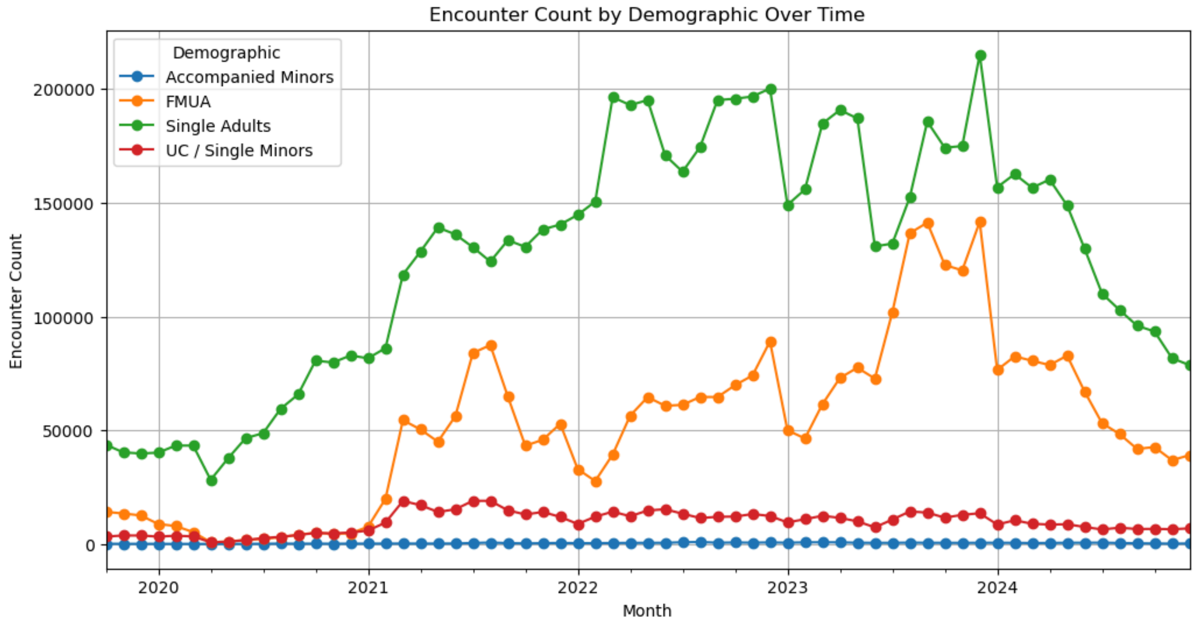
Analyzing and Forecasting U.S. Immigration Trends

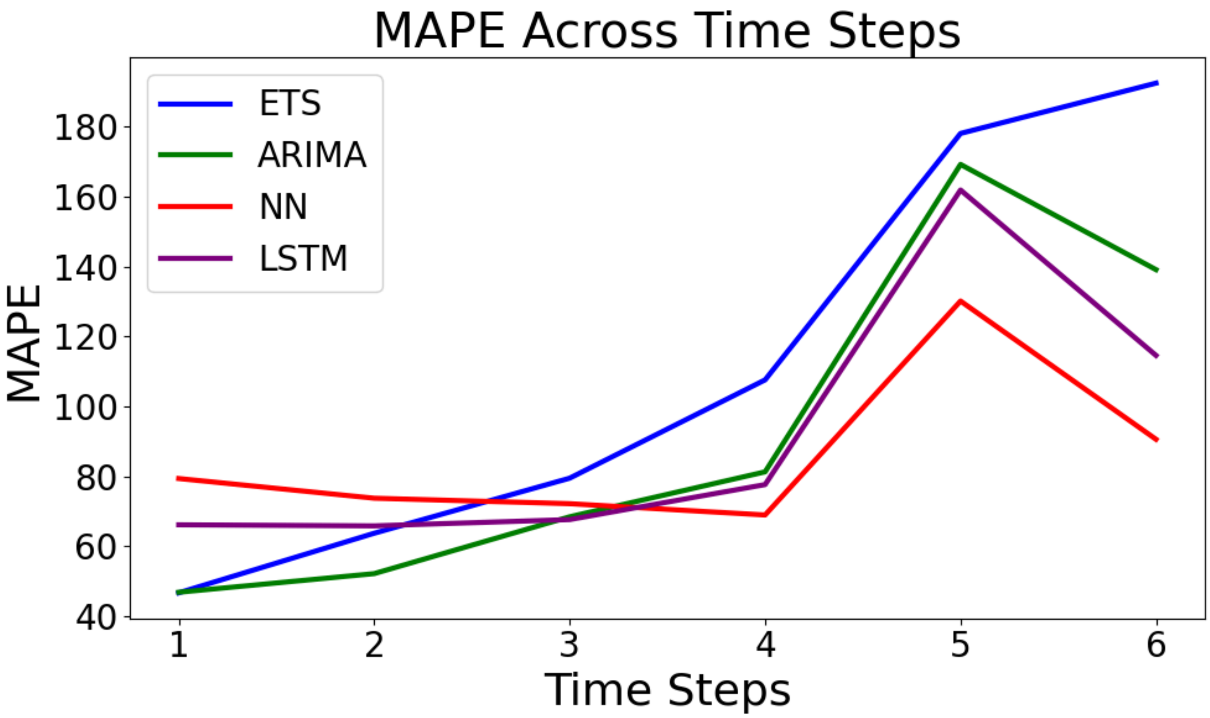
Group 7: Alan Lin, Trevor Petrin, Ganesh Kumar Ramar, Mingyu Chen

Repository: alanklin/AA-Capstone: Boston College Applied Analytics Capstone Project

This week, our main focus was diving into a more data-centric MLOps approach in leveraging trends and other information that would be influential into our model but not yet incorporated. Having the understanding that ARIMA is our current best performer, we shall try this week to optimize the performance of those models by improving our data. However, because ARIMA is at its core a univariate model, we will need to slightly tweak our model built to ARMIAX to ensure we can capture trends beyond the simple ARIMA model.

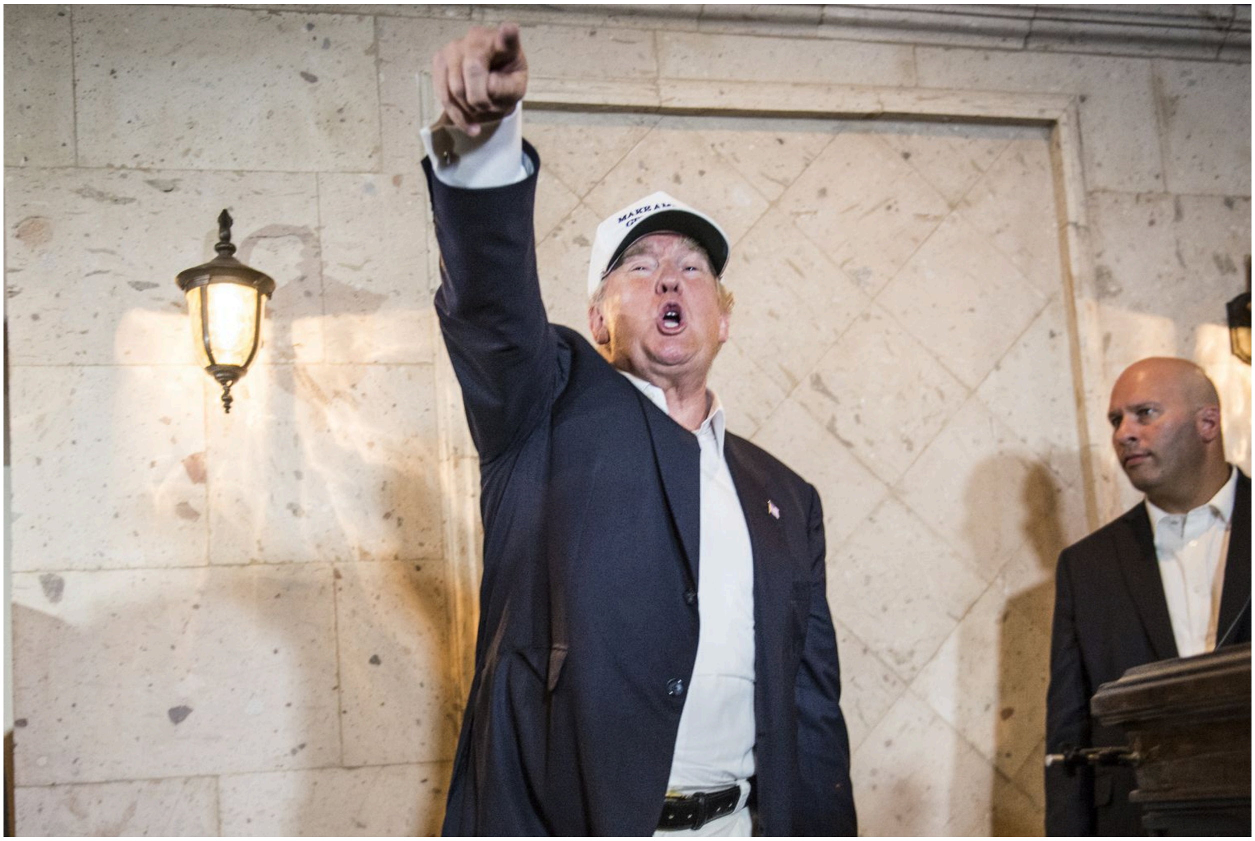


As for what major factors occurred during the duration of the model affecting immigration numbers that were not previously included, those events included the election cycles, the Covid-19 pandemic, and the enactment of Title 42 to expedite rejections at the border. Of these 3 factors, as can be seen in the graph above, we believe it is reasonable to target the change in power between President Donald Trump and former President Joe Biden because of where our models err in the test set. As can be seen in the MAPE graphs below, there is a spike in overall MAPE around month 5 of the test set before slightly dipping in month 6. Those two months are significant to this issue because of the election of President Trump in November 2024, coinciding with our test data month 5.



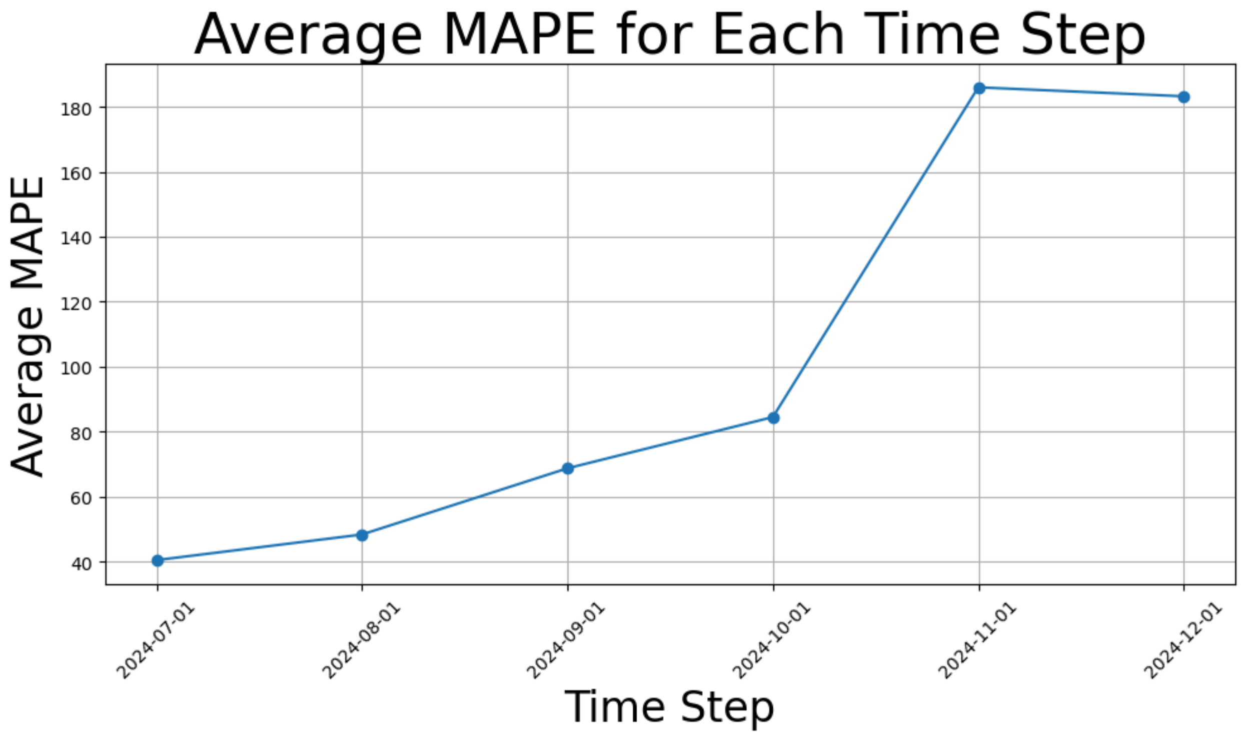
As for how we are going to account for the change in who is in office, because there are only 2 individuals in power during this time, we will use one-hot encoding corresponding with who is in power at the time. The notable exception we will have to this will be the president's lame duck period, thus the binary variable shall change from one to the other from October to November in an election year. This change will allow us in the test set to account for the reelection of President Trump during that time.

We want to take into account with our model the knowledge that Trump would be reelected in November 2024, thus we are slightly changing our model to an ARIMAX forecast, which will make use of the binary variable of which president is in office. ARIMAX is short for AutoRegressive Integrated Moving Average with eXogenous inputs, which basically means it is just an ARIMA variable with external factors.



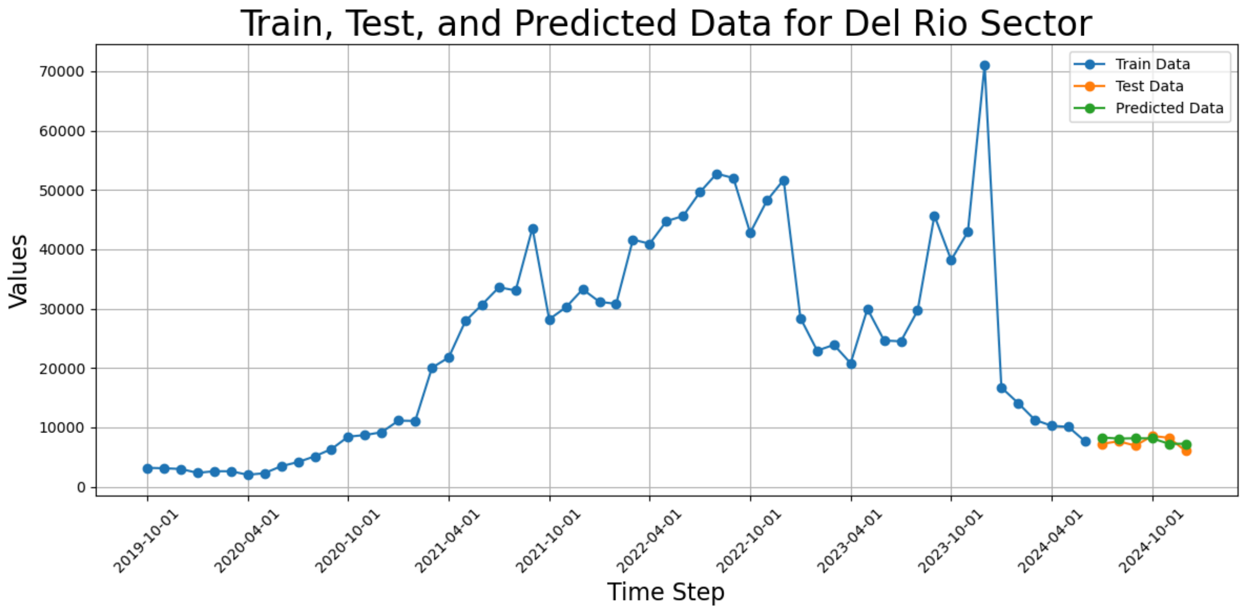
A quick recap as to why who is in office matters. President Trump, during his first term, was notorious for his anti-immigrant rhetoric and use of cruelty to deter migrants through the inhumane detention facilities and the family separation policy upon being apprehended. These policies deterred individuals from trying to enter the country despite the overall promise of a better life in the United States. President Biden, on the other hand, rolled back on Trump's policies leaving the impression that the United States would be more welcoming of migrants. This led to an unsustainable influx of migrants and overwhelmed facilities at the US Southern Border. Though Trump would not have power until January 2025, his rhetoric and previous policies would be enough to deter immigration and Biden would have no power to stop the upcoming policies because he was at that point a lame duck president.

As for the implementation in code, the president at the time was implemented by representing President Trump as 1 and President Biden as 0 for model input. From there, the previous ARIMA code was slightly tailored such that the model was changed to accommodate the ARIMAX model.

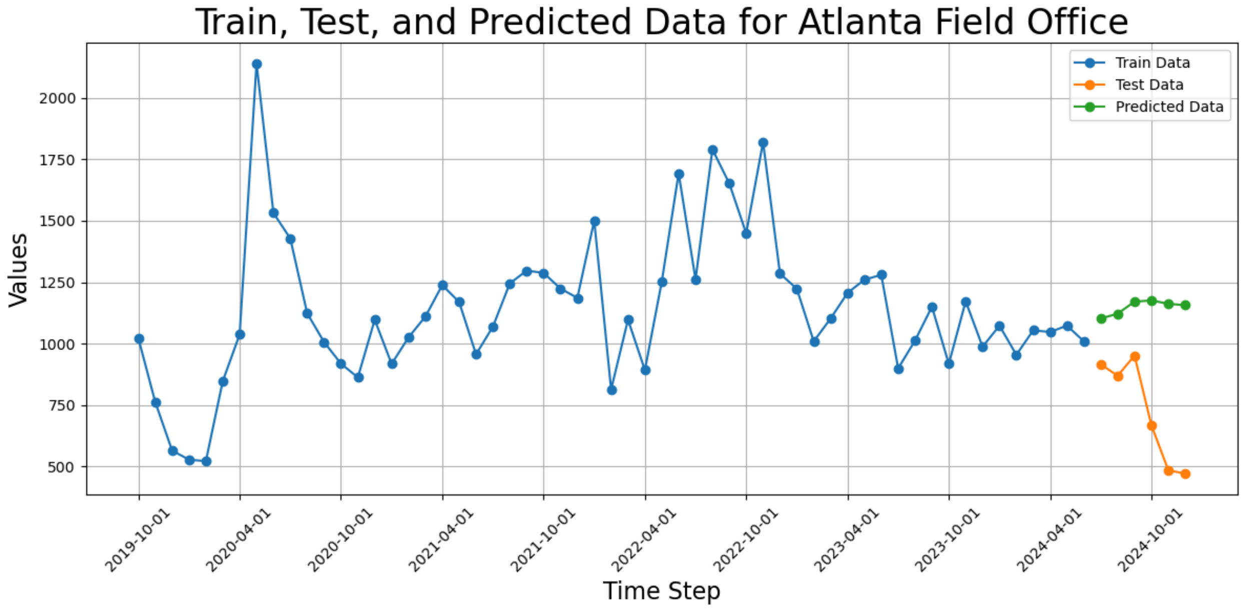


As for the results of changing the model, there was a sizable improvement in forecasting in the short term with the initial 2 month MAPE dropping from 49% in the traditional ARIMA model to 44% in the ARIMAX model. However, as the time went on, the ARIMAX model did not pick up well at all the effect that Trump would have when he got back into office, underestimating the number of migrants who would keep coming into the country.

MAPE Average Results			
	Train	2 Month	6 Month
ARIMA	**53.7%**	**49.6%**	**92.8%**
ARIMAX	**54.7%**	**44.2%**	**101.9%**



As for some of the sectors such as Del Rio, one of the sectors with the highest encounter totals, the model predicted extremely well despite seeing the spike and sudden drop late in the data. This forecast would have proved extremely effective in giving US Customs and Border Patrol (USBCP) the critical information they would need to effectively allocate resources and personnel to handle the rate of encounters.



Other sectors, though, did not show improvement as the Atlanta Field Office. However, when we use the MAPE to calculate this performance it does not tell the whole story while weighting the smaller sectors and field offices the same as the large ones. In other words, it does not tell the full story with our overall metrica that the more volatile sectors with less influx such as Big Bend or Havre being weighted the same as Tucson or El Paso. If we were to have thought of this beforehand, we would go back and understand a better way to weight these such as a weighted sum of the MAPEs or a different formula to better show that we are doing well with the sectors that have a higher influx.

Overall, taking a data-centric approach to our modeling process did have a profound impact upon our first two months of results and gave us better insight for how we could draw in outside factors to expand upon our more basic univariate model. Model building, if done with bad data can be worse than useless, leading to deliberately false conclusions. Data-centric model building will help prevent that by ensuring the refinement of what goes into the model before the process goes awry.

Citations

Lind, D., & Yglesias, M. (2016, August 22). *Donald Trump and immigration, explained*. Vox. https://www.vox.com/2016/8/22/12552082/donald-trump-immigration