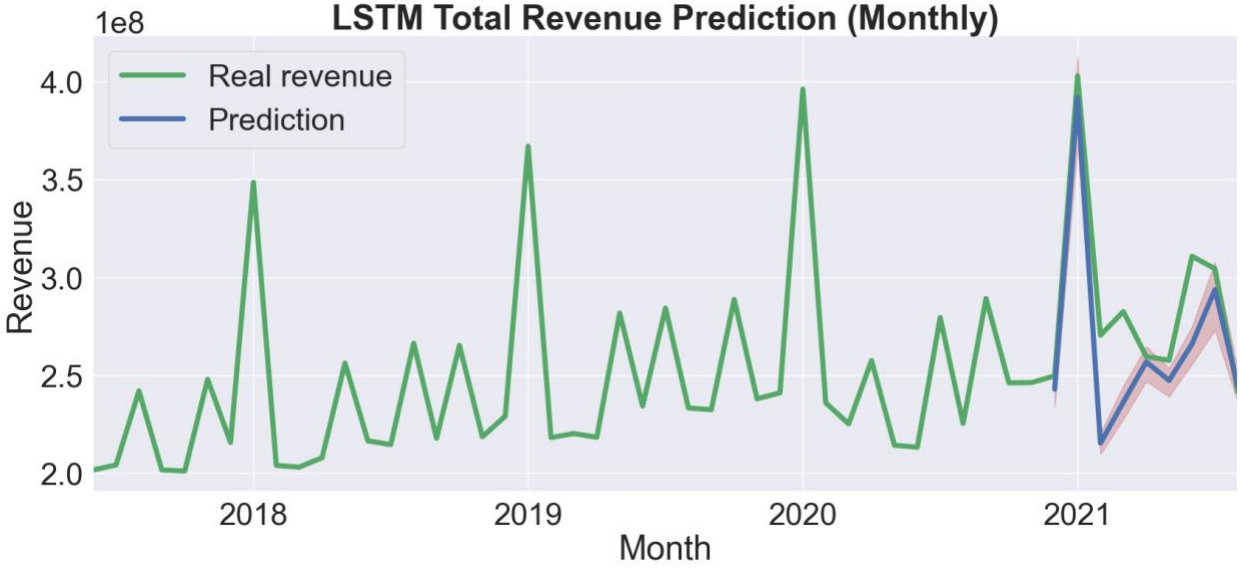

Fig. 4-5-2 LSTM Total Monthly Revenue Prediction

We only use revenue data in LSTM model since including external features worsened the model performance, which is possibly due to the data noises. Due to the small size of the training sample, a single LSTM model has a very large variance in prediction. To mitigate this issue, we construct an ensemble learner combining 100 different LSTM models to predict Paychex's revenue in the testing period. LSTM model performance is shown in Fig. 4-4-2. It achieves 7.4% MAPE, $21.12 million, and $29.00 million in the testing period.

5: Performance and Results

MODEL	Inputs	MAPE	Threshold	Interpretation
FB-Prophet	TS data, external features	4.9%	<5%	Highly Accurate
LSTM	TS data	7.4%	<10%	Good
SARIMA	TS data, external features	8.0%	<10%	Good
ThymeBoost	TS data	9.0%	<10%	Good

Fig. 5 Model Performance Comparison

After choosing the best parameter for each model, we evaluate the performance of our models using revenue data in the testing period. By referring to business insights from the Paychex team and relevant academic literature [14] about model performance benchmark, we have decided that less than 5% MAPE indicates highly accurate performance while between 5% and 10% MAPE is good.

Fig. 5 reports the test MAPE of every model and its corresponding threshold and interpretation. All our models achieve good performance with less than 10% MAPE. Our best model FB-Prophet shows promising results with only 4.9% test MAPE. In the future, more revenue data will potentially boost the LSTM model performance, which is currently the second-best model.

6: Conclusions and Future Works

In conclusion, the FB-Prophet appears to have the best performance with 4.9% of MAPE. In terms of the impact of external activities, we discover that the increase in Covid case, personal income, and job opening rate may indicate an increase in Paychex's internal revenue. There appear to be potential revenue growth opportunities in Texas, Illinois, and the construction industry. Therefore, we suggest Paychex enhance marketing activities in these areas. Some of the

improvements that can be made in the future are first to keep adding revenue data as time goes by will improve the result of learning-based models like LSTM. Secondly, incorporating future marketing peaks into the FB-Prophet model will boost its performance. Lastly, capturing Paychex's client data such as the monthly employee count of each client, and including them as features in all the prediction models will improve accuracy.

Reference:

[1]: Xu, Jin, Jingbo Zhou, Yongpo Jia, Jian Li, and Xiong Hui. "An Adaptive Master-Slave Regularized Model for Unexpected Revenue Prediction Enhanced with Alternative Data." In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 601-612. IEEE, 2020.

[2]: Tong, Howell. "Autoregressive model fitting with noisy data by Akaike's information criterion (Corresp.)." *IEEE Transactions on Information Theory* 21, no. 4 (1975): 476-480.

[3]: Durbin, James. "Efficient estimation of parameters in moving-average models." *Biometrika* 46, no. 3/4 (1959): 306-316.

[4]: Fang, Tingting, and Risto Lahdelma. "Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system." *Applied energy* 179 (2016): 544-552.

[5]: Zivot, Eric, and Jiahui Wang. "Vector autoregressive models for multivariate time series." *Modeling Financial Time Series with S-Plus®* (2006): 385-429.

[6]: Gardner Jr, Everette S. "Exponential smoothing: The state of the art." *Journal of forecasting* 4, no. 1 (1985): 1-28.

[7]: Connor, Jerome T., R. Douglas Martin, and Les E. Atlas. "Recurrent neural networks and robust time series prediction." *IEEE transactions on neural networks* 5, no. 2 (1994): 240-254.

[8]: Hua, Yuxiu, Zhifeng Zhao, Rongpeng Li, Xianfu Chen, Zhiming Liu, and Honggang Zhang. "Deep learning with long short-term memory for time series prediction." *IEEE Communications Magazine* 57, no. 6 (2019): 114-119.

[9]: Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

[10]: Wu, Kai, Jing Liu, Penghui Liu, and Shanchao Yang. "Time series prediction using sparse autoencoder and high-order fuzzy cognitive maps." *IEEE Transactions on Fuzzy Systems* 28, no. 12 (2019): 3110-3121.

[11]: Qin, Yao, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. "A dual-stage attention-based recurrent neural network for time series prediction." *arXiv preprint arXiv:1704.02971* (2017).

[12]: ThymeBoost. Available: https://pypi.org/project/ThymeBoost/

[13]: Facebook Prophet. Available: https://facebook.github.io/prophet/

[14]: Lewis, Colin David. *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann, 1982.

[15]: Tyler Blume, Time Series Forecasting with ThymeBoost: https://towardsdatascience.com/thymeboost-a0529353bf34