Predicting Ukraine's Emerging Humanitarian Needs, Guidehouse
Fall 2023, AI Studio Project Write-Up

Table of Contents

Business Focus and Summary

This project is a response to the humanitarian crisis in Ukraine, which was invaded by Russia in February 2022. The crisis has resulted in tens of thousands of people impacted through casualties and internal displacement. More than 8.2 million civilians fled the country, which by April 2023 created Europe's largest refugee crisis. Besides the mentioned impacts, there was major environmental damage caused by the war which contributed to food crises worldwide and food aid. The ML along with our time series model aims to use the ACAPS Ukraine Master dataset to forecast patterns such as certain needs and the demographics of people internally displaced within different regions known as Oblasts. Additionally, the primary goal of our model besides forecasting needs, was to get any valuable insight that would benefit and guide humanitarian efforts and organizations that aim to provide aid to those affected. Our results for our time series model was an RMSE score of 0.04149 which indicated high accuracy on predicting registered IDPs.

Data Preparation and Validation

DATASET DESCRIPTION

We were originally given 6 datasets provided by the humanitarian data exchange as linked: Ukraine 2023 Humanitarian Needs Overview People in Need, Ukraine Response Activities, Ukraine Flash Appeal, and Humanitarian Needs Overview (2021 - 2023), and ACAPS Ukraine Master Dataset. Some datasets included a variety of categorical and quantitative features such as Oblasts, # of people in need, severity score, amount of people exposed, and number of fatalities while others had the amount of people that received specific humanitarian aid like education.

Since the master dataset had all of the features in the remaining ones, combined we decided to pre-process just the master dataset[1]. We combined both master datasets from 2022 and 2023, removing completely null columns, containing total data for Ukraine or Crimea, as much of Crimea's data was missing. This was the perfect dataset choice as we weren't able to combine the other 6 datasets for the features to be aligned with one another and it was representative of all 24 regions of Ukraine, unlike the others that didn't contain the cities. The master dataset was also diverse in demographics and features that would represent the severity of the crisis:

- Internally displaced people - contextually meaning anyone who has been forced to leave their home as a way to avoid armed conflict but still reside within the Ukrainian borders
- Humanitarian Condition Level (1- 5) - originates from the European's INFORM Severity Index and is based on:
    - Impact of crisis
    - People in need
    - Condition of people
    - Access to humanitarian needs - healthcare, education, shelter, food

DATASET PREPARATION

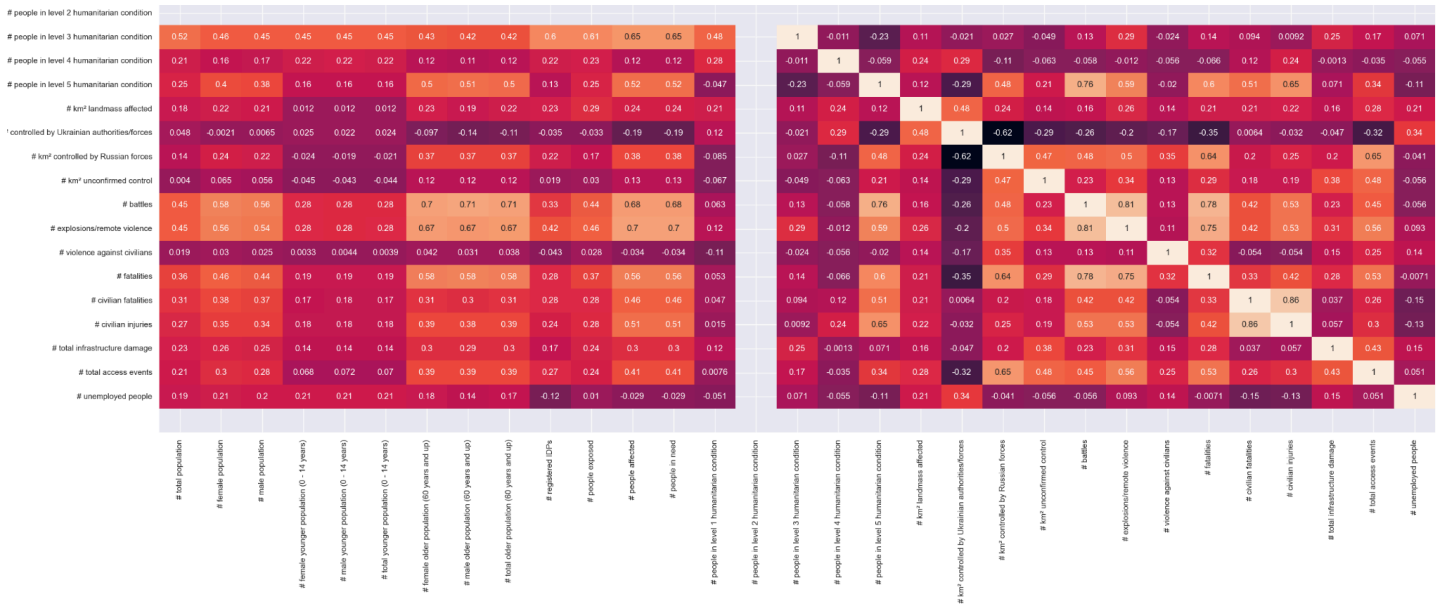
As mentioned, we removed null values but also removed any non-numerical values such as postal code, since we felt that wouldn't be beneficial to location like the Oblast feature. In terms of relevance, we removed data on wages, income, pension, and inflation and also food/fuel cost data as it was too likely for them to be affected by other confounding variables.

EDA (EXPLORATORY DATA ANALYSIS)

We divided exploratory data analysis between each other as we had 526 columns of data. Firstly, looking at the data before finding any patterns or trends, we noticed that most of the outliers were of low values, that being 0 fatalities which was accounted for by the month the data was recorded. Most of the outliers occurred in January 2022 - February 2022, which made sense since that was the beginning of the invasion.

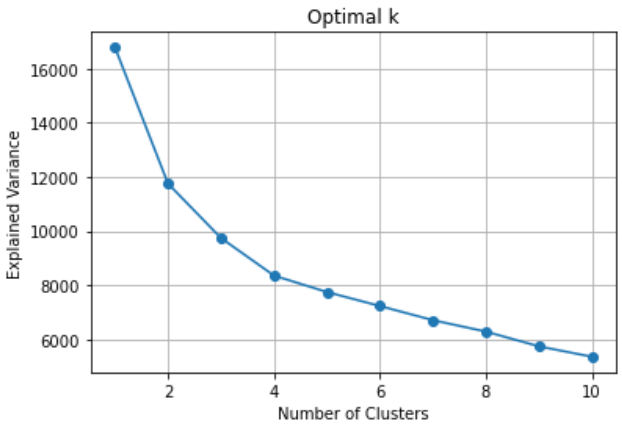
To get an overview of which features had strong or weak correlations, we first created a heat map.

[2] https://github.com/ari-sen/Guidehouse-1D



Since we had a lot of features, we anticipated the heatmap to be big. As we observed, there's a pattern of 1's, a duplicate correlation of each feature, such as a correlation between battles and battles. Despite that, many features like people in level 3 humanitarian conditions and registered IDPs had a strong correlation of .61, and km^2 of Russian forces consistently had lower correlations with IDPs. This gave us great insight into what to expect for our cluster analysis.

Next, we experimented with inertia and distortion to apply the elbow method to find the optimal number of clusters. For distortion, we found that the number of optimal clusters was 5 and inertia resulted in 4. In the end, we decided to expand upon the inertia method and apply PCA to 4 cluster groups using scalar standardization. Before applying PCA, we had to convert the most relevant categorical feature to a numerical value: Oblasts and scale the time column.



For this dilemma, there were two methods: giving each Oblast a number for our time series model or creating dummies for the Oblasts. In the end, thanks to Aaleia, we created dummy variables for the regions which eased our PCA process as we were able to include Oblasts in our analysis and not lose

much information geographically. The heat map and k-means analysis guided us to further evaluate the clusters in depth.

FEATURE SELECTION

In total, our master dataset had 34 features, but we decided to use 32 out of 34 features since it would be representative of reasons why internally displaced people numbers may be higher than other data points. Feature importance was particularly important for our model as we wanted to account for as much inclusivity as we could when addressing the crisis.

We made sure that our model focused on the following features:

- Area controlled by Ukraine and Russia
- Population and people exposed
- Demographics
- Levels of Severity from 1-5 (Minimal, stressed, moderate, severe, extreme)
- Violence
- Fatalities
- Unemployment

Since we wanted to start with the unsupervised learning approach first to find more broad patterns in our data, this was relevant to understanding the territorial dynamics through geographical context. It also aligned with our goal to create an unbiased model since we were including demographics of age and gender, it allowed for our model to capture how the crisis impacted different sectors of the population. Lastly, since we didn't include fuel costs, unemployment was an alternative to representing the correlation between any economic instability and displacement.

Approach

SELECTED MODELS

We selected 2 main candidate models to guide us through this AI Studio project:

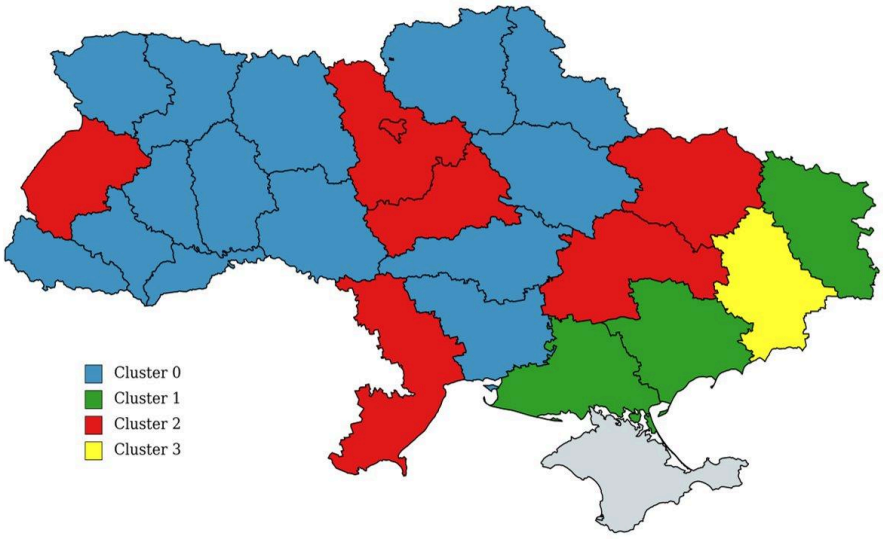
- **Cluster Model using PCA** – we selected this model because it can efficiently handle the high dimensionality our dataset had by reducing the dimensions while capturing the significance variance between the components. Additionally, we considered visualization as it was able to increase simplified interpretability for our audience while maintaining

- **Simple Time series Model** – we selected this model because we wanted to include a mix of supervised and unsupervised especially since we wanted to be able to predict or forecast further displacement in a specific region.

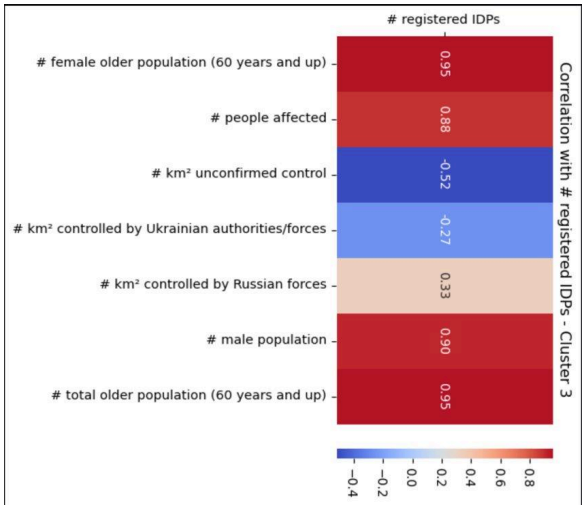
CLUSTER ANALYSIS

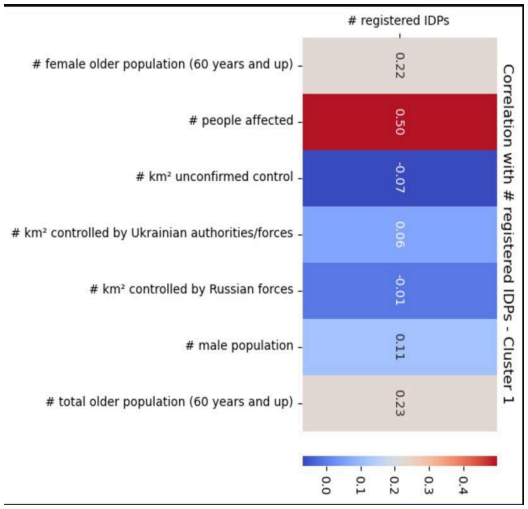
Throughout our project, we wanted to analyze which cluster each Oblast was represented by whether there were any trends within those regions, and the amount of internally displaced people.

Moreover, through PCA we found and defined Cluster 0 and 1 to represent the administrative region: Kyiv, Cherkaska, Chernihivska, Chernivetska, Ivano-Frankivska, Khmelnytska, Kirovohradska, Mykolaivska, Poltavska, Rivnenska, Sumska, Ternopilska, Vinnytska, Volynska, Zakarpatska, Zhytomyrska, Khersonska, Luhanska and Zaporizka. Cluster 1 was found to weigh the importance of unconfirmed control of land while Cluster 0 weighed the importance of older populations that were affected. Cluster 2 was represented by Dnipropetrovska, Kharkivska, Kyiv, Kyivska, Lvivska, and Odeska. Lastly, Cluster 3 was represented by Donetska.



From Cluster 0 and 1, the number of people affected was highly correlated with registered IDPs. Cluster 2 majority of unconfirmed and confirmed regions of Ukraine were correlated with registered IDPs. Cluster 3, which we analyzed had a high correlation between male and female older populations and registered IDPs but had a weaker correlation with the impacted regions in Ukraine and Russia. Lastly, you would think that there would be some type of correlation, moderate or strong between registered IDPs and km^2 of Russian Forces especially in Kyiv, however, there were consistently lower correlations across all 4 clusters as demonstrated below in the correlation map.

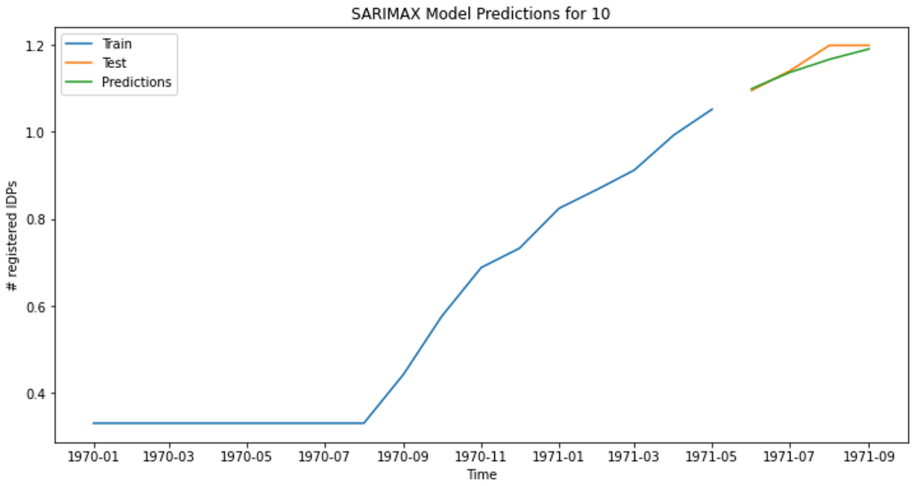




TIME SERIES MODEL - DISPLACEMENT AID

Due to our dataset having a period feature indicating the month and year each data point was recorded, we used a time series model to find systemic patterns over 2 years using this monthly data. The first step in approaching this was deciding whether we used ARIMAX or season ARIMAX. ARIMAX is an automated integrated moving average model with the use of external variables. In the end, we decided to use seasonal ARIMAX as it performed well regardless of seasonality and performed better for certain Oblasts in comparison to ARIMAX.

- Grid search to find optimal and seasonal order for each Oblast
- Optimized model to run grid search 4 times instead of 25 (for each Oblast) to reduce cost and runtime
- We used an 85-15 split to account for any unexpected spikes in IDP numbers

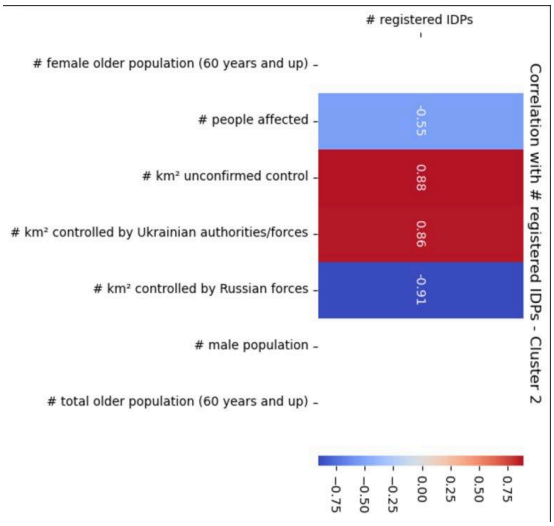


As a result, we found that predicting displacement within Ukraine is possible but to be more accurate, would have to be tweaked and trained with more datasets, especially since our `period` kept being displayed as 1970 - 1971. With this time series model, we were able to achieve an average RMSE of 0.04149 across 25 Oblasts, which implied that it had a high accuracy.

DISPLACEMENT NUMBERS

Below is a snippet of the master dataset represented by Cluster 2, with `Oblast = Kyivska`.

# km² controlled by Russian forces	# km² unconfirmed control
0	0
8440	0
8440	0
0	0
0	0
0	0
0	0
0	0
0	0



As you can see there are two rows where the `km² controlled by Russian Forces` was 8440, but over the months decreased back down to 0. Therefore, we were able to see that that may be why the correlation between the amount of geographic space controlled by Russian forces and `registered IDPs` was negative, specifically -0.91. However, for the red features that had a higher correlation: `km² of unconfirmed control` and `km² controlled by Ukrainian forces`, there are more 0's

which can be explained throughout the months not only by other features but also context, that being how much each region was impacted over time by Russia's invasion.

PREVENTING OVERFITTING

Given the period and sensitivity of our dataset, it introduces the challenging dynamic of ensuring our model can generalize well in the future while maintaining unbiasedness. We were careful with the type of modeling we went with, which is why PCA aligned with that goal as it reduced the amount of noise and features which adapted to the potential of its overfitting. We also experimented with different splits such as 80-20 which performed as well as the 85-15 split given the variance in IDPs. However, a method that we wished we expanded on was regularization on our time series model to see if that would've decreased the RMSE score to indicate a better fit.

VALIDATION STRATEGY

The one strategy that our team used to validate our time series model was splitting the data into two sets: a training and test set to evaluate the model's performance through RMSE, on the test set. We did this, as mentioned in two runs. One being an 80-20 split and the other being an 85-15 split. We would have liked to experiment with more validation techniques that we learned throughout the Machine Learning foundations such as not only finding the RMSE but also the R-squared score. Getting the R-squared score would've given us insight into how well our model was able to approximate the actual data. Additionally, in hindsight, applying cross-validation or random splitting to get more to our time series model would have been interesting to see if there were any other optimal splits besides 85-15 to get the same results we did.

Key Findings And Insights

KEY RESULTS

From our gatherings and efforts, our final model included a PCA visualization of the 4 clusters on a map of Ukraine and 4 correlation maps, each representing the clusters thanks to Sherin. From the 32 features used, we were able to minimize it down to 7 optimal features as presented on our correlation maps: female older population, male population, total older population, people affected, km^2 of

unconfirmed control, km^2 controlled by Ukrainian forces, and km^2 controlled by Russian forces. In addition to those two models, we also included our time series model in our final submission keeping the 85-15 split to get the accuracy score of 0.04149. From our time series model, we found as the time of the crisis increased, the amount of registered IDPs also increased with a predicted value of 1.2 at its peak. Which, according to our cluster analysis was correlated to the amount of geographical space concerned

INSIGHTS

1. **Account for your end goal while also considering your time**. Our team initially didn't know if we wanted to do an unsupervised learning model versus supervised learning since our objective was to predict a value, specifically aid. However, we decided to do both because we still wanted to forecast a value that would be potentially helpful for real-world use in other crises besides the Ukrainian crisis such as the Palestine genocide. We had to allocate time outside of what we learned to learn how to implement a time series model and also experiment with advanced models such as LSTM and SVM. If we were able to get results using the advanced time series models, that would've been good to cross-reference with the simple time series model we created. I'm thankful to have been introduced to the idea of SVM as we have the opportunity to research on our own time and experiment with it throughout the Spring AI Studio.

2. **Experiment with validation strategies even if you think the results won't be satisfying.** The only validation strategy we used was splitting, which although our RMSE indicated our model had high accuracy, there are other factors to consider like the size of the dataset in terms of months. Our dataset had 1 year of data from January 2022 - December 2022 and months of data from January 2023 - September 2023. In the forthcoming years , when the dataset were to be updated this would be helpful for our model to generalize well on future data as there would be more data for it to be trained on. Some validation strategies that could be considered for this project include:
    - Cross-validation using k-folds
    - Rolling forecast origin - this would be helpful particularly for the time series model as it would evaluate the model each time its' forecast origin is updated indicating how our

model would perform on updated data. The question is will the RMSE using rolling forecast origin be less than or more than our result of 0.04149 from splitting? And if so, does that indicate our model would be able to predict well for crises in other countries or would it have to be tweaked to adjust to features that Ukraine may have not had?

3. **Working in teams is encouraging but don't forget about your external resources.** This was an empowering project to gain collaboration and teamwork skills. I was able to not only have the opportunity to work through such challenges with my team but learn from them individually. Though we were making sure to stay on top of communicating with each other , there were times when we were so overwhelmed with deciding which direction to go in, that we overlooked asking our challenge advisor or TA. They were a part of our team also and were there to help us with our confusion, not just to confirm if we were making the right decisions. Our TA, Cindy, was able to help us be open to the fact of using an advanced time series model whether it be used in the final submission or not. This taught many other lessons like dipping your toes in the water when it comes to models.

Acknowledgments

Words can't explain how thankful I am for all of the support surrounding my team and our project! I am so proud of my team and what we've achieved over these 4 months while balancing school and extracurriculars. I've learned and gained so many skills and relationships in these 4 months, and am grateful to everyone at Guidehouse for all that you've shared. To my Challenge Advisor, Karen, TA Cindy, and teammates Aaleia, Sherin, Iman, and Cindy, thank you for your support, and the opportunity to let me learn from and with you. I'm anticipating what the AI Spring Studio will bring for us all.

Additionally, thank you to Break Through Tech, the Cornell Tech AI Program team, and specifically Erika and Abby for their efforts in making this a wonderful program experience for us and Judith for founding this program. I'm excited to see what all our future endeavors bring