Analyzing and Forecasting U.S. Immigration Trends

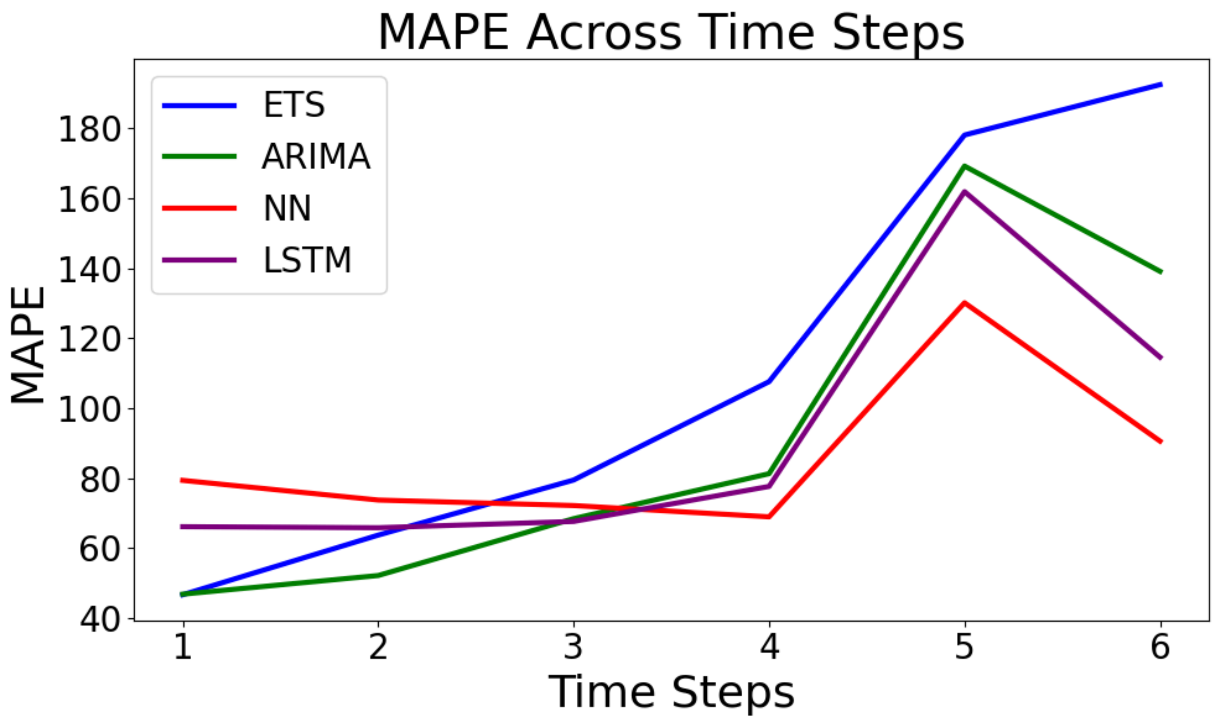
Group 7: Alan Lin , Trevor Petrin , Ganesh Kumar Ramar , Mingyu Chen

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

This week, our efforts were directed towards completing the poster for the Symposium deadline. Since our project utilized both traditional and machine learning forecasting methods, we went back and re-evaluated four types of models this week: ARIMA, ETS, Neural Networks, and LSTMs. We may pick Transformers back up in the coming weeks, as we were running on a bit of a time constraint.

One of the problems from Week 6 was the amount of time, memory, and storage space that was used to generate outputs for every single configuration and model for each Sector/Field Office. I believe there were more than 1000 files created from the Neural Network alone. As promised, I've refined a way to store training loss, validation loss, predictions, and MAPE in a dataframe that can then be used to generate as many plots as we desire. This cut down tremendously on runtime and storage issues. I also incorporated an Early Stopping method on the validation loss, with a patience of 5 epochs to speed up the training process. We had foregone this the last time around in favor of letting each model run to completion so as to prevent any premature stopping leading to suboptimal model training.

Let's put down some of the results we gathered for the Symposium poster.



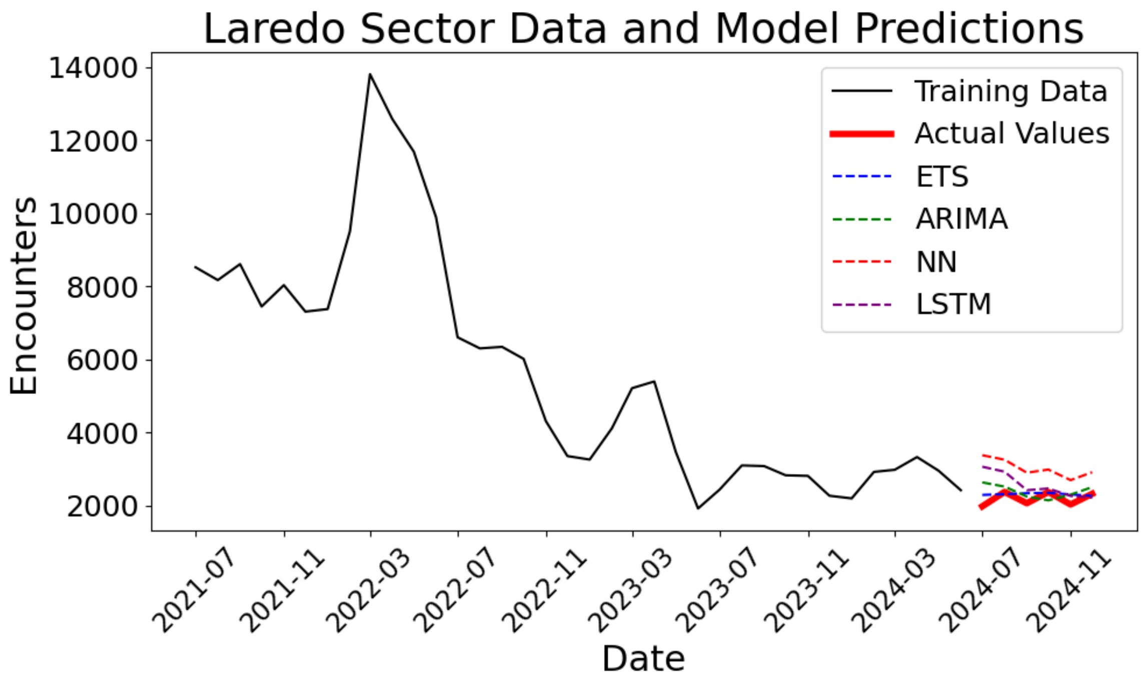
MAPE Average Results		
	2 Month	6 Month
ARIMA	**49.6%**	**92.8%**
ETS	**55.3%**	**111.4%**
LSTM	**66.0%**	**92.3%**
Neural Network	**76.6%**	**85.9%**

Instead of tracking RMSE, we opted for MAPE, or mean absolute percentage error that will give us a better metric of performance for all 41 models that we build for each configuration of hyperparameters. This is due to MAPE being scale-independent; sectors face incredibly different volumes of migrants, from below 100 each month to sometimes over 40,000. MAPE gives us a better way to gauge model performance across multiple models (which we have!). We noticed quickly that ARIMA and ETS performed much better, on average, than Neural Networks or LSTMs, especially in the first two months as noted above in the results table. Across 41 independent sector-level models, ARIMA and ETS were better at forecasting in the short term, which may be due to how predictions were actually made.

As for tuning the ARIMA and ETS, a grid search approach was taken with the parameters to evaluate their best performance upon the train set before being applied to the test set. For ETS, though, there is a problem having 0 as any of the components in a multiplicative model because the model does not allow for multiplication by 0. This became an issue because minmax scaling was used for the data, leading to an inevitable multiplication by 0 problem for the model. This was fixed by adding 1e-6 to any point which was a 0-entry, allowing the model to run smoothly through all data points. Upon completion of the training, the optimal model was used for a 6-month recursive forecast where the output of the model at time t was used as input for time t+1 to avoid data leakage. The data was then inverse-transformed using the unique minmax scaler for the model to be compared to the test data.

As for the functionality of the ARIMA and ETS models, these models recursively made predictions rather than relying upon any kind of new observations. What this means is that the model fed its output into its next step's input in order to continue making predictions past the first time step. This approach was taken to prevent data leakage from the test set into the validation set so the simpler models would not have any kind of information that they should not temporally. This aligns with the LSTM and NN approaches of making predictions 6 time steps ahead without being able to refine their predictions based upon new information.

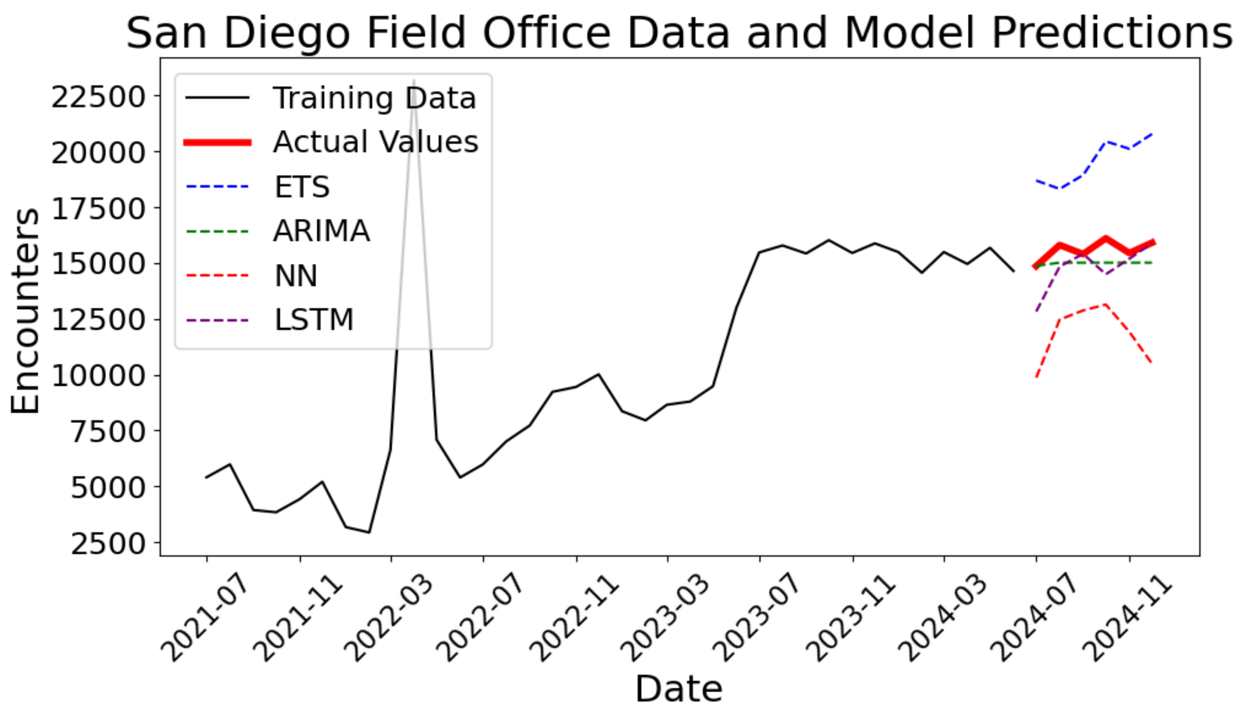
So, we've covered MAPE and how traditional forecasting and machine learning techniques can both be leveraged to predict immigration trends. But we haven't actually seen any examples, yet!



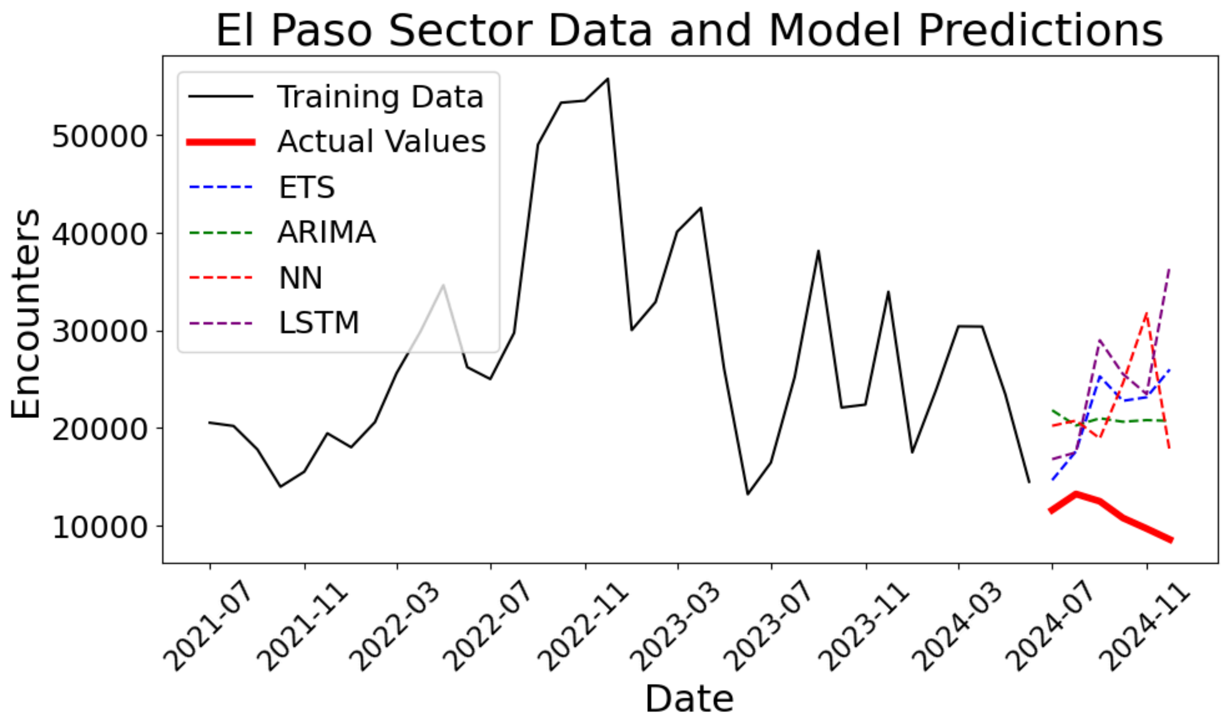
The Laredo Sector in Texas is a key bordering sector at the US-Mexico border. So, it was satisfying to see that all four models performed relatively similarly, even with average MAPE values being around or over 50%. This highlighted the viability of both approaches as tools to help policymakers forecast migration trends at the sector-level just based on monthly data alone. There is so much room for improvement that can be considered moving forward, such as trying to incorporate more features like policy changes or administration changes in key election years. Adding more context and information to what is a pretty basic dataset will be incredibly valuable.

So far, our models have been independent from one another as we didn't want to train a single model that faces sectors with wildly different minimum and maximum encounter counts. Consequently, another improvement I want to take into consideration would be looking into surrounding/neighboring sectors and field offices. Migration trends are generally fluid, where certain sectors may ebb and flow together or even inversely. I believe it may improve predictive accuracy to investigate any correlations and perhaps incorporate features or data from significantly correlated sectors to provide more context to each model. Correlation analysis may help uncover any historical relationships between sectors. For example, if the El Paso Sector in New Mexico is negatively correlated with the Tucson Sector right next door in Arizona, then when we forecast for El Paso, we'll need to dampen or increase the forecast according to the trends in Tucson.

It should be noted for the dataset that the election of current President Trump, who has had a hardline anti-immigration stance in both rhetoric and policy, was elected in November 2024. This event is sure to have an impact upon the number of individuals who are considering applying for asylum in the United States because of the previous policies of family separation, deportations, and dangerous water barricades along the border. This would likely dissuade individuals from making the journey to the United States, hence, a decrease in numbers can be expected either during the transition period from President Biden to President Trump, or after Trump's inauguration when these old policies are likely to be reinstated.

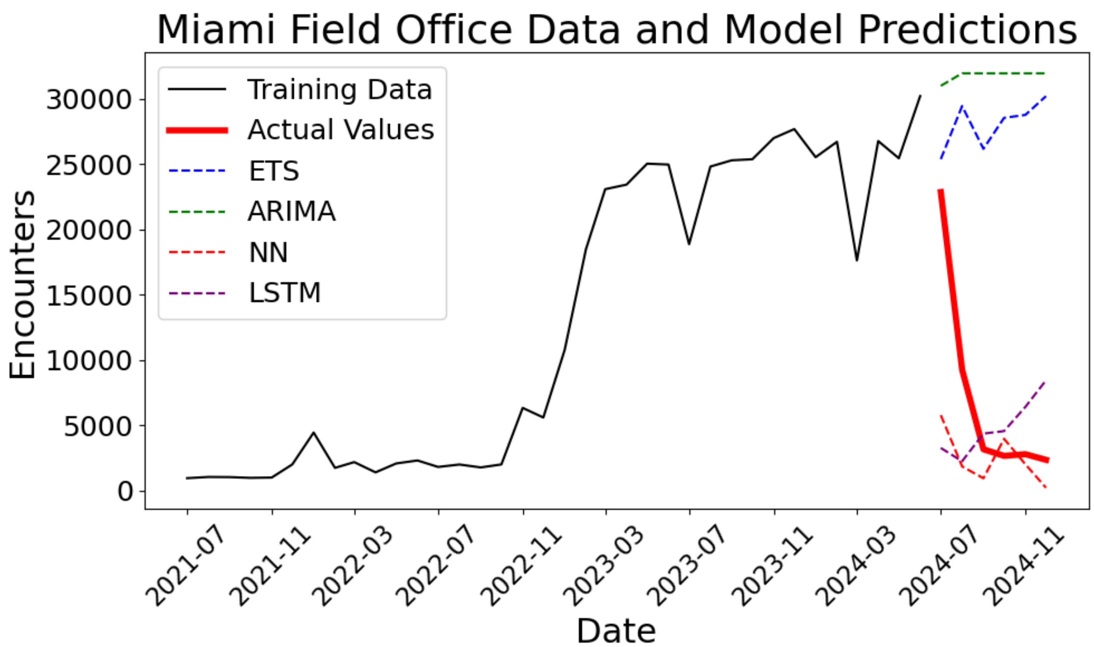


San Diego is an example of a sector where ARIMA adopted a Naive forecasting approach and was still very accurate. LSTM also performed very well here, though it started too low on the number of encounters.

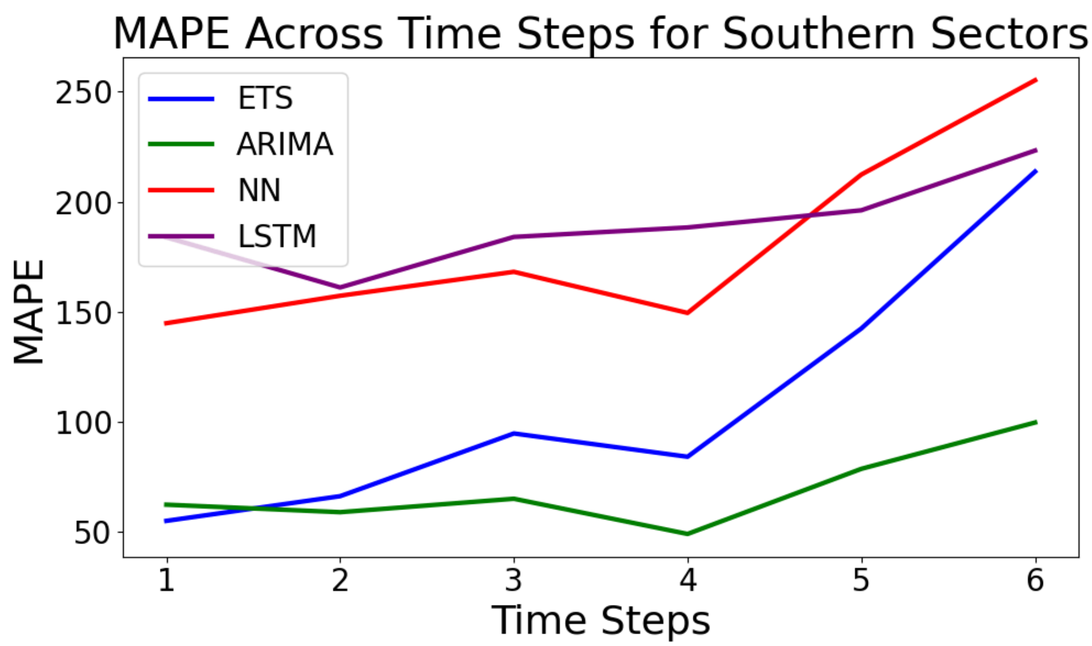


In terms of sheer volume, El Paso Sector had one of the highest rates of encountering migrants because of its southern geographical location and also was where quite a bit of the error in the predictions stemmed from. This forecasting well shows the potential problem with overprojecting the number of encounters at that time period where the funds and resources would be allocated for that sector where they should have been at a sector where the funds were underallocated. This would create a humanitarian issue with the housing facilities being overcrowded and the USCBP officials being stretched thin on personnel as seen during the peaks of encounters from 2021-2022.

Most of the models, though, tended to increase in error as time progressed with the average MAPE increasing for every model between the 2 and 6 month means. This occurrence here can be seen in particular with the LSTM and ETS models as they further miss the trend of the actual data which continues on its slope from August numbers, which cannot be accounted for because it is unobservable from the train data. Because of the more basic nature of the models for ETS and ARIMA, it seems like they, for the Southern Border, are able to better position themselves from their mathematical components to start. This more sure nature of the models leads to the lower error terms near the beginning of the test set though they cannot always capture the correct trend because it changes midway through the test set.



However, it was not a one-size-fits-all solution for the model application, there were some situations where the more complex models marginally outperformed the more traditional methods in areas such as the Miami Field Office. ETS and ARIMA were not able to pick up on the drop of over 20,000 migrants, leading to roughly a 1000% error rate in this sector for both of those models. The Neural Network and LSTM models, though, were able to predict the massive drop in encounters at that region.



As for the methods' predictive capabilities on the Southern Border Sectors, the Neural Network and LSTM were completely put to shame by the simpler ETS and ARIMA models, both having over 150% error most of the time. ETS and ARIMA, though, were able to stay within 100% error for the first 4 months each and within 65% error for the first 2 months. However, because of the vast differences in performance in the overall MAPE performance versus that of the southern MAPE performance, this implies that the Neural Network and LSTM approaches were able to outperform the ARIMA and LSTM Models