EasyVisa Project

Predicting Visa Certification Using Machine Learning

Isaac Gross - Apr 25, 2025

Contents / Agenda

- Executive Summary
- Business Problem & Solution Approach
- Exploratory Data Analysis
- Data Preparation
- Model Building & Tuning
- Model Performance Summary
- Appendix

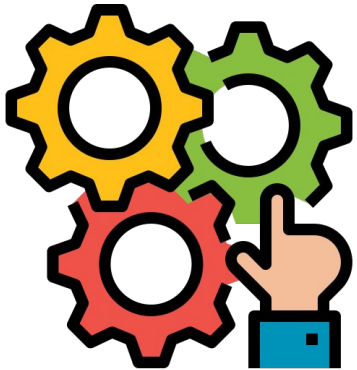
Executive Summary

Our company is spending a lot of time reviewing visa applications manually, and the volume keeps growing.

We built a machine learning model to help them flag the applicants who are most likely to get approved, so they can save time and focus on the ones that need more attention.

After comparing multiple models, the best performer was a **tuned XGBoost model trained on oversampled data**. It gave us a strong balance of precision and recall:

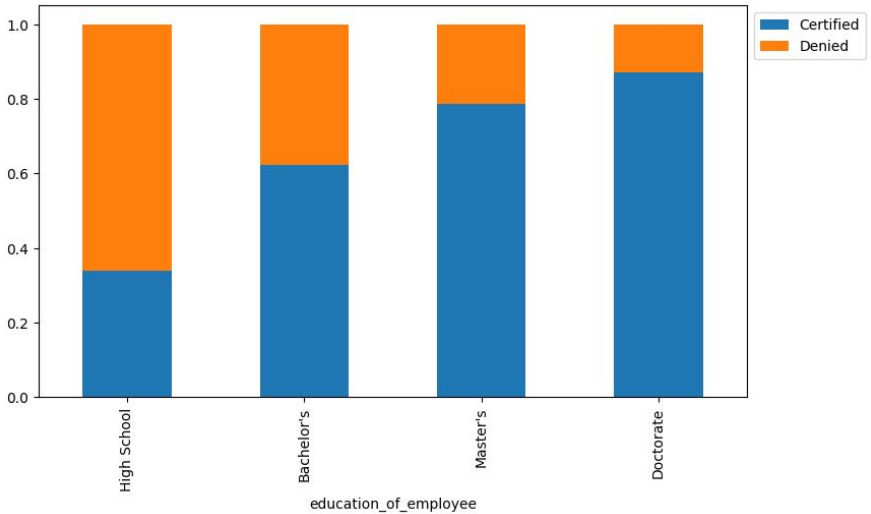
- **F1 Score: 83.7%**
- **Recall: 89.2%**
- **Precision: 78.9%**

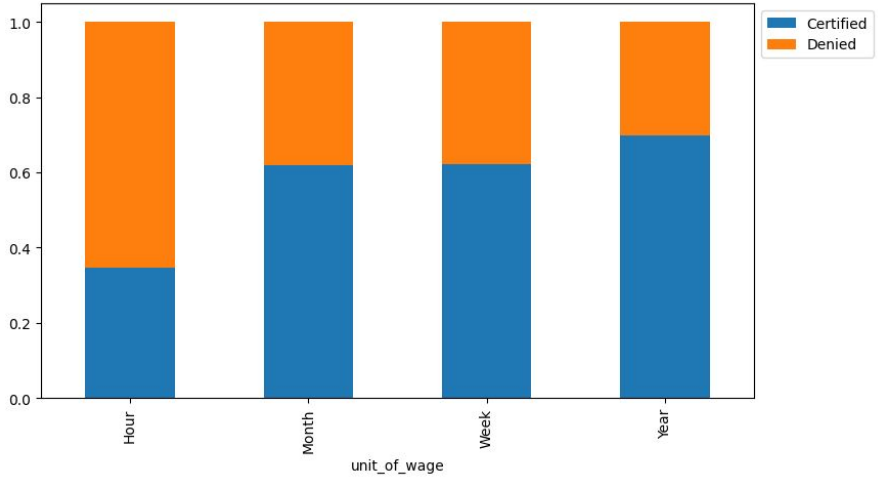


Executive Summary

Actionable Insights

- Applicants with higher education (Bachelor's or above) and prior job experience are much more likely to be certified
- Part-time positions and lower hourly wage roles tend to be denied more often





Executive Summary

Recommendations:

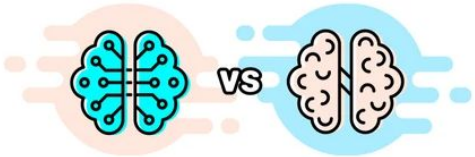
- Use this model to flag likely approvals for faster processing
- Integrate the model's predictions into the review workflow as a priority flag
- Revisit model performance quarterly as new application data comes in

By prioritizing high-recall predictions, we can reduce missed approvals, streamline applicant triage, and focus expert attention where it's needed most — leading to faster, smarter decisions at scale.

Business Problem

Our company processes a large volume of foreign worker visa applications each year. Manually reviewing these applications is time-consuming and resource-intensive.

Our team needed a way to quickly identify strong candidates who are highly likely to be approved — so they could focus their time on edge cases or those requiring closer scrutiny. Without a scalable solution, processing delays and inconsistencies are likely to increase as volume grows.



Solution Approach

We developed a supervised machine learning solution that predicts whether a visa application will be certified or denied based on historical approval patterns.

Our methodology included:

- Exploratory data analysis to uncover approval drivers
- Data cleaning, encoding, and addressing class imbalance using SMOTE and undersampling
- Model training and tuning using cross-validation on four algorithms
- Comparing models on validation and test sets using **recall** and **F1 score**
- Explaining model decisions using feature importance for transparency and stakeholder trust

The final model enables us to **prioritize high-likelihood approvals**, reduce manual review time, and support faster, more consistent decisions.

EDA Results: Key Findings

- **Certified applications made up ~66%** of the data
- Applicants with **Bachelor's and Master's degrees** had significantly higher approval rates than those with a high school education
- **Prior job experience** was a major positive signal — applicants with no experience were much more likely to be denied
- Most certified roles were **full-time**, and part-time positions had a lower likelihood of approval
- **Higher prevailing wages** correlated with more certifications
- **Wage unit mattered**: hourly wage applicants were less likely to be certified compared to those with yearly or monthly wages
- Some patterns were observed by **region and continent**, but education and experience were far stronger drivers

EDA Results: Insights

- ✅ Does education level impact visa approvals?

  Yes — higher education levels lead to significantly more approvals.

- ✅ Does job experience matter?

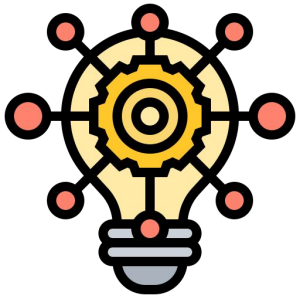
  Yes — applicants with job experience are far more likely to be certified.

- ✅ Are full-time roles more likely to be approved?

  Yes — full-time positions have higher certification rates.

- ✅ Does wage or wage unit affect outcomes?

  Yes — higher wages and non-hourly wage types (yearly, monthly) are tied to higher approval rates.



Data Preprocessing

Before modeling, we cleaned and transformed the dataset to ensure quality, consistency, and compatibility with machine learning workflows.

✔️ **Duplicate Check** - No duplicate records found

✔️ **Missing Values** - Very few missing values and dropped rows with missing data

✔️ **Outliers** - Found negative values in employee count and fixed by converting to absolute values

✔️ **Feature Engineering** - Dropped case_id, one-hot encoded categorical features, and converted target column to 0 = Denied, 1 = Certified

✔️ **Modeling Prep** - Used **SMOTE** and **undersampling** to handle class imbalance, split data: 70% Train, 27% Validation, 3% Test, and applied all cleaning steps consistently across all splits

Model Performance Summary

After testing and tuning multiple models, the final model selected was a **tuned XGBoost classifier** trained on oversampled data using SMOTE.

🔧 Final Model & Parameters

- n_estimators = 200
- learning_rate = 0.1
- subsample = 1.0
- gamma = 5
- scale_pos_weight = 1 (due to SMOTE-balanced data)

Model Performance Summary

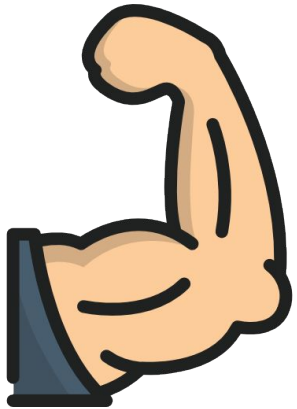
📌 Most Important Features

1. Education level (especially High School, Bachelor's, and Master's)
2. Job experience (Yes/No)
3. Unit of wage (Hourly, Monthly, Yearly)
4. Full-time position (Y/N)
5. Region of employment

Model Performance Summary

📊 Key Performance Metrics

Metric	Training Data	Test Data
Accuracy	81.4%	76.9%
Recall	87.1%	**89.2%**
Precision	78.1%	78.9%
F1 Score	82.4%	**83.7%**



The model generalizes well with **strong recall and F1** on unseen data, making it reliable for use.

APPENDIX

Data Background and Contents

The dataset includes over 32,000 historical visa application records processed by the Office of Foreign Labor Certification (OFLC). The key columns are as listed:

- education_of_employee: Applicant's education level
- has_job_experience: Whether the applicant has prior job experience
- requires_job_training: If the applicant needs additional job training
- prevailing_wage: Wage offered for the position
- unit_of_wage: Wage unit (Hourly, Monthly, Yearly, Weekly)
- full_time_position: Full-time vs. part-time indicator
- region_of_employment: Location of job within the U.S.
- continent: Origin of the applicant
- case_status: Target column (Certified or Denied)

Model Building - Bagging

We first tested a baseline **Decision Tree classifier** and then explored **bagging methods** to reduce variance and improve generalization.

Models Tested

- **Decision Tree:** Basic classifier with limited generalization
- **Bagging Classifier:** Ensemble of decision trees using bootstrap sampling
- **Random Forest:** Extended bagging with random feature selection at each split

Performance Summary (Validation F1)

- Decision Tree: **74.4%**
- Bagging Classifier: **77.4%**
- Random Forest: **79.7%**

The ensemble methods (Bagging and Random Forest) clearly outperformed the standalone decision tree, especially on recall and F1 score.

Model Improvement - Bagging

To improve performance further, we applied **hyperparameter tuning** to the Random Forest model.

Tuning Parameters

- n_estimators (number of trees)
- min_samples_leaf (minimum samples per leaf node)
- max_samples (subsampling rate per tree)
- max_features (number of features per split)

Before vs. After (F1 Score on Validation)

- Untuned Random Forest: **79.7%**
- Tuned Random Forest: **80.2%**

The improvement was modest, but tuning helped stabilize the model and improve consistency without overfitting.

Model Building - Boosting

We also tested two boosting models — **AdaBoost** and **Gradient Boosting** — which build trees sequentially to correct previous errors.

Models Tested

- **AdaBoost:** Focuses on misclassified samples and adds weak learners
- **Gradient Boosting:** Optimizes a loss function at each stage of training

Performance Summary (Validation F1)

- AdaBoost (Untuned): **79.7%**
- Gradient Boosting (Untuned): **81.4%**

Both models outperformed bagging methods, with Gradient Boosting giving the best untuned results.

Model Improvement - Boosting

We used **RandomizedSearchCV** to tune key hyperparameters for both AdaBoost and Gradient Boosting, which significantly boosted performance.

Key Tuning Improvements

- Gradient Boosting:
  - n_estimators, learning_rate, max_features, subsample
  - Validation F1 improved from **81.4%** → **82.3%**
- AdaBoost:
  - Tuned number of estimators and learning rate
  - Validation F1 improved from **79.7%** → **81.5%**

*We also experimented with a **stacking classifier** combining the outputs of multiple models, but **XGBoost still outperformed the stack** in F1 score and generalization, so it remained the final model.*

Great Learning

Happy Learning !