Advanced Machine Learning - Easy Visa Project

Client: Easy Visa Immigration
Project Focus: Automating Visa Approval Predictions through Machine Learning

Advanced Machine Learning - Easy Visa Project

Easy Visa Immigration - Processing Model

- Class: Advanced Machine Learning:
- Name: Daniel Levenstein
- Submission Date: 10/25/2025

Business Context:

The U.S. faces a high demand for skilled labor, which necessitates an efficient process for visa application review. The Office of Foreign Labor Certification (OFLC) processes thousands of applications annually, a task that has become increasingly cumbersome. The goal is to leverage machine learning to streamline candidate shortlisting, improve accuracy, and assist OFLC in making faster, data-driven decisions.

Objective:

Develop a classification model capable of predicting visa approval likelihood, thereby aiding in shortlisting suitable candidates and optimizing the certification process.

Methodology & Key Findings:

- Data Analysis: Performed exploratory data analysis (EDA), handling missing values, and converting categorical features into numerical formats.
- Model Development: Implemented multiple models including Gradient Boosting, AdaBoost, and Decision Trees across different sampling techniques (original, over-sampled, under-sampled).
- Model Tuning: Conducted manual hyperparameter tuning and automated grid search to enhance model performance.
- Best Model: The Over-sampled Gradient Boosting model achieved an F1 score of 80.05%, with minimal performance gains after fine-tuning, making it the most effective for predicting visa outcomes.

Model Validation:

The project began with EDA on the raw data, where I handled missing values and converted categorical features into numerical form. The dataset was then split into training and testing subsets for model evaluation.

Model Building:

Model development was completed in three stages.

1. Initial Models: I trained five different models on the original training data and saved the best-performing one for later analysis.
2. Over-Sampled Models: I applied oversampling to address class imbalance and retrained the same five models, again retaining the best results.
3. Under-Sampled Models: I repeated the process with under-sampled data and compared performance across all sections.

After evaluating all models, I identified which were overfit and which generalized well. The strongest performers were the Gradient Boosting model on the over-sampled data and the AdaBoost model on the original dataset.

Model Tuning:

I conducted two rounds of hyperparameter tuning. In the first round, I adjusted parameters manually to reduce overfitting and achieve balanced performance across models. In the second round, I performed a formal grid search on the top-performing model to further optimize its accuracy and efficiency.

Results:

The final top-performing models were:

- Gradient Boost (Over-Sampled)
- AdaBoost (Over-Sampled)
- Decision Tree (Over-Sampled, manually tuned)
- Gradient Boost (Under-Sampled)

A formal grid search was performed on the top model which was the Gradient Boost Over-Sampled model. The performance differences between the manually tuned model and the grid search were under 1% which I felt was a small enough margin that additional tuning was unnecessary.

Overall, the Gradient Boosting model on over-sampled data provided the best balance between precision and recall, making it the most suitable model for predicting visa application outcomes.

Conclusion

The Gradient Boosting model trained on oversampled data best balances precision and recall, facilitating more reliable visa approval predictions. This approach significantly improves the efficiency of the application review process, reducing manual workload and ensuring a more consistent decision-making framework.