WorldData.io Case Study: Part 1 - Theory

▼ Question 1

> Assume you have **subnational labor market data** for a **country** for **several years** in the past. Which data and methods would you consider to create forecasts of the expected numbers of **employed**, **unemployed** and **inactive** persons in future years?

▼ Clarification

- What are the features associated with the dataset?
- Is the data complete? Are there missing time steps?
- What nation are we considering?
- How far in the future do we want to forecast?
- How often do we want to revisit the model to update it?
- How many years in the past we have (what is meant by “several years”?
- How granular is the data? (monthly, quarterly, yearly)
- Are we concerned with covariates/confounders like age, race, etc.

▼ Understand the Problem

- Assumptions
  - We have 10 years of data at the monthly scale
  - We are dealing with the United States
  - We have complete data for all of the States (subnational)
  - Steps-Ahead
    - We want to forecast one step-ahead (one year)

        - We have several models available
        - We can use traditional models (e.g., Seasonal ARIMA as well as newer ML methods (e.g., XGBoost)
        - W
    - We want to forecast 2 - 3 years (few years)
        - We need to baseline our models
        - We have the option of training the models
    - We want to forecast > 5 years
        - The choice of models will be limited.

▼ Metrics/Features

- As stated in the problem, we want to forecast and begin with the U.S. Bureau of Labor Statistics definitions
    - Employed (people with jobs)
    - Unemployed (People who are jobless, looking for jobs, and available to work)
    - Inactive (all others, people who are neither employed nor unemployed, not in labor force)
- We want to include exogenous variables
    - education enrolment
    - Birth rates
    - stratified population metrics (16 years and over, labor force participants)
    - imprisoned numbers (excluded from labor force)
    - numbers enlisted in Armed Force (excluded from labor force
    - financial events (e.g., recessions)

▼ Models

- I would start with univariate analysis, using techniques to decomposed the seasonal, cyclical and trend components from the time series and fit the

ARIMA model to the remainder. This will serve as the baseline model.

- The next step would be to investigate multivararite time series models. This would be ideal to identify a suitable model which incorporates other features into the prediction. For this task, I would include the Temporal Fusion Transformer model as one of the models for selection as past and future covariates can be incorporated into the model. Additionally, long-term and recent history can be used for future predictions.

▼ Question 2

> Assume you have data on the numbers of employed by sector for **several countries** from **different surveys**, which include different sectors and sub-sectors. How would you combine and structure these data to present comparable numbers for all countries?

▼ Clarification

- How different are the surveys from each country
  - Are there some common surveys which were conducted by common survey outfits?
  - How many common sectors and sub-sector surveys were conducted—even by different surveyors?

▼ Understand the Problem

Assumptions

1. There are few countries that have non-overlapping sectors and sub-sectors.
2. There are cases where sub-sectors are available for some countries
3. We have access to exogenous global covariate data for all countries.
4. Some sectors/sub-sectors are correlated.

▼ Metrics

- I would aggregate all sub-sector data to sector
- I would use a generative model to fill in the missing sector data for countries. This model will use global exogenous data (e.g., goods export/inport data) in addition to the available closely-related sectors. The model used to do this will most likely be a multivarate time-series model
- Assuming that there will be several sectors which are only represented for a small number of countries, I will drop them from the analysis.

▼ Test

- I would focus on the data-rich countries to examine which common features provide the most explanatory power to use for those countries with lower amounts of information. This will all for the examination of a few different models in order to fill in the missing employment data

▼ Conclusions

- The data-rich countries will guide data collection efforts in order to get better employment estimates for the countries with lower amounts of information.