Group Members: Anita Gee, Siwei Guo, Yingzhu Chen, Nemo (Sorachat) Chavalvechakul

Group Name: Mental Health Warriors

Feedback

We did our feedback on AA-Capstone, https://github.com/alanklin/AA-Capstone

Part One: Report Feedback

Week 1 Report Feedback (week1_report.pdf)

- Problem framing is timely and well contextualized, with strong connections drawn between policy events and the proposed forecasting task; tightening the narrative by reducing redundancy would improve focus.
- Data sourcing and structuring are clearly explained, with thoughtful justification for using Area of Responsibility as the modeling unit; briefly outlining intended modeling approaches would better link the data to the forecasting objective.
- Stakeholder impact is effectively quantified, especially through the cost-saving estimate tied to better resource allocation; including preliminary visualizations or trend plots would enhance the communication of key insights.

Week 2 Report Feedback (week2_report.pdf)

- Exploratory analysis is thorough and thoughtful, with strong use of contextual explanations for observed trends; a few summary visuals (e.g., time series plots or stacked bar charts) would quickly highlight the key patterns.
- Variable definitions and data distinctions are clearly documented, especially the comparison between state-level and sector-level data; briefly mentioning which variables are likely model inputs would improve modeling readiness.
- Demographic and geographic insights are valuable, particularly the regional concentration of encounters and discussion of vulnerable groups.

Week 3 Report Feedback (week3_report.pdf)

- Preprocessing steps are clearly explained, especially around filtering and shaping the dataset; summarizing steps in a table could make the process easier to follow.

- The train/test split strategy is appropriate, with good reasoning behind the validation logic; mentioning how seasonality is handled would strengthen the design.
- Exploratory graphs are helpful, particularly in showing data imbalance across sectors; a brief mention of how this impacts model training would be useful.

Week 4 Report Feedback (week4_report.pdf)

- Feature engineering choices are well-justified, especially focusing on Mexico-origin encounters and unaccompanied minors; highlighting how these features will be used in modeling could clarify their role.
- Data structuring is thoughtful, with separate models per AOR to respect regional differences; summarizing this logic in a short diagram or table would aid understanding.
- Scaler selection is carefully reasoned, correctly avoiding global scaling; noting how the scalers will be reused in deployment is a good forward-looking detail.

Week 5 Report Feedback (week5_report.pdf)

- Exploratory data analysis is comprehensive, featuring clear visualizations of variable distributions and missing‑value patterns that establish a solid foundation.
- Discussion of imputation strategies is well articulated; including a comparison table of model performance before and after each method would further strengthen the argument.
- Introduction of engineered features such as text length and sentiment scores is effective; adding a correlation heatmap would quickly illustrate which features most influence the target variable.

Week 6 Report Feedback (week6_report.pdf)

- Implementation of time-series windowing for cross-validation demonstrates solid understanding of Hyndman's method and addresses prior overfitting concerns with clear rationale.
- Justification of hyperparameter choices (learning rate, window size, dropout) effectively ties neural network behavior to forecasting performance, enhancing interpretability of tuning efforts.
- Discussion of sector‑specific model training highlights awareness of regional variance, though inclusion of a summary table comparing key MAPE results across configurations would further strengthen insight into model selection.

Week 7 Report Feedback (week7_report.pdf)

- The explanation of time series validation and windowing is strong and differentiates well from classification tasks.
- Integration of early stopping with a patience of five epochs demonstrates thoughtful control over training time and model convergence.
- Justification for switching to MAPE as a scale-independent metric across 41 sector-level models provides clear rationale and enhances comparability of forecasting results.

Week 8 Report Feedback (week8_report.pdf)

- Thorough justification for choosing XGBoost, with clear discussion of its advantages for sparse, imbalanced text data and regularization benefits.
- Detailed comparison of hyperparameter variants (Conservative, Faster, Fastest) effectively illustrates the bias–variance trade-off, though including a summary table of key metrics would improve clarity.
- Insightful class-wise error analysis highlights minority-class challenges, but adding visualizations could better support recommendations for further data-centric interventions.

Week 9 Report Feedback (week9_report.pdf)

- The explanation of time series validation and windowing is strong and differentiates well from classification tasks.
- The justification for ARIMA's short-term superiority is clear, though a summary graph comparing MAPE by month would help.
- The discussion on model selection and bias-variance tradeoff adds valuable insight to why simpler models like ARIMA were chosen.

Week 10 Report Feedback (week10_report.pdf)

- Transitioning from ARIMA to ARIMAX is well-executed and appropriately justified given the policy changes affecting immigration.
- Short-term MAPE improvement from 49% to 44% is meaningful, but performance decay at month 6 raises questions on exogenous variable effect longevity.
- The inclusion of real-world commentary from non-profit involvement adds ethical perspective and deepens understanding.

Week 11 Report Feedback (week11_report.pdf)

- The ARIMA equation explanation is helpful and supports the concept that feature importance is based on autoregressive structure.
- The section on stakeholder risk is thoughtful and well-balanced, covering CBP, migrants, and political implications.
- The reflection on model misuse and policy risks adds important cautionary insights for how predictions could be interpreted.

Week 12 Report Feedback (week12_report.pdf)

- The packaging of models by sector using .pkl files is efficient and scalable.
- MAPE thresholds for model monitoring are clear; however, MLFlow implementation details could be expanded for future work.
- The retraining strategy is reasonable—monthly retraining or trigger-based based on concept drift is practical and industry-aligned.

Week 13 Report Feedback

- A significant, unforeseen political event—Donald Trump's re-election and subsequent policy rhetoric—caused a sudden and unpredicted decline in migrant encounters, leading to high prediction errors such as high MAPE. While ARIMAX models were considered to account for such external factors, ARIMA was ultimately used. Future work could focus on models better equipped to handle abrupt socio-political shifts.
- The project employed non-traditional validation methods, using recursive predictions on the full training set for ARIMA and ETS models, and windowed validation during training for Neural Network and LSTM models, rather than a standard holdout split. Exploring or comparing performance against a dedicated validation set could offer further insights into the models' generalizability and robustness.
- The real-world problem being forecasted (a high influx of migrants) significantly changed and became less relevant by the time the modeling was complete due to a sharp decline influenced by events the model missed. This highlights the difficulty of applying static forecasting models to highly dynamic situations. Future work might explore faster model adaptation or the incorporation of leading indicators to better anticipate shifts in trends.

PART 2: Notebook Feedback (Week 9–12)

Reviewer:

Ying (Week 1-4 Notebooks)

Siwei (Week 5–8 Notebooks)

Nemo (Week 9–12 Notebooks – feedback focuses on new contributions only)

Anita（Week 13 Notebook）

Week 1 Notebook Feedback

- Week one does not have a notebook. Week one is brainstorming for the topic and ideas.

Week 2 Notebook Feedback (2 notebooks)

- The code is cleanly written and well-structured, with logical cell organization and clear markdown comments that make the EDA easy to follow.
- Visualizations effectively highlight key trends, such as encounter distributions across time, geography, and demographics; consider adding titles and axis labels to all plots for better interpretability.
- State vs. Sector comparison is thoughtfully explored, showing clear understanding of use cases for each; summarizing this comparison in a table or chart could enhance clarity.
- Filtering and grouping logic is robust, ensuring relevant segments (e.g., Title 8 vs. Title 42) are well-separated; adding summary statistics (mean, median) would further support interpretation.
- Notebook reflects strong awareness of modeling implications, particularly regarding seasonality and regional variation; a brief mention of which features may carry forward into model training would be helpful.

Week 3 Notebook Feedback (2 notebooks)

- EDA continues to be thorough and purposeful, with a good focus on encounter patterns and region-specific dynamics; plots support conclusions effectively.

- Sector-level notebook (week3_eda.ipynb) emphasizes encounter spikes and variability well; consider summarizing key takeaways at the end of the notebook for clarity.
- State-level notebook (week3_eda_state.ipynb) correctly identifies overlapping data and provides useful insight into limitations of using state-based aggregation.
- The use of pivot tables and time-based groupings is appropriate, as they prepare the data well for modeling. Adding a few seasonality-focused plots could enhance forecasting readiness.
- Code readability is strong overall, with clean logic and variable naming; minor improvement could be made by consolidating repetitive groupby operations into reusable functions.

Week 4 Notebook Feedback

- Data preparation is well-structured, including clear logic for aggregation, pivoting, and handling missing sector-month combinations.
- Feature engineering is thoughtful, especially the focus on high-impact variables like unaccompanied minors and Mexico-origin encounters.
- High model readiness, with attention to LSTM-specific tensor shaping and appropriate use of individual MinMaxScalers per sector.

Week 5 Notebook Feedback

- Loading and applying pre-trained MinMaxScalers row-wise creates a clear, reproducible preprocessing pipeline, and printing scaler parameters verifies correct setup.
- Wrapping model compilation and @tf.function definitions inside a loop leads to repeated retracing warnings; extracting model-building and predict functions outside the loop would improve efficiency.
- Integrating ARIMA forecasts alongside deep-learning models is a strong comparison, but hard-coded ARIMA order and duplicated loops could be refactored into reusable functions for better modularity.

Week 6 Notebook Feedback

- Automated generation and logging of 27 hyperparameter configurations provides a systematic framework for thorough tuning and reproducibility.
- Clear separation of data scaling, windowing, and model definition enhances readability and makes each pipeline stage easy to debug.

- Consistent plotting of training/validation losses and predictions offers immediate visual insight into model performance and convergence.

Week 7 Notebook Feedback

- The scripts cleanly ingest ETS, ARIMA, NN, and LSTM results into a unified workflow, using clear, descriptive variable names and centralized file‑path handling to streamline comparative evaluation.
- Visualization routines thoughtfully output per‑time‑step MAPE plots formatted for GraphPad Prism export, demonstrating attention to high‑quality, publication‑ready presentation.
- Modular notebook structure—with distinct sections for data loading, model inference, and plotting—paired with concise comments, makes the codebase easy to navigate, maintain, and extend.

Week 8 Notebook Feedback

- Automated ARIMA order tuning via AIC-based grid search ensures each sector's model is optimally selected, demonstrating a rigorous approach to model selection.
- Clear separation of data loading, scaling, train–test splitting, model fitting, and forecasting loops enhances readability and maintainability of the code.
- Visual comparisons of forecasts against actual values for each sector offer immediate, intuitive validation of model performance and facilitate rapid error analysis.

Week 9 Notebook Feedback

- Added time-series validation logic using windowed datasets — effective adaptation for temporal structure.
- Separate model evaluation per sector is smart for managing regional variance, though computationally intensive.
- Use of loss curves for tuning NN hyperparameters adds transparency and supports model selection justification.

Week 10 Notebook Feedback

- ARIMAX integration is a great enhancement — encoding political power shift using binary variables is simple yet effective.

- Clearly labeled input transformations and updated model architecture make replication easy.
- Results comparison between ARIMA vs ARIMAX is helpful; however, could include sector-wise visualizations for more clarity.

Week 11 Notebook Feedback

- Clear explanation of how modifying observations (October–December 2023) affects ARIMA predictions.
- Demonstrated awareness of how recent data has greater weight due to autoregressive lag structure.
- Ethical decisions to exclude protected features from training are implemented in code as described in the report.

Week 12 Notebook Feedback

- Model packaging using sector-based .pkl files is organized and ready for deployment.
- Model monitoring logic using MAPE threshold bands (green/yellow/red) is well-coded and extensible.
- Includes concept drift and data drift detection setup, though external data pipeline/API integration is noted as future work.

Week 13 Notebook Feedback (Anita)

- The project followed a complete model lifecycle: data import/cleaning, splitting, preparation (aggregation, pivoting, scaling), training individual ARIMA models per time series, forecasting, MAPE evaluation, visualization, and saving models/scalers.
- Data cleaning includes a hardcoded "2025 (FYTD)" value requiring manual updates for future use, which is a potential maintenance issue. A more dynamic approach to determine the current fiscal year would improve robustness.
- ARIMA parameter selection used AIC on training data. While MAPE was used for test evaluation, the parameter search didn't directly optimize for forecasting performance on unseen data. Optimizing parameters based on validation set errors or using time series cross-validation could improve generalization.

Part 3: Code OutPut

In the provided repository, there were no code ouputs/pdfs for us to take a look at.

Part 4: Feedback on Github Repository

- The Github Repository did not contain any pdfs of the code outputs of the notebooks, when this was a requirement for each week. Perhaps the team deleted these pdfs in their last commit or they are hard to find.
- This README clearly outlines the model building process within this standalone folder, covering data cleaning, scaling, ARIMA modeling, prediction, evaluation, and saving. It explains the Model Outputs folder with its full and zoomed-in time series plots for performance monitoring. The saving of individual ARIMA models as pickle files for independent tuning is well-documented. The readme file effectively provides project context, including the goal, data source, and intended use by CBP, and acknowledges the authors.
- The README could benefit from more explicit guidance on setting up the environment such as briefly describing the "Functions" file's contents and providing a high-level overview of key data cleaning steps would be helpful. Explaining "sector-level scalers" and specifying the time horizon for MAPE calculation would add clarity. Briefly elaborating on the types of performance issues the visualizations help identify would also be beneficial.

Part 5: Feedback on Visualizations
Full Visualizations folder All Images Feedback

- Each graph effectively presents three key components: the historical "Train Data" used to train the models, the actual "Test Data" representing migrant encounters during the forecasting period (July-December 2024), and the "Predicted Data" which is the model's forecast for that same period. This allows for a direct visual comparison of how well the models performed against real-world outcomes.
- The variety of shapes and scales in the "Train Data" (blue lines) across the different sector and field office graphs is striking. Some sectors show relatively stable encounter numbers, while others exhibit sharp spikes and subsequent declines (e.g., Del Rio Sector, El Paso Sector, Rio Grande Valley Sector). This visual diversity strongly supports the project's decision to train and tune models individually for each sector to account for their unique historical patterns and scales.
- While the divergence between actual and predicted data is evident in many graphs, it is particularly pronounced in sectors and field offices with higher overall encounter numbers (e.g., Miami Field Office, San Francisco Field Office, Swanton Sector, Tucson Sector). In these cases,

the models predict a continuation of high numbers or increasing trends, while the actual data shows a significant drop. This disparity in larger sectors, visible in these graphs, underscores why the project chose MAPE as the evaluation metric, as large absolute errors in these sectors would disproportionately skew other metrics like RMSE or MAE due to their sheer volume. The high MAPE values reported (e.g., 92.8% average MAPE for ARIMA over 6 months) are the quantitative reflection of these large visual divergences in the graphs.

Zoomed Visualizations Folder All Images Feedback

- Across many of the sectors and field offices, the graphs visually demonstrate the significant divergence between the "Test Data" (orange line) and the "Predicted Data" (green line) during the 6-month test period. While the "Predicted Data" often continues the trend observed in the "Train Data" leading up to July 2024, the "Test Data" frequently shows a marked and often sharp decline. This visual discrepancy clearly illustrates the challenge the models faced in accurately predicting the actual numbers.
- The striking drop in the "Test Data" line compared to the "Predicted Data" line in many graphs serves as a visual representation of the impact of the significant political event—the re-election of Donald Trump and subsequent rhetoric/policy shifts—that occurred during the test period. In the report, this event led to a "sudden dip in encounters" which the models were unable to predict because the event fell outside the scope of the historical training data. These individual sector graphs collectively visualize the localized manifestation of this inability to capture that abrupt change, which in turn contributed to the high average MAPE values reported.
- However, we do have Trump's first presidency so there could be some predictor indicators in migrant encounters. Could you have done anything in your modeling approach to reflect this such as perhaps only using data from election years (latter half of 2016 to 2021) to help predict.
- Every graph displays a 6-month forecast horizon, consistently covering the period from July 2024 to December 2024. This uniformity in the prediction window allows for a direct comparison of model performance over the same time frame across all targeted locations. There is also uniformity for each graph showing from July 2022 to November 2024.

Nn_pred_plot Images Feedback

- The blue line, labeled consistently displays the historical migrant encounter data for this specific sector. Across all configurations, this line shows a clear pattern, including a significant surge in

encounters peaking around early 2022, followed by a decline and fluctuations before the forecast period begins. This historical pattern serves as the basis upon which the models were trained.

- The presence of multiple graphs, each showing predictions under different "Config" numbers, visually demonstrates the process of tuning the model for each Sector. The orange dashed line, labeled "Predictions," varies slightly in shape and magnitude across these configurations, reflecting how different tuning parameters (like the p, d, and q parameters for the ARIMA model, which was the best performer and tuned individually by sector) impact the resulting forecast.
- Despite the visible tuning process shown by the different "Config" predictions, none of the tested configurations in each sector's graphs successfully capture the sharp decline seen in the "Actual" data. This underscores the project's finding that standard time series models, trained solely on historical patterns, are fundamentally limited when unexpected, external socio-political events drastically alter those patterns. The tuning process attempted to optimize performance based on historical data, but it could not account for a future event outside that historical scope.

NN_mape_plot Images Feedback

- The exact nature of the "Window 24" calculation and the meaning of the "4-xx-01" x-axis labels in these specific charts is not fully clarified in the reports or the code It could be related to the performance evaluation of the "most recent 24 time steps (24 months/2 years)" but it needs to be clarified and/or mislabeled.
- Many of the charts show a general upward trend in Average MAPE values as the x-axis progresses from 4-01-01 to 4-12-01. There are often significant increases in MAPE towards the later points on the x-axis, such as 4-10-01, 4-11-01, and 4-12-01. Many charts show very high MAPE values, frequently exceeding 100%, 200%, or even 300%, especially towards the end of the depicted period. Some charts reach peaks above 400%, and chart peaks over 800%.
- The overall level and pattern of the Average MAPE varies significantly between the 27 charts, as indicated by the differing y-axis scales and bar heights. This suggests that the model performance (in terms of MAPE) varied across the different sectors or evaluation contexts being visualized.

NN_val_plot Images Feedback

- All configurations show a **decrease in both training loss and validation loss** over the 50 epochs, indicating that the models are learning. Many configurations show a rapid decrease in loss during the initial epochs (e.g., epochs 0-10), followed by a slower decrease or plateauing. The **Training**

Loss tends to fluctuate more than the **Validation Loss** in many configurations, likely due to the nature of batch-wise gradient updates during training.

- These configurations 24-27 also start with **very high initial losses**. However, unlike 19-23, the **training and validation losses track each other more closely** throughout the 50 epochs, both decreasing steadily. The validation loss is often slightly lower than the training loss at the end. While the final loss values are higher (around 0.02-0.055) compared to the better-performing initial configurations, there is **no clear evidence of overfitting** within these 50 epochs. This is great news considering many models are prone to overfitting.
- The consistent decrease suggests training could potentially continue to yield lower losses. Would training for more epochs create better model performance? Consider this while weighing computing restraints.

Average_MAPE for each Time Step Image Feedback

- The image effectively visualizes how the average forecasting error (measured by MAPE) changes over the 6-month test period, from July 2024 to December 2024 and shows the month-by-month progression of the average error.
- The graph clearly shows that the average MAPE generally increases as the forecast horizon lengthens. Starting relatively lower in July 2024 (around 47%), the average error rises steadily through August, September, and October 2024. This is a common pattern in time series forecasting, where predictions further into the future tend to have higher error rates.
- The graph shows a sharp spike in average MAPE occurring around November 2024, reaching a peak well over 160%. This dramatic increase directly illustrates the impact of the unforeseen external event discussed in the reports. The timing of this peak in November strongly suggests it corresponds to when the actual migrant numbers significantly deviated from the models' predictions causing the percentage error to skyrocket for that time step and illustrates the challenge posed by unpredictable socio-political events on the models' performance.