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

Introduction

This week's report documents the end-to-end building and packaging of the US Customs and Border Patrol migrant encounters forecasting model to help address the issue of predicting the flow of migrants into the United States. As it stood, there was a significant spike in the influx of Migrants during the period after the initial shock of the Covid-19 pandemic beginning in January 2021. Several factors contributed to the surge in migrations including a shift in rhetoric about immigrants and the repealing of policies concerning encountered migrants such as family separation as well as the - ever-elusive - promise of better economic opportunities in the US (The American Dream).

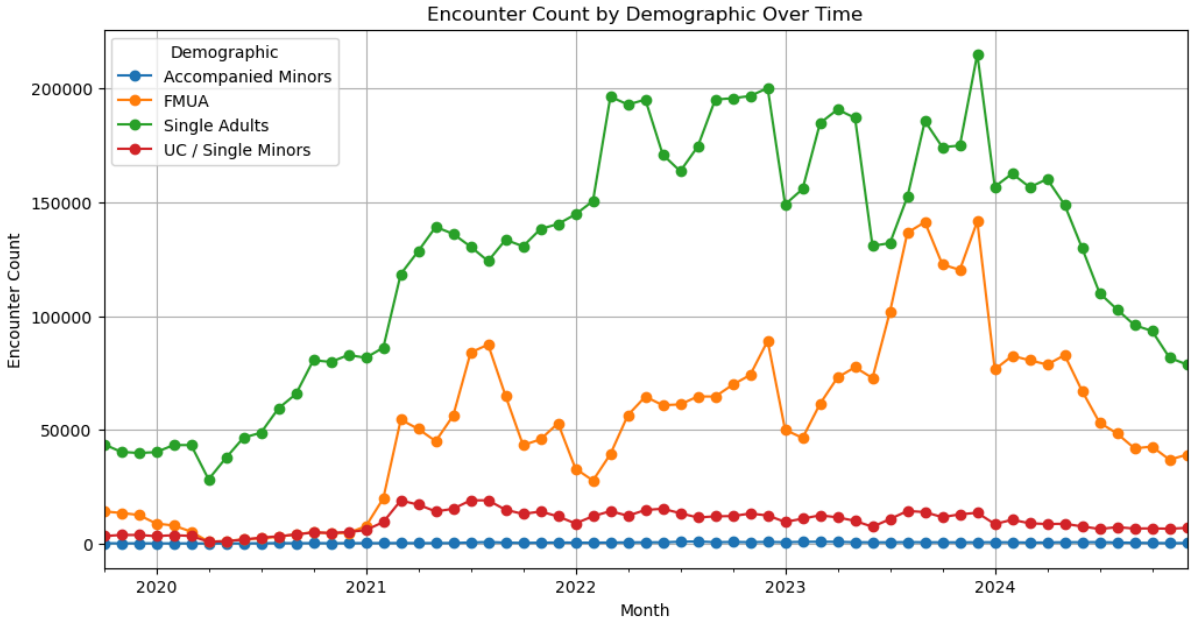


Figure 1: Number of encounters by demographic

For the purposes of this study, an “encounter” is defined as an instance in which a migrant is apprehended by U.S. Customs and Border Protection (USCBP) and requests asylum in the United States. Migrants may be apprehended either at official ports of entry, such as border crossings, seaports, or airports, or between ports of entry, including areas along the U.S.–Mexico border, such as the Rio Grande River. These encounters may be processed under one of two legal authorities: Title 8, which governs the standard procedures for evaluating claims such as political persecution, or Title 42, which, during the COVID-19 pandemic, permitted the rapid expulsion of migrants on public health grounds.

These migration trends led to political turmoil within the United States with issues such as the perceived violence caused by undocumented immigrants, the sudden discourse over sanctuary cities, the idea that too many migrants were granted an asylum hearing, and that the resources along the southern border were too thinly stretched to adequately handle this crisis. As a political result, one of the main consequences was the re-election of Donald Trump to the White House who was keen on changing the more lenient immigration policies to his more hard-line stance in detention and deportation of those who try to cross and those who have been residing in the US already.

Figure 2: Workflow diagram for the project (Generated using Sora by OpenAI)

For a brief overview of the project, our goal is to use the input for the model from October 2019-June 2024 to predict the following 6 months worth of migrant encounters. This report will describe the complete model building process beginning with the data collection and EDA followed by the data preparation, model building, forecasting, and packaging for deployment.

Data Collection

This data was sourced directly from the USCBP data website downloading and aggregating data from two separate flat files, one concerning FY 2019-2022 and one concerning FY2022-FY2025. There was no API available for use to pull this data directly from their database, so the process of pulling the data into the repository was done manually.

Data Aggregation

There were some slight changes that needed to be made to transition this data in preparation for the modeling process. Changes to the dataset included: reassigning of fiscal to calendar year, merging the datasets for continuous flow, aggregating the data by border sector to add value to the forecasts, and aggregating the data by individual demographics such as family status and country of origin. Upon these aggregations, we had a clean-cut 41x57 training dataset with 41 unique sectors to classify a migrant as being encountered at and 57 training time points. Missing entries were classified as a 0, as it implied that there had been no migrants encountered at that sector during that month.

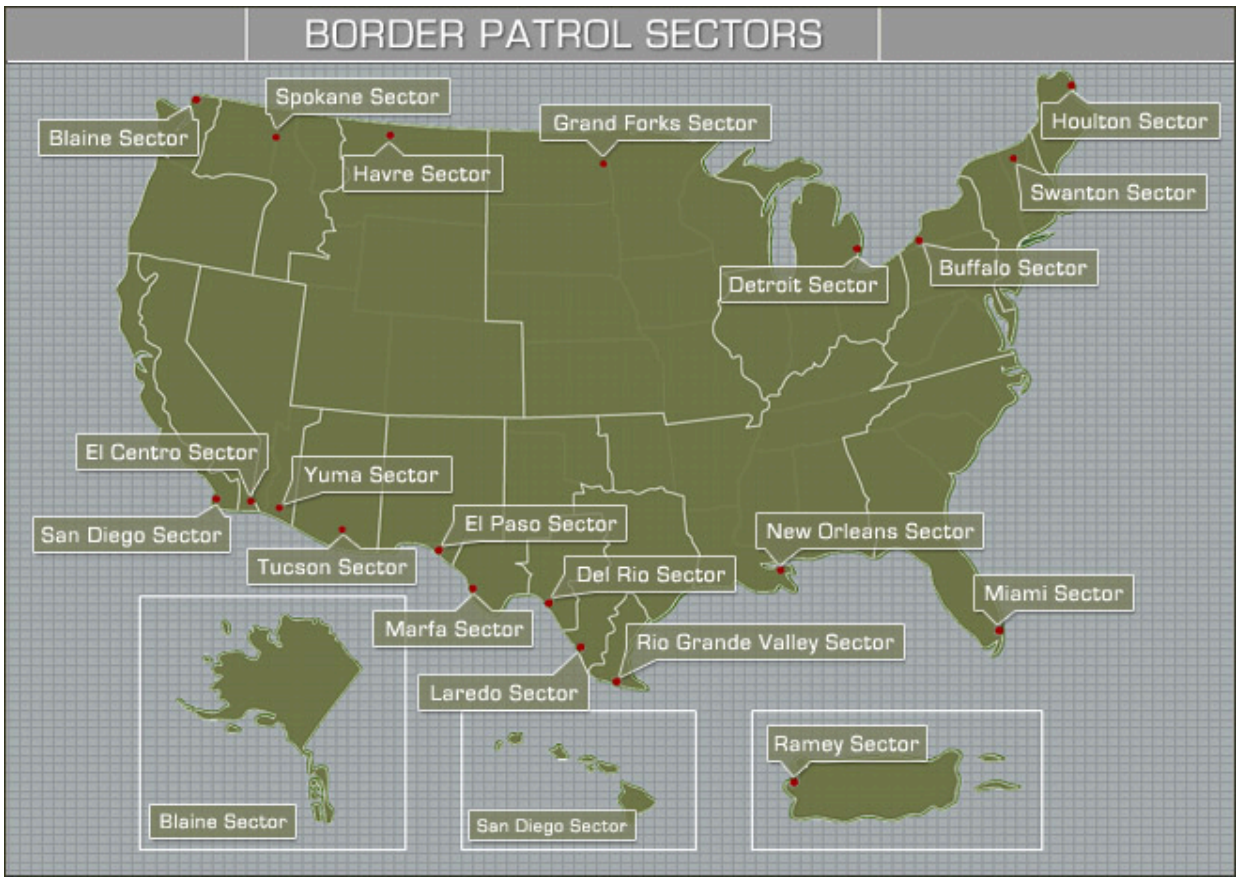


Figure 3: U.S. Border Patrol Sectors

For the model building process, each row was separated such that the individual models were trained upon a sole sector's data to ensure that unrelated trends would not have an impact upon unrelated sectors. For example, the Blaine Sector should not have an impact upon the Miami Sector, thus those two should not be trained nor tested upon the same model.
As for the validation/testing splits, our testing split was the last 6 months of data available from July-December 2024. Initially, that range extended from January-December 2024, though we believed that a more accurate short-term forecast would provide more actionable insights than a longer less accurate reading. We made this change around Week 6 of the project to be done in time for the Analytics in Industry Symposium.

We did not include a traditional validation split because, while feasible, we determined that for our ARIMA and ETS models, recursive predictions on the full training set offered a more realistic assessment of performance. For the LSTM models, we trained on the entire dataset to fully utilize the limited time series data, since random splits would disrupt temporal dependencies. However, for both the neural network and LSTM models, we incorporated a windowed validation approach during training to support iterative model improvement, though this was not used as a holdout set for final performance evaluation.

Model Building

There were several models used to forecast the test set including ARIMA, ETS, Neural Network, and LSTM. Our best-performing model was the ARIMA, so we shall mainly focus on that within the scope of this report. We fine tuned the ARIMA models individually by Sector to achieve the best performance in each region. For the assessment of the model, Mean Absolute Percentage Error (MAPE) was used to determine the overall effectiveness of the model. This metric was used because of the nonstandard size of each sector, where mispredicting some of the larger sectors in the South would disproportionately impact other metrics such as RMSE or MAE.

For the ARIMA model, the data did not need to be scaled for model input, though that step was taken for the Neural network and LSTM.

ARIMA models naturally have three parameters that can be adjusted and tuned for better model performance: *p, d, q*. The *p* parameter is for autoregressive terms, or the number of lag terms included in the equation. In autoregression models, the output is the future data point expressed as a linear combination of the past *p* data points. In general, the more lag terms there are, the better the model is at capturing a broader range of temporal dependencies and forecasting based on past values and patterns. However, there is a delicate balance to be met when choosing *p,* as too many terms can over-complicate the model and lead to overfitting. The *d* parameter is for the number of nonseasonal differences needed to achieve stationarity. This is the degree of differencing needed to make a time series stationary, which simplifies modeling and forecasting by ensuring statistical features such as mean and variance remain constant over time. This can help trends and patterns emerge that would not have been possible with the raw data. The differencing order generally never goes beyond 2, even though there is no theoretical upper limit. So, we limit our *d* values to [0, 1, 2]. Finally, the *q* parameter represents the order of the moving average, or the number of lagged forecast errors included in the model. Including past prediction errors (or residuals) into the model helps to capture the random noise or shocks in time series, which can improve prediction accuracy. We don't want to add too much noise to the model; otherwise the model won't be able to capture the underlying patterns.

ARIMA Tuning Parameters		
Autoregressive (*p*)	Differencing (*d*)	Moving Average (*q*)
(0,4)	**(0-2)**	**(0-4)**

Table 1: Tuning parameters for the ARIMA model

Once the optimal parameters were tuned by region, recursive forecasting was used to predict each time step ahead repeatedly for each of the 6 months in the test set. Tuning was done through a nested for-loop iterating through the parameters because the package used for the ARIMA model was not formatted to have a IterTools input which would make the code cleaner.

One factor we were unable to account for in the model building process was a significant political event during the test period: the re-election of Donald Trump to the U.S. presidency. Following his reelection, public rhetoric and proposed policy shifts—such as threats of stricter asylum procedures and references to deportations to facilities like El Salvador's CECOT prison and Guantanamo Bay—were widely covered in the media. These announcements likely contributed to a marked decline in the number of migrants encountered at the border during the test period. Because this event fell outside the scope of the historical data used to train the model, its impact on migrant encounter numbers was not captured in the model's predictive capacity. As a result, every model failed to predict the sudden dip in encounters, leading to high MAPE values in the 6-month window.

Though we did try to address the issue of the change in presidential administration via ARIMAX models taking into account outside factors, the ARIMA was a better sell for its widespread forecasting use.

MAPE Average Results		
	2 Month	6 Month
ARIMA	**49.6%**	**92.8%**
ETS	**55.3%**	**111.4%**
LSTM	**66.0%**	**92.3%**
Neural Network	**76.6%**	**85.9%**

Table 2: Performance results for all models

Model Packaging

Finally, the models were saved into a singular folder as .pkl files where they can be zipped for deployment. The individual sectors are identifiable by the name of the file, using the naming convention to distinguish them.

For future use, this pipeline can be used to predict the flow of migrants for future encounters through simply incorporating a new csv into the repository to create a batched prediction set. The code has also been updated to be dynamic in the sense that it will now detect the last 6 months of the dataset, instead of being hard-coded for July 2024 as the cutoff date.

Final Thoughts

Unfortunately, by the time our modeling was complete, much of the problem we had forecasted had become obsolete due to a sharp decline in migrant encounters combined with the lag in our data. This trend became a nationwide news story, highlighting how the number of crossings along the U.S. southern border had fallen dramatically — from a peak of around 250,000 per month to approximately 8,000 — effectively reducing the scale of the problem we aimed to address.

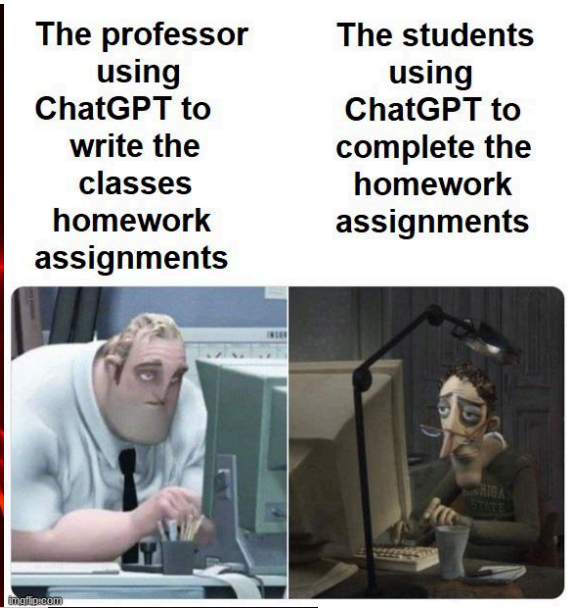
Despite the apparent resolution of the initial issue, we recognize that immigration remains a deeply polarizing and persistent topic in public discourse. Forecasting models like ours must be applied with careful oversight and a commitment to ethical standards, as they carry a real risk of being misunderstood or misused. Through this project, we've developed a deeper awareness of these challenges and the broader responsibilities tied to data-driven work. Ultimately, it is this heightened ethical perspective — the need to always consider the societal impact of our work — that stands out as the most valuable lesson we've taken from this Capstone.

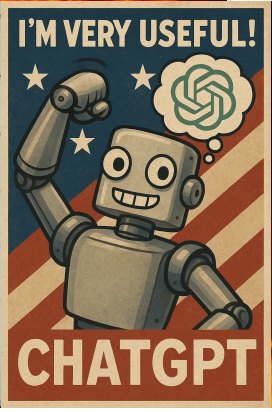
References

ChatGPT was used in the report creation and code generation.

U.S. Border Patrol: Locations & Job Opportunities in the United States, University of North Texas Libraries, 7 Nov. 2008, webarchive.library.unt.edu/eot2008/20081107215339/www.borderpatrol.gov/interactive_map.html.

Humorous References





Thanks for the great semester!