CAPSTONE PROJECT: FINAL REPORT

Paria Elyasi | September 25, 2022

Problem Statement & Background

This project predicts the outcomes of asylum cases. The goal of this project is to look at what features make a difference in outcomes of asylum cases to help immigration attorneys change their approaches accordingly for a favorable outcome. Additionally having a prediction on the outcomes of the case they are representing can better assist them in their legal methodology.

Every year asylum seekers come to the United States requesting protection because they have suffered persecution or fear that they will suffer persecution due to their race, gender, belief, or political opinion. Asylum cases are one of the most difficult and lengthiest immigration cases in the United States. Being denied asylum means being deported or removed from the United States.

I have been a volunteer with two non-profit organizations serving immigrants in Texas. I have witnessed the adversity asylum seekers go through in order to gain legal status in the United States. I decided to explore the data on asylum seekers to gain key insight and help both asylees and immigration attorneys with my findings. With the insight gained from this project, immigration attorneys are able to adjust their approaches and know the key features that make a difference in the outcomes of asylum cases. Also being able to have a case prediction prior to court date can help attorneys adjust their approach for a favorable outcome.

About the Data

Finding a dataset on asylum cases in the United States was difficult. I was unable to find any datasets from United States Citizenship Services (USCIS) as majority were yearly reports on immigration numbers and statistics. After searching extensively, I was able to find a dataset that was originally from Department of Justice (DOJ) and it had cases overseen by immigration judges on different cases.

The dataset included series of .csv files with over 8,000,000 rows and 400 columns combined. It was challenging to import all into MySQL in order to query the data. After many failed attempts to import into MySQL workbench, I found a work around for importing the .csv files into MySQL. I had to import all files into Microsoft SQL Server first and then migrate to MySQL. In my project submission, I have provided a step-by-step instruction on how I did this for others to reference if needed.

Once the data was migrated into MySQL, I noticed that it had a lot of empty cells or missing values. There were also many lookup tables as many of the values were codes that needed to be joined with the lookup table in order to get the actual value. There were many rows without case decision therefore I did not include those in my dataset. Once I queried the data with the features I needed, I exported it as a .csv file.

Data Cleaning and Exploratory Data Analysis

Majority of the data cleaning was for missing values and deciding how to impute them. There were columns with more than 40% missing values and I didn't want to remove them, so I had to strategically impute them with values that would have introduced the least bias. The most challenging ones were date columns which had many missing values. There were also many clerical errors with date entry which had to be fixed.

In order to uncover trends and relationships in my dataset, I did an Exploratory Data Analysis. During Exploratory Data Analysis, I saw that **88%** of cases in my dataset were **rejected** and only **12%** were **accepted**. Average wait time to receive a court date was **2.5 years** from date of entry to the United States. Nationalities such as **Mexico**, **El Salvador**, **Guatemala**, **Honduras**, and **China** had the highest acceptance rate amongst other nationalities. These countries also had the highest number of asylum seekers in our dataset. **Montana**, **Hawaii**, **Alaska** had the highest **acceptance** rate, while **Oklahoma**, **South Dakota**, and **North Dakota** had the highest **rejection** rate.

In the data preprocessing stage, I tried to keep as many features as possible since I wanted to see which ones could be of importance in predicting outcomes of Asylum Cases. However, I did not keep the features that were collinear with other features. For example, Notice Date was collinear with Hearing Date. I also feature engineered the age column as the original data only had Birthdate (month and year). I wanted to see if the age of asylum applicants had any importance in the immigration judges' decision. I also had to convert the numeric features of attorney representation to categorical as it had extreme outliers.

My data had many categorical features. Prior to modeling, I converted all categorical features to numeric. During this process, one value per categorical column was dropped to avoid multicollinearity.

Modeling and Results

For my project, I developed several machine learning models to predict the outcomes of asylum cases. I chose to do classification models as my target was whether the asylum case is **Accepted** or **Rejected**. I fitted a Logistic Regression Model, Random Forest Model, and XGBoost Model.

For all these three classifiers, I started with a basic model to see what the baseline accuracy score was and followed that step by tuning hyperparameters within the models for an attempt to improve the accuracy. Figure 1 shows highest test scores for all three models.

Model	Accuracy
Logistic Regression Model	93.17%
Random Forest Model	90.59%
XGBoost Model	93.55%

Figure 1

Out of all the models, **XGBoost** model was the best performing model with a test accuracy of **93.55%**.

However, besides looking at accuracy, it was important to also interpret the models and look at feature importance for each. For the logistic regression model, I looked at coefficients to find top 20 features most predictive of acceptance and top 20 features most predictive of rejection in asylum case outcomes. For the random forest model, I looked at feature importance. For the XGBoost model it was harder to interpret and look at the features as this model falls into a 'black box model' category. That being said, I used SHAP value to look at top features/predictors in this model.

Choosing an appropriate metric to evaluate the model depends on the business objective. I chose to focus on high precision as incorrectly telling applicant their application will be rejection (false negative) when in actuality it will get accepted is more acceptable. Again, in this case the **XGBoost** Model had the highest precision score (**83%**) out of all other models.

Conclusions

In this project I chose classification models and predicted the outcomes of asylum cases. The goal of this project was to look at what features make a difference in outcomes of asylum cases to help immigration attorneys change their approaches accordingly for a favorable outcome. The features that all three models had in common as strong predictors were: **Nationality** (specifically Honduras and Guatemala), **Absentia** (if the applicant was absent at hearing), **Date of Entry** (year), **Case Description** (specifically Credible Fear Review and Reasonable Fear Case).

There was also a great importance given to the immigration judge who reviews the cases. Again, there has been research that some immigration judges reject cases no matter how strong the case is. At this point I don't have enough evidence or confidence to conclude this however as only the logistic regression model coefficients had immigration judges as features predictive of both acceptance and rejection.

Future of work for this project will be to continue the same steps on a better dataset from the non-profit organization. Having features such as educational background and marital status could assist further into our findings and predictions. I would also like to reduce the number of features (those that didn't show any importance in the models). For example, using only top 5 nationalities instead of all nationalities could work better for our model's accuracy. Taking a forward selection approach as oppose to backward selection will be applied to this project as well. I would also like to add features such as weather and news headlines to see if any interesting trends can be found and how those would change the models' predictions.