INTRODUCTION

CAT receives parts from over 27,000 individual suppliers and 80+ 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

FORECASTING MODEL

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

DATA IMPUTATION

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

FORECASTING SUMMARY

Type of prediction model used per risk indicator type:

- Supply chain disruptions
- Domestic construction labor
- GDP
- Imports
- Exports
- Consumer sentiment
- Producer sentiment
- Eurozone GDP
- Domestic US
- Global economic

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/ usable data.

<table>
<thead>
<tr>
<th>Risk Indicator</th>
<th>Model</th>
<th>Week 1</th>
<th>Week 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply chain disruptions</td>
<td>Linear Regression</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Domestic construction labor</td>
<td>Random Forest</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>GDP</td>
<td>Gradient Boosting</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Imports</td>
<td>Decision Tree</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Exports</td>
<td>Naive Bayes</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>Consumer sentiment</td>
<td>Random Forest</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Producer sentiment</td>
<td>Light Gradient Boosting</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Eurozone GDP</td>
<td>Light Gradient Boosting</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>Domestic US</td>
<td>Light Gradient Boosting</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Global economic</td>
<td>Light Gradient Boosting</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

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 references. 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.

ACKNOWLEDGEMENTS 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
- Our glossary of data sources kept track of what data we had used for future references. 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

References:
- Data Sources
- PyCaret Time Series Forecasting Tutorial