Regional Origin Classification across Administrative Divisions using Surnames

Vipin Gunda, Dr. Aleksandr Michuda

Cornell University

LAIDLAW FOUNDATION

Introduction

The rising availability of smartphones in developing countries directly ties into the growth of internet technologies like ride-share applications. SafeBoda is an example of such a ride-share platform. Based in Uganda, SafeBoda is the leading rideshare platform for boda-bodas in Kenya and Uganda. Done under the supervision of Dr. Aleksandr Michuda, an Assistant Research Professor at the Center for Data Science for Enterprise and Society, this research contributes to his larger study on how the introduction of ride-share applications has transformed the Ugandan labor market.

Specifically, this project uses machine learning and drivers' name data to predict regional origin at different administrative units: SAP region, GAUL (Global Administrative Unit Layers), and district. This research is driven by performance changes for varying administrative units, which could possibly motivate the switch to treat this as a regression problem instead of classification.

Data Processing

The primary dataset for analysis is a compilation of surnames, SAP region, GAUL, and district, along with other demographic and geographic information for voter registrants in Uganda with 14,589,193 rows.

Oftentimes, surnames had "noise," with the primary culprit being miscellaneous symbols, which is expected due to natural errors from user input. Using a function that replaces spaces, hyphens, apostrophes, periods, and zeros (which users commonly mistake for O's), the surname column of the data frame was processed (~58.2 seconds).

```
df_aux.loc[lambda df: df[features]
        .str.replace('-','')
        .str.replace(' ','')
        .str.replace("'",'')
        .str.replace('.','')
        .str.replace('0', 'O')
        .str.isalpha()
        ]
```

Figure 1: Data Processor Code Snippet

The SAP Region, GAUL, and district data were not processed because they were much more structured.

Tools

- XGBClassifier (from XGBoost)
- TF-IDF Vectorizer (from Scikit-Learn)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 2: TF-IDF Weight Calculation

The vectorizer compares the number of times a word appears in a document to how many documents that word appears in to measure the weight of each feature.

Tools (Cont.)

- StratifiedGroupKFold (from Scikit-Learn)

A KFold is a way to split data into k sections, which is useful for splitting our data for training, validation, and testing. A StratifiedGroupKFold, specifically, creates k folds that preserve the percentage of samples for each class and also ensures that the same group will not appear in multiple folds.

Methodology

We train an XGBClassifier on our processed data (X for features and y for labels). We can then run X and y into StratifiedGroupKFold to create a splitter. Using the XGBClassifier and splitter, we can use Scikit-Learn's CalibratedClassifierCV to produce a calibrated model, which can then be used to create a Feature Importance chart.

The F score measures the increase in a model's prediction error after permuting the feature. Since "Abi" has the highest F score, it is the most powerful feature.

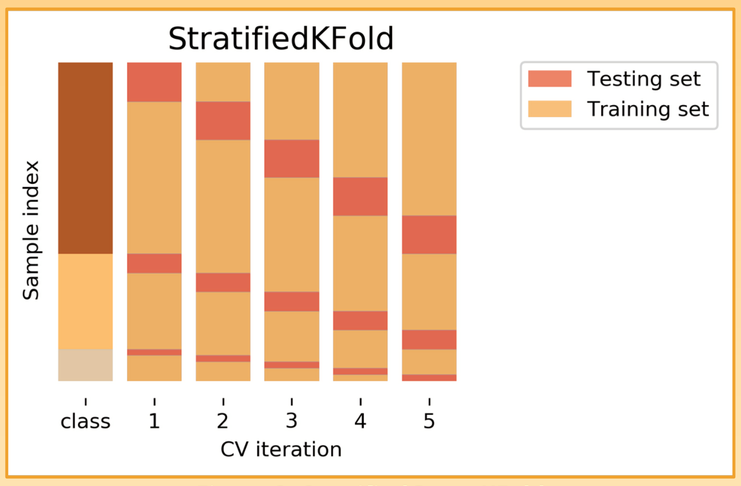

Figure 3: StratifiedGroupKFold

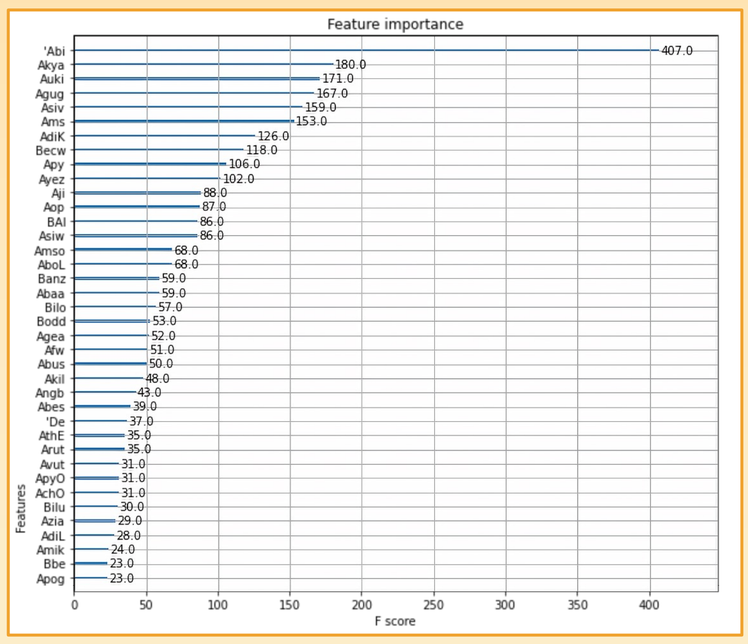

Figure 4: Feature Importance Chart

Confusion Matrices

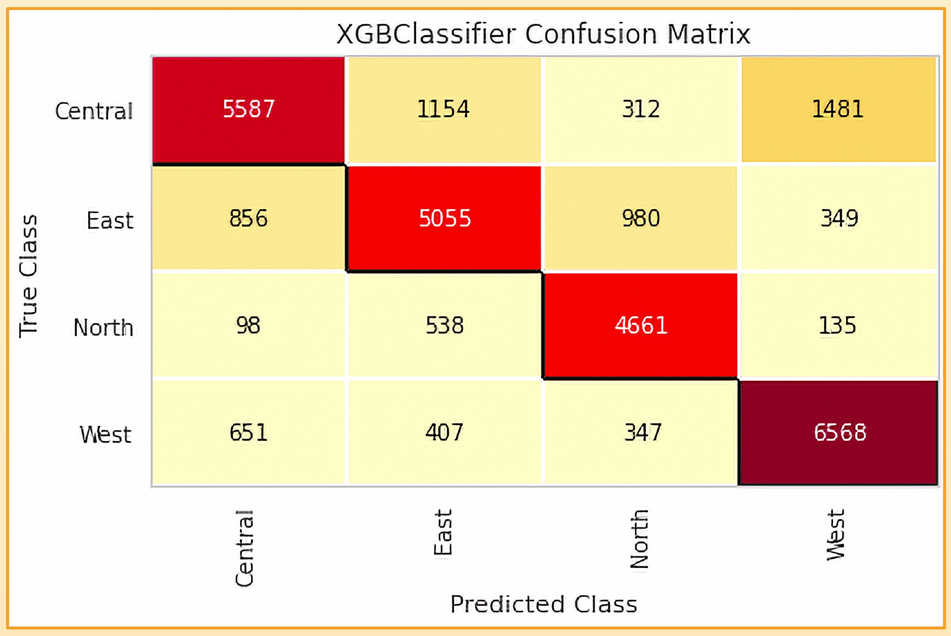

Figure 5: Region Matrix (Frac = 0.001)

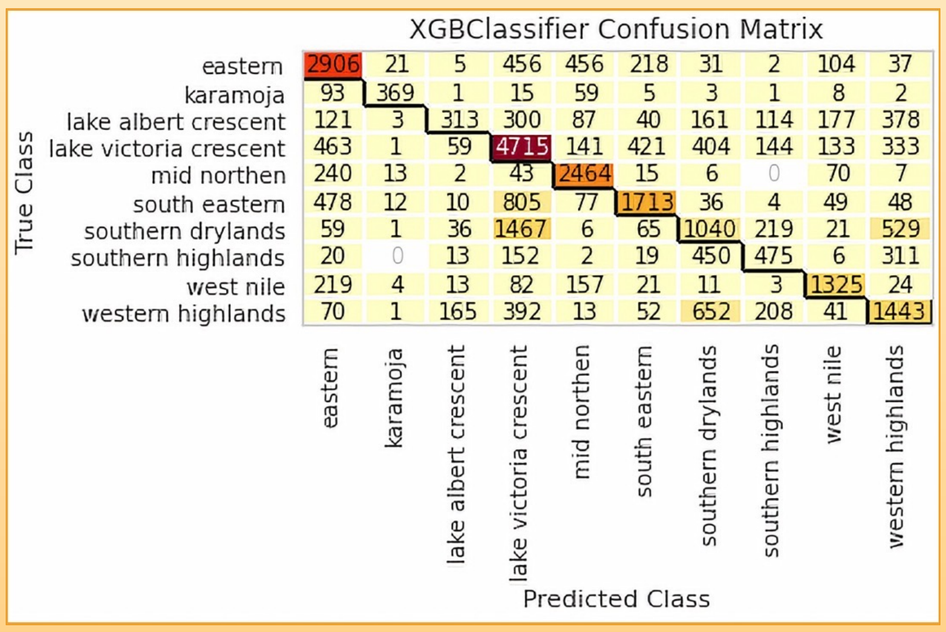

Figure 6: GAUL Matrix (Frac = 0.001)

- Diagonal of region matrix is much more emphasized than the diagonal of GAUL matrix.
- Figure 5: For predicted "West," 6,568 predictions were right, but 1,481 were actually "Central," which is the largest regional misclassification.
- Figure 6: GAUL model did the best with predicting "Lake Victoria Crescent."
- Through these matrices, we can pinpoint which classes our model needs to improve on and discover how our model interprets Uganda by analyzing the matrix trends.

Learning Curves

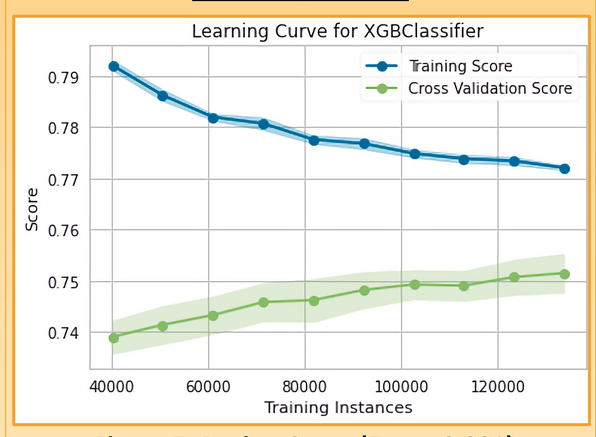

Figure 7: Region Curve (Frac = 0.001)

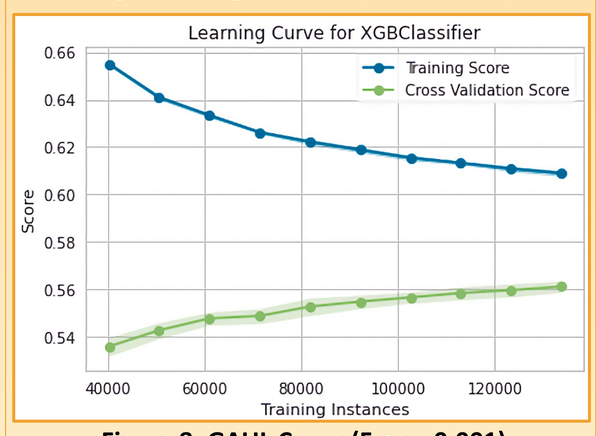

Figure 8: GAUL Curve (Frac = 0.001)

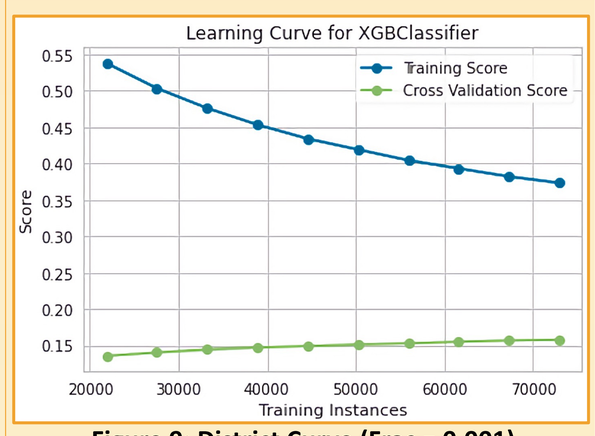

Figure 9: District Curve (Frac = 0.001)

- Convergence
- The SAP region accuracy seems to be converging towards 0.76, while GAUL is heading towards 0.59, and finally, district merging at around 0.27.
- Suggests that the accuracy of the final optimal model from greatest to least will be SAP region, GAUL, and finally district.

This trend is caused by the number of labels for each geographic denomination. Four SAP regions are much easier for the model to grasp than 112 districts. Furthermore, we might expect names to vary more between region to region rather than between neighboring districts.

Conclusion

The next step will be to turn this into a regression problem to predict GPS coordinates rather than geographic denominations. From there, the goal will be to connect weather data for those coordinates and ideally, produce GPS predictions for the drivers. The societal implications come from the possibility of capturing the connection between the rural and urban sectors of the economy more precisely. This could help design policies for workers in cities that need to provide remittances to their family in rural areas, which is crucial when rural households have no means of protecting themselves against shocks such as drought. And so, we will look to expand the scope of the project now that the preliminary analysis is complete.