STA130 Final presentation

By: 99% confidence interval (Selina, Polyna, Flora, Angel, Matthew)

Model Performance Metrics: 3 Different Models

- xgboost
- ffnn
- transformer

Differences Between the 3 Models

- There are 3 models that we are currently using to interpret our data, the FFNN model, XGboost, and the Transformer model. These 3 models each have strengths and weakness when it comes to comparing different indicators using 3 alternative metrics; Accuracy, Sensitivity, and lastly, specificity. Today from our examples we will focus on the indicators of, hdi and income group and look at their accuracy, sensitivity and specificity.

Analyzing our Models: Accuracy

From looking at our maps and tables from the previous slide, this is what we have noticed:

- Maximum accuracy score from FFNN is the highest, with a score of 0.92. while the maximum score for the XGboost and Transformer are tied with a score of 0.72.
- Judging from the increments of the scales used in each model, we find that:Ffnn has the greatest range of 0.6, XGboost is a close second with 0.35, Transformer has the smallest range of 0.15
- Transformer with a medium hdr_hdicode is greater than its high hdr_hdicode, while XGboost with a low, medium hdr_hdicode is roughly the same
- The accuracy is quite spread out throughout the countries, although It seems that countries higher up in the maps have a high accuracy. From this analysis, we can assume that FFNN is the best model for determining the accuracy score for the hdi, since it has the highest accuracy score.

Analyzing our Models: Sensitivity

From our analysis, we have noticed:

- The sensitivity score for the FFNN and XGboost model is 1, and the Transformer score is 0.667
- In terms of sensitivity in our maps, we find that countries focused in Northern Africa and the Middle East have the highest sensitivity scores.
- From the data table, we have identified that overall the XGboost model has the highest score in terms of sensitivity (judging based on the average).
- Finally, we can assume that the XGboost model is the best for measuring the sensitivity for HDI. The XGboost model has the highest sensitivity score, meaning that it is the model that provides the least 'false negative' results. The XGboost model has the least falsely indicated tests for when something is not present, when it is present.

Analyzing our Models: Specificity

From our analysis, we have noticed:

- FFNN model has the highest score for specificity for HDI
- The highest score for XGboost is 0.741, and the maximum score for transformer is 0.739. the 2 scores are very similar with only a 0.002 difference.
- The FFNN model is also the one with the highest range from the scores, with a range of around 0.9, followed by XGboost with a range of 0.5, and Transformer with a range of around 0.2.
- The maps of specificity for the three models are similar to the maps of accuracy, with the countries higher up and in the corners of the map having the highest specificity scores.

Analyzing our Models: Specificity

- From the maps and table we have noticed that the sensitivity and specificity could be inversely related. If one of the models have a high sensitivity score, it seems that the specificity score is lower. Which makes sense, since we know that specificity tests for any 'true negatives' while sensitivity tests for 'false negatives' in a set of data. We have assumed that the FFNN is the best for measuring the specificity, as it has the highest score.

Analyzing our Models: Accuracy

- From this data we have determined that depending on the income group, the accuracy varies. In the low income group, Transformer has the highest accuracy of 0.552. Lower middle income group, FFNN has the highest score of 0.667. Upper middle income, FFNN has the highest score of 0.874. High income, FFNN has the highest score of 0.920.
- From the maps we find that the XGboost and Transformer models look very similar, with the same countries having the highest accuracy scores. And with the FFNN model the accuracy scores are very spread out, and generally more leveled. There are a lot more countries with higher accuracy scores compared to XGboost and Transformer. We Assumed that FFNN is the best model for accuracy in terms of income groups, since FFNN has the highest accuracy score out of 3 of the 4 different income groups.

Analyzing our Models: Sensitivity

- For determining sensitivity, the scores between each model varies depending on the income group.
    - For the low income group in our data, XGboost and FFNN ties between the highest score for sensitivity with both their scores being 1.000. while Transformer has a score of 0.786.
    - The lower middle income group, XGboost has the highest sensitivity score of 0.909.
    - Upper middle income, the highest is XGboost with a score of 0.467.
    - High income group, XGboost and Transformer ties at being the highest. Both with a score of 0.375, while FFNN is 0.000.
- After analyzing the maps, we have found that all 3 maps look very similar, with the highest sensitivity scores focused in Africa. Lastly, we have determined that XGboost is the model for determining sensitivity. As it has the highest sensitivity score for all the income groups. even after tying with some of the other models.

Analyzing our Models: Specificity

- Using the table to determine our specificity, we can see that our score between each model varies depending on the income group.
    - In the lower income group, the model with the highest specificity score is the Transformer model, with a score of 0.477.
    - Lower middle income group, FFNN has the highest specificity score of 0.711.
    - Upper middle income group, FFNN has the highest specificity score of 1.000.
    - In the high income group, FFNN once again has the highest specificity. With a score of 1.000.

Analyzing our Models: Specificity

- From our analysis of the maps, we find that the XGboost and Transformer models have very similar maps. The highest specificity scores are usually focused in more "Eurocentric" places like Europe, North America and Australia. However, Chilly and a small region of Africa also have high specificity score. The FFNN map is very different from the 2 other models. With no specific focus on certain countries. The FFNN map is lot more "evened" out through the map.
- We have previously assumed that sensitivity and specificity have an inverse relationship. And we can see that this also the case for the income group models. With models scoring high in sensitivity having low scores in specificity. And from this data we have determined that FFNN is the best model for measuring specificity, as it scored the highest in terms of specificity for 3 out of the 4 different income groups.

Making assumptions

Following the analysis from the maps and models, we will now assume and hypothesize that determining a model that is best is quite difficult, since each model are best at measuring different metrics; sensitivity, accuracy, and specificity, depending on the category we want to measure. To further prove this hypothesis, we did a linear regression model and hypothesis testing.

Proving our Assumption from Data

From the linear regression model there are a few factors that we can look at interpret. Starting with the R-squared value, after the adjustment we get a value of 0.285. this suggests that around 28.5% of the variability of the dependent variable is associated with the independent.

Looking at the coefficients, we get different values, ranging from some being negative, and others being positive. Negative coefficients typically represent an inverse relationship where as the independent variable increase, dependent will decrease. Whereas positive coefficients mean if independent increases, so does the dependent.

Now focusing on the null hypothesis, which states that the independent variable has no association with the dependent variable. In this case, our null hypothesis would be, "The Categories an observation belongs to does not matter". Which is incorrect. If we look at our p-values, they are all extremely small. We know that when the p-value is less than 0.001 it means strong evidence against the null hypothesis. The final conclusion would be that the null hypothesis is incorrect and the categories an observation belongs to does matter when determining which metric works best. This also ties into the proof of our hypothesis.

Ethical consideration in "Conflict escalation predictions"

	Predict conflict	Predict no conflict
Conflict occur	(TP, true positive) Prepared for conflict	(FN, false negative) Unprepared for conflict
Conflict didn't occur	(FP, false positive) Waste of prepared resources	(TN, True negative) No waste of resources

TP: The country would be prepared to distribute resources and deal with the conflict

TN: The country wouldn't waste resources to prepare for a potential conflict

FP: The country would waste resources to prepare for a non-existent conflict

FN: The country would be unprepared to deal with the conflict and aid its citizens during this time

Conclusion

In conclusion, The null hypothesis is proven to be false, since it is evident that all 3 models have their own strengths and weaknesses depending on the category they are measuring. All three models serve as tools for forecasting the escalation of conflicts. The first graph utilizes a Feed Forward Neural Network. The second graph employs the XGBoost algorithm (Extreme Gradient Boosting). Lastly, the third graph uses a transformer model. It is remarkable that all three methods mentioned above are neural networks, which use text data alone to identify potential conflict zones. Despite using different data sampling methods and modeling approaches, these neural network-based graphs show remarkably similar results.

Upon thorough analysis of the provided information, it becomes evident that these neural network-based graphs appeared to be great at analyzing geographical escalations. XGBoost, ffnn and transformer can potentially be adapted for forecasting conflict escalation as a binary classification task, their direct comparison in terms of sensitivity and specificity might be challenging. However, despite being similar, none of these graphs can be used as the best model for conflict escalation prediction. These results can be used by researchers on this topic in the future.