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

For this week's capstone report, our task is to examine the most relevant features of the Immigration Trends ARIMA model described in the Week 9 report. Our team will vary slightly from the others with the focal points of this report, as our model concerns time series analysis as opposed to a classification or regression problem. Protected Categories that may be incorporated in the model will also be examined, as the original dataset is separated by national origin and familial status, both protected categories in the eyes of US federal law. Additionally, model bias is covered as well as bias removal and stakeholder risks.

Feature Importance

For the feature importance of the model, this is relatively straightforward because we are not working with a machine learning model here such as a Neural Network, rather a mathematical model, ARIMA. As a quick recap, the formula for ARIMA is as follows:

$$\Delta^d y_t = c + \sum_{i=1}^{p} \varphi_i \Delta^d y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Where:

- y_t is the value at time t
- Δ^d denotes differencing applied d times to remove trend
- φ_i are the autoregressive (AR) coefficients
- θ_i are the moving average (MA) coefficients
- ε_t is white noise at time t

In this formulation:

- The autoregressive terms (AR, governed by p) capture the importance of prior observations.
- The moving average terms (MA, governed by q) capture the impact of prior forecast errors.
- The integrated term (I, governed by d) adjusts the series to stationarity by differencing.

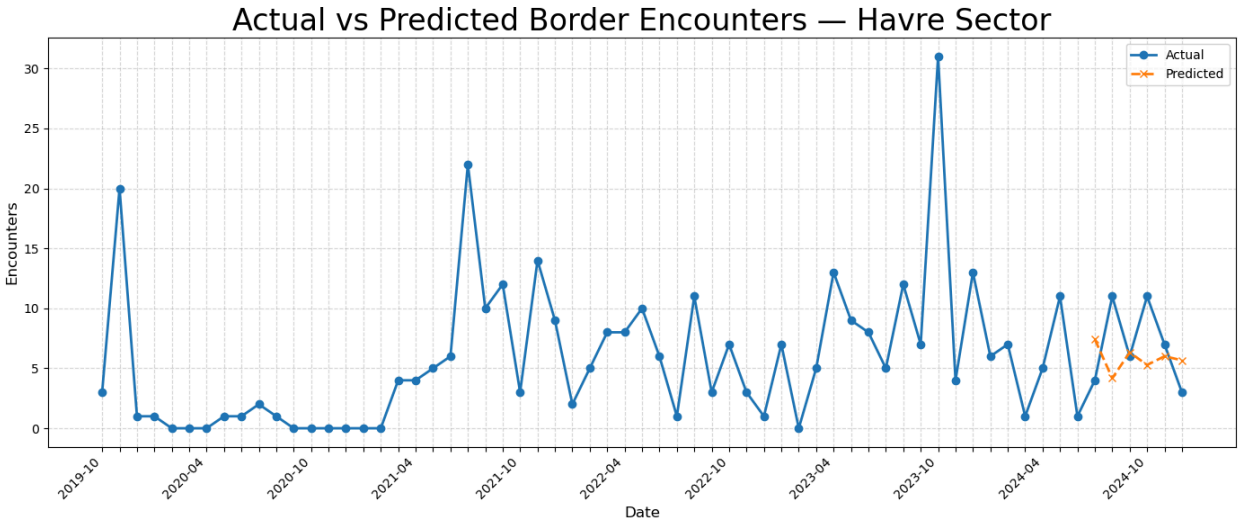
So unlike in machine learning, where feature importance is often derived from model weights or decision paths, in ARIMA the "features" are essentially the lags of the target variable and its past errors, and their importance is directly reflected by the magnitude and statistical significance of the coefficients φ_i and θ□. As for how this answers the question about the most important features of the model, long story short, the most important features for each of the 41 models created are the most recent observations.

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

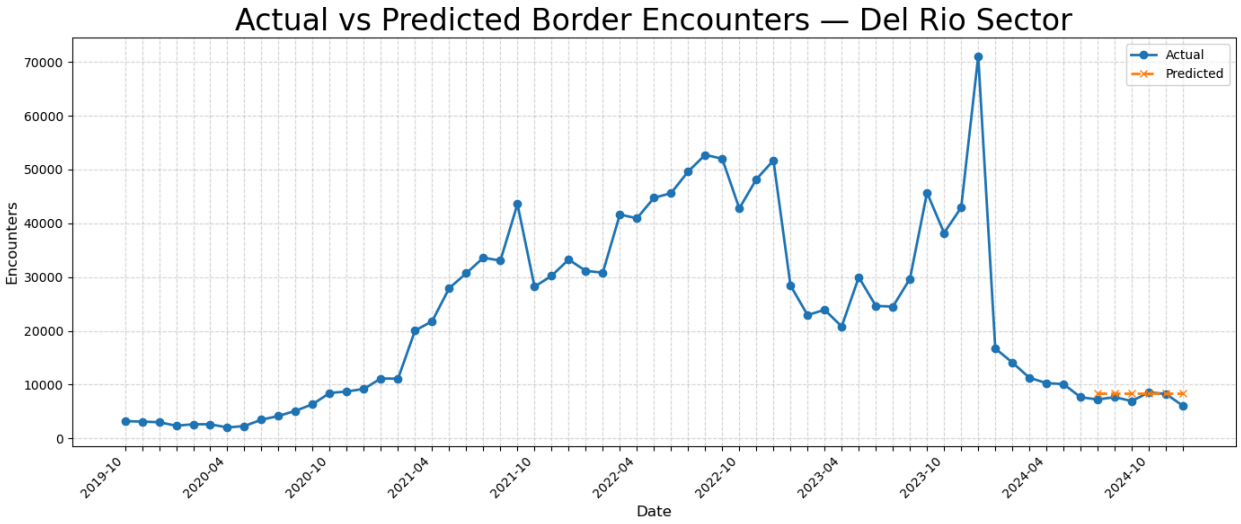
How recent the observations would be for their relevance, though, would be dependent upon the maximum of the p or q values regarding either the autoregressive or the moving average terms. A higher value of either of those two hyperparameters would imply a longer historical window for the values or errors.

To satisfy the assignment specification of selecting five predictions at random and altering them to produce a different model output, it would be straightforward to modify any of the observed data points within each sector for the period of October to December 2023 — as these directly influence the forecast values.

However, altering any of these observations would impact the model's structure by changing the fitted values associated with the autoregressive (φ_i) and moving average (θ□) components. In other words, even a small adjustment to the input data can propagate through the ARIMA model and lead to recalibrated coefficients, which in turn affect the predicted outcomes. Because each individual sector was tailored such that the (p,d,q) value pairings were optimized for performance, the number of months that would be immediately impacting the prediction would vary from model to model.



As for how the model would have changed with an impact upon the points observed for the Havre Sector, a spike in the encounters would have certainly driven the predicted values upwards because of the autoregressive component of the model to deal with trend. However, with such a small number of encounters in this sector, another example will be provided.



As for what would push the Del Rio Sector one way or the other, it would push the predictions the same because of the autoregressive term being a function of the previous observation. It would depend upon the error of the model on the past term, though for how measured the response would be as it does not want to overreact if it was previously forecasting well.

Protected Categories

The final iteration of the model does not include protected categories as input features, although variables such as country of origin and familial status were present in the original dataset. These attributes were used during exploratory data analysis (EDA) to better understand the dynamics and patterns of migration flows. However, by the time the model was constructed, the data had been aggregated, and such individual-level features were no longer used in model training or forecasting.

Importantly, the use of these protected categories during EDA was legally and ethically appropriate, as the analysis focused on population-level trends rather than individual-level decision-making. In fact, it would also be both lawful and ethically justifiable to use disaggregated attributes—such as predicting the number of migrants arriving by family status—in order to help U.S. Customs and Border Protection (USCBP) allocate resources more effectively. For example, forecasting the number of family units encountered at the border would help ensure that adequate food, shelter, and medical resources are available to meet the specific needs of children and parents. Predicting the number of migrants arriving by country of origin

could also provide USCBP the foresight to have translators or fluent speakers ready on standby, helping to ease the stress of migrants coming to a foreign country facing a language barrier. This would help ease tensions between officials, officers, and migrants as a sense of familiarity and cooperation could speed up processing times, thus cutting down on the time that migrants need to spend in makeshift camps awaiting their appointment.

As for broader moral questions surrounding the legal treatment of migrants—such as the disproportionate rejection of claims from Mexican nationals under Title 42—these are beyond the scope of the model. Title 42 is no longer in effect, and decisions about asylum hearings fall under the jurisdiction of policymakers and immigration courts. Our objective is strictly operational: to ensure that frontline USCBP personnel are equipped with the information and resources necessary to treat all individuals humanely and fairly upon arrival.

Model Bias

The major point where bias could be found in our model is a bias within the data that we have known of repeated encounters, especially during the peak years such that individuals attempted several times to cross the border. However, because the policy then was such that USCBP agents were to just tell the individuals to turn around instead of facing other consequences, it commonly resulted in the same individuals just trying again. This likely led to some inflation in the actual number of migrants, which may have introduced some inaccurate spikes into our training data. This problem would probably be more relevant for our prior models such as the neural network and LSTM, which hold onto more long-term patterns in its predictions. ARIMA and ARIMAX are a bit more hardy to these spikes as it happened over 2 years ago (or 24 time steps in months) as there is much more emphasis placed on recent data. As for how we can remedy this issue, the truth is we can't. USCBP did their jobs well in these instances where the individual was apprehended again after they tried to cross, which is what needed to be done. A change to this would require a change in policy such as the one occurring right now where migrants are detained and deported.

A second point of emphasis which needs to be noted are the undocumented migrants who were not apprehended by USCBP. It would need to be added in using a sector-by-sector basis where

the number of migrants is increased proportionally based upon further research. Redetermining the number of total migrants as opposed to the number of encountered migrants would make a huge difference in the USCBP ability to encounter and apprehend migrants avoiding detection.

Bias Removal Strategies

As for removing the bias to become the total number of crossings as opposed to the number of encounters, we would find the estimated proportion of migrants missed by sector and add that into the training data to reflect how many total migrants to expect within the test data. It is admittedly a very naive and straightforward approach, but because there is not a truly known number of undocumented immigrants, the best thing to do is estimate. As for the implementation in code, it would be relatively straightforward to just multiply the original pre-scaled numbers by the proportion of increase then continue with the model building process.

Model Stakeholder Risks

Identifying the stakeholders, here they include USCBP agents, political leaders, migrants, and US Citizens. Let’s go deeper into each group at a time.

First, from the perspective of USCBP agents, it's essential that they feel adequately prepared for the demanding and often high-risk nature of their work. Access to accurate and timely information about who is attempting to enter the country—such as expected demographics, locations, and timing—can enhance the safety of both agents and migrants. Just as migrants are seeking the opportunity to plead their case for asylum, agents want to return home safely at the end of the day. The more informed and prepared they are, the greater the likelihood of avoiding conflict, minimizing risk, and ensuring that encounters are handled with clarity and professionalism. However, it is important to highlight potential risks and limitations that our final product may introduce to these individuals: one, that misinterpretation and misuse of the model can lead to poor resource allocation and two, our forecast model is not well-designed to handle sudden changes in migration patterns from policy changes, natural disasters, and geopolitical events. It’s important that forecasting outputs are treated strictly as probabilistic, not

deterministic, such that overconfidence in the model's outputs may put USCBP agents and officials at risk if they take the predictions completely at face value.

Next, from the standpoint of political leadership, immigration policy represents both a logistical challenge and a political flashpoint. Leaders seek to avoid the kind of public backlash that emerged during the Trump administration, when inadequate resource allocation and controversial enforcement tactics—such as family separation and the detention of children—sparked national and international criticism. These policies became a focal point of public discourse and were widely cited as contributing factors in the political shift that occurred in the 2020 election.

However, the political dynamic shifted again by the 2024 election cycle. This time, criticism was directed at the Biden administration, with opponents arguing that immigration enforcement had become too lenient. The narrative focused on the perception that border controls were insufficient to manage the volume of migration, leading to public safety concerns. One such flashpoint was the tragic murder of Laken Riley in Georgia, which was cited as emblematic of broader concerns around migrants who had allegedly evaded detection or processing by USCBP. While such incidents are complex and require careful contextualization, they underscore how immigration can quickly become a symbolic issue with far-reaching political consequences.

For policymakers, then, the goal is not only to manage immigration effectively but also to demonstrate that the system is both humane and under control—balancing the need for compassion with the imperatives of security, order, and public trust. Still, it remains critical to point out the risks of using forecasting models to inform policy-making. For example, taking enforcement and political action in advance based on predictions from a model can have unintended consequences later down the line, as the conditions that led to these forecasts have now been shaped by the present. In some cases, the quote, "one often finds his destiny on the path he takes to avoid it" may ring true to policymakers looking to pre-emptively act decisively on mere predictions. At the same time, forecasting models can be misused for political narratives, where predictions are intentionally disguised as reality to drive home the point that immediate policy changes must be taken in order to avoid a certain outcome.

From the perspective of migrants, the stakes are often matters of life, liberty, and family unity. Many are fleeing violence, persecution, poverty, or political instability in their home countries and view the United States as a place of refuge and opportunity. For them, clear, consistent, and humane immigration policies are essential—not only for ensuring due process and the right to seek asylum, but also for preserving dignity during what is often a traumatic and uncertain journey. Predictive models that help streamline processing and resource allocation can ultimately contribute to fairer treatment and reduced wait times, ensuring that those with legitimate claims are not lost in bureaucratic bottlenecks. These positives contrast sharply with the unintended risks of such a forecasting model: for example, if high waves of migrants are forecasted, pre-emptive crackdowns may be initiated which may result in heightened tensions between agents and migrants or migrants may be further pushed to employ more desperate tactics, such as choosing more dangerous routes or receiving help from cartels. Misallocation of resources from misforecasts can force migrants to face longer detention in overcrowded holding facilities and slower processing times from understaffed centers. The very essence of our project might be turned on its head if our model is misused or inaccurate.

Finally, for U.S. citizens, immigration is often a prism through which broader concerns about national identity, economic security, and rule of law are projected. Some view immigration as a strength, vital to the country's diversity and economic growth, while others see it as a challenge to social cohesion and public safety. Regardless of viewpoint, most Americans want to see a system that functions transparently, protects borders effectively, and treats all individuals—citizens and non-citizens alike—with fairness and humanity. When predictive tools are used responsibly, they can support this vision by informing smarter policy, improving operational readiness, and building public confidence in the government's ability to manage complex issues at the border. When they aren't, and this is an unfortunate reality to consider, then the reverse of all the good we want to see may be terrifying to think about. Based on our work with sentiment analysis on tariffs, it was clear that news sources can hold significant power in determining what the public thinks about important issues. Statistics, although powerful in its applications, can be equally dangerous when used to support a destructive and dangerous narrative. It can be hard to deny the numbers, especially if it has been painted in a certain bad light. Predictions, even from accurate forecasting models, can be used to fabricate a story that

feeds into many fears of US citizens. We've mentioned before that our project doesn't solve the underlying issue that is the US immigration system. We certainly don't need more resistance against reforming a highly contentious and complicated system that requires more than just fear-driven policy dictating who can and cannot come into our country. This issue requires a great amount of thought, empathy, compassion, and willingness to try even against the odds from US citizens along with our representatives in the legislative branch.

In short, each stakeholder—agents, leaders, migrants, and citizens—has legitimate interests, and a well-designed forecasting model should aim not only to inform policy, but also to serve the broader goal of fostering a safe, just, and orderly immigration process. On the other side of the coin, it's imperative that these models are accurately reported and represented to each group aforementioned in order to ensure that models are used not as tools of fear or control, but as instruments of clarity, preparedness, and compassion. As we develop and deploy forecasting tools, we must remain vigilant in how their results are interpreted and communicated. Transparency, humility, and ethical foresight must guide our efforts—recognizing that behind every data point is a human story, and behind every prediction lies a potential policy decision that can shape lives and futures. The true power of these models lies not in their ability to predict the future with certainty, but in their potential to support more informed, humane, and effective policymaking. In a space as charged and consequential as immigration, we owe it to every stakeholder—from the agents on the ground to the migrants at the border and the citizens watching from afar—to get this right.