SIADS 593: Milestone I
Team Project Proposal

version 2022.07.27.1.CT

Proposal Title: Predicting Recession: A Multiple Discriminant Analysis of Economic Indicators

1. Team members

Please list your team members (2-3 max).

- Avery Cloutier
- Matthew Grohotolski
- Sruthi Rayasam

2. Project summary

Summarize your proposed project in a few sentences.

- What is your proposed project and why are you proposing it?
- What are the question(s) you want to answer, or goal you want to achieve?

Our project aims to develop a predictive model to identify potential recessions by analyzing key economic indicators using Multiple Discriminant Analysis (MDA). Recognizing the profound impact recessions have on societies and economies, early detection is crucial for policymakers, businesses, and individuals to make informed decisions.

We propose to investigate the following questions:

- Which economic indicators are most significant in predicting recessions?
- How effectively can MDA classify periods as recessionary or non-recessionary based on these indicators?

By integrating diverse datasets and employing MDA, we seek to uncover patterns that precede recessions, offering a tool for early warning and decision-making support.

3. Datasets

Describe one primary dataset and at least one secondary dataset. If other secondary datasets will be used please describe them as well.

The proposed datasets should exhibit different features/columns and/or different access methods, e.g., *.csv file, *.json file, API retrieval, web scraping, etc. Different time periods, for example, with the same features/columns are not considered a different dataset. Remember, the focus of the project in this Milestone course is to give you the opportunity to practice your data manipulation skills, so feel free to challenge yourself.

If you're unsure if your data sets are "different enough" describe the datasets and request a review via the *#siads593_[semester]_001_project* Slack channel.

Please note: all proposed datasets ***MUST*** be publicly available to all members of the class (students, instructors, course support personnel, etc.). Use of proprietary datasets for this project is ***not*** permitted.

3.1 Primary dataset description

Describe your primary dataset. How is the data collected and how will you access it? Please share what features in the dataset are relevant to your topic. At a minimum, include the following information:

- Short description (i.e., 1-3 sentences) of its key features
- Estimated size (in records and/or bytes)
- Location (give the URL or other access method)
- Format (CSV, JSON, etc.)
- Access method (download, web scraping, API, etc.)

- **Short description:** The Federal Reserve Economic Data (FRED) repository offers a comprehensive collection of U.S. economic indicators, including real GDP, unemployment rates, Consumer Price Index (CPI), federal funds rate, and consumer sentiment indices.
- **Estimated size:** ~50,000 records
- **Location:** FRED - Federal Reserve Economic Data
- **Format:** CSV (exportable)
- **Access method:** API and direct CSV download (via Python fredapi or manual download)

3.2 Secondary dataset(s) description

Describe your secondary dataset(s). How is the data collected and how will you access it? Please share what features in the dataset(s) are relevant to your topic and describe the data types you're expecting. At a minimum, for each secondary dataset include the following information:

- Short description (i.e., 1-3 sentences) of its key features
- Estimated size (in records and/or bytes)
- Location (give the URL or other access method)
- Format (CSV, JSON, etc.)
- Access method (download, web scraping, API, etc.)

- **Short Description:** The National Bureau of Economic Research (NBER) provides official dates for U.S. business cycle peaks and troughs, identifying periods of recession and expansion.
- **Estimated Size:** Approximately 100 records, covering monthly data since 1854.
- **Location:** NBER Business Cycle Dating
- **Format:** CSV
- **Access Method:** Direct download

- **Short Description:** The University of Michigan's Surveys of Consumers offer insights into consumer sentiment, capturing expectations about personal finances and the economy.
- **Estimated Size:** Around 1,000 records, with monthly data available since the 1950s.
- **Location:** University of Michigan: Consumer Sentiment Index
- **Format:** CSV
- **Access Method:** Direct download or API access via FRED

- **Short Description:** % of labor force unemployed and actively seeking work.
- **Estimated Size:** ~900 records (monthly from 1948–present)
- **Location:** FRED Unemployment Rate
- **Format:** CSV
- **Access Method:** Direct download from FRED

3.3 [YES] Affirm: datasets are public.

Please write YES in the above box to confirm that your primary and secondary datasets are accessible and available to your classmates and the instructional team.

4. Cleaning and manipulation

Describe how you will need to manipulate your datasets: how will you handle missing or anomalous data? How will you join your primary and secondary datasets? What cleaning and manipulation challenges, if any, do you anticipate?

To prepare the datasets for analysis, we will undertake the following steps:

- **Time Alignment:** Standardize the time frames across datasets, converting all data to a consistent monthly frequency to facilitate accurate merging and comparison.
- **Handling Missing Data:** Employ interpolation methods or remove records with missing values, depending on the extent and significance of the gaps.
- **Feature Engineering:** Create lag variables to capture the delayed effects of economic indicators on recession onset.
- **Normalization:** Standardize variables to ensure comparability and to meet the assumptions of MDA.
- **Data Integration:** Merge datasets on the date field, ensuring that each record represents a comprehensive snapshot of the economic indicators for that month.

Anticipated challenges include resolving discrepancies in data frequency and addressing any structural breaks in the time series data due to changes in data collection methodologies over time.

5. Analysis

Describe any analyses you plan to undertake. For each, please give the technique or approach and briefly explain what you expect to learn from it.

Our analytical approach will encompass:

- **Feature Importances:** Display the contribution of each economic indicator to the discriminant functions and dropout irrelevant features not contributing to final MDA results.
- **Principal Component Analysis (PCA):** Apply PCA to reduce dimensionality and multicollinearity among predictors, enhancing the robustness of the MDA model.
- **Multiple Discriminant Analysis (MDA):** Utilize MDA to identify linear combinations of economic indicators that best differentiate between recessionary and non-recessionary periods.
- **Model Evaluation:** Assess model performance using metrics such as classification accuracy, sensitivity, specificity, and confusion matrices.
- **Cross-Validation:** Implement k-fold cross-validation to ensure the model's robustness and generalizability.

Through this analysis, we aim to uncover the most influential economic indicators in predicting recessions and to understand their combined effects.

6. Visualizations

Describe in 1-3 sentences at least **two** data visualizations that you plan to create. Include the chart type (e.g. bar chart, scatterplot, SPLOM, etc.) as well as the variables (features) you intend to plot.

We plan to create the following visualizations:

- **Correlation Heatmaps:** Illustrate the relationships among economic indicators to identify potential multicollinearity issues.
- **Time Series Plots:** Display trends of key economic indicators (e.g., GDP, unemployment rate, initial claims) over time, highlighting recession periods for contextual understanding.
- **Discriminant Function Plots:** Visualize the separation achieved by the MDA model between recessionary and non-recessionary periods.

These visualizations will aid in interpreting the model's findings and in communicating insights effectively.

7. Ethical considerations

Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

While the data used is publicly available and aggregated at the national level, we acknowledge the importance of ethical considerations in our analysis:

- **Data Interpretation:** We will exercise caution in interpreting the results, avoiding overreliance on the model's predictions for policy decisions without considering broader economic contexts.
- **Transparency:** All methodologies and assumptions will be documented clearly to ensure transparency and reproducibility.
- **Limitations:** We will explicitly state the limitations of our model, including potential biases and the historical nature of the data, to prevent misapplication of the findings.

8. Contributions

Indicate the contribution that each team member will make to the project.

- **Avery Cloutier:** Responsible for data cleaning, feature engineering, and visualization creation.
- **Matthew Grohotolski:** Lead on data acquisition and preprocessing, development of the analytical framework, and coordination of team activities.
- **Sruthi Rayasam:** Focused on model implementation, evaluation, and documentation of findings.

All team members will collaborate on the final report and presentation, ensuring a cohesive and comprehensive project outcome.

Changelog

(2022.07.27.1.CT) Update for 593
(2021.07.24.1.AW) Adjust title, number sections, simplify section headings, edit text