Caterpillar Supply Chain Resiliency

Team Members – Angel Avila Gonzalez, Nihar Kodkani, Dev Panchal, Netra Joshi, Ramya Rajaram, Puwei Peng, Huy Vu, Arnav Daryani, Aaryan Wadhwani, Christina Lee, Jorge Mendoza, Harsh Srivastava

INTRODUCTION

CAT receives parts from over **27,000** individual suppliers and **50+** countries for its various industries. Supply chain disruptions are prone to occur when there are so many suppliers and many risk variables present. Supply chain disruptions can cause tremendous loss of time and money for CAT and its many customers.

GOALS

Use open-source temporal data to predict over 100+ supply chain risk variables to forecast future supply chain disruptions.

RESEARCH METHODOLOGY

DATA ACQUISITION
- Acquired 110+ datasets
- 30+ supplier countries
- 7 risk categories

DATA PREPROCESSING
- Standardized column names
- Standardized data granularity
- Combined into one master table

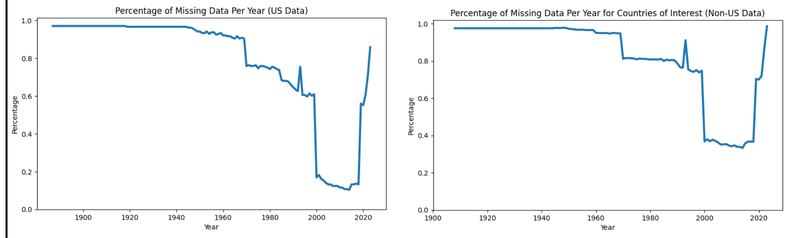
DATA IMPUTATION
- Filled in missing values
- Filtered out workable data
- Imputed data through statistical models

FORECASTING MODEL
- Analyzed data trends
- Compared analysis to PyCaret best models
- Performed prediction using PyCaret

SAMPLE MASTER TABLE DELIVERABLE

Country	Year	Inflation	Population	...	PPI Metal Mining
United States	1940	...	...	...	...
United States	1941	...	...	...	...
...	...	...	...	...	...
United States	2019	...	...	...	...

POST CLEANING DATA ANALYSIS



The percentage of missing data graphs gave us an idea about the years with the majority of the missing data in the US and Non-US master tables. This allowed us to identify the time-frame with the least amount of missing/unusable data.

TOOL IMPLEMENTATION : PYCARET MODELLING

PyCaret is an open-source machine learning library in Python that provides a streamlined workflow for building, comparing, and deploying machine learning (ML) models.

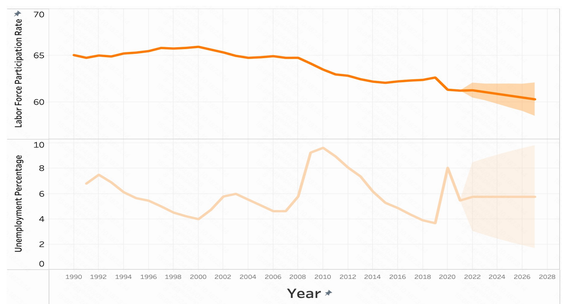
Our business use case for PyCaret is to automate the development of ML models. The library was useful to compare several models and choose the optimal one with the best predefined accuracy metrics

	Model	MASE	RMSSE	MAE
naive	Naive Forecaster	0.4671	0.3715	100.2660
auto_arima	Auto ARIMA	0.4745	0.3771	101.5906
theta	Theta Forecaster	0.4749	0.3772	101.5307
ets	ETS	0.4871	0.3871	104.2719
exp_smooth	Exponential Smoothing	0.4881	0.3878	104.4450
omp_cds_dt	Orthogonal Matching Pursuit w/ Cond. Deseasonalize & Detrending	0.5796	0.4604	123.9498
ridge_cds_dt	Ridge w/ Cond. Deseasonalize & Detrending	0.5802	0.4610	124.0824
lr_cds_dt	Linear w/ Cond. Deseasonalize & Detrending	0.5802	0.4610	124.0824
llar_cds_dt	Lasso Least Angular Regressor w/ Cond. Deseasonalize & Detrending	0.5802	0.4610	124.0832
lasso_cds_dt	Lasso w/ Cond. Deseasonalize & Detrending	0.5802	0.4610	124.0832
en_cds_dt	Elastic Net w/ Cond. Deseasonalize & Detrending	0.5802	0.4610	124.0836
br_cds_dt	Bayesian Ridge w/ Cond. Deseasonalize & Detrending	0.5842	0.4641	124.9084

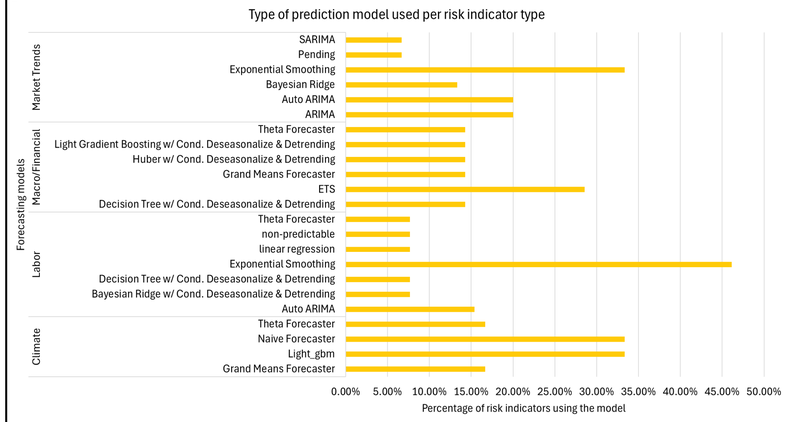
FORECAST MODELLING EXAMPLE

Risk Factor Prediction Example: Labor Force Participation Rate & Unemployment Percentage

- The modelling approach we took was a blend of manual and software-based modelling to forecast the risk factor. We studied the trend and volatility of each risk factor by visualizing the data and assigned a forecasting model based on the trends we observed. Then, we used PyCaret to choose the best forecasting model based on error metrics.



FORECASTING SUMMARY



This graph illustrates the frequency distribution of various models applied to different risk factors. By analyzing this distribution, we can determine if a particular risk factor is predicted with greater accuracy by a specific model.

CONCLUSION

We ended with 5 final deliverables. Our glossary of data sources kept track of what data we had used for future reference. The non-imputed master table and the imputed master table was used for PyCaret modeling. The forecast summary sheet documented our understanding of models that were used for the workable dataset. Our final deliverable was the final forecasting master table.

Lessons learned:
- Pattern recognition
- Dealing with scope creep
- Data imputation and processing
- Scraping from open-source databases
- Data visualization and PyCaret modeling

FUTURE WORK

In the future, it is valuable to consider opportunity cost relationship between scraping and processing open-source data and obtaining proprietary data.

Furthermore, the project can be expanded by using our cleaned and forecasted data to create a risk-predictive probability model. The team can validate the forecasting data through simulation testing.

Additionally, an internal risk indicator model, incorporating Caterpillar data can be visualized through a Power BI Dashboard.

ACKNOWLEGEMENTS AND REFERENCES

We would like to thank:
- Corporate TA: Angel D Avila Gonzalez
- Corporate Mentors: Somesh Mohapatra
- Purdue Data Mine Team: Nathan Ramquist, Jill Gough, Emily Hoeing, Cai Chen

References:
- Data Sources
- PyCaret Time Series Forecasting Tutorial