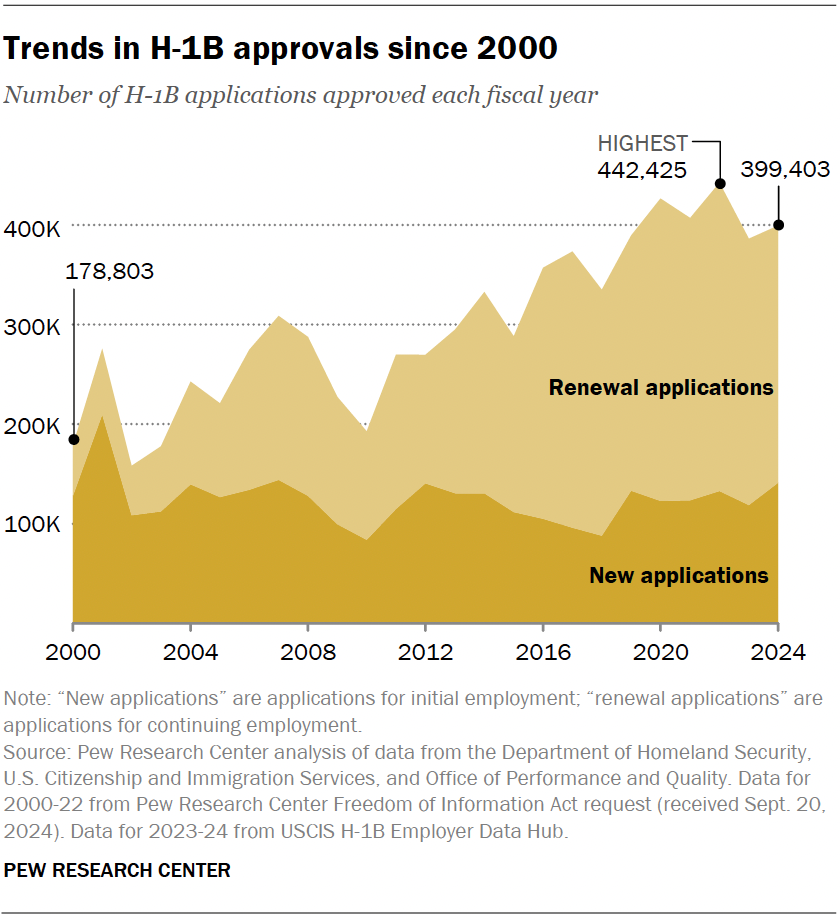
**Predicting H1B Visa Application Outcomes Using LCA Data (2020-2024)**

**Delphin Kaduli**

[kaduli@cua.edu](mailto:kaduli@cua.edu)

## Background

The H-1B visa is a non-immigrant program that allows U.S. companies to employ foreign workers in specialty occupations requiring specialized knowledge and at least a bachelor’s degree or higher in a related field. Since 2004, approximately 85,000 H-1B visas have been granted annually by the U.S. immigration authorities. In recent years, the demand for H-1B workers has steadily increased. However, the complexity and uncertainty of the selection process have raised concerns among employers and applicants regarding the value and risks associated with pursuing sponsorship.

In this project, I developed a predictive model to estimate the outcome of H-1B visa petitions. The model achieved strong performance, with an F1-score and recall of 0.98, and a precision of 0.97. This predictive tool can serve as a valuable resource for future applicants and sponsoring employers, helping them make more informed decisions in a highly competitive and uncertain process.

## Problem Statement﻿

The H-1B visa petition process is highly competitive and often unpredictable, creating significant uncertainty for both employers and prospective employees. Despite the critical impact that approval or denial can have on workforce planning and career opportunities, there is limited transparency into the factors that influence petition outcomes.

This project addresses the question: **Can we develop a machine learning model capable of accurately predicting the outcome of H-1B visa petitions based on historical application data?**  
 By leveraging data-driven techniques, the goal is to identify patterns and insights that could provide valuable guidance to employers and applicants navigating the H-1B process.

## Methodology

## This project was conducted following the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a widely used methodology that structures data science projects into six clearly defined phases: 1. Business Understanding

## The primary objective of this project was to develop a machine learning model capable of predicting the outcome of H-1B visa petitions (Certified or Denied) based on historical Labor Condition Application (LCA) data. This predictive tool aims to provide greater transparency and assist employers and applicants in assessing the likelihood of petition approval

## Data Understanding

The dataset used in this project was collected by scripting downloads directly from the official U.S. Department of Labor (DOL) website.  
Quarterly LCA disclosure files from fiscal years 2020 to 2024 were scripted with python, ensuring the use of authentic and up-to-date records.  
Initial exploration included:

* Reviewing data attributes (employer, job title, salary, location, case status, etc.)
* Identifying data types (categorical, numerical, dates)
* Detecting missing values and inconsistencies
* Analyzing class imbalance between approved and denied cases

This phase provided critical insights into the structure, quality, and challenges within the collected data.

## 3. Data Preparation

Comprehensive data cleaning and preparation steps were undertaken:

* Filtering to retain only H-1B petitions
* Dropping irrelevant or redundant columns
* Handling missing values appropriately
* Engineering new features (e.g., processing time, annual wage normalization)
* Encoding categorical variables
* Splitting the dataset into training, validation, and test sets
* Scaling numerical features where necessary Special care was taken to address class imbalance and preserve the integrity of important features.

## 4. Modeling

Two primary machine learning models were trained and evaluated:

**Logistic Regression**

Logistic regression uses the sigmoid function as its hypothesis as given by equation (1), The key assumption behind logistic regression for binary classification problems is illustrated using equation (2). When training the model, logistic regression attempts to find theta parameter that maximizes the log-likelihood of the data as given by equation (3). Logistic regression thus gives a conditional distribution of response Y given predictor variables represented in matrix x. Furthermore, Logistic regression can be thought of as a generalized linear model in which the link function given by equation (3) relates the response variable Y to the linear model.

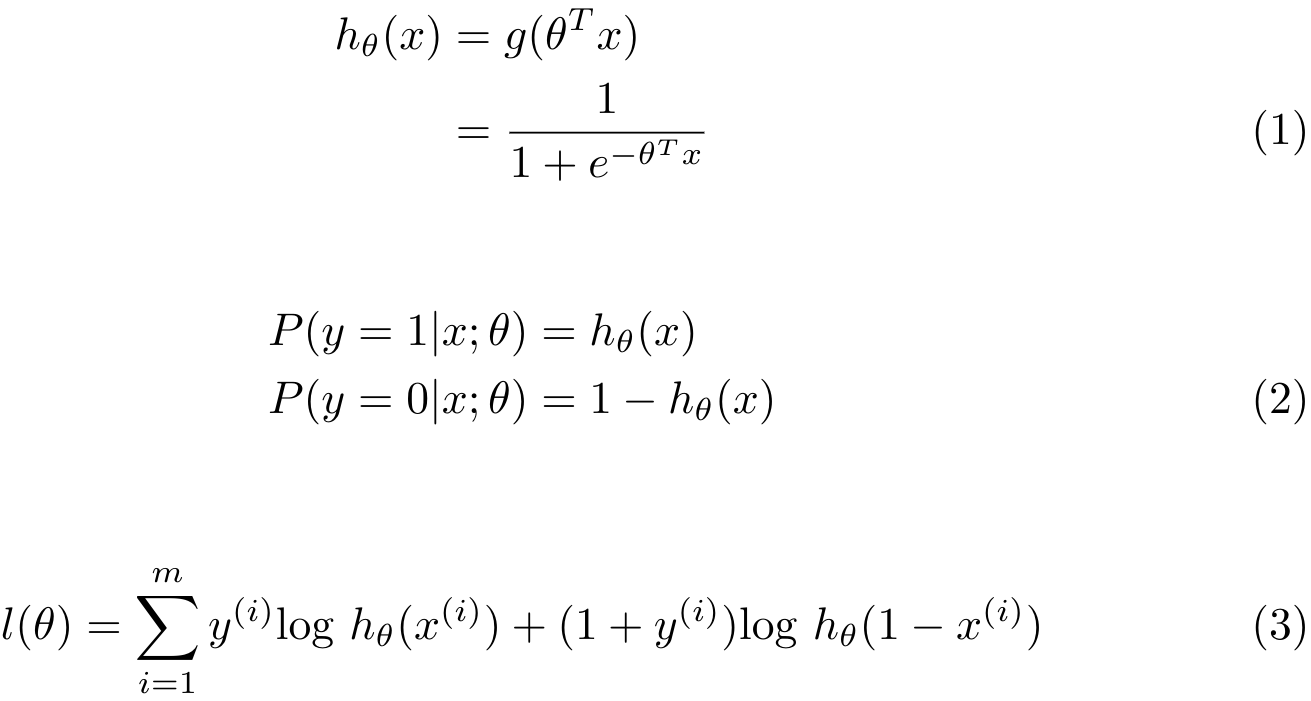


Figure : Logistic Regression formulas

**Gradient Boosting Classifier**

A white sheet with black text

AI-generated content may be incorrect.As the name suggests, gradient boosting is an ensemble boosting technique in which predictions are made from a collection of individual so-called, “weak learners”. The individual models are iteratively trained on a subset of the training set and gradient descent is used to minimize a loss function sequentially. This allows the model to learn from past mistakes and help improve the generalizability of the model. Following is the general gradient boost algorithm:

Figure : Gradient boosting formulas

Initial training was performed with default parameters to establish baseline performance.  
Hyperparameter tuning was subsequently applied to the Gradient Boosting model using GridSearchCV to optimize its predictive power while controlling training time.

**Evaluation Metric**

Since the objective of the model is to correctly predict the outcome of the H-1B visa petition a high true positive rate and low false negative is desired in our prediction. This is to ensure that an applicant is provided with a likelihood of the true or accurate outcome of their application rather than being given false outcome. **recall**, commonly known as a sensitivity, which measures the true positive rate and **precision** which minimizes the false negative, are suitable evaluation metrics.

To capture both recall and precision in a single metric, the F1 score is utilized to optimize my models. A high F1 score indicates that the model has high recall and high precision. It is important to note that accuracy as an evaluation metric when dealing with high-class imbalance can be miss-leading due to inherent bias present in the dataset.

Since the dataset contains high class-imbalance characteristics, the average prediction outcome cannot be used as a baseline metric. As such, I used Logistic Regression without regularization as my baseline model, with a 0.74 F1 score. The success of my model was evaluated based on whether they can outperform the baseline model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Baseline Models** | | | | |
| **Classifier** | **F1-Score** | **Recall** | **Precision** | **AUC-ROC** |
| **LogisticRegression** | **0.74** | **0.60** | **0.99** | **0.57** |
| **GradientBoosting** | **0.91** | **0.84** | **0.99** | **0.71** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuned Model** | | | |
| **Classifier** | **F1-Score** | **Recall** | **Precision** |
| **GradientBoosting** | **0.96** | **0.97** | **0.97** |

## 

## Results

## 

Performance of baseline models Performance of tuned best model

After performing hyperparameter tuning and addressing class imbalance through under sampling (adjusting the Certified-to-Denied ratio from 99:1 to approximately 75:25), the performance of the models improved noticeably.

Among the classical models, the **Gradient Boosting Classifier** demonstrated the best results, achieving an **F1-score of 0.96**, a **Recall of 0.97**, and a **Precision of 0.97** on the test set. Compared to the baseline model, the tuned Gradient Boosting model showed a clear improvement in both minority (Denied) and majority (Certified) class predictions.

**Conclusion**

In this project, it was shown that machine learning models can effectively predict the outcome of H-1B visa petitions based on application attributes. Classical models including Logistic Regression, and Gradient Boosting were trained, tuned, and evaluated. After addressing class imbalance and optimizing hyperparameters, the Gradient Boosting model demonstrated the strongest overall performance.

These results suggest that with careful data preprocessing, balancing, and tuning, classical models can achieve strong predictive power in this domain, providing valuable decision support for employers and applicants navigating the H-1B visa process.

## Future Recommendation

To further enhance the predictive performance and practical value of the H-1B visa outcome model, the following future directions are recommended:

* **Expand the Dataset**: Incorporate additional predictors such as applicants’ educational background, years of work experience, and employer sponsorship history to improve the richness and predictive power of the dataset.
* **Integrate Recent Data**: Extend the dataset by including more recent application records beyond 2024 as they become available, allowing the model to remain up-to-date with current visa approval trends.
* **Develop a Web Application**: Build an interactive web application where future H-1B applicants or employers can input their petition details and receive real-time predictions based on the trained model, making the solution accessible and actionable.

These enhancements will not only improve model accuracy but also support broader real-world usage for applicants, employers, and immigration consultants.

## References

1. Pew research center, [“what we know about the U.S. H-1B visa program”](https://pewrsr.ch/4h8Dxvj), CarolybeIm, Sahra M.
2. “Data mining for business analytics concepts techniques and-applications in python”, Galit S, Peter C. B, Peter G, NItin Patel.