PAYCHEX®

Payroll | Benefits | HR | Insurance

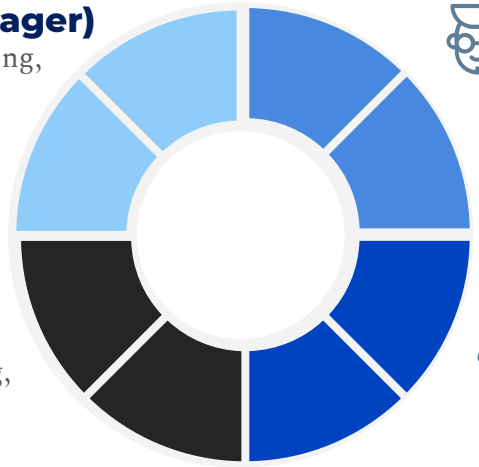
Final Presentation

December 6, 2021

Team member

Shijing Li (Project Manager)
Data Collection/Preprocessing, Visualization

Yangxin Fan
Data Collection, EDA, Modeling



Yuan Wang
EDA, Feature Engineering, Modeling

Lingyu Ye
Data Collection/Preprocessing, Visualization

Agenda

01 Introduction

02 Data Overview

03 Modeling

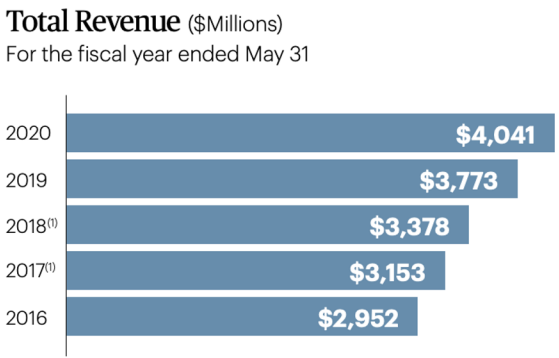
04 Future Work

1. INTRODUCTION

Who is Paychex?

#1 Ranked

401(k) Plan Recordkeeper



- ✓ Leading provider of integrated human capital management solutions for payroll, benefits, human resources, and insurance services
- ✓ **680,000** business clients
- ✓ **41%** Return on Equity

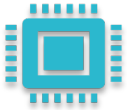
Resource:https://www.paychex.com/

Goal & Vision

Goal:

Identify

Identify correlations between Paychex revenue and external features by time



Build

Build various time-series, hybrid, and deep learning-based models to predict Paychex revenue by time

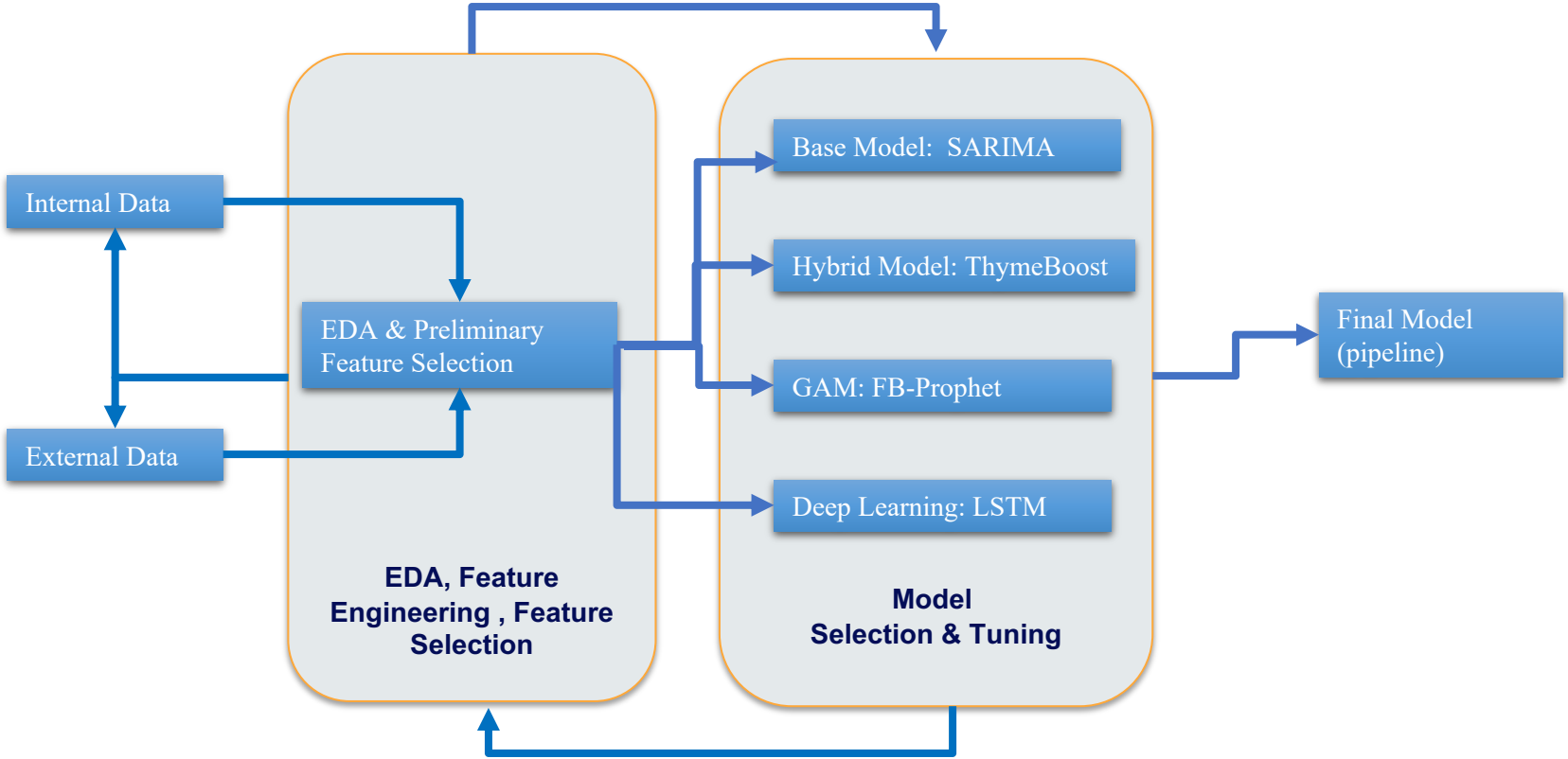
Deliver

Deliver constructive suggestions for Paychex to increase revenue by identifying new "growth"

Vision:

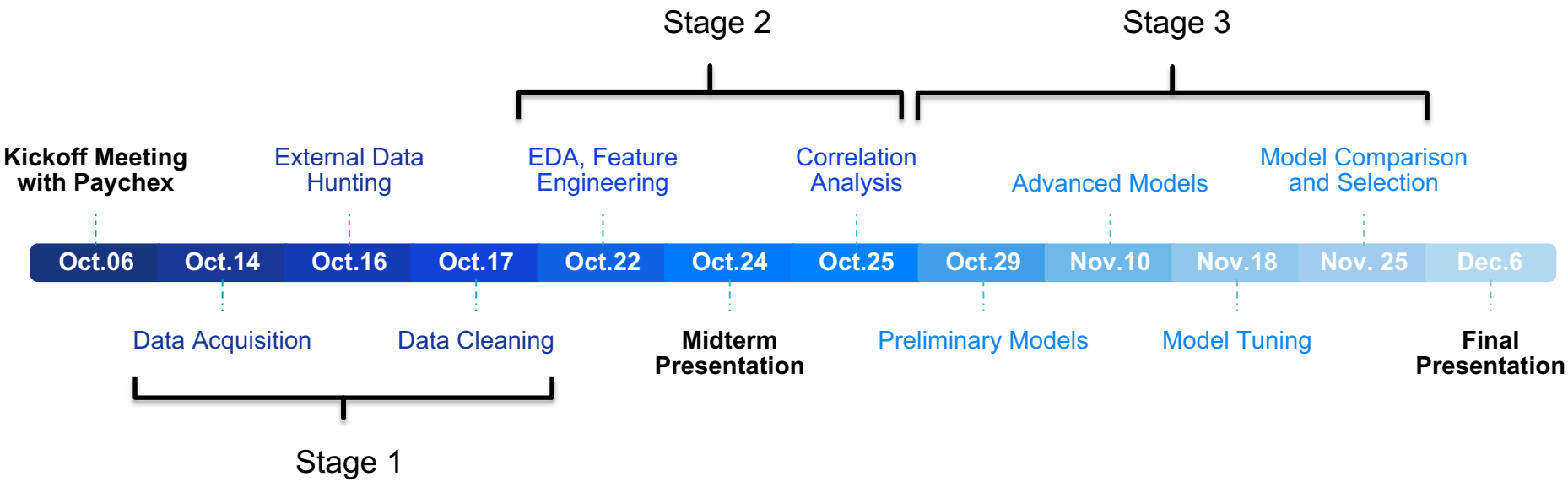
Identify the external activities that have a significant impact on Paychex's internal activity to build prediction models that will forecast Pyachex's future monthly revenue gain or loss. Produce the best fit model by using performance measurements.

Method Overview



GAM: General Additive Model

Milestones


Stage 2
Stage 3
Kickoff Meeting with Paychex
External Data Hunting
EDA, Feature Engineering
Correlation Analysis
Advanced Models
Model Comparison and Selection
Oct.06
Oct.14
Oct.16
Oct.17
Oct.22
Oct.24
Oct.25
Oct.29
Nov.10
Nov.18
Nov. 25
Dec.6
Data Acquisition
Data Cleaning
Midterm Presentation
Preliminary Models
Model Tuning
Final Presentation
Stage 1

2. Data Overview

Internal Data

Monthly Revenue Dataset:

Column Name	Description	
UNIQUESYSTEMCLIENTID	Client ID	Client Information
EECOUNT	Employee count	
BUSTYPELEVEL1	Industry type (NAICS2)	
STATECODE	State Code	
TOTALREV	Total monthly revenue	Monthly total revenue and product count of all Paychex's service products
PRODUCTCOUNT	Total monthly product count	
PAYROLLREVENUE	Total monthly revenue in Payroll	Monthly revenue of 7 different service products
PEOREVENUE	Total monthly revenue in PEO	
ASOREVENUE	Total monthly revenue in ASO	
TIMEREVENUE	Total monthly revenue in Time	
RETIREMENTREVENUE	Total monthly revenue in Retirement	
HNBREVENUE	Total monthly revenue in Health & Benefit	
WRKSCOMPREVENUE	Total monthly revenue in Work Compensation	

Dataset Description

FY17-FY21
823,963 rows, 247 columns
(No missing values & duplicates)

External Data

BUREAU OF ECONOMIC ANALYSIS

Monthly
- GDP per capita
- Person income per capita
- Inflation rate (quarterly)

U.S. BUREAU OF LABOR STATISTICS

Monthly
- Job opening Rate
- Hiring rate
- Separation rate
- Unemployment rate

Centers for Disease Control and Prevention

Monthly
- Case number
- Death number
- Vaccination number

GDP Per Capita

State	2018-Q1	2018-Q2	2018-Q3	2018-Q
Alabama	40963.28	41091.58	41063.68	41
Alaska	71920.29	72419.56	72300.30	72
Arizona	43801.17	43874.85	43792.51	43
Arkansas	38205.09	38494.65	38597.23	38
California	66231.62	66856.95	67369.28	67
Colorado	60117.41	60159.96	60089.51	60
Connecticut	69444.61	69778.74	70191.01	70

Job Tunrover and Unemployment Rate

Month	Job openings rate	Hiring rate	Total seperations rat
June-17	4.00%	3.70%	3.60%
July-17	4.00%	3.80%	3.70%
August-17	4.00%	3.70%	3.60%
September-17	4.00%	3.60%	3.60%
October-17	4%	3.80%	3.60%

COVID-19 Case Number

State	20-Jan	20-Feb	20-M
Alabama	0	0	6
Alaska	0	0	
American Samoa	0	0	
Arizona	6	23	6
Arkansas	0	0	3

Data Preprocessing

Data cleaning

1. Check missing and duplicate values
2. Standardize data granularity

Data transformation

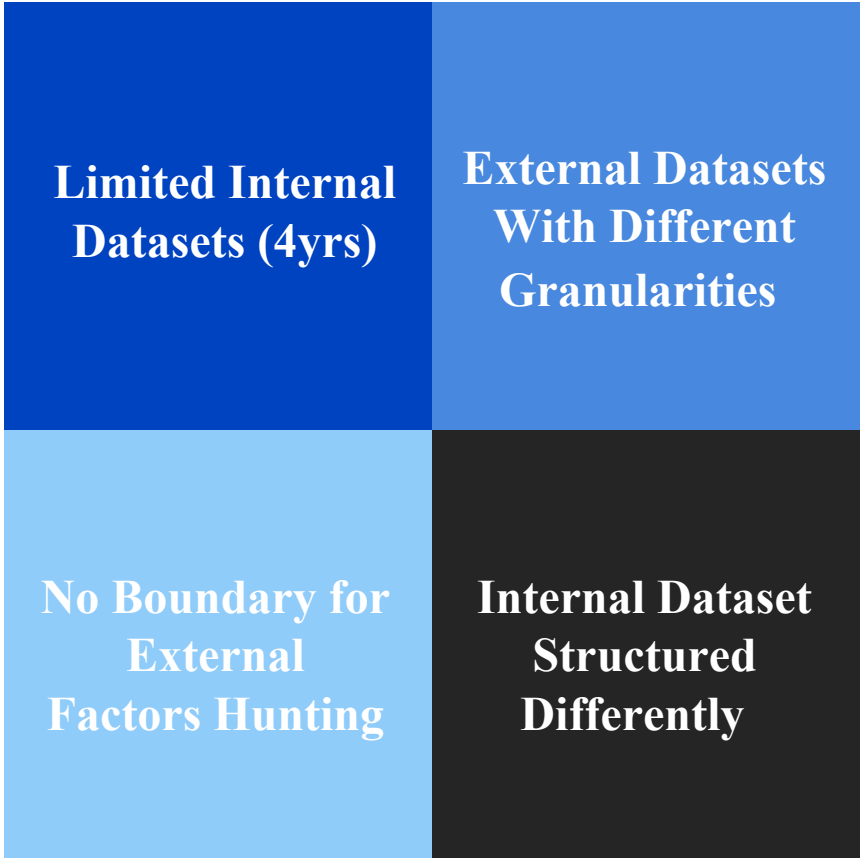
1. Transform internal & external data to usable formats
2. Merge internal & external datasets

Data reduction

1. Aggregate monthly revenue of all industries
2. Feature selection

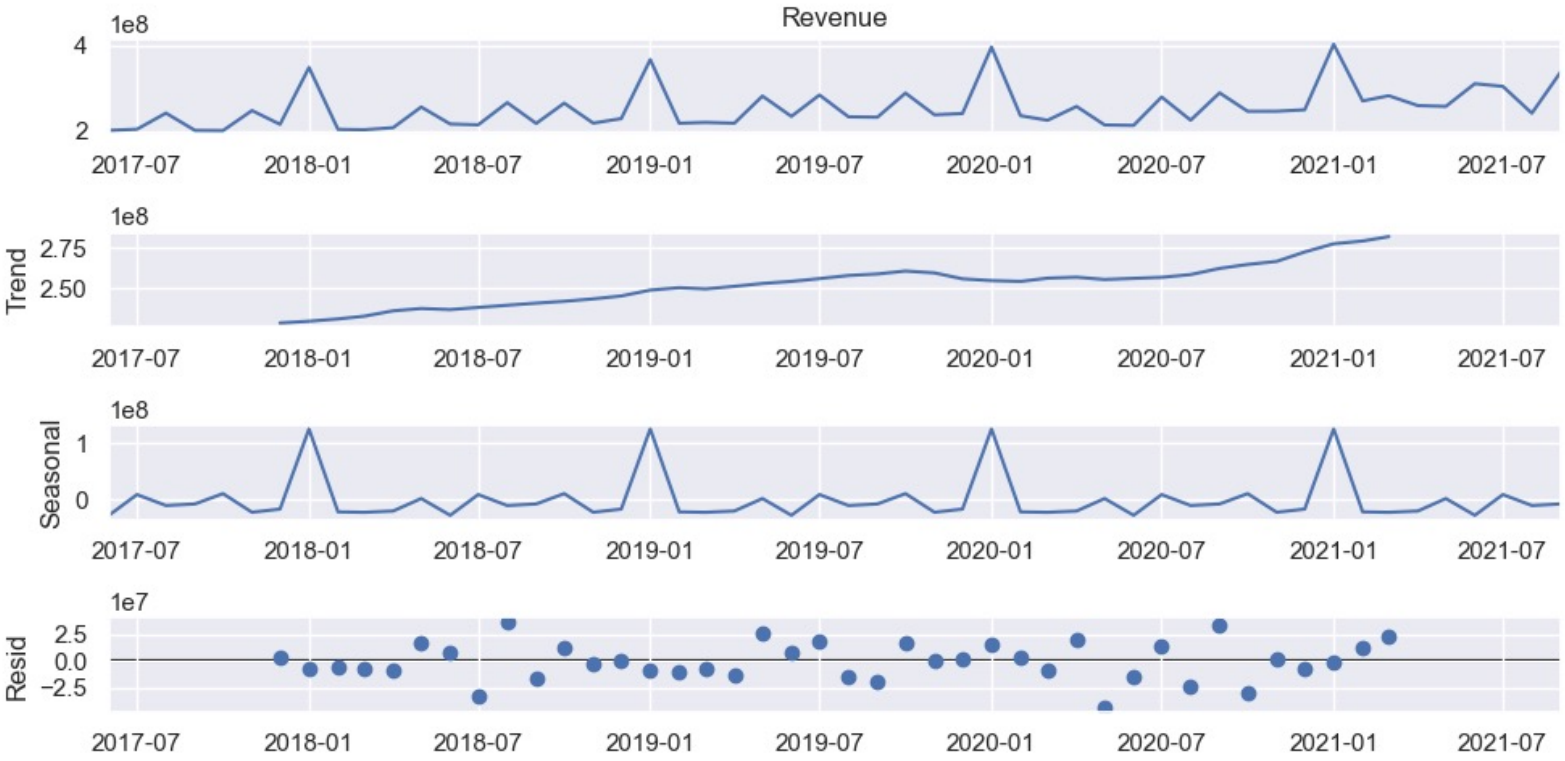
```
[9] def revenue_table(df):
      strings = df.columns.tolist()
      #Total Revenue
      total_cols = [string for string in strings if "TOTALREV" in string]
      #PAYROLLREV
      payroll_cols = [string for string in strings if "PAYROLLREV" in string]
      #PEOREV
      peo_cols = [string for string in strings if "PEOREV" in string]
      #ASOREV
      aso_cols = [string for string in strings if "ASOREV" in string]
      #TIMEREV
      timer_cols = [string for string in strings if "TIMEREV" in string]
      #RETIREMENTREV
      retire_cols = [string for string in strings if "RETIREMENTREV" in string]
      #HNBREV
      hnb_cols = [string for string in strings if "HNBREV" in string]
      #WRKRSCOMPREV
      wrkscomp_cols = [string for string in strings if "WRKRSCOMPREV" in string]

      other = df[total_cols].sum().sum() - (df[payroll_cols].sum().sum()+df[peo_cols].sum().sum()
       +df[timer_cols].sum().sum()+df[retire_cols].sum().sum()+df[hnb_cols].sum().sum()
       +df[wrkscomp_cols].sum().sum())
      rev_by_category = [df[total_cols].sum().sum(), df[payroll_cols].sum().sum(),
                         df[peo_cols].sum().sum(), df[aso_cols].sum().sum(),
                 df[timer_cols].sum().sum(), df[retire_cols].sum().sum(),
                 df[hnb_cols].sum().sum(), df[wrkscomp_cols].sum().sum(),
                 other]
      label = ['Total', 'Payroll', 'PEO', 'ASO', 'Time','Retire', 'H&B', 'WorkComp', 'Others']
      return rev_by_category
```

```
table_rev = pd.DataFrame(
    columns=['Total', 'Payroll', 'PEO', 'ASO', 'Time','Retire', 'H&B', 'WorkComp', 'Others'],
     index=[2018, 2019, 2020, 2021, 2022])
table_rev.loc[2018] = revenue_table(px_18)
table_rev.loc[2019] = revenue_table(px_19)
table_rev.loc[2020] = revenue_table(px_20)
table_rev.loc[2021] = revenue_table(px_21)
table_rev.loc[2022] = revenue_table(px_22)
# Add percentage for convenience
table_rev['Payroll %'] = table_rev['Payroll']/table_rev['Total']
table_rev['PEO %'] = table_rev['PEO']/table_rev['Total']
table_rev['ASO %'] = table_rev['ASO']/table_rev['Total']
table_rev['Time %'] = table_rev['Time']/table_rev['Total']
table_rev['Retire %'] = table_rev['Retire']/table_rev['Total']
table_rev['H&B %'] = table_rev['H&B']/table_rev['Total']
table_rev['WorkComp %'] = table_rev['WorkComp']/table_rev['Total']
table_rev['Others %'] = table_rev['Others']/table_rev['Total']
pd.options.display.float_format = '{:,.2f}'.format
table_rev
```

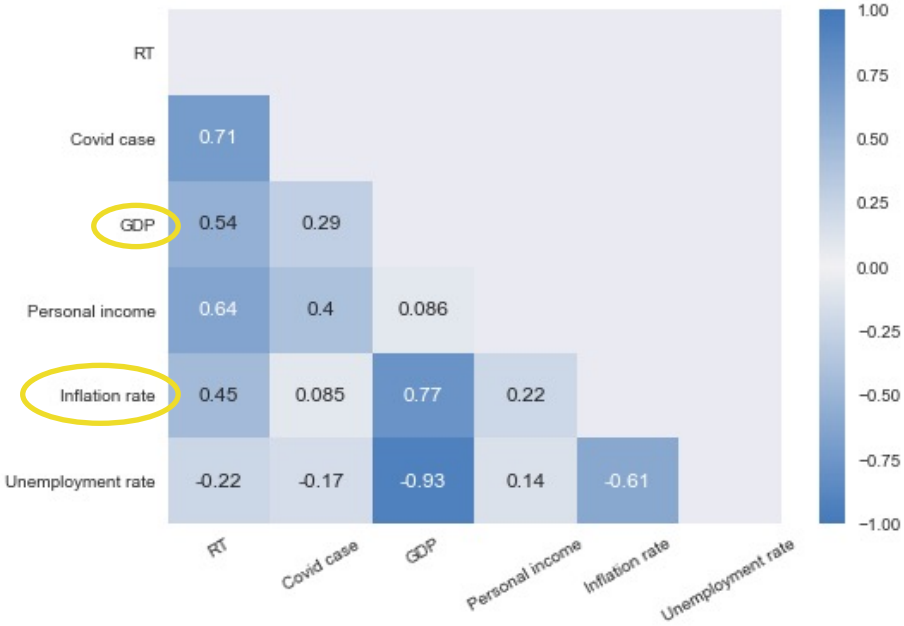

Limited Internal Datasets (4yrs)
External Datasets With Different Granularities
No Boundary for External Factors Hunting
Internal Dataset Structured Differently

Challenges

Exploratory Data Analysis – Data Decomposition



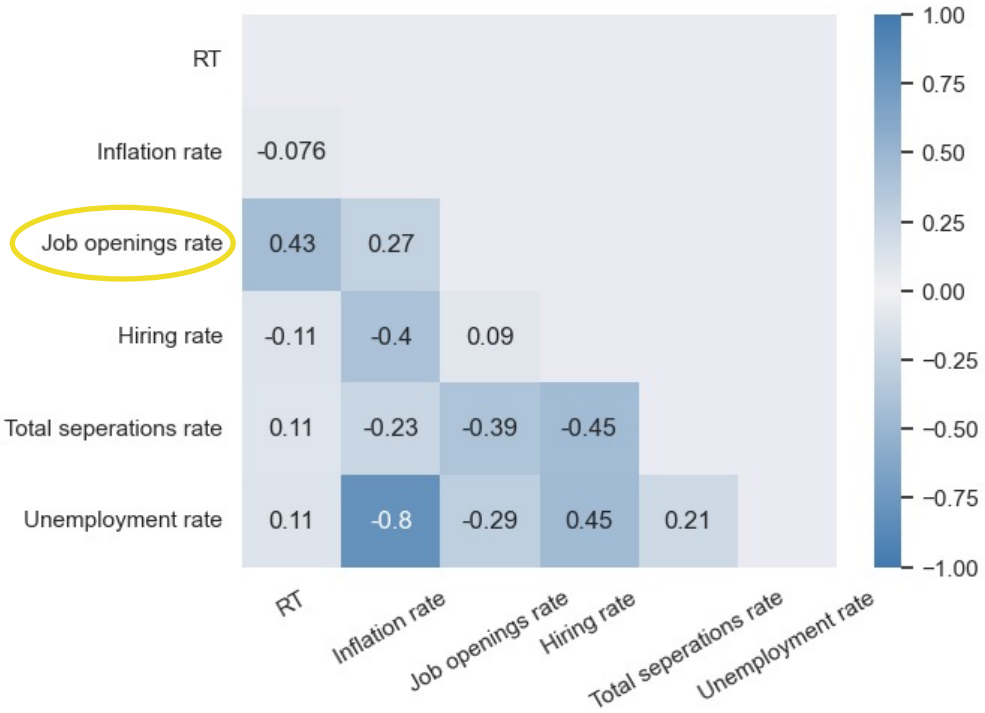
Exploratory Data Analysis – Covid Correlation Analysis



- Correlation between RT (residual + trend) **Seasonally Adjusted Revenue**, **COVID-19** and **Econimic** external factors
- Quarterly data from 2020 Q1 to 2021 Q3

Conclusion: GDP & Inflation Rate are **highly** correlated with Seasonality Adjusted Revenue. (r = 0.81)

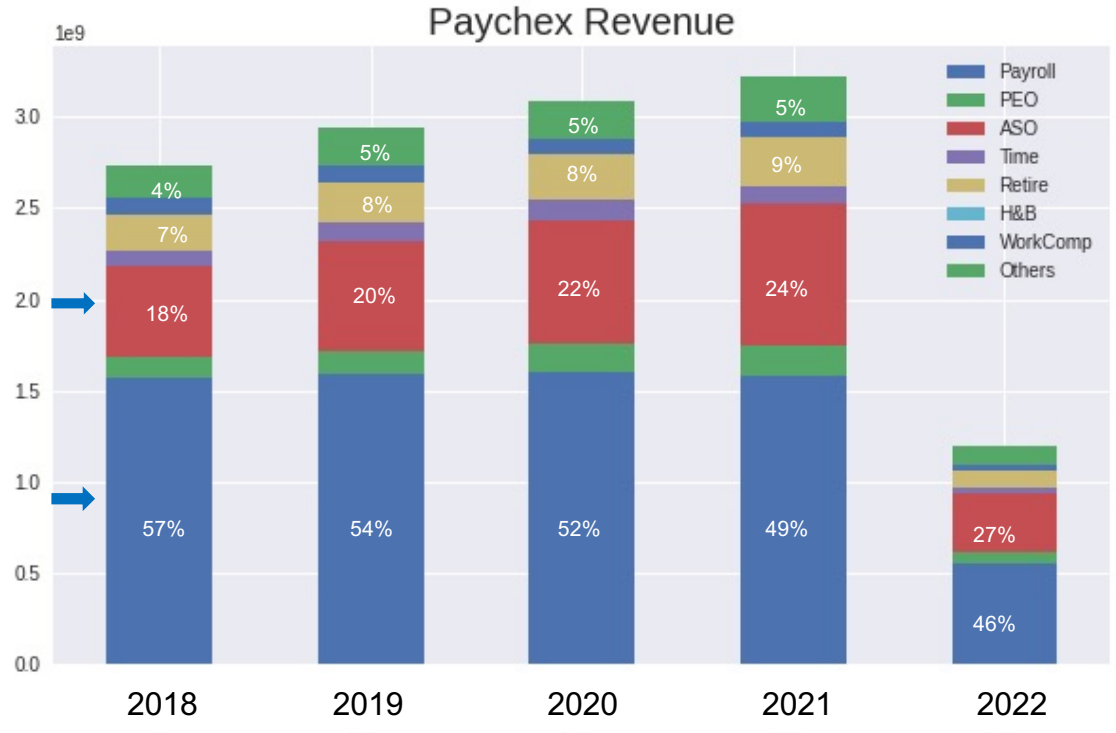
Exploratory Data Analysis – Correlation Analysis



- Correlation between RT (residual + trend) **Seasonally Adjusted Revenue** and **BLS** external factors
- Monthly data from 06/2017 - 08/2021

Conclusion: Job Opening Rate is **moderately** correlated with Seasonality Adjusted Revenue (r = 0.43)

EDA – Data Visualization

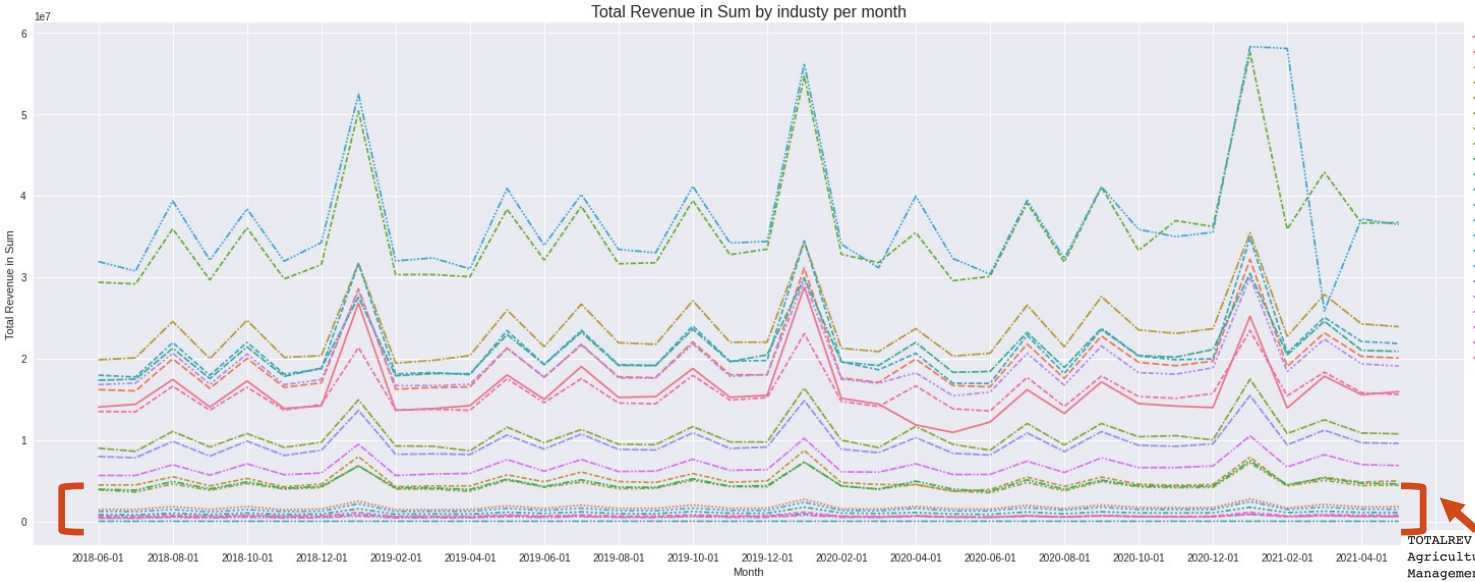


PEO: Professional Employer Organization
ASO: Administrative Services Organization
H&B: Health & Benefits

- The **total revenue** is increasing over the years.
- The revenue percentage for the **Payroll** service product is decreasing.
- The revenue percentage for the **ASO** service product is increasing.

* 2022 fiscal year is incomplete (3 months data), but the trend of each portion stays consistent.

EDA – Data Visualization



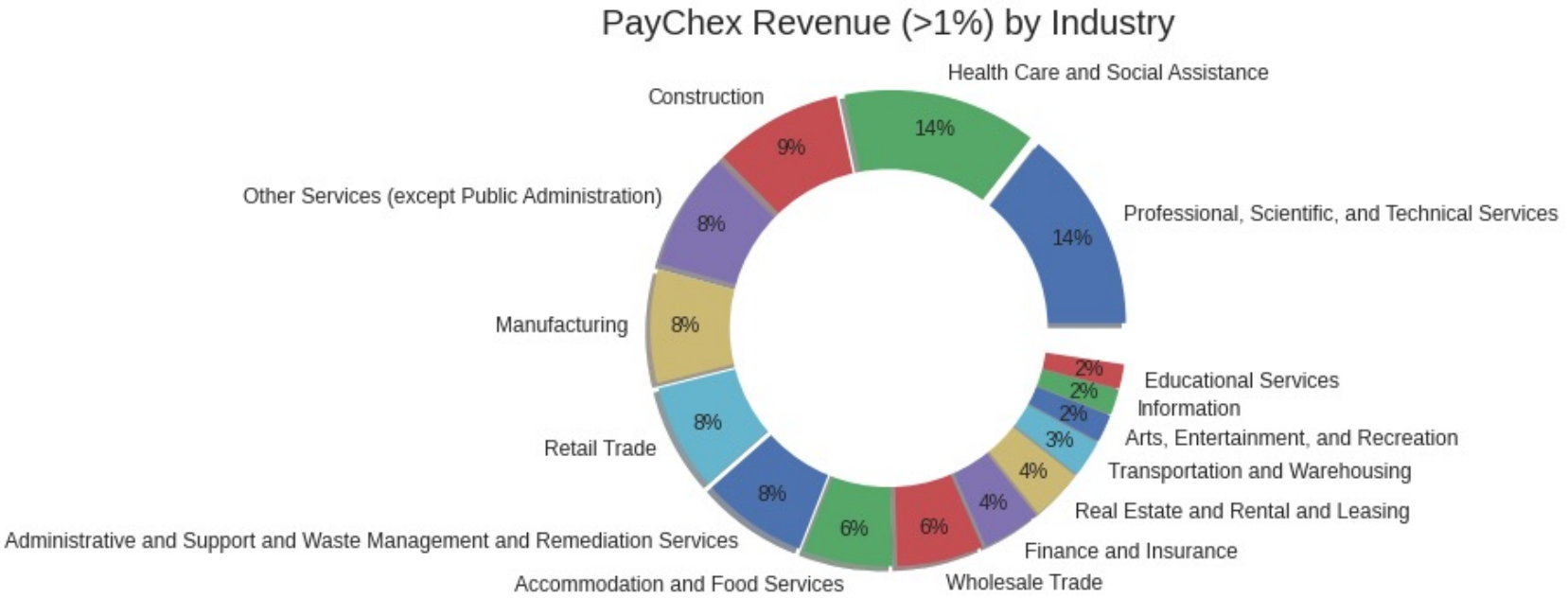
TOTALREV

Agriculture, Forestry, Fishing and Hunting	1.745665e+06
Management of Companies and Enterprises	1.052043e+06
Mining, Quarrying, and Oil and Gas Extraction	6.105092e+05
Not Specified	1.264843e+03
Public Administration	1.474732e+06
Unclassified	7.086776e+05
Utilities	5.264096e+05

dtype: float64

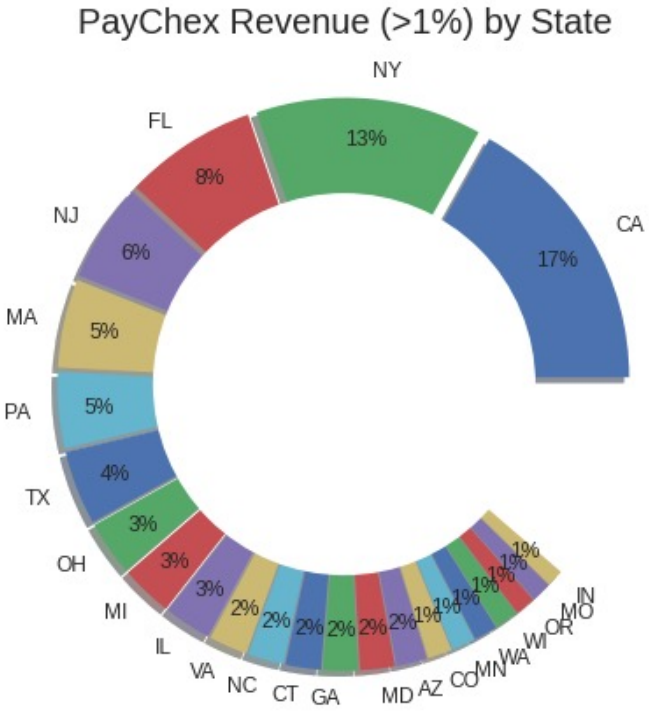
- Over the years, there is no significant change in revenue contribution among different industries.
- We can use Total Revenue instead of each industry for the future revenue prediction.

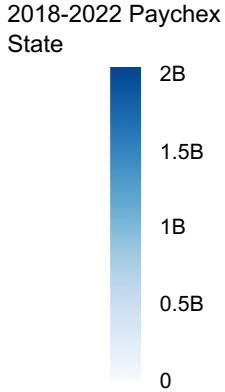
EDA – Data Visualization



- The top-3 industries are Professional services, Health care, and Construction.
- One trillion infrastructure bill passed will lead a high potential opportunity in construction industry.

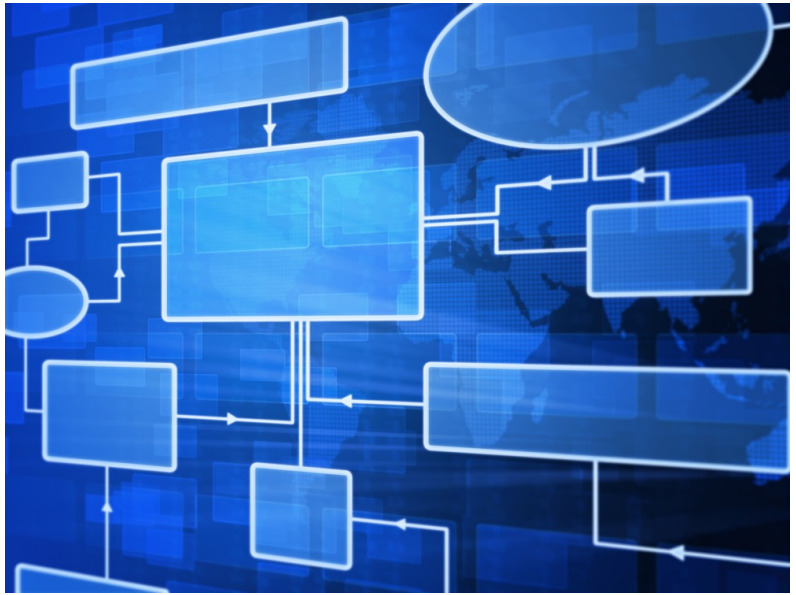
EDA – Data Visualization





- Top-5 states with the highest revenue**:** **CA**, **NY**, **FL**, **NJ**, and **MA**
- **TX**, **IL** have large potential to obtain more revenue since they are 2nd and 5th largest economy state in the U.S.

3. Modeling



Performance Measurement

- **MAE**: Mean absolute error (MAE) is the average of the absolute errors.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

- **MAPE**: Mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics.

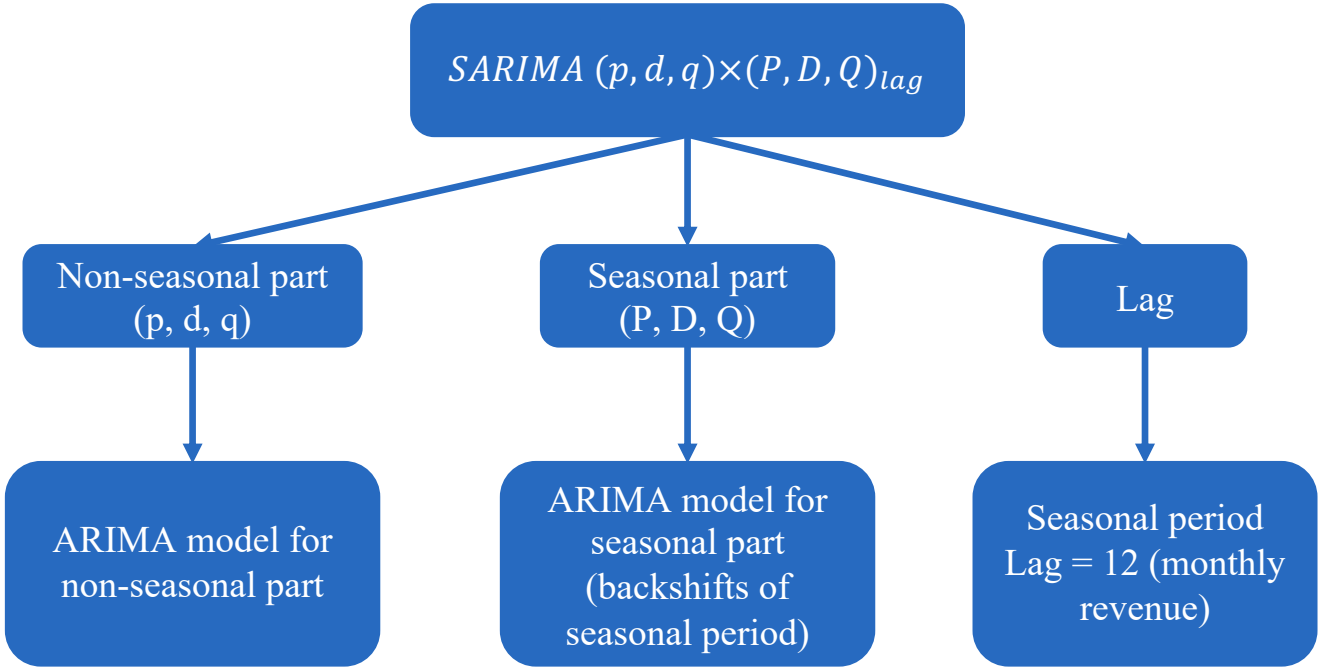
$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{V(t) - P(t)}{V(t)} * 100\right|$$

- **RMSE**: Root Mean Square Error (RMSE) is the standard deviation of the errors/residuals.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

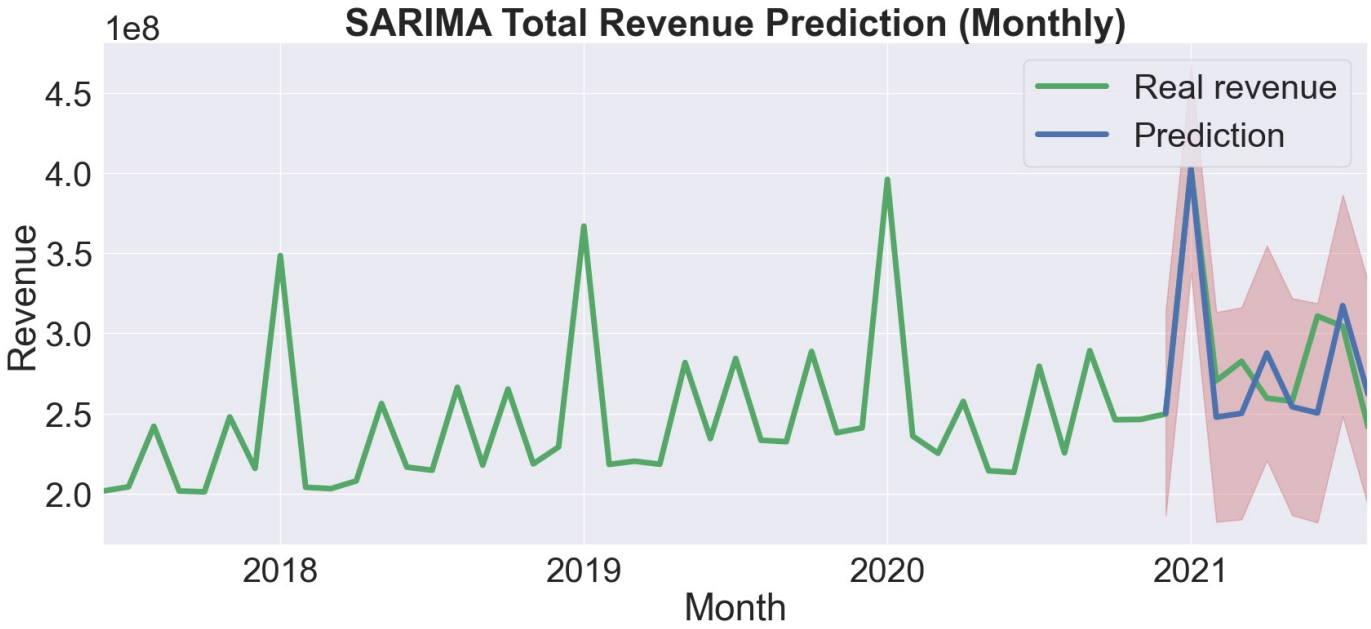
Testing period: 12/2020 to 08/2021

Model - SARIMA



- (p, d, q) (P, D, Q) are integers ≥ 0 and refer to the order of the autoregressive, integrated, and moving average of the model respectively

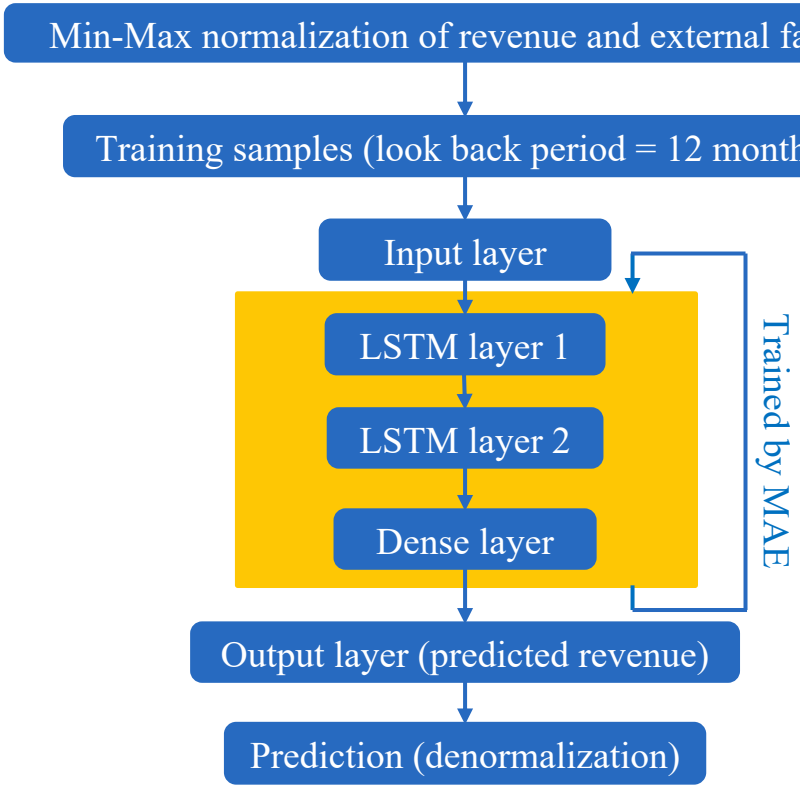
Model - SARIMA

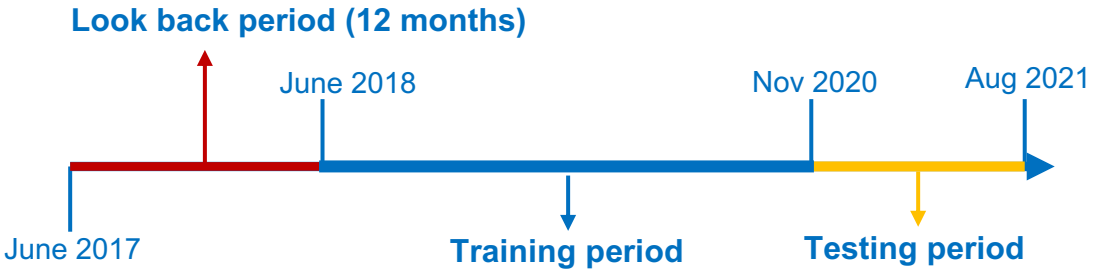


Test MAPE	0.08 (**8.0%**)
Test MAE	$22.28 million
Test RMSE	$28.71 million

- External feature used: **Job openings rate** (correlated with seasonally adjusted revenue)
- Final model chosen by **AIC cost**: $SARIMA\ (0,1,1)\times(1,1,1)_{12}$ + Job Openings Rate

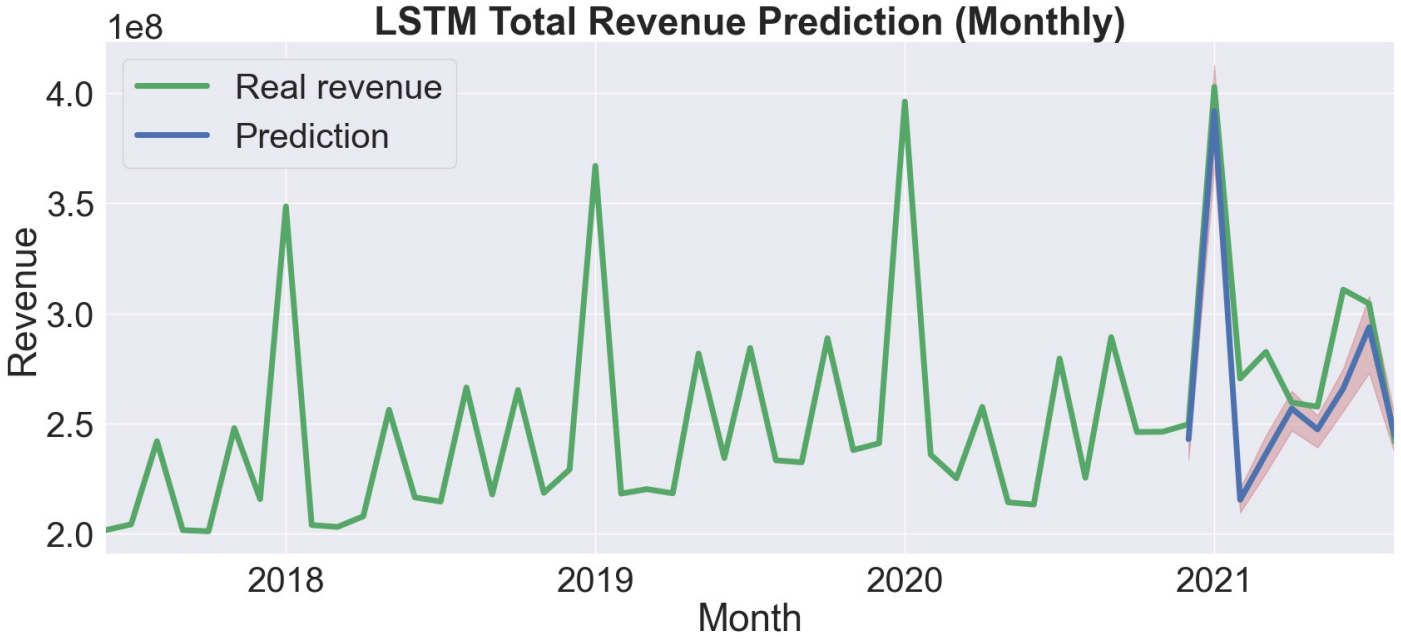
Model - LSTM





- **Training period:** 06/2018 - 11/2020 (30 months)
- **Testing period:** 12/2020 - 08/2021 (9 months)

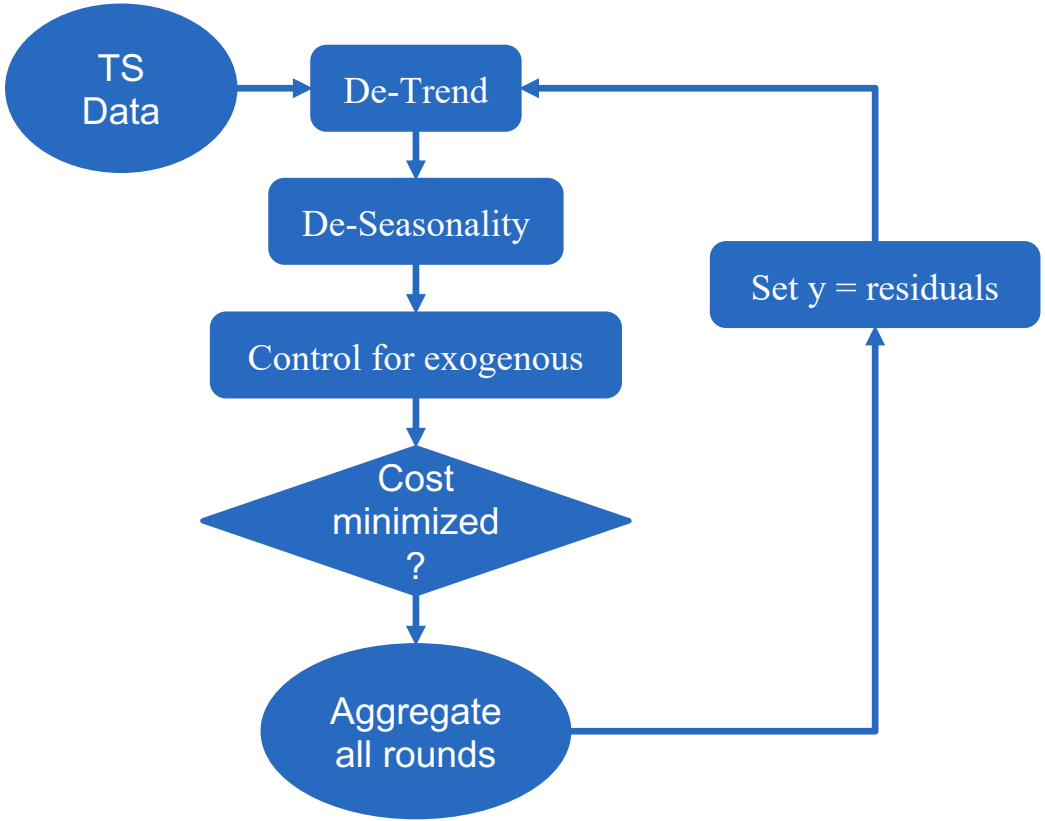
Model - LSTM



Test MAPE	0.074 (7.4%)
Test MAE	$21.12 million
Test RMSE	$29.00 million

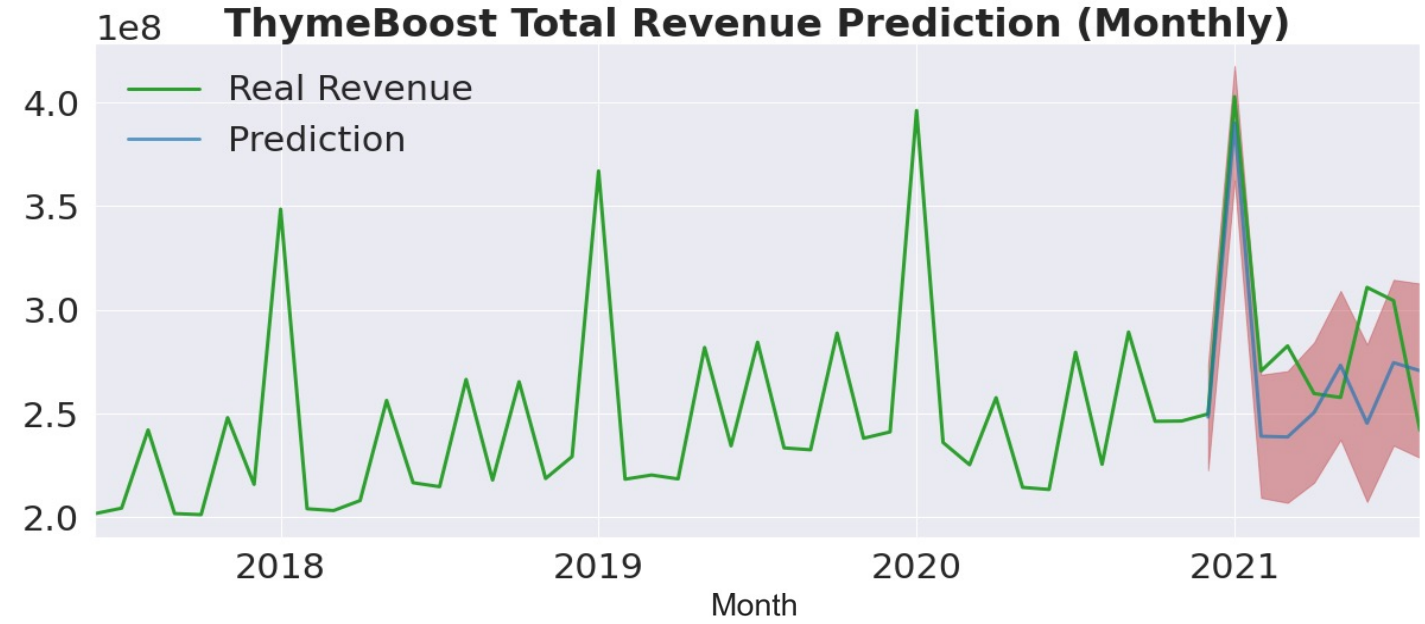
- No external features since including external features worsens model performance
- Ensembled LSTM model (100 LSTM models)

Model - ThymeBoost



- Traditional time series decomposition for trend, seasonality
- **Gradient boost model for residual**
- Cost minimize

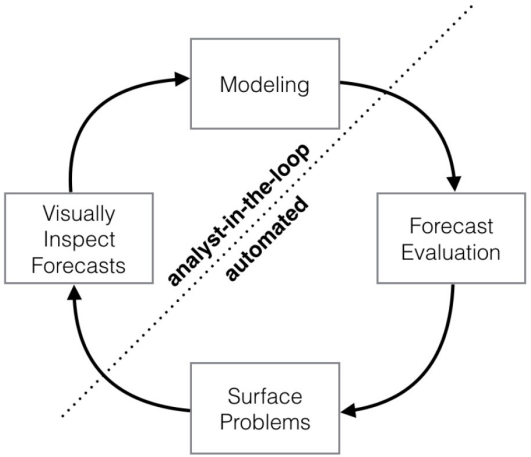
Model - ThymeBoost



Test MAPE	0.09 (9.0%)
Test MAE	$25.73 million
Test RMSE	$31.81 million

- No external features
- Trend estimator: Linear
- Seasonal estimator: Fourier
- Cost: AIC

Model – FB-Prophet

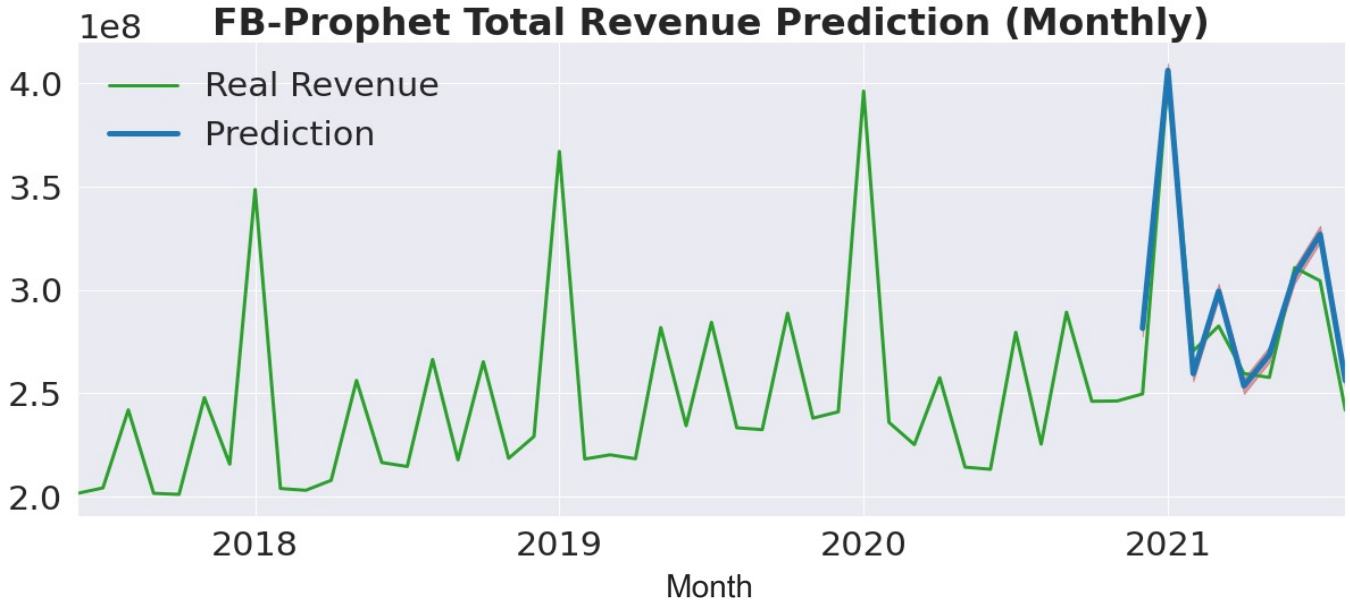


$$y(t) = g(t) + s(t) + h(t) + e(t)$$

- **g(t)** is a trend function which models the non-periodic changes.
- **s(t)** represents a periodic changes
- 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.

https://peerj.com/preprints/3190.pdf#section.1

Model – FB-Prophet + peak indication



Test MAPE	0.049 (4.9%)
Test MAE	$13.33 million
Test RMSE	$15.98 million

- External features used : Job Openings Rate, Hiring Rate, Total Separations Rate, Unemployment Rate, Inflation Rate
- **Holidays/Peak events** indication applied both on **train** and **test periods**

Model Comparison

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

Conclusion

- **The increase in GDP, Inflation rate and, Job opening rate may indicate an increase in Paychex revenue**

- **FB-Prophet has the best performance with 4.9% of MAPE**

- **Potential revenue growth opportunities in Texas and Illinois. (enhance marketing activities)**

- **Potential opportunities in construction industry**

4. Future Work

Future works

01	Future Revenue Data	→	LSTM (Deep Learning)
02	Future Marketing Peaks	→	FB-Prophet
03	Client Data (EEcount etc.)	→	Prediction Model

Acknowledgement

- Paychex Client

  Ryan Hayden
  Joel Page
  Oseoba Airewele
  Mike Childress

- Advising Professor

  Prof. Cantay Caliskan
  Prof. Ajay Anand


Thanks!

Appendix – Project Directory