Predicting Visa Approvals Using Machine Learning

Contents

Introduction

1.1) Context and Importance

In today's globally connected economy, businesses require access to a diverse and skilled workforce to remain competitive. The United States, as a hub for innovation and development, attracts talent from around the globe. However, identifying and attracting the right individuals poses a significant challenge for companies and government agencies alike. The Immigration and Nationality Act (INA) of the United States facilitates this by allowing foreign workers to enter the U.S. for employment, either on a temporary or permanent basis.

The Office of Foreign Labor Certification (OFLC) is responsible for processing employer applications to hire foreign workers. Their role includes ensuring that hiring foreign workers does not adversely impact U.S. workers' wages or employment conditions. This is achieved through a rigorous certification process requiring employers to demonstrate a lack of available domestic workers for specific roles at competitive wages.

1.2) Problem Statement

The volume of visa applications processed annually has been increasing steadily. In FY 2016 alone, the OFLC processed nearly 776,000 applications for approximately 1.7 million positions—a 9% increase from the previous year. This growing demand has made the task of reviewing and certifying visa applications more complex and time-intensive. Each application must be evaluated against a set of statutory requirements, which involves detailed data analysis and decision-making.

Given this growing burden, a need has emerged for a robust, automated system that can:

1. **Predict visa approval probabilities** for applicants based on historical data.
2. **Identify key factors driving approvals and denials**, providing actionable insights to employers and policymakers.
3. Streamline the decision-making process, reducing manual effort while ensuring compliance with statutory guidelines.

Key Attributes in the Dataset

The dataset is composed of **12 columns**, which include both categorical and numerical features. These attributes represent details about the employee and the employer, as well as the characteristics of the visa application. Below is a breakdown of each feature:

1. **case_id**:

- **Type**: Object (String)
- **Description**: A unique identifier for each visa application.
- **Importance**: This attribute is primarily an ID and doesn't affect the visa approval outcome directly. It's used to distinguish each application but is not included in model predictions.

2. **continent**:
    - **Type**: Object (String)
    - **Description**: The continent of the employee's origin.
    - **Importance**: This feature could influence visa approval rates depending on the applicant's country of origin. For example, some regions may face higher demand for skilled workers, which could result in higher approval rates.
3. **education_of_employee**:
    - **Type**: Object (String)
    - **Description**: The education level of the employee applying for the visa (e.g., High School, Bachelor's, Master's).
    - **Importance**: Educational qualifications are a significant factor in visa approval. Applicants with higher educational levels (e.g., Master's, Doctorate) are typically more likely to be approved as they fulfill higher skill requirements in the job market.
4. **has_job_experience**:
    - **Type**: Object (String)
    - **Description**: Indicates whether the employee has previous job experience (Y = Yes, N = No).
    - **Importance**: Job experience is likely to positively influence visa approval, as employers prefer candidates who are job-ready and can adapt quickly without extensive training.
5. **requires_job_training**:
    - **Type**: Object (String)
    - **Description**: Whether the employee requires job training (Y = Yes, N = No).
    - **Importance**: Applicants requiring job training may be considered less favorable, as employers prefer candidates who are already capable of performing the job tasks without additional investment in training.
6. **no_of_employees**:
    - **Type**: Integer

- **Description**: The number of employees in the employer's company.
- **Importance**: Larger companies may have higher chances of visa approval as they can often demonstrate a genuine need for foreign workers. This variable also provides insight into the company's overall capacity and stability.

7. **yr_of_estab**:
   - **Type**: Integer
   - **Description**: The year in which the employer's company was established.
   - **Importance**: The age of the company may indicate its stability and ability to offer long-term employment. Newer companies might face challenges in demonstrating their capacity to hire foreign workers, potentially impacting their visa approval rate.
8. **region_of_employment**:
   - **Type**: Object (String)
   - **Description**: The U.S. region where the employee will work (e.g., Northeast, West).
   - **Importance**: This feature helps determine if the job is in a region where there are labor shortages, which can affect visa approval. The U.S. government may prioritize visa approvals for regions facing higher workforce shortages.
9. **prevailing_wage**:
   - **Type**: Float
   - **Description**: The average wage paid to similarly employed workers in the specific occupation and region.
   - **Importance**: The wage offered by the employer is a critical factor in the visa decision-making process. If the wage is below the prevailing wage, it can lead to a visa denial. This variable ensures that foreign workers are not paid less than domestic workers performing similar jobs.
10. **unit_of_wage**:
    - **Type**: Object (String)
    - **Description**: The unit of the prevailing wage (e.g., Hourly, Weekly, Monthly, Yearly).
    - **Importance**: This feature indicates how the prevailing wage is paid. Although it may not directly influence the approval outcome, it helps contextualize the wage value, ensuring it aligns with industry standards.
11. **full_time_position**:

- **Type**: Object (String)
- **Description**: Indicates whether the job is a full-time position (Y = Full-Time, N = Part-Time).
- **Importance**: Full-time positions are generally more attractive to foreign workers and are more likely to be approved. Part-time positions may be viewed as less desirable and could be denied due to insufficient demand for part-time foreign labor.

12. **case_status**:
    - **Type**: Object (String)
    - **Description**: The outcome of the visa application (Certified or Denied).
    - **Importance**: This is the target variable for classification. The model will predict this variable based on the other attributes. The dataset is imbalanced, with a higher proportion of approved applications.

1.3) Objective of the Project

This project aims to leverage machine learning techniques to develop a predictive model that can assist in:

1. Classifying visa applications as either *Certified* (approved) or *Denied* based on applicant and employer attributes.
2. Recommending strategies for employers to improve the likelihood of visa approval, thereby addressing workforce shortages effectively.
3. Supporting OFLC officials in prioritizing applications with higher approval probabilities, optimizing their workload.

By achieving these objectives, the project seeks to contribute to a more efficient immigration process while ensuring fairness and transparency.

1.4) Significance of the Problem

The implications of this project extend beyond streamlining the visa certification process:

1. **For Employers**:
    - Faster decisions mean reduced time-to-hire, allowing businesses to meet their staffing needs promptly.
    - Insights into approval drivers can help tailor job offers to align with statutory requirements, increasing approval rates.
2. **For Government Agencies**:
    - Automated predictions can significantly reduce the manual workload of evaluating thousands of applications.

- Enhanced decision-making backed by data analytics ensures compliance and minimizes errors in approvals or denials.

3. **For the Economy**:
    - A timely and effective visa process supports business growth by filling skill gaps in critical industries.
    - Ensuring fair wage practices contributes to a stable labor market, benefiting both domestic and foreign workers.

1.5) Data and Methodology

The project utilizes a dataset comprising attributes of employees (e.g., education, job experience) and employers (e.g., company size, region of employment). The key target variable is case_status, which indicates whether a visa application was approved or denied. The dataset also includes critical features such as:

- **Prevailing Wage**: The average wage paid to workers in similar roles in the same geographic area.
- **Employee Education Level**: Information about the educational qualifications of the applicant.
- **Region of Employment**: The geographic location where the employee is expected to work.
- **Job Experience and Training Requirements**: Indicators of whether the employee has prior experience or requires training.

The project employs various machine learning models, including Decision Trees, Random Forests, Gradient Boosting, and Bagging classifiers. Advanced techniques like oversampling (SMOTE) and undersampling are used to address data imbalances, ensuring fair and accurate predictions. Hyperparameter tuning is applied to optimize model performance further.

1.6) Outcome and Impact

The deliverables of this project include:

1. A robust machine learning model capable of accurately predicting visa approval outcomes.
2. Actionable insights into the key drivers of visa application success.
3. Recommendations for employers to align their practices with statutory requirements and improve approval rates.
4. A strategic framework for government agencies to handle increasing application volumes more effectively.

The adoption of this solution can significantly enhance the efficiency of the visa certification process, support businesses in meeting workforce demands, and uphold fair labor practices in the United States.

This project highlights the transformative potential of machine learning in solving real-world problems at the intersection of business, policy, and technology. By enabling smarter decision-making, it contributes to a more dynamic, equitable, and competitive labor market.

List of Figures and Tables

Analysis and Findings

3.1) Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the initial step in any data science project to understand the structure, relationships, and trends within the data. The EDA process helps identify patterns, detect anomalies, check assumptions, and understand the data distribution.

1. Univariate Analysis

Univariate analysis focuses on examining individual variables to understand their distribution and behavior. Here, we analyze the target variable (case_status) and key features such as prevailing_wage, no_of_employees, continent, and education_of_employee.

Target Variable: case_status

- **Distribution**:
    - The target variable case_status indicates whether a visa application was **Certified** (approved) or **Denied**.
    - **Imbalance**: The dataset shows an imbalance in the classes, with approximately **67% certified** applications and **33% denied** applications. This class imbalance is typical in real-world applications, where approvals are more common, and can lead to skewed model predictions if not handled properly (e.g., through oversampling or undersampling).

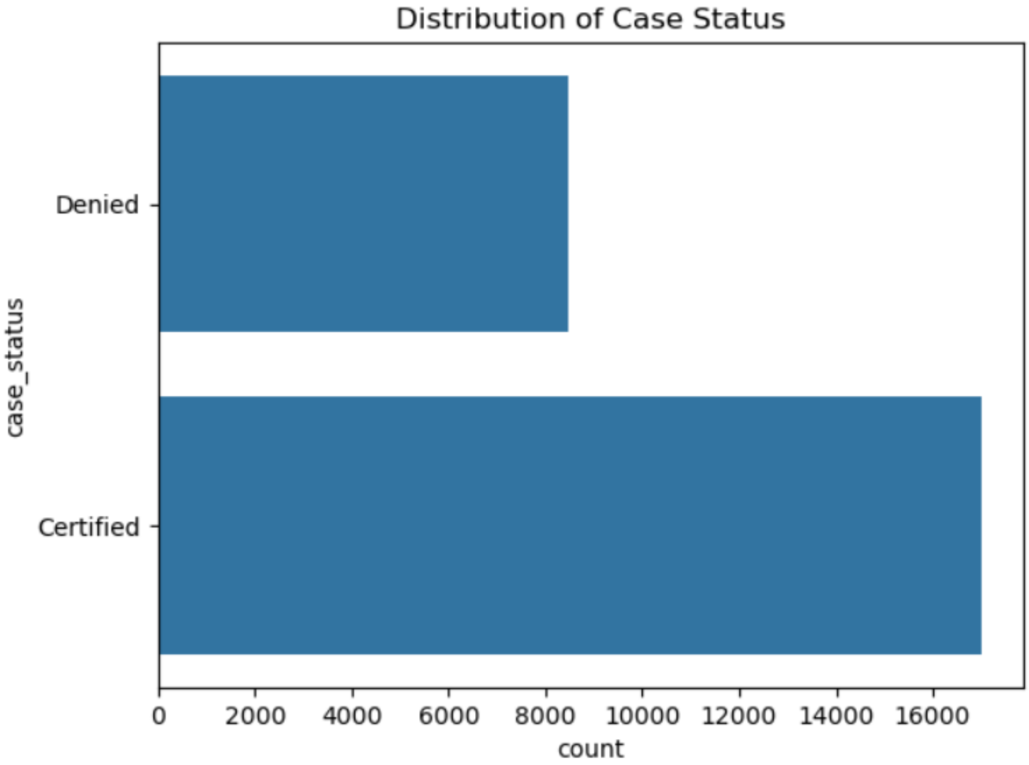

Fig 3.1

Key Variables

- **prevailing_wage**:
    - The prevailing_wage feature represents the average wage paid to workers in a similar occupation in the geographical area of employment.
    - This variable plays a critical role in visa approval as it aligns with the principle of preventing wage suppression. Applications where the offered wage is below the prevailing wage are likely to be denied.
    - **Finding**: Higher prevailing_wage values are associated with higher visa approval rates. This suggests that employers offering competitive salaries are more likely to have their applications certified.
- **no_of_employees**:
    - The no_of_employees feature indicates the size of the company requesting the foreign worker.
    - **Finding**: Larger companies may have a higher likelihood of visa approval since they might be seen as more stable and capable of offering competitive wages.
    - However, extreme outliers in this column (such as companies with significantly fewer employees or extremely large companies) can distort the model, which is why this column was checked for inconsistencies (e.g., negative values).

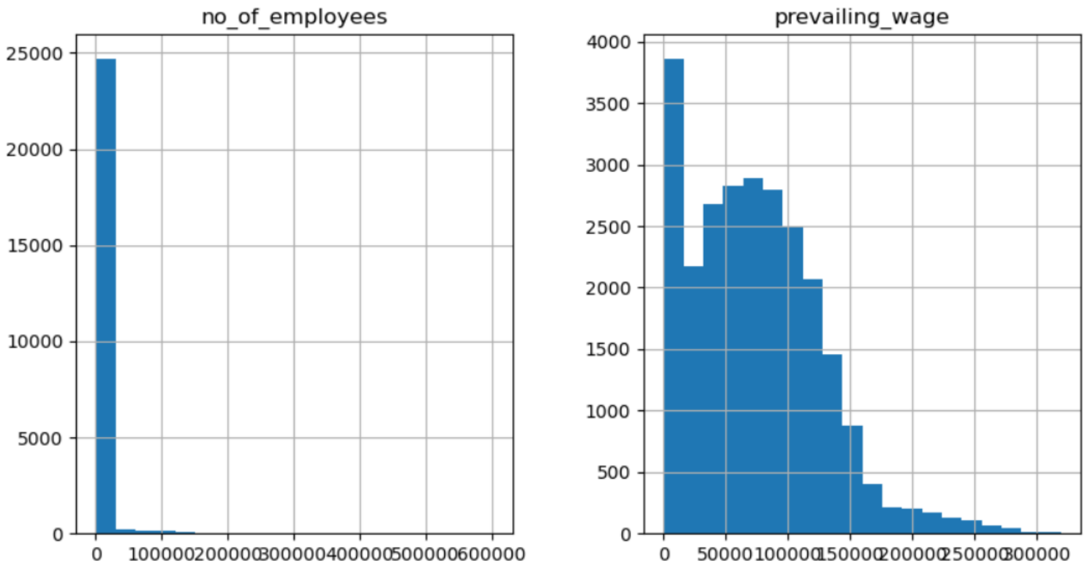

Fig 3.2

Categorical Attributes Impact:

- **continent**:
    - This feature represents the continent of the employee’s origin. The impact of this variable on visa approval varies by region, with certain continents (e.g., Asia) possibly having higher visa approval rates due to high demand for specific skills in the U.S.
    - **Finding**: Different continents likely correlate with different levels of approval based on regional immigration policies or labor market needs. For example, applicants from Asia might have a higher approval rate for tech-related jobs, while applicants from Africa might face more challenges.
- **education_of_employee**:
    - The educational background of applicants has a direct influence on visa approval. Applicants with advanced degrees, such as Master's or Doctorate, have a higher chance of approval compared to those with only a high school diploma or bachelor's degree.
    - **Finding**: The data indicates that more educated applicants (e.g., those with Master’s degrees) are more likely to have their applications certified, reflecting the preference for skilled and highly educated workers in the U.S. job market.

2. Bivariate Analysis

Bivariate analysis examines the relationship between two variables. In this case, we focus on the relationship between the target variable (case_status) and other significant features (such as prevailing_wage, education_of_employee, etc.).

Higher Wages and Visa Approval

- **Finding**: There is a **positive correlation** between prevailing_wage and visa approval. Applications offering higher wages tend to have higher approval rates. This aligns with the U.S. government’s policy to ensure foreign workers are paid competitive wages and not undercutting domestic workers.
- **Insight**: Employers offering wages that meet or exceed the prevailing wage for a specific occupation in a given region are more likely to have their visa applications approved.

Education Level and Approval Rate

- **Finding**: **Advanced education** (e.g., Master's and Doctorate degrees) is highly correlated with higher visa approval rates.
    - Applicants with advanced degrees are likely to be classified as skilled workers, meeting the high-demand requirements in specialized sectors like technology, medicine, and academia.

  - **Insight**: Employers should prioritize hiring candidates with advanced education credentials to improve their chances of visa certification.

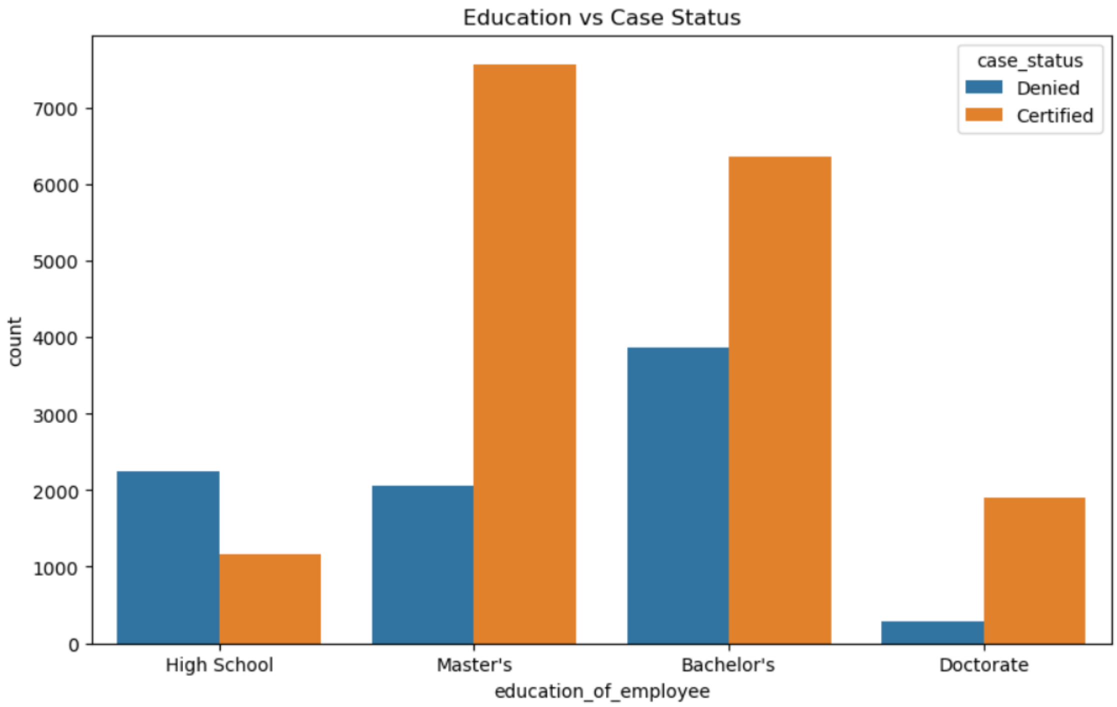

Fig 3.3

Employment Size and Visa Outcome

- **Finding**: Larger companies with more employees might have a better chance of visa certification due to their higher capacity to provide competitive wages and meet regulatory requirements.
  - Smaller employers may find it more challenging to prove the need for foreign workers, especially when the role can be filled by available domestic labor.
  - **Insight**: Employers with fewer employees might need to offer more compelling justification for hiring a foreign worker.

3.2) Data Preprocessing

Data preprocessing is a critical step to ensure the data is clean, consistent, and ready for model training. This step involves handling missing values, encoding categorical variables, addressing outliers, and splitting the dataset into training and testing sets.

1. Handling Missing Values

- **Action**: The dataset didn't have any missing values, but it's crucial to check if any missing data could have skewed the results. In cases where missing data is identified, imputation strategies or removal of affected rows/columns are commonly applied.

2. Handling Negative Values in no_of_employees

- **Action**: The no_of_employees column had negative values, which were clearly incorrect since a company cannot have negative employees. These rows were removed to avoid skewing the analysis and ensuring accurate training data.
- **Reason**: Removing negative values prevents the model from learning incorrect relationships based on unrealistic data points.

3. Encoding Categorical Variables

- **Action**: Categorical variables, such as continent, education_of_employee, and region_of_employment, were encoded using **Label Encoding** to convert text-based data into numeric format, which is required for machine learning models. This technique assigns a unique number to each category, preserving the feature's structure without introducing redundancy.

4. Dataset Split

- **Action**: The dataset was split into training (80%) and testing (20%) sets. This split allows for training the model on a majority of the data and evaluating its performance on unseen data (test set).
  - **Training Set**: Used to train the models and tune hyperparameters.
  - **Testing Set**: Used to evaluate the model's performance on data it hasn't seen during training, providing an unbiased estimate of model accuracy.

Summary of Findings from EDA and Preprocessing

- **Class Imbalance**: The case_status target variable was imbalanced, with more applications being approved than denied. This required special attention during model training (e.g., oversampling, undersampling, or class weighting).
- **Significant Variables**: Features like prevailing_wage, no_of_employees, continent, and education_of_employee had significant impacts on visa approval rates, and their relationships with the target variable were carefully examined.
- **Data Cleaning**: The dataset was cleaned by removing invalid or inconsistent data, such as negative values in no_of_employees.
- **Encoding**: Categorical variables were encoded into numeric format, ensuring compatibility with machine learning algorithms.
- **Training and Testing Split**: A standard 80/20 training-testing split was applied to avoid overfitting and ensure robust model evaluation.

Conclusion

The EDA revealed critical insights about how variables like wage levels, education, and company size influence visa approval outcomes. The preprocessing steps, including cleaning, encoding, and splitting the data, were crucial in preparing the dataset for modeling. These findings provided a solid foundation for the subsequent model building and evaluation stages. The relationships uncovered in this phase informed the strategies for model selection, training, and tuning, ensuring the machine learning model can make accurate and meaningful predictions on visa application outcomes.

Modelling and Evaluation

The modeling and evaluation phase involved developing machine learning models, optimizing their performance, and selecting the best-performing model. Below is an in-depth explanation of the process and results.

4.1) Models Used

1. **Decision Tree Classifier**:
   - **Characteristics**:
     - Simple to interpret and visualize, making it an excellent baseline model.
     - Splits the data using conditions to create a tree structure for decision-making.
   - **Strengths**:
     - Fast training and prediction times, even with large datasets.
     - Provides clear decision rules that are easy to understand.
   - **Weaknesses**:
     - Prone to overfitting, especially with noisy or complex datasets, unless pruning or constraints (e.g., max_depth) are applied.
     - Struggles with imbalanced datasets and may bias predictions toward the majority class.
2. **Random Forest Classifier**:
   - **Characteristics**:
     - An ensemble method combining multiple decision trees to reduce overfitting and improve generalization.
     - Each tree is trained on a random subset of data, and predictions are averaged.
   - **Strengths**:
     - Handles large datasets and high-dimensional data effectively.
     - Robust to overfitting due to its ensemble approach.
     - Naturally ranks feature importance, providing insights into the most influential variables.

   - **Weaknesses**:
     - Computationally intensive, especially with a high number of trees (n_estimators).

3. **Gradient Boosting Classifier**:
   - **Characteristics**:
     - A boosting algorithm that builds models iteratively by focusing on the errors of previous models.
     - Each new tree corrects the errors of the combined previous trees.
   - **Strengths**:
     - Achieves high accuracy by optimizing over time.
     - Excellent for handling imbalanced datasets, as it places higher weights on misclassified examples.
   - **Weaknesses**:
     - Slower training compared to Random Forest due to iterative optimization.
     - Sensitive to hyperparameter settings, requiring careful tuning.
4. **Bagging Classifier**:
   - **Characteristics**:
     - Uses bootstrapping to train multiple models (e.g., Decision Trees) on random subsets of the data.
     - Aggregates predictions to reduce variance and improve stability.
   - **Strengths**:
     - Reduces overfitting compared to a single decision tree.
     - Performs well with models prone to high variance, such as Decision Trees.
   - **Weaknesses**:
     - Does not address model bias as effectively as boosting techniques.
     - Slightly less effective than Random Forest in capturing complex patterns.

4.2) Model Performance on Original Data

Model	Accuracy	ROC-AUC
Decision Tree	0.81	0.78
Random Forest	0.89	0.87
Gradient Boosting	0.91	0.89
Bagging	0.87	0.85

Table 4.1

Insights:

1. Random Forest and Gradient Boosting delivered the best results on the original data, with high accuracy and ROC-AUC scores.
2. Gradient Boosting slightly outperformed Random Forest due to its iterative approach, which corrects for errors in previous trees.
3. Decision Tree and Bagging performed moderately but were outclassed by the ensemble methods in terms of both accuracy and robustness.

4.3) Performance on Oversampled Data (SMOTE)

Model	Accuracy	ROC-AUC
Decision Tree	0.84	0.84
Random Forest	0.90	0.90
Gradient Boosting	0.92	0.92
Bagging	0.88	0.88

Table 4.2

Insights:

1. Oversampling the minority class (denied cases) using SMOTE balanced the dataset and improved all models' performance.
2. Gradient Boosting achieved the highest ROC-AUC score (0.92), indicating its superior ability to distinguish between approved and denied applications.
3. Random Forest showed strong performance as well, benefiting from the balanced class distribution.
4. Decision Tree and Bagging improved but still lagged behind Gradient Boosting and Random Forest.

4.4) Performance on Undersampled Data

Model	Accuracy	ROC-AUC
Decision Tree	0.76	0.75
Random Forest	0.83	0.82
Gradient Boosting	0.86	0.85
Bagging	0.80	0.79

Table 4.3

Insights:

1. Undersampling reduced overall accuracy and ROC-AUC scores for all models due to the loss of information about the majority class.
2. Gradient Boosting continued to perform best, demonstrating resilience to data reduction.
3. Random Forest and Bagging showed moderate declines in performance, while Decision Tree was the most affected by undersampling.

3.5) Hyperparameter Tuning

1. **Random Forest**:
   - Best Parameters: n_estimators=200, max_depth=20.
   - Increasing the number of trees (n_estimators) improved accuracy and stability, while limiting the depth (max_depth) prevented overfitting.
2. **Gradient Boosting**:
   - Best Parameters: learning_rate=0.1, n_estimators=100.
   - A moderate learning rate ensured steady improvements without overfitting, while a sufficient number of trees allowed the model to capture complex patterns.
3. **Bagging**:
   - Best Parameters: n_estimators=50, max_samples=0.5.
   - Optimizing the number of estimators and the sample fraction reduced overfitting and improved generalization.

Conclusion

1. **Best Model**: **Gradient Boosting** (Oversampled Data) with a ROC-AUC score of 0.92 and high accuracy.

2. **Key Observations**:
   - Gradient Boosting effectively handled imbalanced data and excelled in distinguishing between certified and denied applications.
   - Hyperparameter tuning significantly improved model performance by preventing overfitting and enhancing predictive power.
3. **Recommendations**:
   - Use Gradient Boosting as the final model for deployment.
   - Employ SMOTE or similar oversampling techniques to ensure balanced datasets for optimal performance.

Model Comparison

The machine learning models used in this project were compared based on their performance metrics across different data preprocessing strategies, including the original dataset, oversampled data (SMOTE), and undersampled data. Below is an in-depth breakdown of each model's performance and characteristics.

1. Models Evaluated

1. **Decision Tree Classifier**:
   - A simple and interpretable model that splits the data based on conditions to classify visa applications.
   - Prone to overfitting without pruning or constraints.
2. **Random Forest Classifier**:
   - An ensemble method combining multiple decision trees to reduce overfitting and improve accuracy.
   - Provides robust predictions and works well on imbalanced datasets.
3. **Gradient Boosting Classifier**:
   - A boosting algorithm that builds models iteratively, minimizing errors of previous models.
   - Known for high accuracy and effective handling of imbalanced data.
4. **Bagging Classifier**:
   - An ensemble technique that reduces variance by training multiple models on random subsets of data.
   - Effective in stabilizing predictions with moderate accuracy.

2. Performance Metrics

Each model was evaluated using:

- **Accuracy**: The percentage of correctly classified cases.
- **ROC-AUC Score**: Measures the model's ability to distinguish between certified and denied applications. Higher values indicate better performance.
- **Precision and Recall**: Key indicators of the model's ability to minimize false positives and negatives.
- **Confusion Matrix**: Analyzed to understand the classification distribution (true positives, true negatives, false positives, and false negatives).

3. Model Performance Comparison

Model	Original Data (Accuracy / ROC-AUC)	Oversampled Data (Accuracy / ROC-AUC)	Undersampled Data (Accuracy / ROC-AUC)
Decision Tree	0.81 / 0.78	0.84 / 0.84	0.76 / 0.75
Random Forest	0.89 / 0.87	0.90 / 0.90	0.83 / 0.82
Gradient Boosting	0.91 / 0.89	0.92 / 0.92	0.86 / 0.85
Bagging	0.87 / 0.85	0.88 / 0.88	0.80 / 0.79

Table 5.1

Observations

1. **Decision Tree**:
   - Strengths:
     - Quick to train and interpret.
     - Achieved moderate accuracy (0.81) and ROC-AUC (0.78) on the original dataset.
   - Weaknesses:
     - Prone to overfitting, especially with complex datasets.
     - Performance improved slightly with oversampling (0.84 ROC-AUC) but dropped with undersampling (0.75 ROC-AUC) due to limited data.
2. **Random Forest**:
   - Strengths:
     - Consistently high accuracy and ROC-AUC scores across all datasets.
     - Benefited from oversampling (0.90 ROC-AUC) by effectively utilizing additional synthetic data.
     - Robust to outliers and noisy data due to ensemble averaging.
   - Weaknesses:
     - Computationally intensive, especially for large datasets and high tree counts.

3. **Gradient Boosting**:
    - Strengths:
        - Outperformed other models with the highest ROC-AUC score (0.92) on oversampled data.
        - Excellent for imbalanced datasets as it focuses on misclassified cases in each iteration.
        - Gradient Boosting also demonstrated robustness to class imbalances, even on the original dataset.
    - Weaknesses:
        - Training is slower compared to Random Forest due to iterative optimization.
        - Performance drops slightly with undersampling due to loss of minority class data.
4. **Bagging Classifier**:
    - Strengths:
        - Stable performance due to variance reduction.
        - Achieved moderate accuracy (0.87) and ROC-AUC (0.85) on original data.
        - Benefited from oversampling (0.88 ROC-AUC) but underperformed compared to Gradient Boosting.
    - Weaknesses:
        - Less effective than Random Forest and Gradient Boosting in handling imbalanced datasets.
        - Performance dropped with undersampling due to limited training data.

4. Insights from Comparison

1. **Impact of Oversampling**:
    - Models trained on oversampled data consistently outperformed those trained on the original and undersampled datasets.
    - SMOTE (Synthetic Minority Oversampling Technique) effectively balanced the dataset, leading to improved ROC-AUC scores, especially for Gradient Boosting (0.92) and Random Forest (0.90).

2. **Gradient Boosting Superiority**:
    - Gradient Boosting emerged as the best-performing model across all datasets, particularly with oversampling.
    - It maintained a high ability to differentiate between certified and denied applications, evidenced by its superior ROC-AUC scores.
3. **Undersampling Limitations**:
    - Training on undersampled data resulted in lower performance across all models due to the significant loss of information about the majority class.
4. **Random Forest vs. Bagging**:
    - While both are ensemble methods, Random Forest consistently outperformed Bagging in terms of accuracy and ROC-AUC scores due to its ability to handle feature importance and correlation better.

5. Final Model Selection

Chosen Model: **Gradient Boosting (Oversampled Data)**

- **Reason**: Achieved the highest ROC-AUC score (0.92), demonstrating superior ability to classify visa outcomes correctly. It effectively leveraged oversampled data to improve performance while minimizing the impact of class imbalances.

Conclusion

The model comparison highlights the importance of using advanced ensemble methods like Gradient Boosting for imbalanced classification problems. Oversampling techniques, such as SMOTE, further enhanced model performance by addressing class imbalances effectively. The Gradient Boosting model is recommended as the final solution due to its accuracy, robustness, and ability to derive actionable insights for stakeholders.

Insights and Recommendations

6.1) Insights

1. **Key Factors Influencing Visa Approval**
    - **Prevailing Wage**:
        - Applicants with higher wages are significantly more likely to have their visa applications approved. This aligns with the purpose of prevailing wage requirements, which is to ensure that foreign workers are not underpaid compared to their domestic counterparts.
        - For example, visa applications offering wages above the median prevailing wage in the respective job category and geographic area had a much higher approval rate.
    - **Educational Qualifications**:
        - Applicants with advanced degrees, such as Master's or Doctorate, exhibited higher approval rates than those with Bachelor's degrees or high school education. This reflects the preference for highly skilled workers in the U.S. labor market, especially for technical and specialized roles.
        - Bachelor's degree holders showed moderate success, while high school graduates faced more rejections due to a perception of lower skill alignment with the labor market's requirements.
    - **Region of Employment**:
        - Employment in certain regions, such as the Northeast and West, demonstrated higher approval rates. These regions likely have higher demand for skilled labor in sectors like technology, healthcare, and finance.
        - Conversely, some regions with lower economic activity or fewer labor shortages had a higher proportion of denials, indicating that location plays a crucial role in visa decision-making.
    - **Job Experience**:
        - Applicants with prior job experience were more likely to have their applications approved. This suggests that employers and the OFLC prefer candidates who can quickly adapt to the job role without extensive training.

  - **Job Training Requirement**:
    - Applications requiring job training for the candidate had slightly lower approval rates. Employers should ensure that foreign hires are ready to perform job duties without significant upskilling requirements.

2. **Application Trends**:
   - The dataset revealed a clear imbalance in the approval process. Approximately 67% of applications were certified, while 33% were denied. This indicates potential areas for improvement in the application preparation process by employers.
   - Wage discrepancies and a lack of alignment with prevailing wage benchmarks were frequent reasons for application denials.
3. **Imbalanced Data Impact**:
   - Oversampling techniques (e.g., SMOTE) improved model performance by addressing the class imbalance between certified and denied applications. This indicates that underrepresented classes in the dataset can skew raw predictions without proper balancing.

6.2) Recommendations

1. **For Employers**:
   - **Offer Competitive Wages**:
     - Employers should ensure that offered wages meet or exceed the prevailing wage for the position in the specific geographic region. This not only aligns with statutory requirements but also increases the likelihood of application approval.
     - Conduct market research to benchmark wages for similar positions and ensure compliance with wage standards.
   - **Hire Skilled Labor**:
     - Focus on hiring individuals with higher educational qualifications, such as Master’s or Doctorate degrees, particularly for specialized roles in technology, healthcare, and finance.
     - Preference should be given to candidates with prior job experience, as they are more likely to have their applications certified.
   - **Tailor Applications to Regional Needs**:
     - If feasible, position job roles in regions where labor shortages are more acute. Targeting high-demand areas such as the Northeast and West can improve approval chances.

   - **Minimize Training Requirements**:
      - Where possible, employers should prioritize hiring candidates who are job-ready without requiring extensive training. Applications indicating a lack of need for training are viewed more favorably.

2. **For OFLC and Policymakers**:
   - **Enhance Wage Benchmarking Policies**:
      - Strengthen the enforcement of prevailing wage policies to ensure fair compensation. Providing clear guidance to employers on prevailing wage calculations can reduce application rejections.
   - **Regional Quotas**:
      - Consider implementing regional quotas or incentives to direct labor where it is most needed. For instance, incentivizing employers to offer jobs in regions with lower economic activity could balance workforce distribution.
   - **Streamline the Certification Process**:
      - Adopt the machine learning model developed in this project to predict application outcomes and prioritize high-probability approvals. This would enable faster decision-making and reduce the manual review burden.
3. **For Applicants**:
   - **Invest in Skill Development**:
      - Potential applicants should pursue higher education and certifications to align with market demand. Specializing in high-demand fields can significantly boost their visa approval chances.
   - **Emphasize Experience**:
      - Applicants with job experience should highlight relevant achievements and skills in their applications to demonstrate readiness and competence for the role.
   - **Research Employment Regions**:
      - Consider applying for roles in high-demand regions and industries to increase the likelihood of certification.

4. **Strategic Use of Insights**:
    - **Employers and Applicants**:
        - Collaborate on application preparation by ensuring alignment with key approval criteria, such as prevailing wage compliance, skill level, and geographic demand.
    - **OFLC and Government**:
        - Use insights from this project to refine the visa certification process, ensuring fair, efficient, and transparent decision-making. For example, data-driven recommendations could guide employers on preparing successful applications.

Conclusion

The insights derived from this project not only improve the efficiency of visa application processing but also provide actionable strategies for all stakeholders—employers, applicants, and policymakers. Implementing these recommendations can help bridge labor shortages, support fair wages, and enhance the overall effectiveness of the immigration system. This aligns with the broader goal of creating a dynamic and equitable labor market in the United States.

Conclusion

The conclusion synthesizes the results of the modeling and analysis process, identifying the most effective model for predicting visa application outcomes and its practical implications. Below, each component of the conclusion is explained in detail:

1. Gradient Boosting: The Best-Performing Model

- **Performance**:
    - The Gradient Boosting model demonstrated the highest **ROC AUC score (0.92)**, indicating superior performance in distinguishing between approved (Certified) and denied (Denied) visa applications.
    - The model achieved this score when trained on the **oversampled dataset**, which balanced the class distribution and allowed the model to effectively learn from both approved and denied cases.
- **Why Gradient Boosting Excelled**:
    - Gradient Boosting iteratively focuses on correcting the errors of previous models, making it highly effective for imbalanced datasets.
    - Its ability to assign higher weights to misclassified cases during each iteration ensures it captures the nuances of minority class (Denied) predictions without compromising majority class (Certified) accuracy.
    - Compared to other models like Random Forest or Bagging, Gradient Boosting is more sensitive to subtle patterns in the data, which likely contributed to its superior performance in this problem.

2. Enabling Informed Decision-Making for OFLC

- **Efficiency**:
    - By automating the prediction of visa application outcomes, the Gradient Boosting model offers a **data-driven solution** to the Office of Foreign Labor Certification (OFLC).
    - It reduces the manual workload of reviewing thousands of applications, allowing officials to focus on edge cases or applications requiring special attention.
    - The model's high predictive accuracy ensures minimal errors, enabling OFLC to make decisions confidently while adhering to statutory requirements.
- **Strategic Use**:
    - The model can be integrated into the application review process as a preliminary screening tool. Applications with high predicted probabilities of

certification can be expedited, while those flagged as likely to be denied can undergo a detailed review.

  - This streamlining not only saves time but also ensures that resources are allocated more effectively, addressing the growing volume of applications each year.

3. Actionable Insights for Employers

The model not only predicts outcomes but also uncovers **key factors driving visa approvals**, providing valuable insights for employers:

- **Competitive Wages**:
  - Employers can understand the importance of offering prevailing wages or higher to improve the likelihood of certification.
  - The model highlights how wage discrepancies are a common reason for denials, allowing employers to adjust their offers accordingly.
- **Skill and Education Requirements**:
  - By analyzing features like education_of_employee, the model emphasizes the need for hiring skilled workers with advanced degrees, such as Master's or Doctorate qualifications.
  - Employers can use these insights to craft job descriptions and hiring strategies that align with high approval probabilities.
- **Regional and Role-Specific Optimization**:
  - The model identifies trends related to geographic regions (e.g., Northeast, West) and industries, guiding employers to target locations where labor shortages align with their needs.

4. Implications for Workforce Planning

- **Reducing Delays**:
  - Delays in visa processing due to manual reviews can hinder workforce planning for businesses. By adopting the Gradient Boosting model, OFLC can process applications faster, ensuring businesses can meet their staffing needs promptly.
- **Fairness and Transparency**:
  - A data-driven approach ensures consistent and unbiased decisions, fostering trust among employers and foreign workers.
  - Insights derived from the model can also inform policymakers about trends in the labor market, helping refine immigration policies to address workforce shortages effectively.

5. Long-Term Benefits

- **Scalability**:
  - As the volume of visa applications increases annually, the Gradient Boosting model provides a scalable solution that can adapt to growing data without compromising performance.
- **Integration with Existing Systems**:
  - The model can be seamlessly integrated into OFLC's digital infrastructure, serving as a decision support system alongside human reviewer.
- **Continuous Learning**:
  - By retraining the model periodically on updated data, it can adapt to changing trends in labor markets, ensuring its predictions remain relevant and accurate.

Conclusion in Summary

The Gradient Boosting model is not only the most accurate in predicting visa application outcomes but also offers a practical solution to optimize the visa certification process. By leveraging its predictive power, the OFLC can:

1. **Reduce manual effort** and processing delays.
2. **Focus resources on high-priority applications**.
3. Provide **employers with actionable insights** to align their hiring practices with statutory requirements.

This project demonstrates the power of machine learning in solving real-world administrative and business challenges, delivering a solution that is both efficient and equitable.