Group Project Check-in

submitted by Radhika Kaul, Odiche Nwabuikwu, & Ruochen Wang

Question of Interest: Predicting the number of unemployment claims filed per U.S. state in in the second week of May (when project is due) using data from various sources but the major source of what we want to predict comes from the Unemployment Insurance Weekly Claims data from the U.S. Department of Labor (DOL)'s Employment & Training Administration.

Project Progress: In the wake of the COVID-19 pandemic and responses set up by different states to this crisis, we have finally decided on pursuing the following predicted model for our analysis:

What we are trying to predict: Either New Initial Claims filed per week OR percentage of the labor force filing for unemployment claims per week. Initial claims is one of the most-sensitive, high-frequency official statistics used to detect changes in the labor market (Source: Aaron Soujourner and Paul Goldsmith Pinkham UI trends data.

Predictors: Structural predictors include population, per capita income,top industry contributing to GDP, working population, region, and percentage population with a highschool degree. Some concerns of using these variables is high correlation in population variables and the need to pre-process (standardize/normalize/log) them. Real-time predictors include Number of COVID cases per state, number of days since the stay-at-home orders started, peaks predicted for each state (although we haven't found a particular data source), Democratic or Republic governor, number of days till stay-at-home orders will end, and days since the protests started per state.

- **Pre-processing data**:
- **Predictive model**: Since we are using weekly time series data by state (and we have 15 weeks worth of data on both initial and continiuing uneployment claims) in addition to having features that are structural as well as real time, we propose to use a **Random Forest** model. This example gives an idea of why not only initial claims (UI claims people applied for the first time) but also continuing claims (UI claims people who remain on unemployment benefits) give us a real-time impact on unemployment as the US started observing its first cases of COVID-19 cases. Random Forest is known to be robust to correlated predictors because in the aformentioned section, some structural variables are highly correlated (ex. population and working population by state). Also given the volatility in the weekly claims data which we plan to explore, this method has relatively better chance of making 'out-of-sample' predictions, and better interpretability.
- **Error Metric**:

Data Cleaning

```
library(tidyverse)
```

```
topindustry_bystate <-
  read.csv("data/raw/gdp_bystate_byindustry_2019.csv")

topindustry_bystate <- topindustry_bystate %>%
  filter(str_detect(string = Description,
                    pattern = "    ")) %>%
  filter(!str_detect(string = Description,
                     pattern = "     "))
```