Final Project Proposal (DATS-6103)

Submission Deadline: **Oct 26**

Name(s): Will Kennedy, Ilgaz Kuscu, Alejandro Gomez

Objective

[Describe the overall objective of the project. If necessary, explain the context and motivation.]

The goal of this project is to create a very simple model to estimate current GDP before it's officially released. We will use a few basic economic indicators and easy machine learning models to make a first nowcast. This is not meant to be a perfect forecast, just a simple, working example of how nowcasting works.

Impact

[In a few sentences, delineate the potential impact of the project]

A nowcasting model can provide valuable early information:

- Timely insight: It gives an early signal of how the economy is performing, instead of waiting weeks for official reports.
- Better decision-making: Businesses, analysts, and policymakers can use early estimates to plan ahead, adjust forecasts, or react to changes.
- Useful for smaller economies: In countries where official data is often delayed, this kind of model can offer valuable early information with minimal cost.
- Educational value: This project is also a learning exercise, showing how data science can be applied to real economic problems without needing complex or expensive infrastructure.

Our goal is to build a simple, clear, practical and understandable nowcasting pipeline that could be improved over time.

Dataset(s)

[List your data sources with links to them. If you have already uploaded them to your repository

on GitHub, please mention the location. In addition, briefly discuss the datatypes and the reliability of the data.]

We will use public and free data from:

- Federal Reserve Bank of St. Louis (FRED) — GDP, unemployment, and retail sales
  https://fred.stlouisfed.org/
- Bank for International Settlements (BIS) — payments data
  https://data.bis.org/search
- We will work with simple time series (monthly or quarterly).

All datasets are numeric and reliable, and will be cleaned and used to train a basic regression model (e.g., linear regression or random forest) to produce a GDP nowcast.

Approach

[Talk about how you plan on approaching this project through several steps. List the steps below.]

We will begin with a review of the existing literature on GDP nowcasting. This will give us an idea of the best model types, model specifications, data sources and predictors that have already been identified by the existing data science and econometrics community. GDPNow, produced by the Atlanta Federal Reserve Bank, is a popular model with a publicly released methodology. Additionally, the paper "Lessons from Nowcasting GDP across the World" by Cascaldi-Garcia et al. and recent papers on GDP nowcasting like "Nowcasting GDP with a pool of factor models and a fast estimation algorithm" by Eraslan and Schroder will give us a stronger idea of where the field is currently, and how we can mimic and potentially even build on what has already been done.

After gaining a stronger understanding of GDP nowcasting as it is currently being done, we will collect the appropriate data. This will be a mix of the best predictors as identified by previous data scientists, and some new, more novel predictors that will set our analysis apart.

We will then construct a series of nowcasting models using this data, evaluating it via back-testing on historical data. The models we consider will be informed by the ones we are learning in DATS-6103, which include linear regression, decision tree regression/random forests, and support vector regression. Some models from class aren't suited for predicting a continuous value (KNN, Discriminant Analysis, Naive Bayes), so we'll supplement these models with ones from the literature, probably time-series models. We will calculate MSEs for these models to be able to compare them in terms of their accuracy.

Once the best model and specification is selected, we produce a final model and submit our best estimate of current GDP, ahead of the announcement.

Timeline

[Edit the following example timeline]

This is a rough timeline for this project:

- (1 Week) Literature Review:
  Goal: gain a solid understanding of the current state and efficacy of GDP nowcasting models.

- (1 Week) Data Acquisition:
  Goal: Have a list of data sources and pipelines set up to integrate them into a code + modelling environment (R, Python). Also will want to have timelines for when this data updates – there could be a data update between this week and the end of the project, and completely up-to-date data is essential for the best nowcasting.

- (1 Weeks) Data Wrangling:
  Goal: Have a cleaned version of all the data acquired during the previous result. Missing/inconsistent values handled, and a reference data dictionary for our key predictors and tables.

- (2 Week) Modeling:
  Goal: Have final model specifications with predicted values and MSEs + AICs/BICs for each to compare.

- (0.5 Week) Compiling results:
  Goal: Clean up modelling code, combine into one, easy-to-run replication package, decide on which model is "best".

- (1.5 Week) Writing up the report
Goal: Have a final, concise but informative deliverable on our work.

Possible Issues

[List some of the prospective challenges and issues, and discuss how you envision overcoming them]

Data availability could be a challenge – the best economic data is often behind a paywall or a government security clearance, meaning even if our model is ideal, we probably won't be able to compete with models from econometricians with higher budgets and access to regulatory data. We might offer more of a value-add from this project via testing out an unorthodox model or group of predictors to see if that improves our performance.

The data we use will likely come from a variety of sources, and will therefore likely have different frequencies. Bridging the gap between daily, weekly, and monthly data will be non-
trivial – we could just convert everything to the largest frequency to match, but we would lose a lot of variation in doing so. Additionally, nowcasting often suffers from a 'ragged edge' problem, where data is released at various lags and thus some data points are much more up-to-date than others. Solving these problems will require some creative data wrangling and modeling – perhaps interpolating values for the lower frequency data, or using complex dynamic factor models.

Computational power could prove to be a constraint, considering the high demands of some of the newest, most powerful big-data models. We might need to see what GW resources are available to optimize the limited time we have for this project.