Capstone Three- Project Proposal

https://github.com/allen44/endo-us-econ-analysis

Problem Statement

Create a model of the textual data in the monthly ISM ROB reports, and predict the (binary) direction of change in GDP growth in the present and future.

Context

There an aphorism in finance and economics circles that goes something like:
“The ISM report predicts GDP with a time lag of three to six months”. There is plenty of numerical data in the report and lots of academic research has already been done on the numerical data in the report showing the ISM report is a useful leading indicator of the future of the business cycle. However, I’m not aware of any research on using the textual data from the report to predict any future economic data.

As GDP is reported quarterly while the ISM report is produced monthly, determining a relationship between the textual data in the ISM report and the GDP report could be useful in predicting GDP earlier than the official report, and could indicate the usefulness of using the ISM report text to predict other macroeconomic indicators.

Criteria for Success

When testing the model on unseen data (forecasting), the model must correctly predict the binary direction of rate of change of GDP growth with better than 50% accuracy.

Scope of Solution Space

For this project, we will only focus on building a model that can accurately predict the direction of change in GDP growth, while using only the textual data from the ISM report. There is plenty of numerical data in the report, but it will not be used for this analysis.

A high-performing model will have many broader applications (all of which are outside the scope of this project) such as,

- forecasting the state or change of other macroeconomic data,
- forecasting the time to next change in state,
- and identifying individual features that best forecast other macroeconomic data.

Constraints within Solution Space

- Data from the one data source only- We decide to not use numerical data from the ISM reports or economic data from other sources.
- Forecast must use only data from before the GDP report. If the ISM report is released on the same day as the GDP report, the most recent earlier data will be used for the model.
- Future text will have vocabulary not present in the training text. The selected model will have to be robust to these unexpected new words. For example, the word “COVID-19” was invented in 2019, and in high use in 2020, but non-existent in training data before 2019.

Stakeholders to provide key insight

The full multi-decade archive of text is not available for us to model,but the ISM may allow their members to access archival reports.

Key data sources

The sole data source for this analysis is the archive of PRNewsWire, a news website that still has older copies of the ISM Reports as well as the most recent copies. There is no public API for data from this website, so web scraping tools will be used.

Main URL: https://www.prnewswire.com/news/institute-for-supply-management

Link to a PDF of this Project Proposal on Github