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

For Week 5, we focused on data augmentation, dimensionality reduction, and feature engineering of the dataset but also went more into detail with creating basic models for forecasting immigration by US Customs and Border Patrol (USCBP) Sector. Working ahead was done to stay consistent with the poster deadlines for the MSAA/MSAE Symposium initial poster draft submission due 2/20/2025. Initial models created included ARIMA, Neural Network, LSTM, and Transformer models with the best performer being the Neural Network. Hyperparameter tuning has yet to be performed to optimize each of the models, but the base models will allow for this to be implemented more quickly within the upcoming weeks.

For the data aggregation and feature reduction, to quickly recall the operations performed last week, the data needed to be aggregated by Border Sector to make the predictions less granular than predicting by every individual section of the population at each Border Sector individually. To accomplish this, the pivot function was used to disregard features irrelevant for building a time series model such as the migrant's country of origin, their family status, or under which title of the law they were encountered under. Other redundant features such as the individual year, month, and abbreviated Area of Responsibility (AOR) columns were also dropped. The dimensionality of the data transformed from (57344, 13) to (41, 39) for the training data set, and was done so to better format the data for input into a time series model such that every row had the time step in each column. This resulted in a dataframe containing panel data.

	Component	Land Border Region	Area of Responsibility	Demographic	Citizenship
0	Office of Field Operations	Northern Land Border	Boston Field Office	FMUA	BRAZIL
1	Office of Field Operations	Northern Land Border	Boston Field Office	FMUA	OTHER
2	Office of Field Operations	Northern Land Border	Boston Field Office	Single Adults	BRAZIL
3	Office of Field Operations	Northern Land Border	Boston Field Office	Single Adults	CANADA
4	Office of Field Operations	Northern Land Border	Boston Field Office	Single Adults	CHINA, PEOPLES REPUBLIC OF

This initial dataframe was then converted to a reshaped array with dimensions (41, 39, 1) for input to the model using TensorFlow for the Machine Learning models.

Area of Responsibility ···	# 2024-01-01	# 2024-02-01	# 2024-03-01
Atlanta Field Office	1073	954	1054
Baltimore Field Office	1565	1586	1440
Big Bend Sector	324	568	436
Blaine Sector	223	235	226
Boston Field Office	4508	3611	4736

Now where feature engineering comes into this process, as detailed in our Week 4 notebook, we created a dataframe representing the proportions of migrants whose country of citizenship was Mexico. We chose to highlight Mexico because the majority of total U.S. migrants come from there and the relevance of recent political rhetoric from US politicians. Creating a parallel time series with this proportion at the same time steps as the overall aggregated Sectors will give a broader insight into how to make these numbers more manageable for USCBP. Additionally, a variable for the overall Unaccompanied Minor demographic was created with the same methodology behind the proportion of individuals from Mexico. This will add value in trying to allocate resources properly for the Border Sectors, as this specific demographic needs to be dealt with using utmost caution as to not ignite a political or humanitarian crisis. In April 2024, Judge Dolly M. Gee of the United States District Court of Central California ruled that the federal government must provide shelter to migrant children. The ruling requires the Department of Homeland Security (DHS) and CBP to process children quickly, place them in safe facilities, and stop directing minors to open-air sites/migrant camps situated right at the border. It reaffirmed that children are entitled to safe and sanitary facilities where they will be given adequate food, water, shelter and medical care. Although this has led some families to disband before approaching the border to take advantage of this ruling, it has been a major victory for humanitarians and other advocates of children's rights. Still, it's a shame that it required a federal

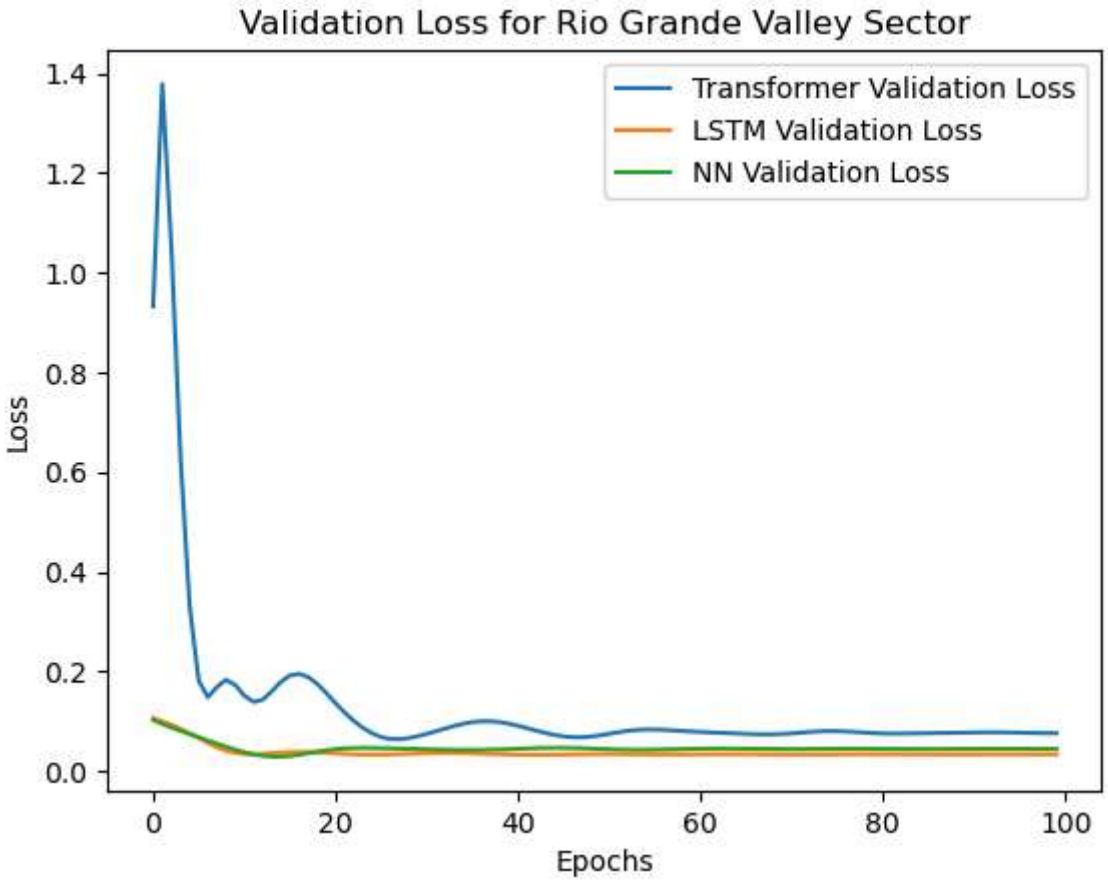
court to direct the government to do what basic human decency and the law clearly require. USCBP does not need a media firestorm nor a disruption in their operations because they are unable to adequately deal with unaccompanied minors, and it will also have a negative impact in public perception for other related organizations such as DHS and ICE.

As for the implementation of the new features, the tensor will eventually be updated such that it becomes a (41, 39, 3) tensor for model input, accounting for the new features. This will be implemented in the dataframe by having an array of values at every time series entry and converting into a Tensor accordingly.

Concerning the model building, we shall touch on it here briefly, but will leave most of the details of the implementation and hyperparameter tuning to future weeks. The models considered this week included ARIMA, Neural Network, LSTM, and Transformer models. The only principle that was changed of this building process from the previous weeks' plan was that the validation set was changed to be the last 12 steps of the training data to reflect the 12 months of the test step. However, going forward, this will need to be changed, and the results will reflect the reason as to why very well.

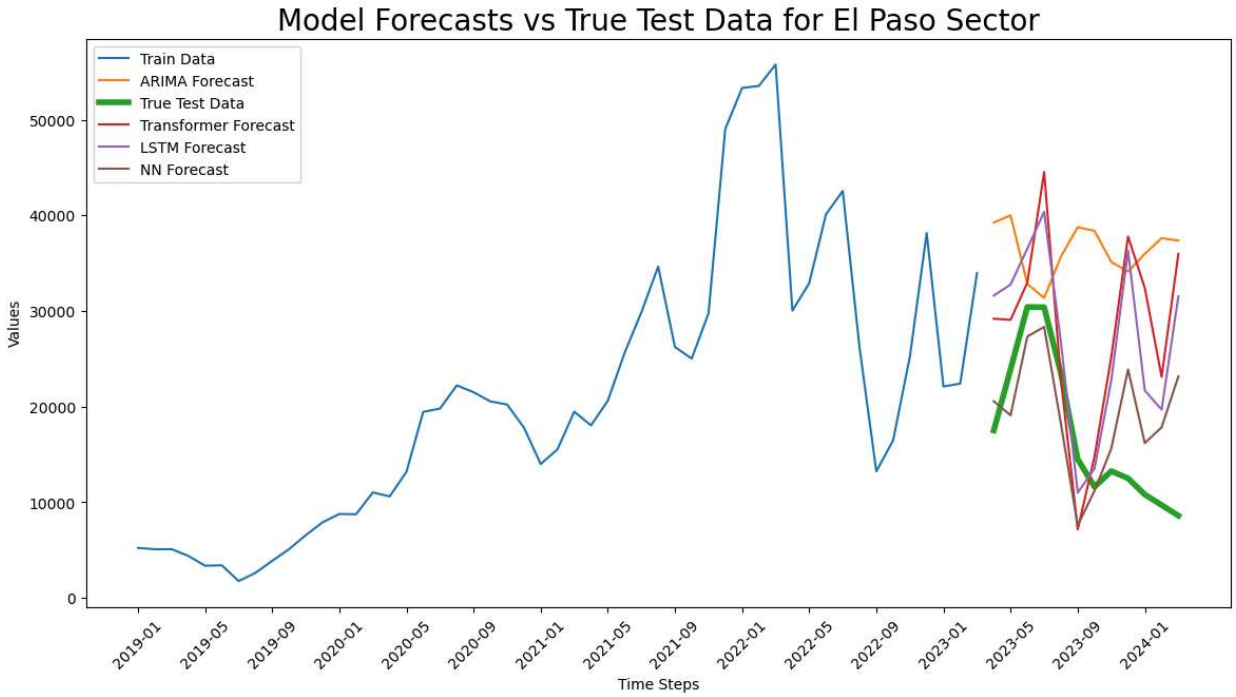
Model Architecture		
Neural Network	LSTM	Transformer
- 128 Nodes - Dropout 0.1 - 3 Layers	- 64 Nodes - Dropout 0.1 - 3 Layers	- 8 heads - Dropout 0.1 - 3 Layers

For the validation loss, the Rio Grande Valley Sector was used to determine the number of epochs to be used because it is one of the busier, more important sectors.



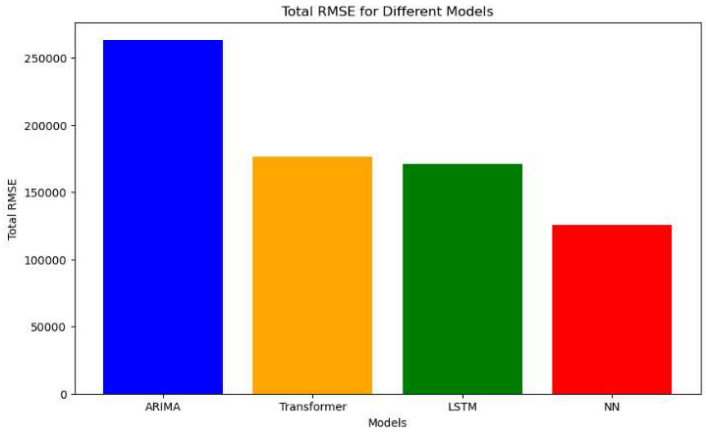
The ‘bend in knee’ principle for choosing the number of epochs to train the models on suggested that roughly 10 iterations be used for the LSTM and NN models. For the Transformer, though it was lossier to begin with, the validation loss of the Rio Grande Sector suggested it takes roughly 20 epochs to find proper values. Alpha for the models was not tuned here, kept at the default for all models while using adam as the optimizer and measuring loss using Mean Squared Error. Each model was trained and used to predict on each sector individually, as it would not be reasonable to train the model on scaled data from the Northern and Southern Borders with their different trends.

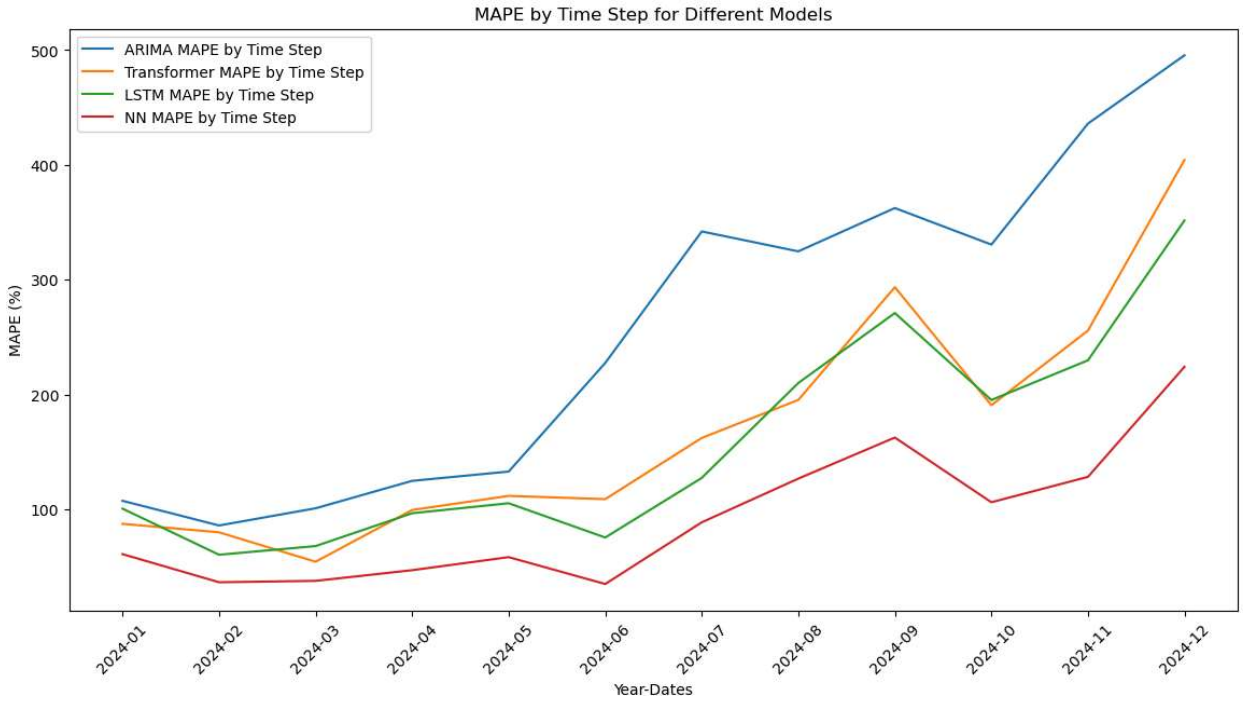
As for the ARIMA model, the StatsForecast package was used separately on each Sector, as the package will automatically tune the sector for the best fit.



The example plot shows the models' performance on the El Paso Sector of the Southern Border, one of the busiest Sectors overall. It is clear that the Neural Network was the best performer of the bunch with ARIMA not picking up on the trends at all. Towards the end of the test set, though, there is a clear divergence of all of the models from the true data. An educated guess why this occurs would be the model overfits on the validation set with the patterns in the last 6 months of the predictions mirroring too closely to what is happening in the last 6 months of the validation. We plan to address this issue by using the cross-validation performed in the Hyndman book, mentioned in the Week 3 report, which will provide iterative feedback for the model as opposed to the hard and fast 12 points in 2023.

Overall, the best model before serious hyperparameter tuning was the Neural Network based upon Root Mean Squared Error, marginally outperforming the ARIMA, Transformer, and LSTM when summed across all models.





Using Mean Absolute Percentage Error will provide a better context for the model's true performance. Overall, for the first 6 months of the data, the Neural Network had 3 months below a MAPE of 40%, and 4 months below 50%, which though not perfect, is impressive considering the overall plummet in migration beginning January 2024. As the time series goes on, though, there is a significant increase in error in the back 6 months of the data. This issue likely stems from not yet performing cross validation upon the models as was mentioned in the previous section and will be addressed going forward.

Though there is still much to be done with regards to redoing the validation and implementation of hyperparameter tuning, much progress was made this week for having the infrastructure available to more easily and readily create the models. The first 6 months of predictions using the Neural Networks were also extremely encouraging, especially with such solid performance without any kind of tuning. These results leave the team in a position to improve upon them in the coming weeks.