Julian Suarez

Sep 17, 2020

Introduction

This is a personal project to develop a machine learning algorithm to create a forecast of the unemployment rate of Australia resulting from the government lockdown measures taken due to Covid-19.

Data Exploration

For the development of this project it was necessary to obtain data regarding two main aspects. Firstly the unemployment figures of Australia and secondly data categorizing government lockdown measures.

Australian government response data was taken from the web page Our World in Data, an organization who has compiled a vast range of data related to the world pandemic due to Covid-19. Specifically I made use of the COVID Government Response data set where there is a daily detailed description of several government measures taken as well as a numerical scale measuring the severity of the restriction. Let's take school closure policies, for countries requiring school closure they get assigned the max calcification that being a three, while countries with no school closure measures will be assigned a 0.

With the government response data found I was interested in making use of the large daily data as it will better inform the algorithm of the effect of lockdowns on unemployment. However I couldn’t find any reliable data with the daily unemployment figures of Australia (unemployment rate or total unemployment). On the other hand the Australian Bureau of Statistics although it has a comprehensive description of the unemployment in the country, it is updated monthly. This presents a major problem as my working data set is constrained by the months passed on 2020, as I am only interested in seeing the dependence of the unemployment on the australian government lockdown measures.

To combat this problem I decided to increase the data set pool by including other countries' unemployment to the data set. This approach nonetheless comes with its own pros and cons. In its favor we increase the data set so the algorithm has more data to model its predictions. On the other hand I am training the model to predict unemployment according to global tendencies not particularly how lockdowns affected Australia directly. Although this might be the case, it is still valuable to observe global unemployment tendencies as the lockdowns do not only affect a specific country, more generally measures taken by neighbor countries and superpowers will have an impact on any specific country's unemployment. Now that I’ve discussed the challenges faced on our unemployment data, I used the data provided OECD data on the world's unemployment rate.

The last step necesar to discuss is the discrepancy between the government response and the world's unemployment rate data set. There are two problems that need to be addressed. Firstly not all of the countries on the government response are present on the world’s unemployment data, this happens from discrepancies on information reported to foiregin entities which is out of our control, as such for the model training the two data sets are joined by countries present. Secondly as mentioned previously the government response data has a daily rollup, as such I had to group the data to create a monthly data set. Here the decision was made that a good way to represent a countries lockdown measure was by reporting the monthly average as the government's response each month. This approach for reporting monthly government responses may lead to a biased data set, favoring earl policy changes. However by the nature of the situation I believe that favoring early lockdown measures follows the global government's approach regarding their response against Covid-19.

Finally all of these considerations were implemented on python under the file dataExploration.py. This script opens the raw data from downloaded csv files formatted and joins the two datasets and saves the final table to a file on the computer as well as the individual formatted data tables.

XGBoost Model

For this project I choose to use the eXtreme Gradient Boosting machine learning algorithm (XGBoost) because of the scalability and flexibility inherent to the model. Specifically, due to the nature of the data being utilized for the project I used the supervised learning regression model XGBRegressor algorithm. SInce the data being used represents general unemployment trends in the world I decided to separate the training of the data into two parts.

First I would stack three XBGRegressors, each with a different objective function for minimization, then use the output of each model using the government response of Australia to train a second stage model with Australian unemployment data. The purpose of this staged training is to initially make use of global unemployment data to inform the trends expected to be seen by Australia, then calibrate those predictions with the Australian unemployment to better predict the magnitude of the predictions according to the unemployment rates observed in Australia.

This approach in hand will allow me to make use of other countries' lockdown measures from the government response data set to simulate how a second or third wave may affect the unemployment rate in Australia.

Model Parameter Tuning

For a step by step description of the parameter tuning method it's better to read the documentations on auFunctionProjects.py. However the main objective of this process is to optimize the parameters of the model to improve the predictions. This is achieved by iteratively measuring the predictions errors when there is change in any one parameter value. The code then picks the parameter value which minimizes the predictions error and sets it to the model.

First Stage Model Output

Squared Log Error Model

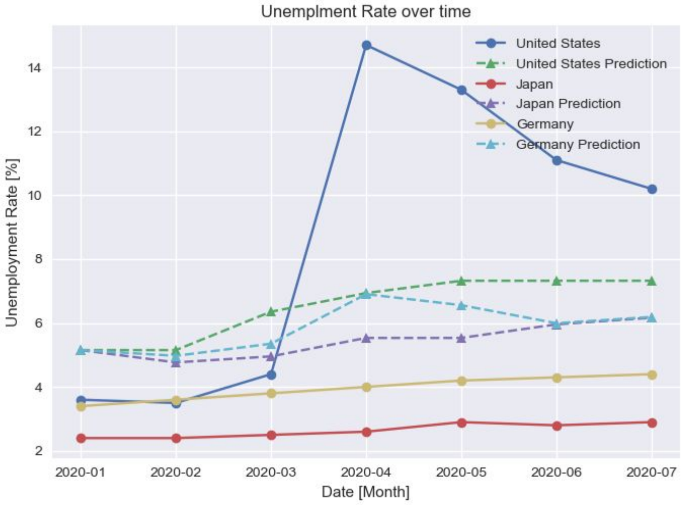

Figure 1. Unemployment rate over time for squared log error model, solid lines represent real data and stripped lines represent predicted data.

Squared Error Model

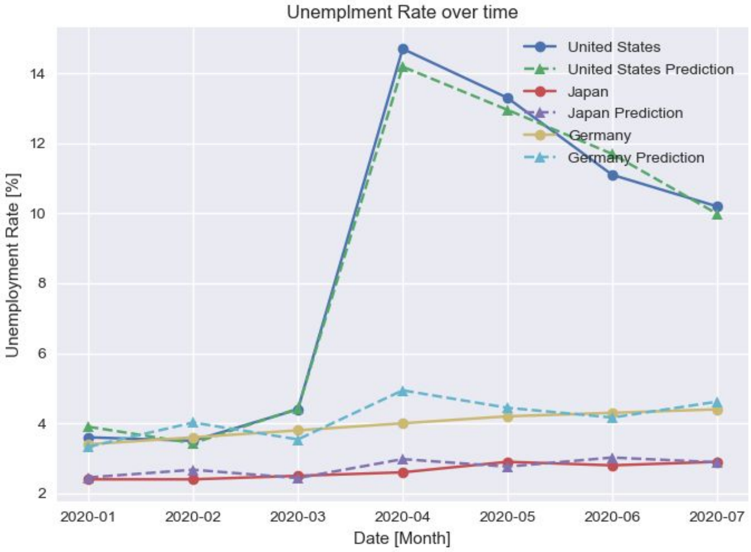

Figure 2. Unemployment rate over time for squared error model, solid lines represent real data and stripped lines represent predicted data.

Pseudo Huber Error Model

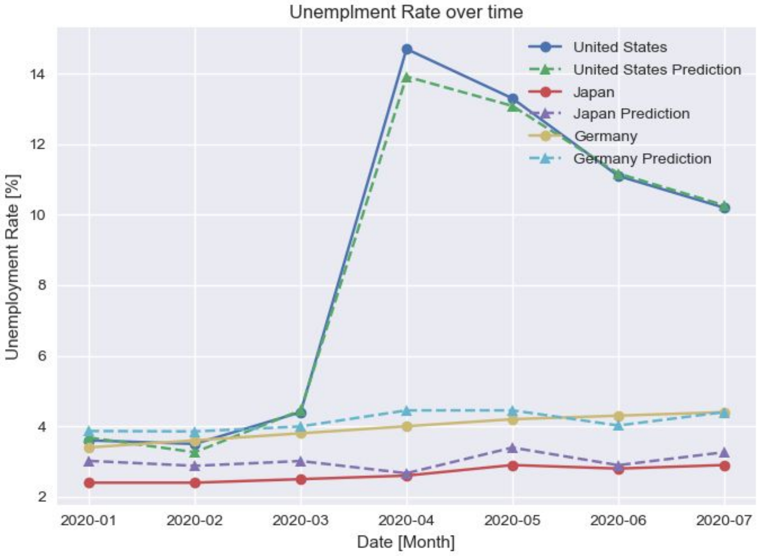

Figure 3. Unemployment rate over time for Pseudo Huber Error Model, solid lines represent real data and stripped lines represent predicted data

Final Model predictions

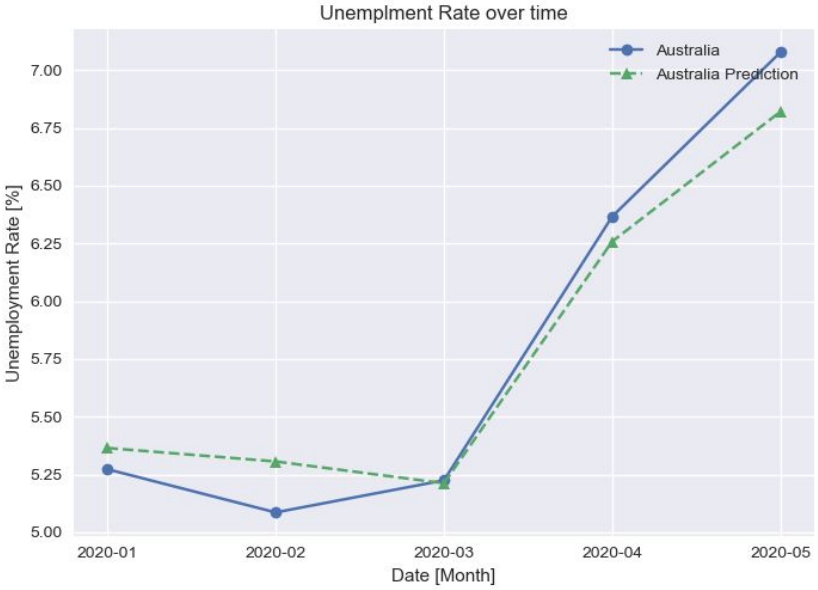

Figure 4. Unemployment rate over time for Final Model, solid lines represent real data and stripped lines represent predicted data

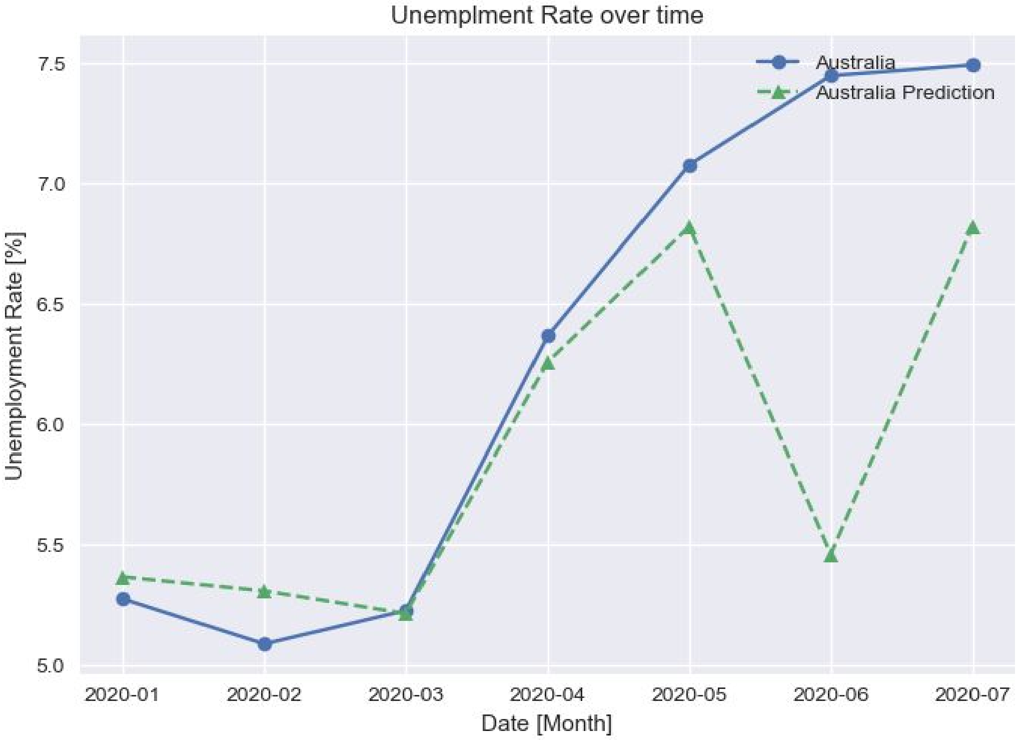

Figure 5. Graph of models predictions for unemployment data including the test data.

Something to mention about our predictions, as seen on the unemployment rate in June it suddenly decreases which is not what we expect the behavior to be. This may be a result that some of the countries used for training the first layer of algorithms do not have data after June, as such the algorithm sees an unexpected drop in the unemployment rate after June.