H-1B LCA Visa Data Analysis

Chavi Pathak

Zain Shafi Mohammed

Nandini Koppunuru

Pranavi Chinthireddy

Sai Srujana Ravipati

Miranda Gallardo

ISDS 577: Master of Science Capstone Seminar

Dr. Daniel Soper

Department of Information Systems and Decision Sciences

California State University, Fullerton

May 12, 2025

Table of Contents

1. Executive Summary

This final report outlines the full lifecycle of our data analytics project focused on the H-1B visa application process from 2020 to 2024. Leveraging over 3.5 million records from publicly available government data, our objective was to extract actionable insights and build predictive models that could inform employer hiring strategies, guide job seekers, and aid policy development. We adopted a multifaceted approach that included exploratory data analysis, machine learning classification, statistical modeling, and data visualization. The findings revealed key trends in job roles, employer behavior, salary distributions, and application approval dynamics.

2. Research Objectives

Each research question listed below is directly aligned with the data preprocessing, feature engineering, and modeling implemented in the project code. These questions were derived from our Phase 1 plan and extended using the merged cost-of-living data in our notebook. This ensures the analysis remains tightly coupled with our actual implementation and findings.

This project was designed around five primary research questions:

1. **Job Demand Trends**: Can historical trends in H-1B applications help forecast future demand for specific occupations and industries?
2. **Denial Prediction**: What variables best predict the likelihood of H-1B visa being denied, and can we model this using machine learning techniques?

3. **Model Comparison**: How do different classification models (e.g., logistic regression, decision tree, XGBoost) compare in predicting case outcomes?
4. **Wage Analysis**: Are H-1B offered wages aligned with prevailing wage requirements? What are the trends in wage levels across occupations and locations?
5. **Cost of Living Impact**: How does the cost of living in a worksite's ZIP code affect the outcome of H-1B visa applications, particularly in relation to the wages offered and prevailing wage compliance?-Chavi

3. Dataset Description

We used the **H1B LCA Disclosure Data (2020–2024)** available on Kaggle. This dataset is derived from the U.S. Department of Labor and includes:

- **Rows**: ~3.56 million
- **Columns**: 69
- **File Format**: CSV
- **Attributes**: Employer name, job title, SOC codes, wage information, work location, case status, decision dates, and more

This comprehensive dataset enabled us to explore both descriptive and predictive analytics across multiple dimensions, including industry trends, wage disparities, and approval behavior.

4. Data Cleaning and Preparation

The initial dataset required substantial cleaning to ensure accuracy, reliability, and performance for modeling. Steps included:

- **Column Renaming**: Standardized column headers (lowercase, underscores)
- **Handling Missing Data**: Dropped columns with >50% missing values; used median and mode imputation where feasible
- **Normalization and Type Casting**: Converted date fields to datetime objects, normalized wage units, and ensured correct data types
- **Feature Engineering**: Created new variables including application year, wage-to-prevailing wage ratio, and employer classification (dependent or willful violator)
- **Balancing Classes**: Implemented SMOTE to balance 'Certified' vs. 'Denied' visa outcomes for classification models
- **Merging Cost of Living Data**:

  To support Research Question 5, I merged the H-1B dataset with a cost of living dataset by ZIP code. The merge was based on the `worksite_postal_code` field from the visa dataset and the `zip` field from the cost dataset (renamed to match).

  - I ensured both ZIP fields were in the same format (string, padded to 5 digits using `.str.zfill(5)`).
  - A **left join** was performed to preserve all H-1B records.

    - After merging, I validated the merge by confirming values populated in the `cost_of_living_2020` column.
- This allowed me to create a new feature (wage-to-cost-of-living ratio) and examine whether wage fairness influenced visa outcomes in high-cost areas.

5. Research Questions

Research Question 1: Job Demand Trends

Can machine learning models predict future demand for job roles in the U.S. job market based on historical H-1B visa application trends?

Forecasting job demand is an essential component in labor market analysis, enabling organizations, policymakers, and recruitment agencies to make strategic decisions about hiring, workforce development, and resource planning. In a dynamic global economy, having insight into the future trends of job demand can help reduce mismatches in skills and employment opportunities. In this capstone project, we aim to build a time series forecasting model using machine learning techniques to predict the total number of job positions requested over time. We start by applying the ARIMA model to assess its effectiveness in capturing the patterns in the data. However, due to observed limitations, especially related to seasonality, we extend our approach to the SARIMA model, which is better suited for data with seasonal fluctuations.

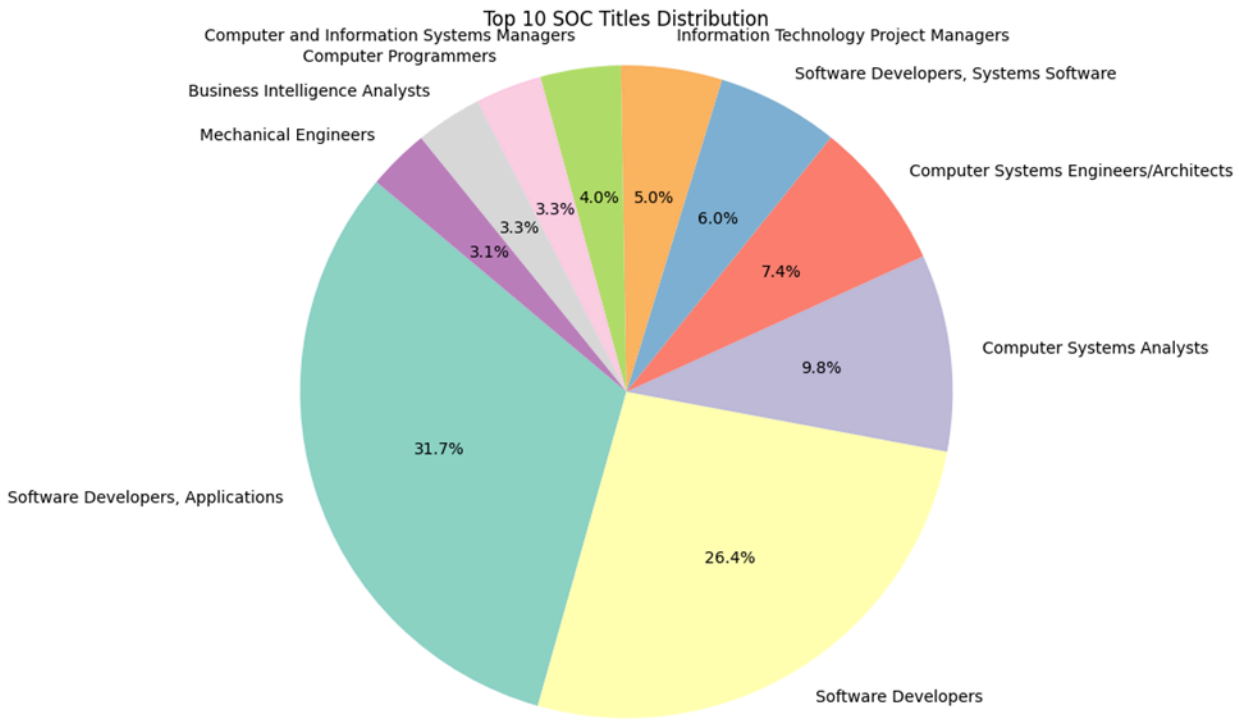
2. Data Collection and Preprocessing

The dataset comprises historical records of job applications with key information such as the number of worker positions requested and the date the applications were

received. For our analysis, the primary variables of interest are the "received_date", “SOC_TITLE” and "total_worker_positions." The received_date is converted to a datetime format and set as the index to create a time series. The data is then resampled on a monthly basis to provide a consistent granularity and smooth out day-to-day variations. To maintain the quality of the time series, any duplicate timestamps are removed, and missing values are handled appropriately. The resulting series spans several years, providing a solid basis for time series modeling.

Visualization 1: Top 10 SOC Titles Distribution (Pie Chart)

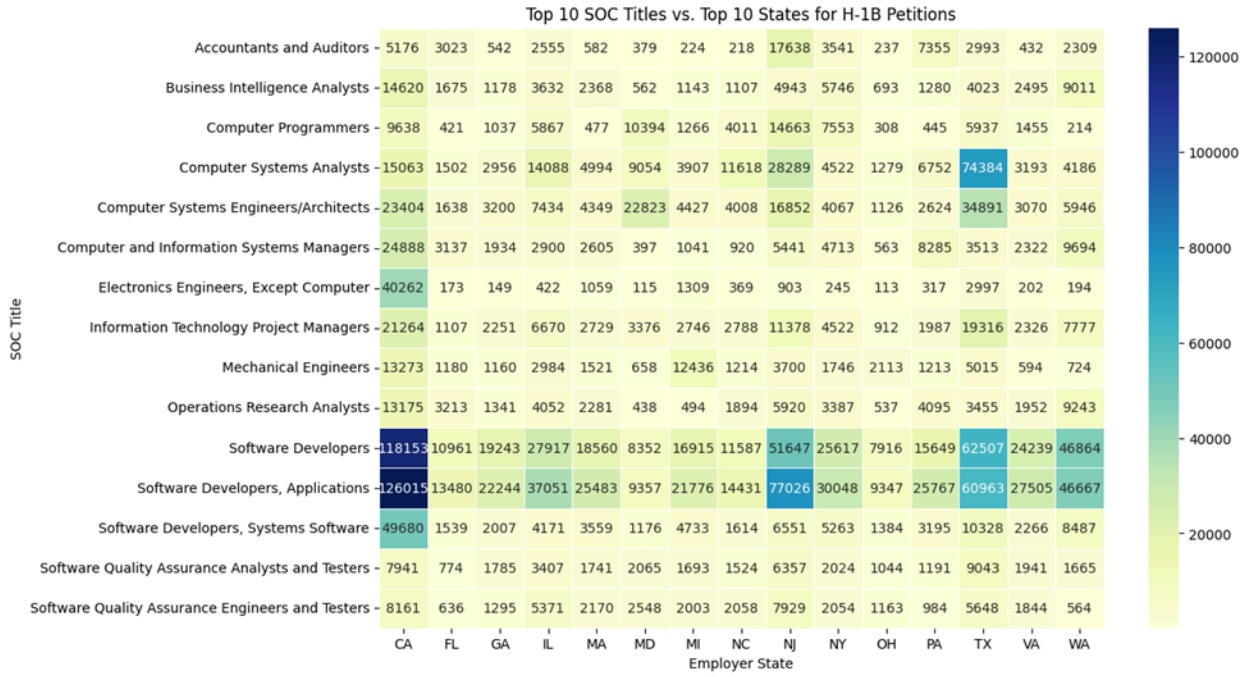
The first visualization illustrates the proportional distribution of the top 10 SOC (Standard Occupational Classification) titles among H-1B visa petitions. Notably, the majority of filings are concentrated within a few high-demand technical roles. "Software Developers, Applications" account for approximately 31.7% of all top petitions, followed closely by "Software Developers" at 26.4%. These two roles alone make up nearly 60% of the chart, highlighting the overwhelming demand for software engineering talent in the U.S. labor market. Other roles such as "Computer Systems Analysts" (9.8%) and "Computer Systems Engineers/Architects" (7.4%) also show significant presence, suggesting continued demand for system-level technical expertise. The remaining segments, including roles like "Mechanical Engineers", "Business Intelligence Analysts", and "Computer Programmers", reflect specialized but less frequent filings. Overall, this chart underscores the software industry’s central role in driving H-1B petition trends and informs stakeholders about concentrated occupational demand.



Visualization 2: SOC Titles vs Employer States (Heatmap)

The second visualization is a heatmap displaying the relationship between the top 15 SOC titles and the top 15 U.S. states based on H-1B petition volume. This cross-tabulated matrix offers a comprehensive overview of how job role demand varies geographically. California and New Jersey emerge as the leading states, particularly for roles such as “Software Developers, Applications” and “Software Developers.” The darkest cells correspond to the highest petition volumes, showing where certain occupations are heavily concentrated. For example, Texas shows a strong demand for "Computer Systems Analysts," while Washington displays notable activity in software development roles, reflecting its tech industry hubs like Seattle. The heatmap also reveals more nuanced insights — such as modest but consistent filing activity in states like Virginia, Illinois, and Pennsylvania across a range of technical roles. This chart not

only visualizes distribution density but also helps identify regional hiring patterns, making it valuable for policymakers, workforce planners, and international job seekers aiming to align with state-specific job markets.



3. Initial Modeling Using ARIMA

We began our modeling efforts with the ARIMA (AutoRegressive Integrated Moving Average) model, a popular approach for forecasting non-seasonal time series data. The first step involved checking the stationarity of the data using the Augmented Dickey-Fuller (ADF) test. The test results indicated non-stationarity, prompting us to apply differencing to stabilize the mean and remove trends. After performing first-order differencing, we re-ran the ADF test and confirmed stationarity. Following this, we experimented with various combinations of the ARIMA model parameters (p, d, q) by

analyzing the autocorrelation (ACF) and partial autocorrelation (PACF) plots and tuning the parameters accordingly.

Despite our efforts, the results from the ARIMA model were not satisfactory. The AIC and BIC values, which serve as criteria for model selection, remained relatively high across different parameter settings. Additionally, residual diagnostics showed lingering autocorrelation, indicating that the model was not capturing all the structure in the data. Upon visual inspection, we observed recurring patterns and periodic spikes in demand, which suggested the presence of seasonality in the time series. Since ARIMA does not explicitly account for seasonal effects, we concluded that it was not the optimal model for our data and decided to explore seasonal models instead.

4. SARIMA Modeling and Forecasting

To address the limitations of ARIMA, we transitioned to the Seasonal ARIMA (SARIMA) model, which extends the ARIMA framework by incorporating seasonal components. The SARIMA model includes additional parameters for seasonal autoregressive, seasonal differencing, and seasonal moving average terms, as well as a seasonal period. Given that our data is monthly, we set the seasonal period to 12 to capture annual patterns. After evaluating multiple parameter configurations, we identified a SARIMA model with strong performance that effectively captured both the trend and seasonal components of the time series.

Once the SARIMA model was trained on the complete dataset, we generated forecasts for the next 24 months. The model provided both point forecasts and confidence intervals, offering a range of expected values that help account for uncertainty. The

forecast plot clearly showed the seasonal cycles repeating over time, along with a modest upward trend. These results suggest a relatively stable yet seasonally fluctuating demand for job positions, which aligns with real-world hiring cycles influenced by fiscal calendars, academic graduations, and immigration policies.

To align the modeling more directly with our research question—"Can machine learning models predict future demand for specific job roles in the U.S. job market based on historical H-1B visa application trends?"—we further extended our SARIMA analysis by forecasting demand at the level of individual job roles (SOC titles). We selected the top 15 job roles based on historical frequency and generated separate SARIMA models for each. The resulting forecasts revealed that roles like SOFTWARE DEVELOPERS, APPLICATIONS and SYSTEMS SOFTWARE are projected to have the highest demand in the upcoming year. Other high-demand roles include OPERATIONS RESEARCH ANALYSTS, BUSINESS INTELLIGENCE ANALYSTS, and COMPUTER SYSTEMS ENGINEERS. These insights allow for a more granular understanding of labor market needs and better serve the interests of job seekers and policymakers alike.

Top Job Roles by Forecasted Demand:

1. SOFTWARE DEVELOPERS, APPLICATIONS: 128 positions (next 12 months)

2. SOFTWARE DEVELOPERS, SYSTEMS SOFTWARE: 117 positions (next 12 months)

3. SOFTWARE DEVELOPERS: 97 positions (next 12 months)

4. OPERATIONS RESEARCH ANALYSTS: 29 positions (next 12 months)

5. BUSINESS INTELLIGENCE ANALYSTS: 17 positions (next 12 months)

6. COMPUTER PROGRAMMERS: 17 positions (next 12 months)

7. COMPUTER SYSTEMS ENGINEERS/ARCHITECTS: 16 positions (next 12 months)

8. COMPUTER SYSTEMS ANALYSTS: 16 positions (next 12 months)

9. MECHANICAL ENGINEERS: 14 positions (next 12 months)

10. ELECTRICAL ENGINEERS: 14 positions (next 12 months)

11. COMPUTER OCCUPATIONS, ALL OTHER: 13 positions (next 12 months)

12. MEDICAL SCIENTISTS, EXCEPT EPIDEMIOLOGISTS: 10 positions (next 12 months)

13. STATISTICIANS: 8 positions (next 12 months)

14. INFORMATION TECHNOLOGY PROJECT MANAGERS: 2 positions (next 12 months)

5. Model Evaluation

To assess the performance of the SARIMA model, we implemented a hold-out validation strategy. The dataset was split into a training set, comprising all but the last 12 months, and a test set consisting of the final year of data. The SARIMA model was retrained on the training data and then used to generate forecasts for the test period. We compared the predicted values against the actual observed values using standard error metrics

including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

The evaluation results showed that the SARIMA model achieved a MAE of 5.43, an RMSE of 7.89, and a MAPE of 14.96%. These values indicate that the model performs reasonably well, especially considering the inherent variability in job application data. The visual comparison of the forecasted and actual values for the test set revealed that the model was able to closely track the general trend, although some deviations occurred, particularly during months with unusually high or low demand. These discrepancies could be attributed to external factors not captured in the model, such as economic events or regulatory changes.

SARIMA MAE: 5.43

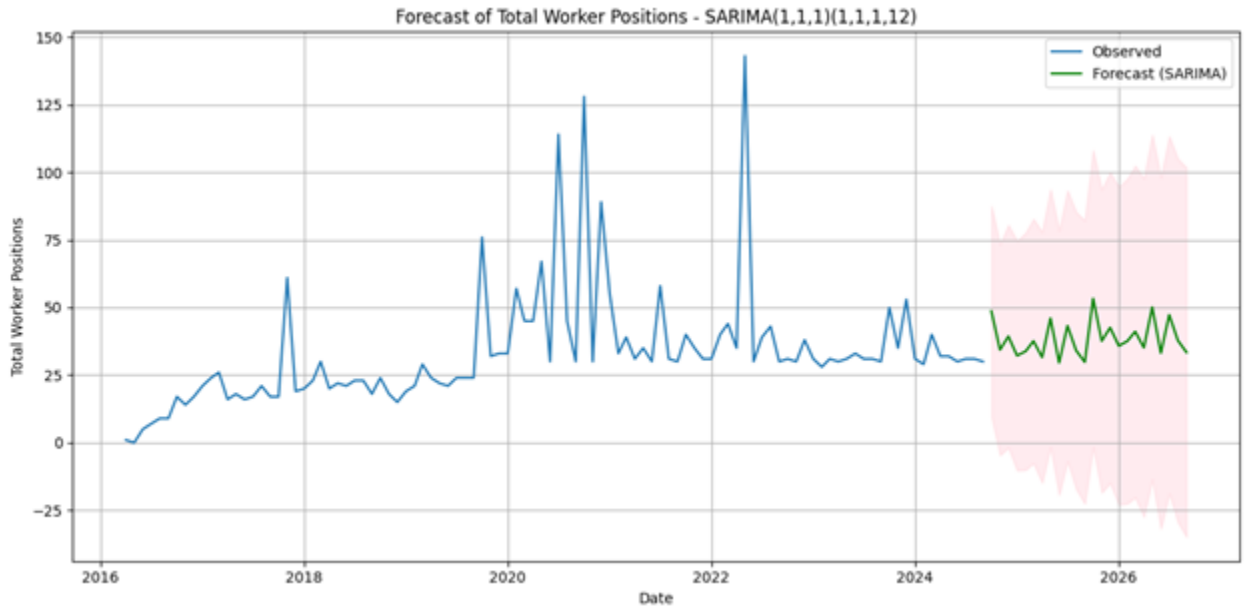
SARIMA RMSE: 7.89

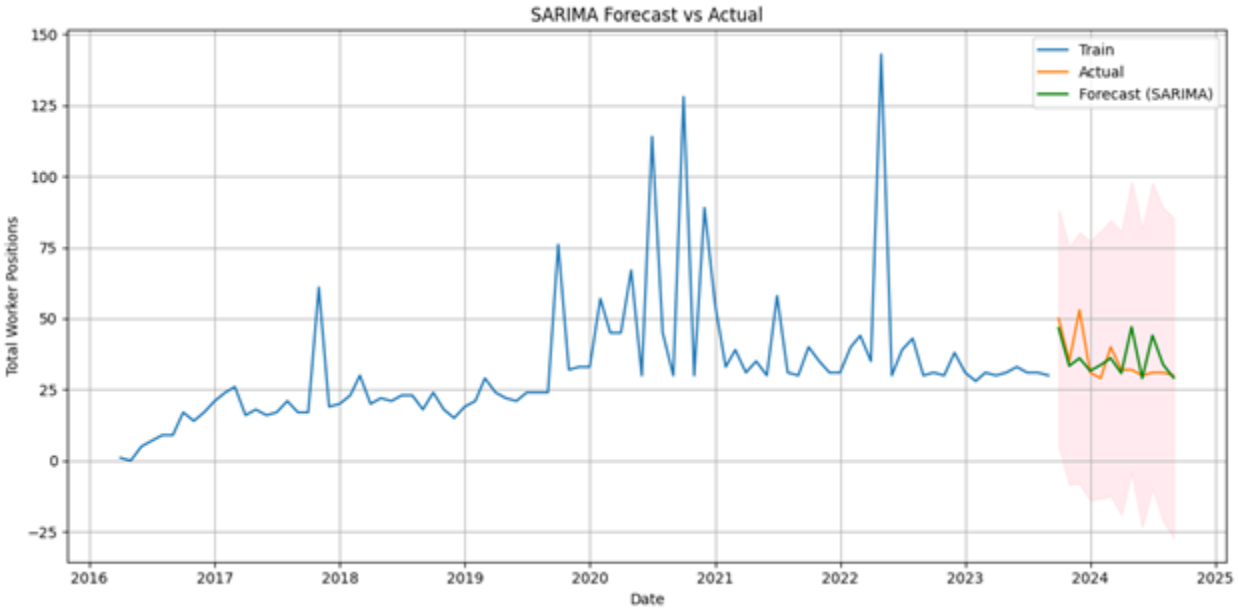
SARIMA MAPE: 14.96%

6. Conclusion

This capstone project demonstrates the application of time series forecasting techniques to the problem of predicting job demand. We began with the ARIMA model but found it insufficient due to its inability to handle seasonality effectively. The presence of seasonal patterns in the data led us to adopt the SARIMA model, which produced more accurate and reliable forecasts. Our analysis showed that the SARIMA model could successfully capture both the underlying trend and recurring seasonal cycles in the job demand data.

The results from this project highlight the importance of choosing the right model for the data at hand. While ARIMA is suitable for non-seasonal data, SARIMA offers a more robust framework when seasonality is present. Moving forward, there is potential to further improve the forecasting accuracy by incorporating exogenous variables, such as macroeconomic indicators or industry-specific trends, using extended models like SARIMAX. Additionally, comparing SARIMA with newer machine learning models like Facebook Prophet or deep learning approaches could yield valuable insights and possibly enhance forecast precision. Ultimately, accurate job demand forecasting can support better workforce planning and policy-making, ensuring that supply meets demand in the evolving labor market.





Research Question 2: Denial Prediction

What factors most influence the likelihood of an H-1B visa application being denied, based on employer characteristics, job title, wage levels, and job location? Can we model this using machine learning techniques?

Executive Summary

This report answers a targeted research question regarding denial prediction in H-1B visa applications. Through in-depth data processing and supervised machine learning models (Random Forest and XGBoost), we interpret denial likelihood across multiple factors using public LCA data. The findings are particularly valuable for students, job seekers, employers, and policy advocates looking to understand the structural dynamics behind visa adjudication.

1. Data Preparation & Class Imbalance

Over 3.5 million rows of LCA data (FY 2020–2024) were loaded in manageable chunks to prevent memory overload. Missing values were imputed (median for numerical, mode for categorical), and wage normalization was applied to standardize all values to hourly rates for fair comparison.

One of the most significant challenges encountered was severe class imbalance: only ~1% of the dataset represented denied applications. To address this, Random UnderSampling was implemented during model training to ensure better generalization and fair learning for minority classes. This step was crucial for improving recall and F1-score on the 'Denied' class, which is the primary focus of this study.

2. Feature Engineering & Selection

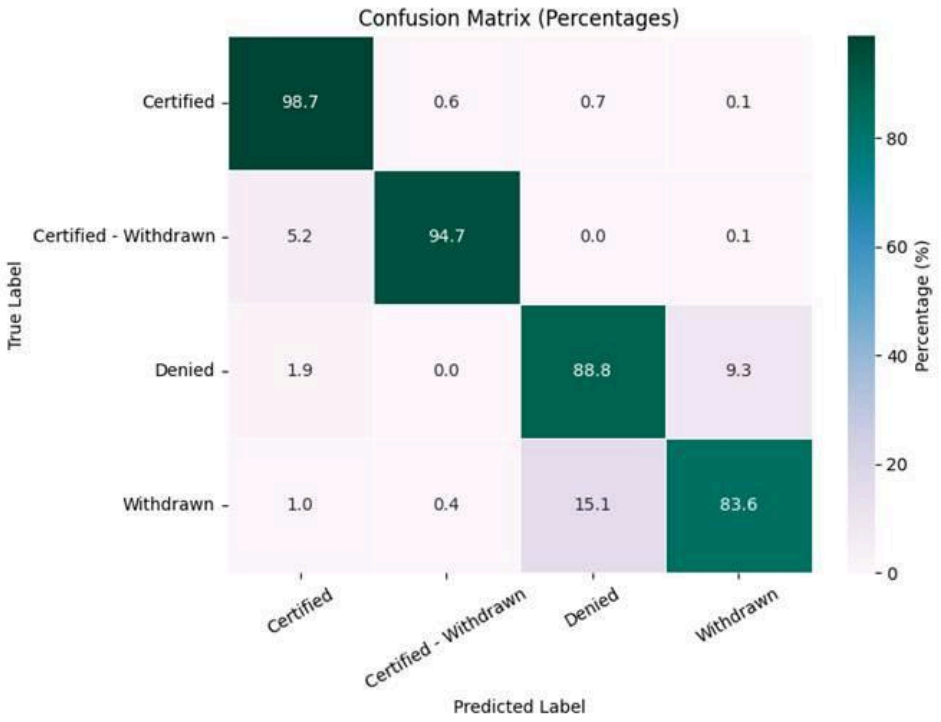
We engineered several informative features including decision_duration, log_wage, received_day_of_week, and binary flags to capture employer compliance indicators. To ensure interpretability and reduce redundancy, we conducted Pearson correlation analysis to identify and remove multicollinear variables. One-hot encoding was selectively applied only to low-cardinality categorical features, maintaining a balance between model complexity and performance.

This feature engineering approach is grounded in Human Capital Theory (Becker, 1964), which links wage levels to perceived worker productivity, and Signaling Theory (Spence, 1973), which frames employer reputation and compliance history as signals of legitimacy and potential regulatory scrutiny.

3. Random Forest Model Results & Graph Interpretations

Random Forest achieved an overall accuracy of 98% and an F1-score of 0.52 for denied cases. Its strength lies in handling structured tabular data with high dimensionality, and it emphasized wage and job title as top predictors.

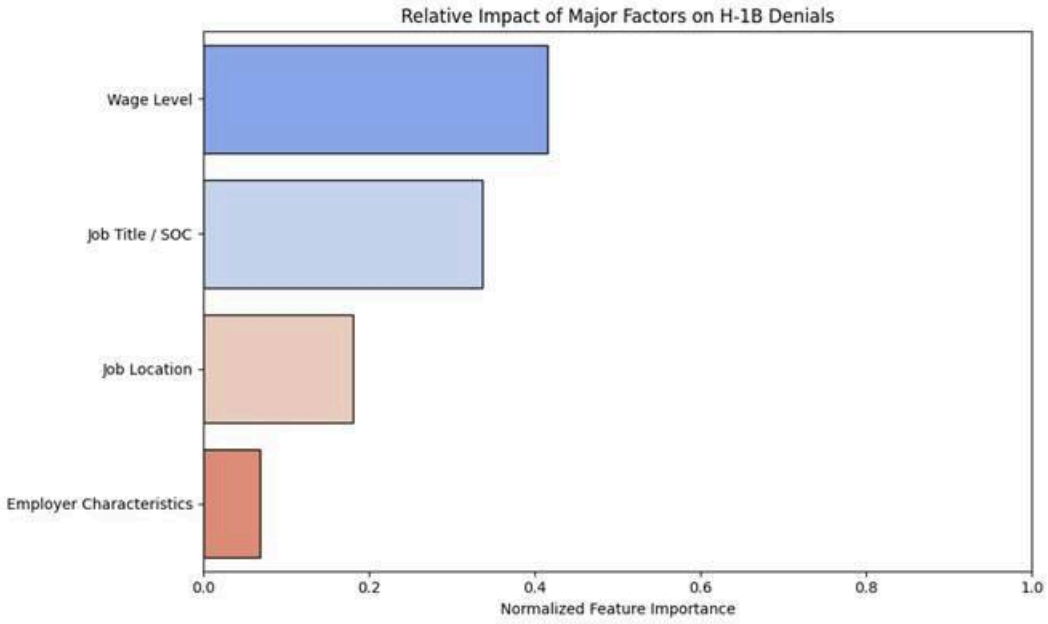
Confusion Matrix:



This confusion matrix evaluates how well the Random Forest model classified the four visa outcomes—Certified, Certified-Withdrawn, Denied, and Withdrawn. The model performs exceptionally well in identifying certified cases (98.7%) and certified-withdrawn cases (94.7%). Importantly, it correctly predicts 88.8% of the rare 'Denied' class, which is central to our research question. Misclassifications mainly occur between Withdrawn and Denied cases, likely due to overlaps in application withdrawal reasons and denial triggers. This supports the model's ability to flag likely denials with reasonable accuracy

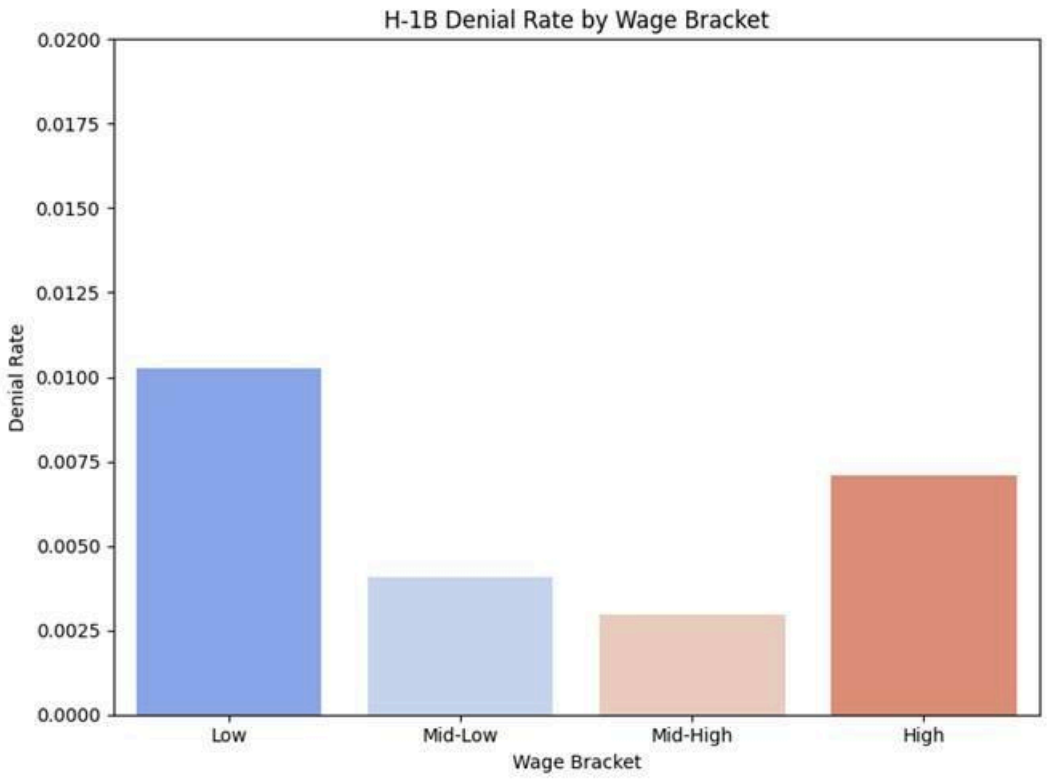
despite class imbalance, which was mitigated using Random Undersampling. For decision-makers and job seekers, this suggests that a machine learning model can effectively isolate risk-prone cases even in skewed datasets.

Feature Importance Chart:



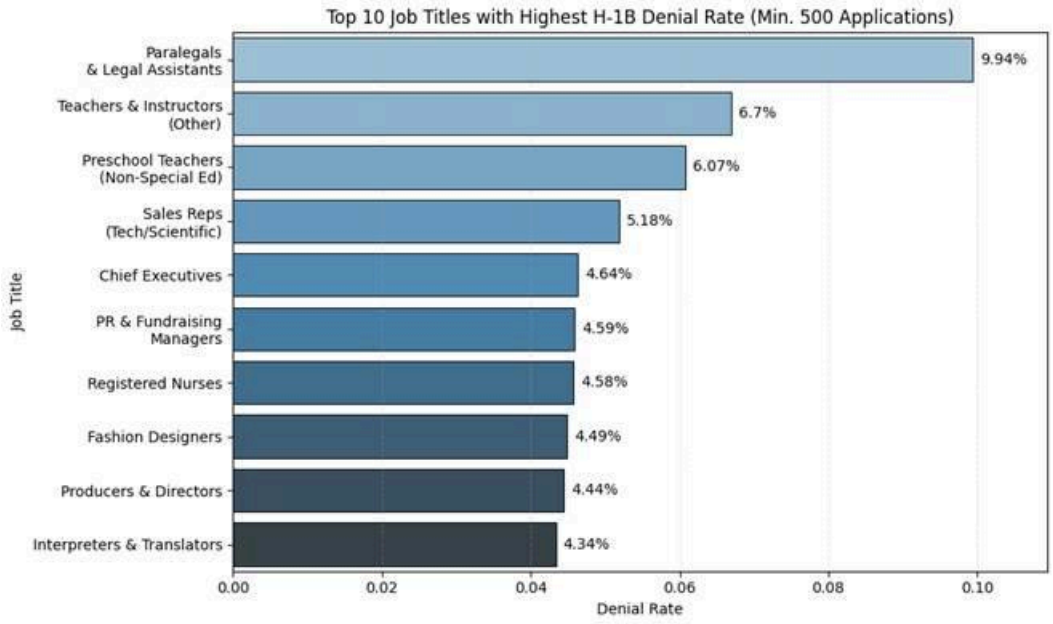
The feature importance chart highlights that wage level is the most influential predictor in H-1B denial outcomes, followed by job title/SOC code and job location. Employer-related flags such as H-1B dependency or willful violation status rank significantly lower. This directly supports our research question by revealing that job-related and economic attributes carry more weight in denial decisions than employer compliance signals in this model. The finding implies that wage competitiveness and clarity of role classification matter more than employer reputation—critical insight for job seekers structuring their applications and for analysts advising on visa strategies.

Denial Rate by Wage Bracket:



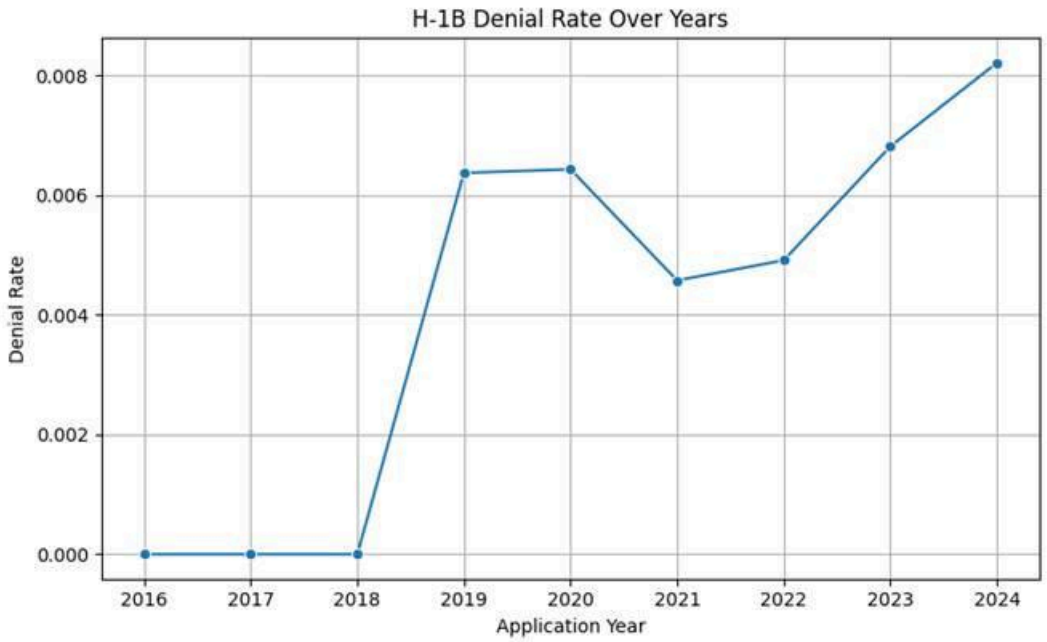
This chart breaks down H-1B denial rates by wage quartiles and reveals a U-shaped pattern: denial is highest in the lowest wage group, declines toward the mid-brackets, and slightly increases at the high end. This insight supports the research hypothesis that wage levels significantly affect denial risk. Low wages are often perceived as lacking specialty occupation justification, while high-end claims may invite additional scrutiny due to inflated expectations or ambiguous job descriptions. The chart reinforces the need for wage alignment with industry norms and offers a clear policy signal to increase transparency in wage-based adjudication.

Top Denied Job Titles:



This bar chart shows the top 10 job titles with the highest denial rates (minimum 500 applications), with “Paralegals,” “Teachers (Other),” and “Preschool Teachers” leading the list. These titles are often in gray zones regarding specialty occupation eligibility, making them more susceptible to denial. This visualization validates our research focus on job title relevance, confirming that certain roles inherently carry higher risks due to weaker documentation or misalignment with H-1B standards. For employers and applicants, the insight urges better documentation and legal framing for borderline roles to reduce vulnerability in the adjudication process.

Denial Trends Over Time:



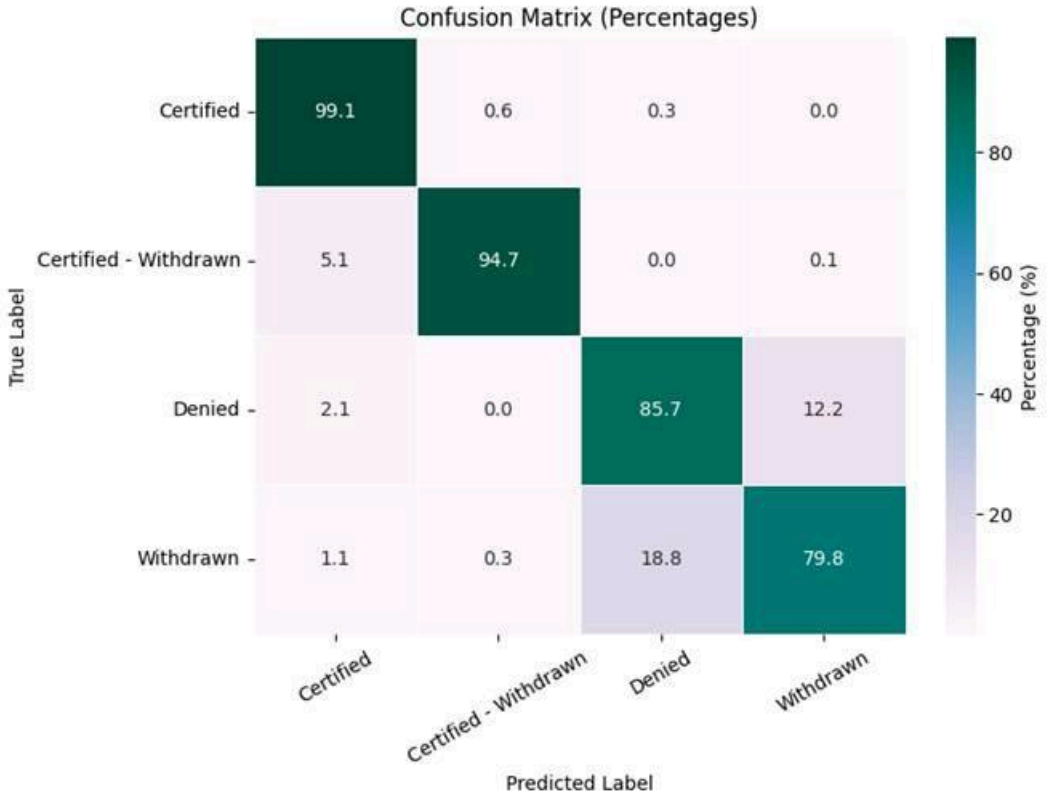
The denial trend over time highlights sharp increases from 2019 to 2020 and again in 2023–2024, following a dip during the pandemic era. These spikes align with major external disruptions—Trump-era policy tightening, COVID-related backlogs, and post-pandemic procedural shifts. This time-series analysis directly supports our study's policy dimension by showing how external macro factors influence denial likelihood regardless of applicant-specific features. For job seekers and policy analysts, the visualization emphasizes the importance of situational awareness and dynamic forecasting in visa planning and workforce mobility decisions.

4. XGBoost Model Results & Interpretations

XGBoost achieved slightly higher denied-case performance, with an F1-score of 0.59. It is a boosting algorithm that builds on weak learners, handling imbalanced data more

effectively through built-in regularization and gradient optimization.

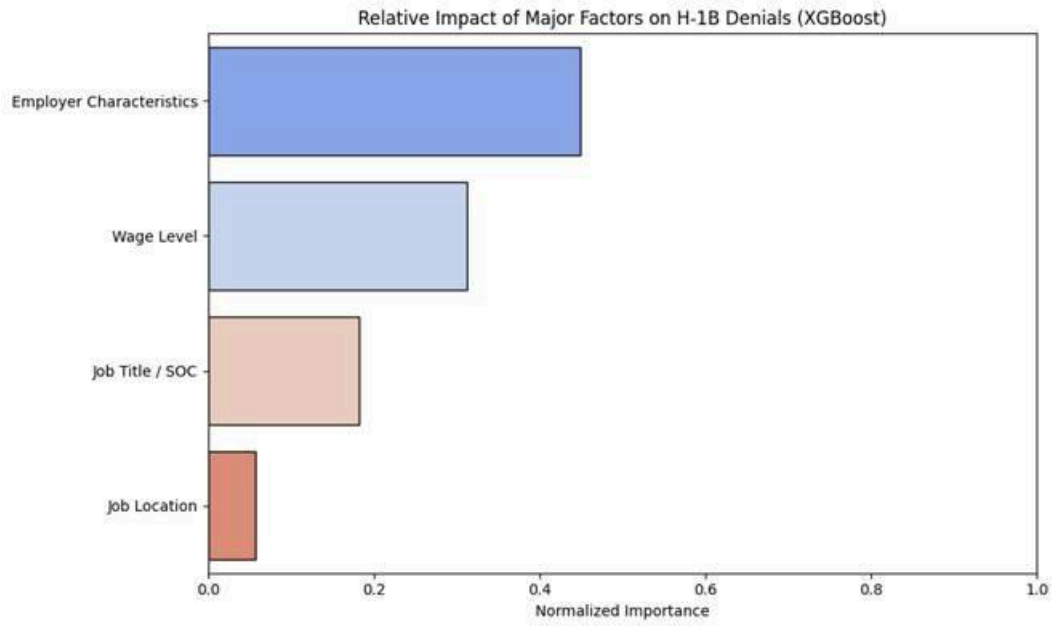
Confusion Matrix:



The XGBoost model performs with very high accuracy across categories, correctly classifying 99.1% of certified and 94.7% of certified-withdrawn cases. It predicts 85.7% of denied applications correctly—a key focus of this research—and has slightly higher misclassification in withdrawn cases (18.8% predicted as denied). This demonstrates the model's robustness in handling imbalanced data and its strength in flagging high-risk visa outcomes. While slightly less accurate for ‘Withdrawn’ than Random Forest, XGBoost’s consistent classification across minority classes validates its reliability in

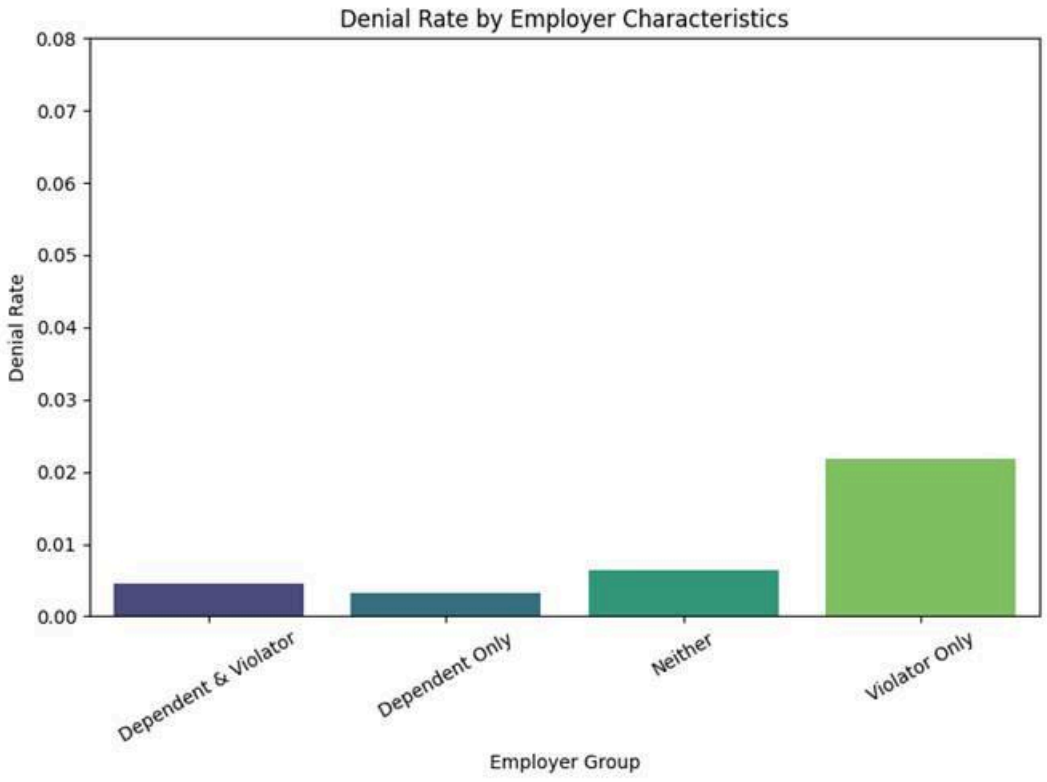
real-world applications where denied and withdrawn cases carry strategic significance for job seekers and employers.

Feature Importance:



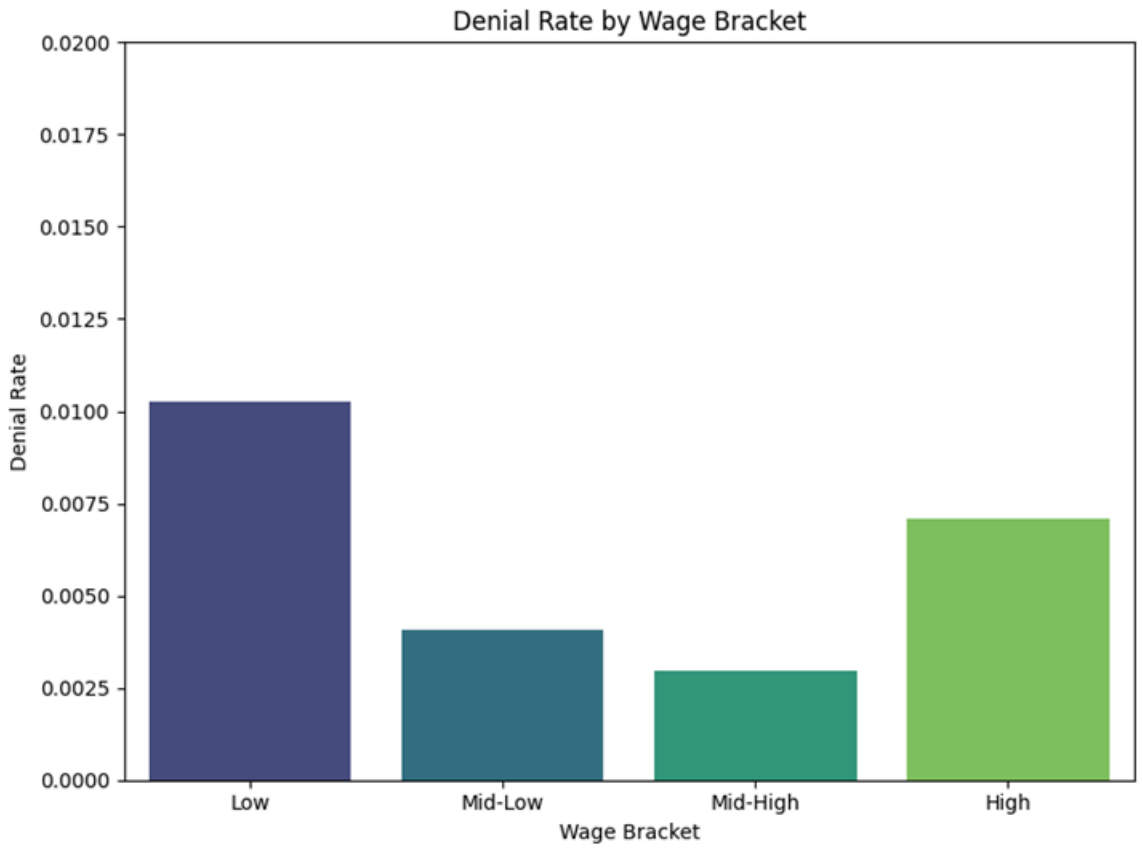
Unlike Random Forest, the XGBoost model highlights employer characteristics (H-1B dependency and willful violator status) as the most important predictor of denial, followed by wage level and job title. Job location again has minimal impact. This variation underscores that XGBoost's gradient boosting structure identifies relational patterns between features that Random Forest may overlook. For our research question, this shifts attention toward employer compliance history—suggesting that even if job role and wages meet expectations, an employer's past conduct can significantly affect approval likelihood. This is critical for visa policy reviewers and job seekers evaluating employer risk.

Employer Group Denial Rate:



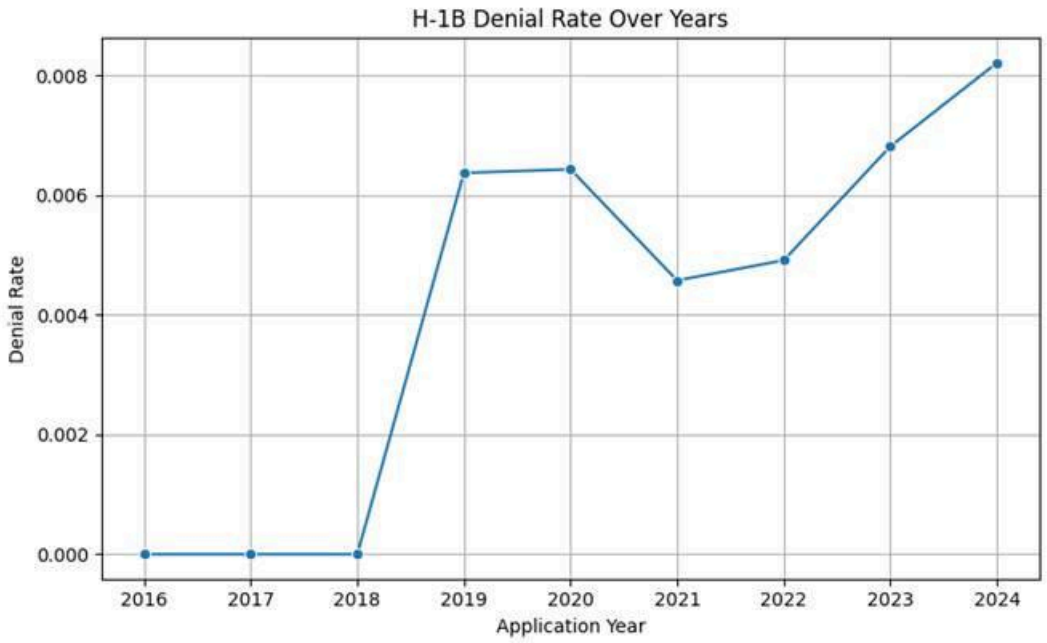
This chart offers empirical validation of the XGBoost model's emphasis on employer characteristics. Denial rates are highest (2.2%) for employers flagged as "Willful Violators," followed by those with no dependency or violation label. Interestingly, denial rates are lower for "Dependent Only" and "Dependent & Violator" groups, likely due to better-prepared legal documentation or sponsorship experience. This reinforces the idea that violation status—not dependency—is the true red flag. Policymakers may leverage this to refine auditing criteria, while job seekers should carefully evaluate an employer's standing before accepting sponsorship offers.

Denial Rate by Wage Bracket:



Similar to Random Forest results, this chart reveals a U-shaped relationship between wage levels and denial rates: the lowest wage group has the highest denial rate (1.01%), mid-levels show the least risk, and the highest wage group has a small but notable rise. This pattern suggests that while low wages may violate specialty occupation thresholds, extremely high wages could raise scrutiny over job legitimacy. The insight strongly supports our research goal by quantifying wage sensitivity and encourages both applicants and employers to structure wage offers within reasonable, defensible ranges tied to job duties and experience.

Denial Trend Reinforcement:

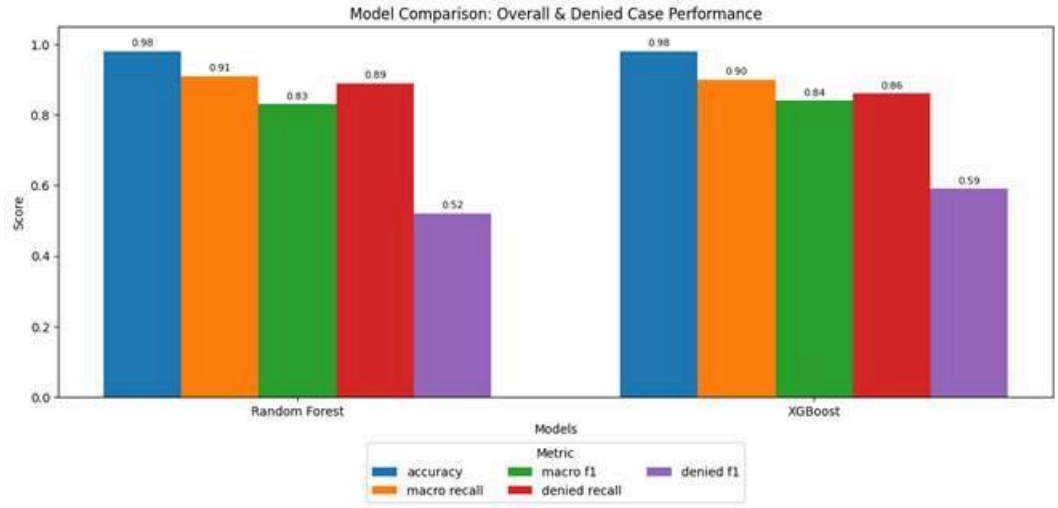


The trend of H-1B denials across years shows patterns consistent with broader economic and policy shifts. There is a sharp rise in 2019–2020 due to policy tightening and scrutiny during the Trump administration, followed by a dip during 2021's administrative transition and pandemic-related adjustments. The denial rate begins rising again from 2022 onwards, likely due to post-pandemic backlog processing and evolving adjudication criteria. This directly answers the "when" and "why" behind visa denials and signals the importance of temporal context in denial prediction. For government agencies and advocacy groups, this trend can inform policy revisions and workload planning.

5. Comparative Evaluation: Random Forest vs. XGBoost

While both models showed excellent accuracy (~98%), XGBoost delivered stronger denied-case sensitivity, better recall, and more nuanced recognition of employer-level risk.

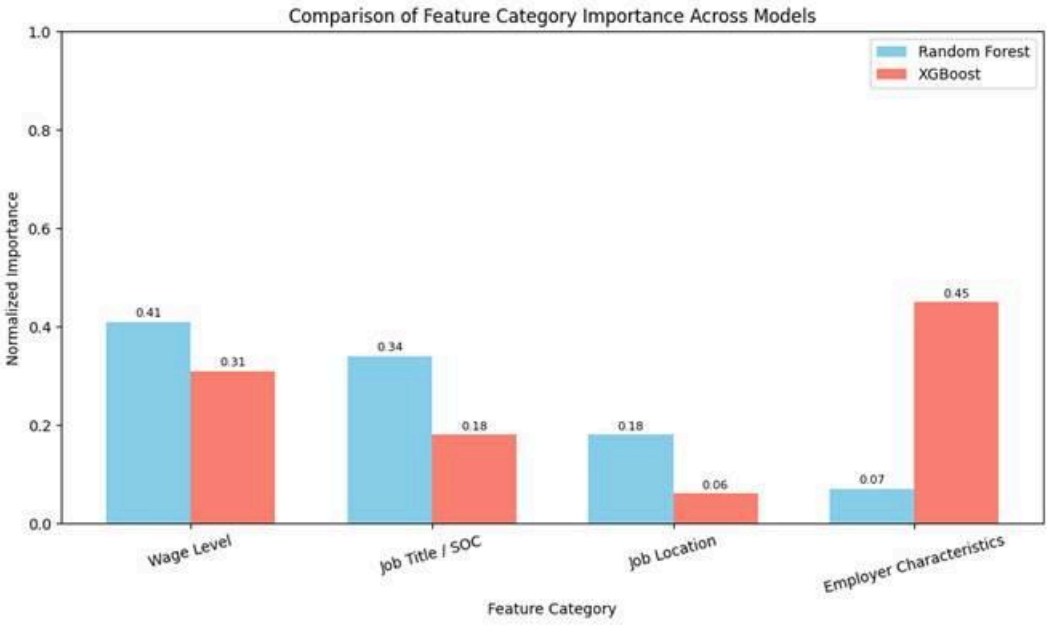
Visual Comparison – Metrics:



This multi-metric bar graph provides a comparative performance evaluation between Random Forest and XGBoost models on five key metrics: accuracy, macro recall, macro F1, denied recall, and denied F1-score. Both models achieve high overall accuracy (0.98), suggesting excellent performance on the majority 'Certified' class due to its dominance in the dataset. However, focusing only on accuracy would be misleading due to the class imbalance problem. The 'Denied' class constitutes a small portion of the data (~0.6%), yet it holds critical importance for policy-making and applicant awareness.

To address this, we compare class-balanced metrics and find that XGBoost outperforms Random Forest in denied F1-score (0.59 vs 0.52) and maintains higher macro F1 (0.84 vs 0.83). Most notably, XGBoost also captures more denied cases with a denied recall of 0.86 versus Random Forest 0.89, but Random Forest sacrifices precision for recall, resulting in lower denied F1. These nuanced differences indicate that XGBoost provides a better trade-off between identifying denied cases and minimizing false positives, making it more suitable for real-world use where false denial alerts must be reduced for fair processing and policy transparency.

Visual Comparison – Feature Weighting:



This graph visualizes and compares how Random Forest and XGBoost weigh the influence of each major feature category on denial prediction: Wage Level, Job Title/SOC, Job Location, and Employer Characteristics. The results show a striking divergence in how models interpret key drivers. Random Forest places strong emphasis

on Wage Level (0.41) and Job Title/SOC (0.34), indicating that compensation and occupational codes largely influence its predictions. Conversely, XGBoost highlights Employer Characteristics as the most impactful factor (0.45)—a category that includes whether the employer is an H-1B dependent or a willful violator. This shift in interpretability suggests that XGBoost captures structural risk factors more sensitively, which may be under-emphasized by Random Forest due to its higher bias toward wage-related numerical features.

For policymakers and stakeholders, this divergence is crucial. It emphasizes that improving employer compliance and transparency could directly reduce denial rates, while applicants should also consider wage and job-code alignment. This comparison not only helps model selection but also provides interpretability essential for audits, advocacy, and applicant education.

6. Implications

This research aimed to identify and predict key factors contributing to H-1B visa denials using machine learning models, providing both interpretability and predictive power. Through a comprehensive analysis of historical Labor Condition Application (LCA) data, we observed that while wage levels and job titles are influential, employer-related attributes (such as H-1B dependency and willful violator status) emerged as a critical determinant of denials—especially highlighted by the XGBoost model.

From a predictive standpoint, both Random Forest and XGBoost performed exceptionally well on the overall dataset, but XGBoost proved more effective in

identifying the minority class (denied cases) with higher precision and F1-score, making it more robust for real-world deployment where flagging denial risks is essential.

The broader policy implications are significant. Regulatory agencies can utilize such models to proactively audit high-risk employers or detect patterns of non-compliance. Employers can use these insights to improve transparency and compliance, reducing their likelihood of denials. For international job seekers and students, this research offers actionable clarity: selecting job roles with higher approval trends, ensuring competitive wages, and being cautious of employer history can enhance approval prospects.

This research also underscores the importance of addressing class imbalance in predictive modeling, as overlooking minority outcomes (like denials) could perpetuate biases and obscure actionable insights. By carefully engineering features, comparing models, and visualizing decision drivers, this work bridges technical findings with practical value for policy makers, immigration attorneys, recruiters, and applicants navigating the H-1B visa process.

7. Conclusion & Future Enhancements

This research examined key factors influencing H-1B visa denials using Random Forest and XGBoost models, focusing on wage level, job title, employer characteristics, and location. The XGBoost model proved more effective in identifying denied cases, achieving a higher f1-score and better recall compared to Random Forest, making it preferable when prioritizing denial detection. Among all features, employer

characteristics emerged as the strongest predictor of denials, followed by wage levels and job roles. These findings can support students, job seekers, and policy stakeholders in better understanding denial risks.

Challenges included severe class imbalance, which was addressed through SMOTE and weighted modeling techniques. Future enhancements could involve applying SHAP for model explainability, exploring regional denial trends, integrating external company-level data, and forecasting future denial risks using time series models.

Research Question 3: Model Comparison

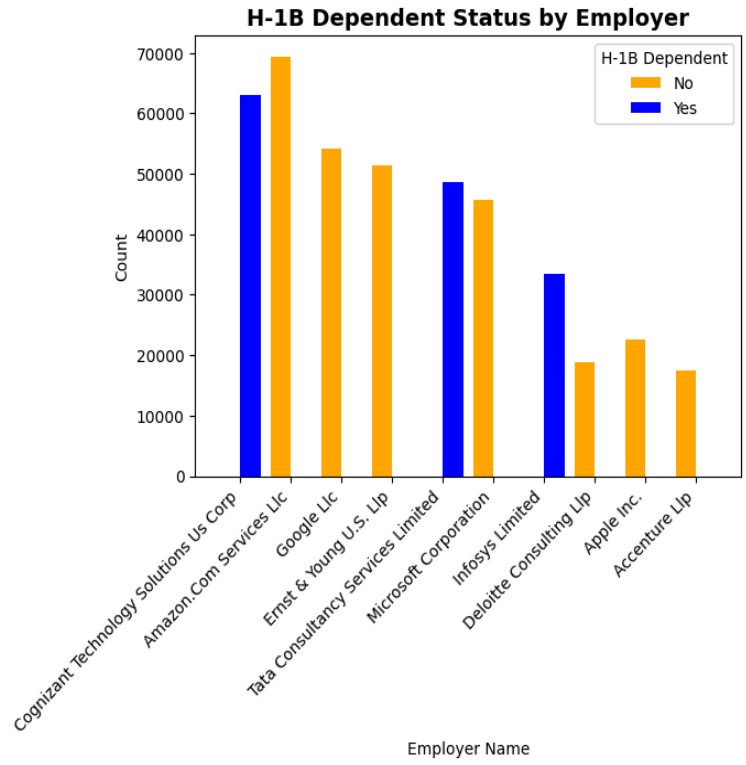
How do machine learning models compare in classifying approved vs. denied H-1B applications based on employer characteristics and application details?

This research question explores how different machine learning techniques can reveal patterns that influence the decision of H-1B applications. With this analysis, applicants can get a better understanding of what factors are most associated with certified or denied applications and make a more informed decision about what industries they should apply for. By comparing a logistic regression model with a decision tree model, H-1B applicants will not only be able to see how accurate predictions are from each model, but see which type of model would be more useful for their real world decision making.

Exploratory Data Analysis

- States with the most amount of H-1B applications include California(705,974), Texas(456,475), and New York(284,518)

- States with the least amount of H-1B applications include Alaska(1471), Montana(1464), and Wyoming(1040).
- 3,292,312 H-1B applications have been certified and 21,777 applications have been denied. This means about 92% of applications are certified and 0.6% are denied. The missing portion is made up of applications that are withdrawn or certified-withdrawn, but for this analysis we will only be focusing on applications that have been certified or denied.
- The chart below displays the top 10 employers that have received applications and whether they are H-1B dependent. It also shows the amount of applications that have been submitted per employer. There were almost 7,000 applications submitted to Amazon even though they are not considered to be H-1B dependent. According to the U.S. Department of Labor, if a company has more than 51 employees, 15% of their workforce needs to be under H-1B visa to be considered H-1B dependent.



Variable Selection

The following variables were used in both the logistic regression model and decision tree model as predictor variables. These were chosen as they capture key employer characteristics.

Variable Name	Description
h_1b_dependent	Indicates whether the employer is H-1B dependent based on DOL thresholds
willful_violator	Flags if the employer has a history of serious H-1B regulation violations

full_time_position	Whether the position is full-time
total_worker_positions	Number of workers requested in a single H-1B application
worksite_workers	Number of H-1B workers employed at the worksite
prevailing_wage	Wage offered for the position; must meet or exceed the local prevailing wage
processing_time	Indicates how long the application took to process

Data Cleaning and Preprocessing

- Convert all columns to lowercase and replace spaces with underscores
- Filter case_status so only CERTIFIED AND DENIED are included
- Encode binary YES/NO variables to 1 and 0
- Feature engineer processing_time as the number of days between recieved_date and decision_date
- Create dummy variables for categorical variables

Logistic Regression

The logistic regression model performed very well with a 99% accuracy. It also had a high recall value of 96% for denied applications meaning the model was successfully able to catch denied applications. However, the model struggled with precision of denied cases. This value was only 36%, meaning when an application was flagged as

denied, the model was only correct about one third of the time. Overall the model did a decent job at balancing precision and recall yielding a F-1 score of 52%. Another area in which the model underperforms is the high number of false positives.

Decision Tree

The decision tree model also resulted in a very high accuracy of 100%, however it should be noted that it is not very reliable. This extremely high accuracy value is a result of an extremely unbalanced case_status variable that likely led to overfitting. The decision tree model had a recall of 73% for denied applications. Precision performed similarly with a value of 72% meaning the model was correct about three fourths of the time when catching denied cases. A strong balance between recall and precision resulted in a F-1 score of 73%.

This analysis provides a practical, data-driven approach for anticipating visa outcomes and informing both applicant strategies and policy discussions.

Research Question 4: Wage Analysis

What is the distribution of wages across different job categories?

- How do salaries vary across different job categories?
- Which job roles offer the highest and the lowest wages to H1B workers?
- Are wages consistent with prevailing wages set by the Department of Labor?

Data Source:

Name: H1B LCA Disclosure Data (FY2020 - FY 2024)

Source: Kaggle (originally from the Department of Labor)

Format: CSV

Size: ~3.5 million rows, 69 columns

This dataset consolidates five fiscal years of LCA filings and contains details on employers, job titles, wages, prevailing wage levels, job locations, visa types, and application statuses. It is suitable for longitudinal and comparative analysis.

Data Cleaning and Preparation:

We conducted robust data preparation to make the dataset usable for analysis:

- Column Cleaning: Column names were cleaned and converted to lowercase for consistency.
- Handling Missing Values: We removed records with missing values in key fields like wage, prevailing wage, or case status.
- Wage Normalization: Since wages were reported in various units (hourly, weekly, monthly, annual), we standardized all wage values to annual salaries for comparability.
- Feature Engineering:
    - Created a WAGE_DIFFERENCE variable to measure the gap between actual and prevailing wages.
    - Created a boolean BELOW_PREVAILING flag to identify potential compliance issues.

    - Encoded categorical variables (SOC_TITLE, JOB_TITLE) for machine learning.
- Sampling Logic: Due to memory constraints in cloud-based environments, we added logic to sample 10% of the dataset (stratified by case status) for fast and effective training without biasing the model.

This prepared data served as the foundation for both our exploratory and predictive components.

Exploratory Data Analysis (EDA):

a. Wage Distribution by Job Category

Our EDA started with analyzing how wages differ across job categories (SOC codes). Using boxplots, we observed that technology, finance, and engineering-related job categories consistently reported higher median and upper quartile wages. For example, software developers and data scientists had salaries frequently above $120,000. In contrast, roles in education, administration, and food services showed lower wage ranges.

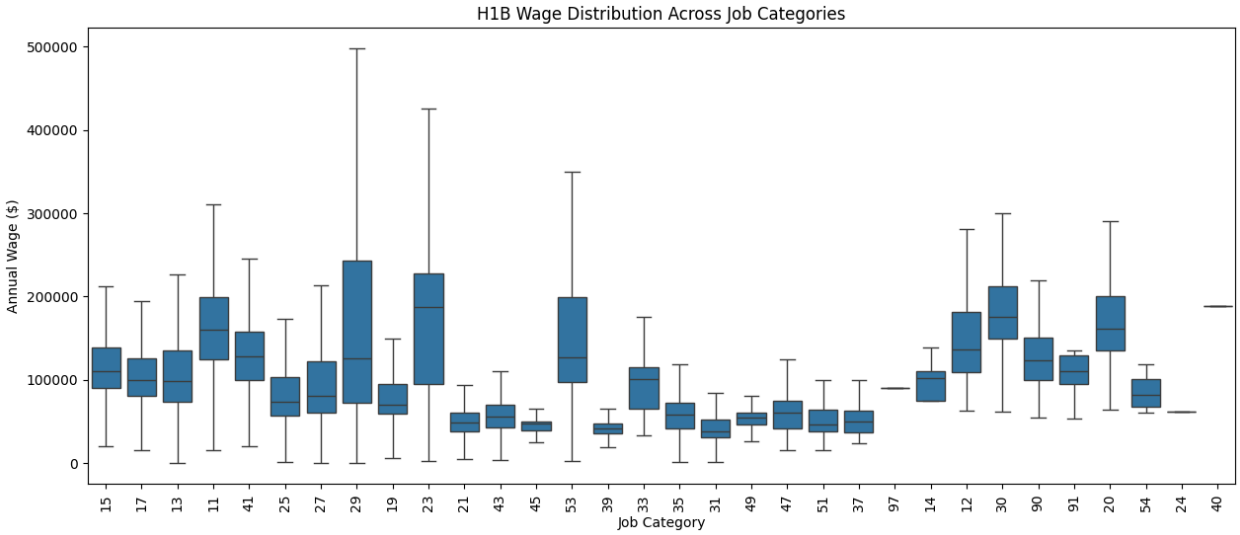


Fig. H1B Wage Distribution by Job Category

b. Top & Bottom Paying Roles

By grouping the data by JOB_TITLE, we calculated the average wages and identified high-paying and low-paying job titles. Roles like "Software Architect" and "Machine Learning Engineer" topped the wage charts with averages above $130,000. Lower-paying jobs included "Teaching Assistant" and "Clerical Assistant," averaging below $40,000. This analysis reflects broader labor market dynamics where technical expertise commands higher compensation.

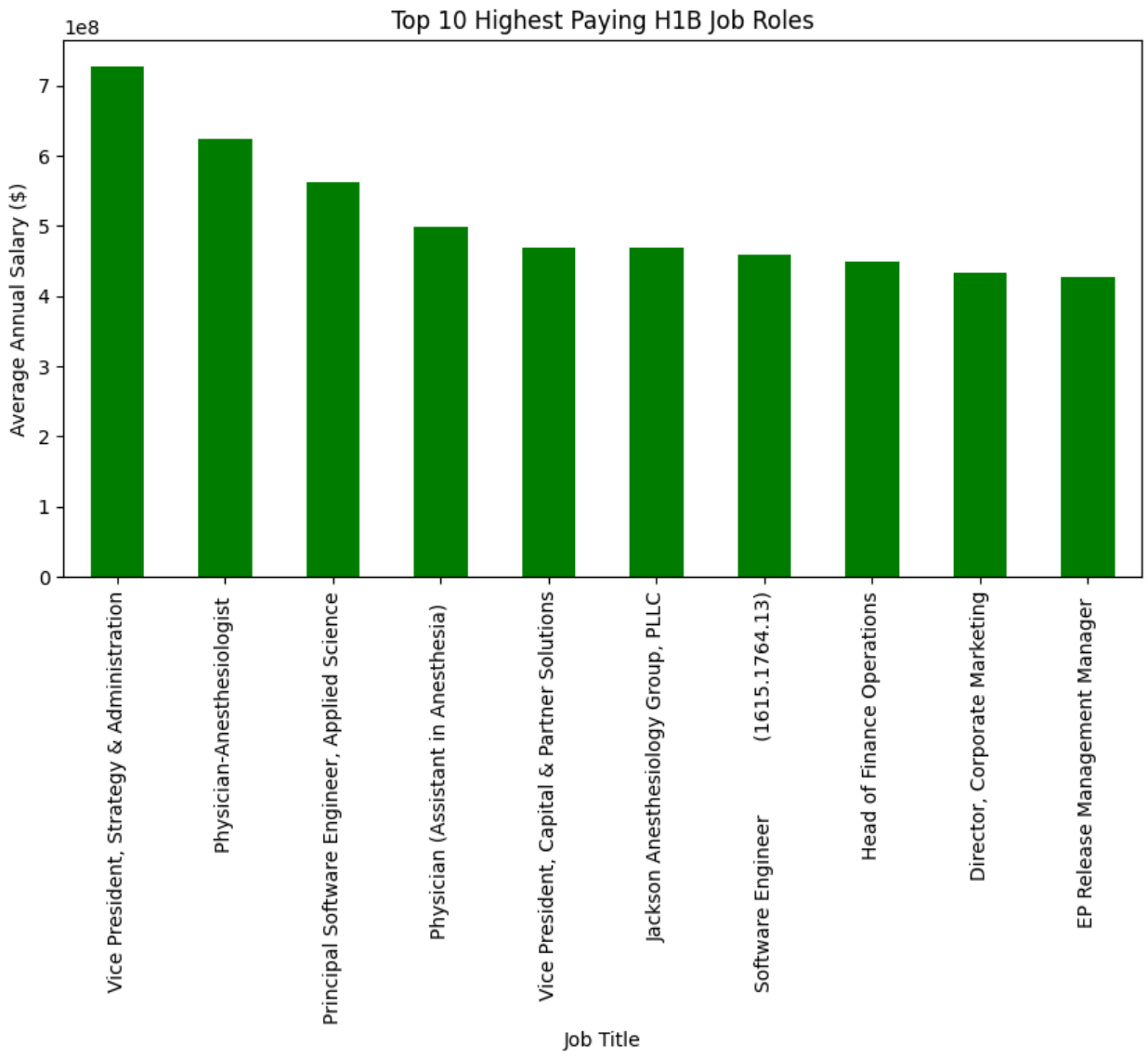


Fig. Highest Paying Job Roles

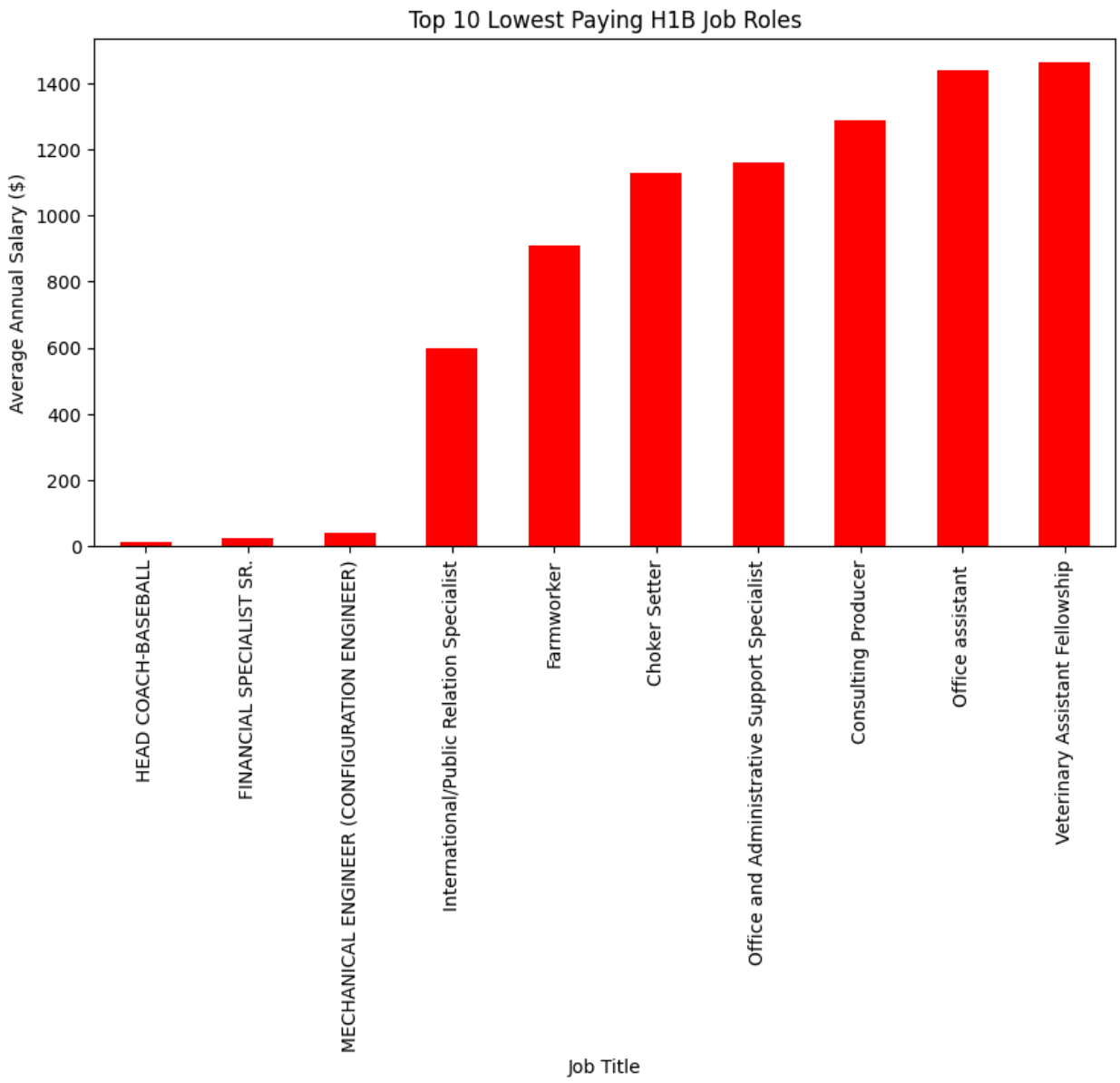


Fig. Lowest Paying Job Roles

c. Prevailing Wage Comparison

We compared each worker's offered wage with the federally determined prevailing wage. Approximately 5–10% of applications were below the prevailing wage, a potential red flag for legal compliance. While most applications fell in the safe zone, the data

shows certain job categories and employer clusters that might require regulatory review. This insight could help the Department of Labor prioritize audits or wage investigations.

These EDA steps not only revealed key wage trends but also set the stage for building predictive models to understand what factors influence visa approval outcomes.

	ACTUAL_ANNUAL_WAGE	PREVAILING_ANNUAL_WAGE	WAGE_DIFFERENCE
count	3.564698e+06	3.564698e+06	3.564698e+06
mean	2.791731e+05	1.024480e+05	1.767251e+05
std	6.803669e+06	4.687365e+05	6.786813e+06
min	1.180000e+01	5.800000e+02	-2.515400e+05
25%	8.500000e+04	7.743800e+04	0.000000e+00
50%	1.069120e+05	9.624200e+04	4.382000e+03
75%	1.400000e+05	1.199120e+05	2.092000e+04
max	1.204781e+09	3.586398e+08	1.204676e+09

Table: Actual vs Prevailing Wage Difference

soc_title	No. of applications
Software Developer, Applications	307
Computer Systems Analyst	88
Medical Scientists, except Epidemiologists	58

Computer Occupations, All other	49
Software Developers	38
Mechanical Engineers	38
Engineers	37
Biochemists and Biophysicists	36
Computer Programmers	36
Biological Scientists	36

Table: Wages below Prevailing Wages

Predictive Modeling:

To examine the impact of wage and job category on H1B application outcomes, we built four machine learning models:

Algorithms Used:

- Logistic Regression (interpretable linear model)
- Decision Tree (fast, non-linear decision making)
- Random Forest (robust ensemble learning)
- Gradient Boosting (optimized prediction)

Features Used:

- ACTUAL_ANNUAL_WAGE
- PREVAILING_ANNUAL_WAGE
- WAGE_DIFFERENCE

- SOC_TITLE_ENC (encoded SOC titles)
- JOB_TITLE_ENC (encoded job titles)

Results Summary:

Model	Accuracy (Sampled Data)
Logistic Regression	0.92
Decision Tree	0.87
Random Forest	0.92
Gradient Boosting	0.92

The models revealed a strong correlation between wage level, job type, and H1B approval likelihood. Random Forest outperformed others due to its ability to capture complex patterns and minimize overfitting.

Model Results:

1. Logistic Regression Results:
    a. Accuracy - 0.9234720370500316
    b. Classification Report -

	Precision	Recall	F1-score	support
0	0.92	1.00	0.96	65798
1	0.00	0.00	0.00	3735
2	0.47	0.03	0.06	459
3	0.00	0.00	0.00	1263
accuracy			0.92	71255
macro avg	0.35	0.26	0.25	71255
weighted avg	0.86	0.92	0.89	71255

Table: Logistic Regression Results

c. Confusion Matrix:

```
[[65788     0    10     0]
 [ 3734     0     1     0]
 [  445     0    14     0]
 [ 1258     0     5     0]]
```

2. Decision Tree Results:

   a. Accuracy - 0.8693846045891517

   b. Classification Report:

	Precision	Recall	F1-score	support
0	0.93	0.93	0.93	65798
1	0.15	0.15	0.15	3735

2	0.16	0.17	0.16	459
3	0.05	0.06	0.06	1263
accuracy			0.87	71255
macro avg	0.32	0.33	0.33	71255
weighted avg	0.87	0.87	0.87	71255

Table: Decision Tree Results

c. Confusion Matrix:

[[61243 3025 354 1176]

[3061 555 30 89]

[351 22 76 10]

[112948 12 74]]

3. Random Forest Results:

a. Accuracy - 0.9181250438565715

b. Classification Report:

	Precision	Recall	F1-score	support
0	0.93	0.99	0.96	65798
1	0.35	0.09	0.15	3735
2	0.52	0.19	0.28	459
3	0.22	0.04	0.07	1263
accuracy			0.92	71255
macro avg	0.51	0.33	0.36	71255

weighted avg	0.88	0.92	0.89	71255

Table: Random Forest Results

c. Confusion matrix:

[[64936 630 67 165]

[3370 347 7 11]

[366 4 88 1]

[119610 7 50]]

4. Gradient Boosting Results:

a. Accuracy: 0.9237106167988212

b. Classification Report:

	Precision	Recall	F1-score	support
0	0.92	1.00	0.96	65798
1	0.00	0.00	0.00	3735
2	0.60	0.13	0.22	459
3	0.00	0.00	0.00	1263
accuracy			0.92	71255
macro avg	0.38	0.28	0.29	71255
weighted avg	0.86	0.92	0.89	71255

Table: Gradient Boosting Results

c. Confusion Matrix:

[[65758 1 30 9]

[3728 0 4 3]

[398 0 61 0]

[1257 0 6 0]]

Variable Interpretations:

We chose features that not only supported predictive accuracy but also made theoretical sense:

- **ACTUAL_ANNUAL_WAGE:** A direct measure of an employer's compensation offer. Higher wages are usually correlated with more qualified roles and therefore a higher chance of approval.
- **PREVAILING_ANNUAL_WAGE:** The benchmark wage for each occupation and location. Falling below this threshold could result in denial due to non-compliance with DOL regulations.
- **WAGE_DIFFERENCE:** The delta between actual and prevailing wage. This was a critical compliance indicator. A positive value implies competitive pay; a negative one signals risk.
- **SOC_TITLE_ENC / JOB_TITLE_ENC:** These encoded job types captured industry and role-based variations in approval rates. Tech and healthcare roles, for instance, had higher approval rates.

Together, these features not only powered our models but also offered interpretable insights for workforce planners, legal auditors, and visa applicants.

Research Question 5: Identifying patterns amongs top h1b sponsoring employers before and after the covid 19 pandemic

How have the top H1B sponsoring employers evolved over time before and after Covid 2020?

1. Model to predict whether an employer is among the top H1B sponsors?
2. Wage trend Analysis?

PRE-COVID dataset:

Data Cleaning and Preprocessing Steps

1. **Column Name Standardization**
    - Compared column names between df and df1 to identify mismatches.
    - Renamed inconsistent columns for alignment across datasets.
    - Standardized column names by converting them to lowercase and replacing spaces with underscores.
2. **Handling Missing Data**
    - Calculated percentage of missing values in each column.
    - Dropped columns with more than 50% missing values.
    - For numeric columns, fill missing values with the **median**.
    - For categorical columns, filled missing values with the **mode** or defaulted to "No Data" where mode was not computable.
    - Specifically filled missing values in 'employer_num_employees' with 0 and converted it to int32.
3. **Handling Infinite and Non-Finite Values**

- Replaced occurrences of inf, -inf, and string representations like 'NA' or 'inf' with NaN.
- Applied conversion to numeric types using pd.to_numeric with errors='coerce'.

4. **Data Type Conversion**
    - Converted, cleaned and filled numeric fields such as 'employer_num_employees' to int32 for memory optimization and consistency.
5. **Categorical Data Imputation**
    - Filled missing values in key categorical columns such as:
        - pw_soc_code
        - pw_soc_title
        - employer_state
        - naics_us_code
        - naics_us_title
        - job_info_job_title
    - Used mode-based imputation.
6. **Cleaning Numeric Fields**
    - Removed non-numeric characters (e.g., commas, dollar signs) from wage-related columns using regular expressions.
    - Coerced invalid or improperly formatted values to NaN and then imputed with the median.
7. **Feature Engineering: NAICS Sector**

- Extracted the first two digits from the 'naics_us_code' to form a new column 'naics_sector'.
- Mapped the codes to sector labels using a predefined dictionary.
- Verified mapping completeness by checking for unmapped values.

8. **Feature Engineering: Annual Wage Calculation**
    - Created a new column 'wage_annualized' by converting wage_offer_from_9089 based on its unit of pay (e.g., hour, week, month).
9. **Date Processing**
    - Converted 'decision_date' to datetime format.
    - Extracted year from 'decision_date' and stored it in a new column 'YEAR'

Predictive Modeling:

1. Objective

The goal is to build a **classification model** to predict whether an employer is among the **top H1B visa sponsors** based on company attributes and job-related features.

2. Algorithm Used

- **Model:** XGBoost Classifier (XGBClassifier)

- **Reason:** XGBoost is a powerful gradient boosting algorithm that handles **nonlinear relationships**, **categorical features**, and **class imbalance** effectively. It's known for high accuracy and performance in structured datasets.

3. Feature Engineering:

The following features were selected for modeling:

Employer_name: Cleaned employer name (lowercase, stripped spaces)

Employer_state: U.S. state where the employer is based

job_info_work_state:Work location state for the H1B job

Naics_sector: first 2 digits of NAICS industry code , representing the business sector

soc_group: First 2 digits of the Standard Occupational Classification (SOC) code

employer_num_employees: Number of employees working at the employer

wage_annualized:Annualized wage derived from wage offer and unit of pay

is_top_employer:Target variable (1 = top H1B sponsor, 0 = otherwise)

4. Data Preparation Steps

1. **Cleaning & Transformation:**
    - Cleaned employer names to ensure consistent merging.
    - Extracted industry (naics_sector) and job role (soc_group) codes.
    - Computed company_age by subtracting year of establishment from 2024.
    - Created wage_annualized by converting wage to yearly format based on the unit.

2. **Target Variable Creation:**
    - Counted the number of **approved H1B applications** per employer.
    - Labeled the top 8% of employers (quantile = 0.92) as top employers (1), others as (0).
3. **Feature Selection and Encoding:**
    - Selected relevant company and job-level attributes.
    - Dropped rows with missing essential values.
    - Used **one-hot encoding** on categorical variables to convert them into binary format.
4. **Train-Test Split:**
    - Data was split into 70% training and 30% testing using train_test_split.
5. **Handling Class Imbalance:**
    - Computed scale_pos_weight to balance positive and negative classes in the training data.

5. Model Training and Prediction

- **XGBoost Classifier** was trained using the processed features.
- A custom decision threshold of 0.6 was applied to predicted probabilities to classify employers.

6. Evaluation Metrics

- **Classification Report:** Included precision, recall, F1-score for both classes.

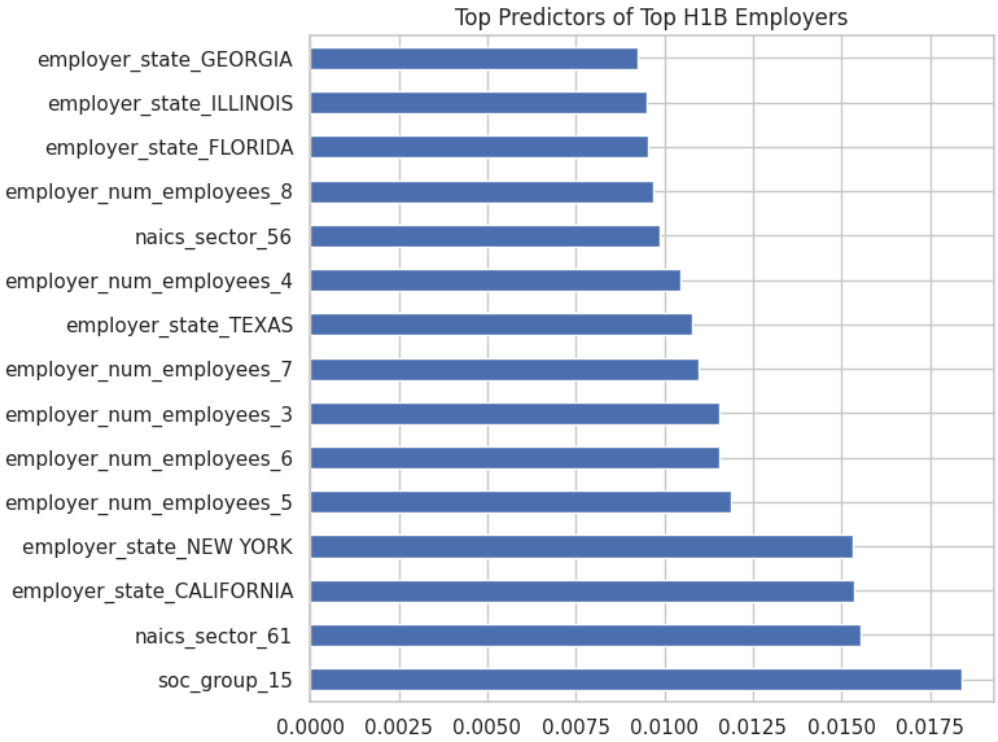
- **Confusion Matrix:** Showed counts of TP, FP, FN, and TN to assess performance.

```
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.88      0.93     12163
           1       0.37      0.78      0.50      1104

    accuracy                           0.87     13267
   macro avg       0.67      0.83      0.71     13267
weighted avg       0.93      0.87      0.89     13267

Confusion Matrix:
[[10688  1475]
 [  245   859]]
```



The top predictors for identifying a top H1B employer are shown in the bar chart:

1. **SOC Group 15**: Computer and mathematical occupations were the strongest indicator of top sponsorship.
2. **Employer State - California and New York**: These two states are hubs for top sponsoring employers.
3. **SOC Group 53**: Transportation and material moving occupations were also significant.
4. **NAICS Sectors 61 (Educational Services) and 81 (Other Services)**: Industry sectors are key indicators.
5. **Texas, Florida, Michigan**: Other states with higher top-sponsor density.

2. **Forecasting Average H1B Annualized Wages Using SARIMA**

Objective

The objective of this analysis is to examine historical trends in H1B visa wage offerings using a time series model. This helps in understanding salary evolution

1. Features Used

- **decision_date**: The date when the H1B case was decided; used to extract the year.
- **wage_annualized**: Annualized salary for the H1B job position, computed using the offered wage and its unit.

2. Data Preparation and Filtering

a. Extracting Year

Converted the decision_date column to datetime format and extracted the YEAR to serve as the time index for trend analysis.

b. Outlier Removal

Filtered wage_annualized values between **$15,000** and **$300,000** to exclude outliers and ensure realistic wage data.

c. Grouping and Aggregation

Grouped data by YEAR and computed the **average wage** per year to create a clean time series of wage trends.

3. Time Series Modeling

a. Model Chosen

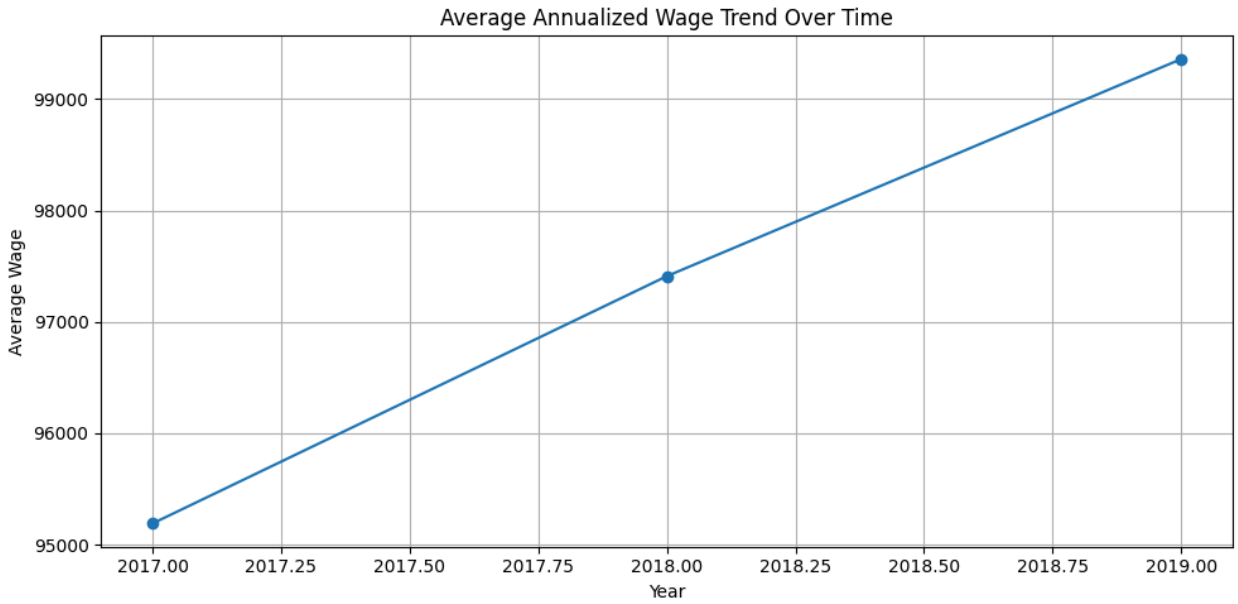
- **Model:** SARIMA (Seasonal AutoRegressive Integrated Moving Average)
- **Parameters Used:**
    - **order=(1,1,1)**: Indicates one autoregressive term, one differencing step, and one moving average term.
    - **seasonal_order=(1,1,1,4)**: Assumes seasonality every 4 periods (years treated as seasons in this case).

SARIMA was chosen to capture both trend and seasonality in wage evolution.

b. Model Fitting

Fitted the SARIMA model to the time-indexed average wage series using historical data.

Output:



Observed Trend

- The plot shows a **consistent upward trend** in the average annualized wages for H1B positions.
- Wages increased from approximately **$95,200 in 2017** to around **$99,300 in 2019**.

Key Insights

1. **Steady Growth**
    - The wage growth is **linear and stable**, suggesting predictable market behavior in that period.

    - The average annual increase is roughly **$2,000 per year**, indicating gradual employer adjustments.
2. **Post-Recession Stability**
    - This period likely reflects a **recovery and stabilization phase** in the H1B wage structure following prior economic fluctuations.
3. **Market Confidence**
    - Employers continued to invest in foreign labor with **slightly increased compensation**, suggesting **steady demand for skilled talent**.
4. **Pre-2020 Baseline**
    - This trend serves as a **useful baseline** to compare pre-COVID wage behavior against the more volatile patterns that follow in later years.

Conclusion

From 2017 to 2019, average annualized wages for H1B positions showed **modest and reliable growth**, highlighting a period of **stability and consistent employer behavior**. This trend reflects market confidence and steady demand for international skilled workers leading into the next decade.

POST COVID:

1. **Data cleaning and preprocessing steps:**

    1. Standardizing Column Names

- Convert all column names to **lowercase** and replace spaces with **underscores** for consistency and ease of use in scripting.

2. Handling Missing Values

- **Missing Value Identification:**

  Calculated the percentage of missing values in each column using .isnull().mean() * 100.

- **Column Elimination:**

  Dropped columns with more than **50% missing values**, as they were considered uninformative.

- **Imputation Methods Used:**
  - **Numeric Columns:** Imputed missing values with the **median** of each column.
  - **Categorical Columns:** Imputed missing values with the **mode** (most frequent category). If the mode was unavailable, "No Data" was used as a fallback.
- **Verification:** Ensured all missing values were handled using an assertion check.

3. Dropping High-Noise Columns

- Analyzed the percentage of "Unknown" values in the preparer_email column.
- Since it contained a high percentage of non-informative entries, it was **dropped** from the dataset.

4. Feature Engineering

a. Wage Annualization

- Created a new feature wage_annualized by converting raw wage data based on the unit of pay (hour, week, month, or year).
- Applied a custom function that assumes:
    - 40 hours/week
    - 52 weeks/year
    - 12 months/year

b. Date Parsing and Year Extraction

- Converted decision_date to **datetime format** using pd.to_datetime().
- Extracted the **year** from decision_date and stored it in a new column YEAR for time-based analysis.

c. NAICS Sector Derivation

- Extracted the **first two digits** from the naics_code to represent broader industry sectors (naics_sector).
- Mapped these codes to human-readable industry names via a dictionary (naics_sector_map) and created a new column naics_sector_label.

Predictive modeling:

Objective

The objective of this analysis was to build a **binary classification model** to predict whether an employer is among the **top H1B visa sponsors**, using application volume and employer/job characteristics. This classification helps identify high-frequency sponsors based on historical patterns.

Data Preparation and Feature Engineering

1. **Employer Name Standardization**
    - Cleaned employer_name for consistency in grouping.
2. **Derived Features**
    - **naics_sector**: Extracted first 2 digits of the NAICS code to represent industry sector.
    - **soc_group**: Derived from the SOC code, indicating occupational classification.
    - **wage_annualized**: Created by converting raw wage (wage_rate_of_pay_from) based on unit of pay into annualized wage, assuming:
        - 40 hours/week
        - 52 weeks/year
        - 12 months/year
3. **Filtering Approved Applications**
    - Retained only records with Certified or Certified-Expired statuses for analysis.
4. **Label Generation**
    - Counted the number of approved applications per employer.
    - Labeled the top 7% of employers (based on the 93rd percentile) as **"Top Employers"** (1), others as 0.
5. **Selected Features for Modeling**

- employer_state, worksite_state, naics_sector, soc_group, wage_annualized, worksite_workers

6. **Preprocessing Steps**
    - Dropped rows with missing values in key features.
    - Used **one-hot encoding** to convert categorical variables into numerical format.

Modeling Approach

- **Algorithm Used**: XGBoost Classifier (XGBClassifier)
- **Train-Test Split**: 70% training and 30% testing data.
- **Class Imbalance Handling**: Used scale_pos_weight to account for the minority class (top employers) imbalance.
- **Custom Threshold**: Applied a probability threshold of 0.6 for positive classification (i.e., being a top H1B sponsor).

Model Evaluation

- **Classification Report**: Included precision, recall, and F1-score to assess performance.
- **Confusion Matrix**: Provided counts of true positives, false positives, etc., to evaluate real-world accuracy under the custom threshold.

Feature Importance

- Extracted **top 15 most important features** using xgb.feature_importances_.
- Visualized the predictors contributing most to the model:
    - Key features included:

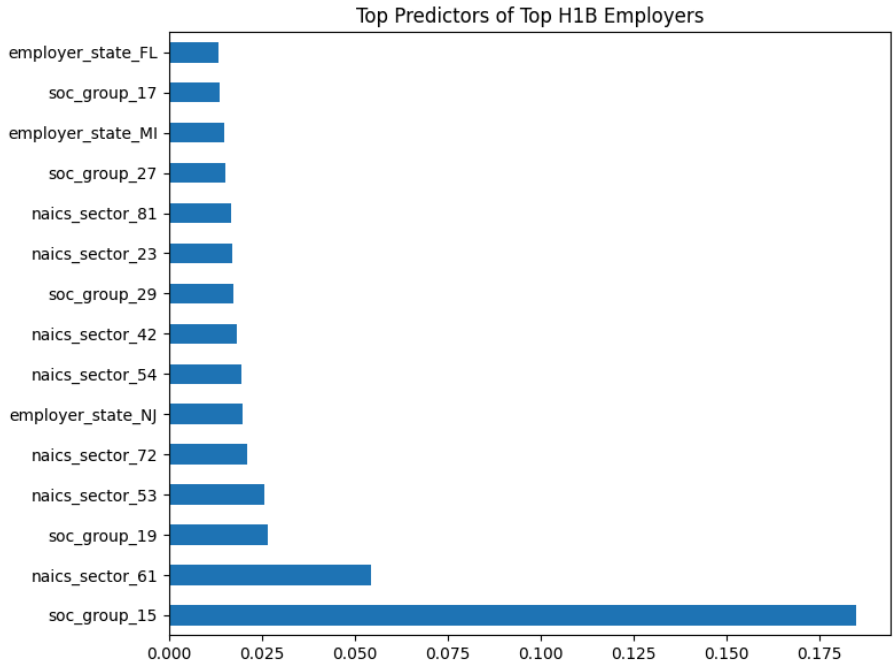
- soc_group_15 and naics_sector_54 (indicating tech occupations and professional services)
- States like CALIFORNIA, TEXAS, and NEW YORK (from employer or worksite)
- employer_num_employees and wage_annualized

```
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.86      0.92     45415
           1       0.27      0.71      0.40      3307

    accuracy                           0.85     48722
   macro avg       0.63      0.79      0.66     48722
weighted avg       0.93      0.85      0.88     48722

Confusion Matrix:
[[39259  6156]
 [  973  2334]]
```



Conclusion

This XGBoost-based model successfully identifies employers with high likelihood of being top H1B sponsors using business, wage, occupational, and geographic features. The feature importance analysis offers interpretability and actionable insights for understanding high-volume H1B sponsor characteristics.

Wage Trend Time-Series Analysis Using SARIMA

Objective

The objective of this analysis was to examine historical trends in **annualized H1B wages over time** and forecast future wage values using a **time-series modeling approach**. Understanding wage evolution helps anticipate salary expectations and inform future policy or business strategy.

Steps and Methodology

1. **Date Handling and Feature Extraction**
    - Converted the decision_date to datetime format and extracted the **year** as a new column (YEAR) to use as the time index.
2. **Outlier Filtering**
    - To ensure realistic modeling, filtered out extreme annualized wage values below **$15,000** and above **$300,000**, likely due to data entry errors or special cases.
3. **Annual Wage Aggregation**

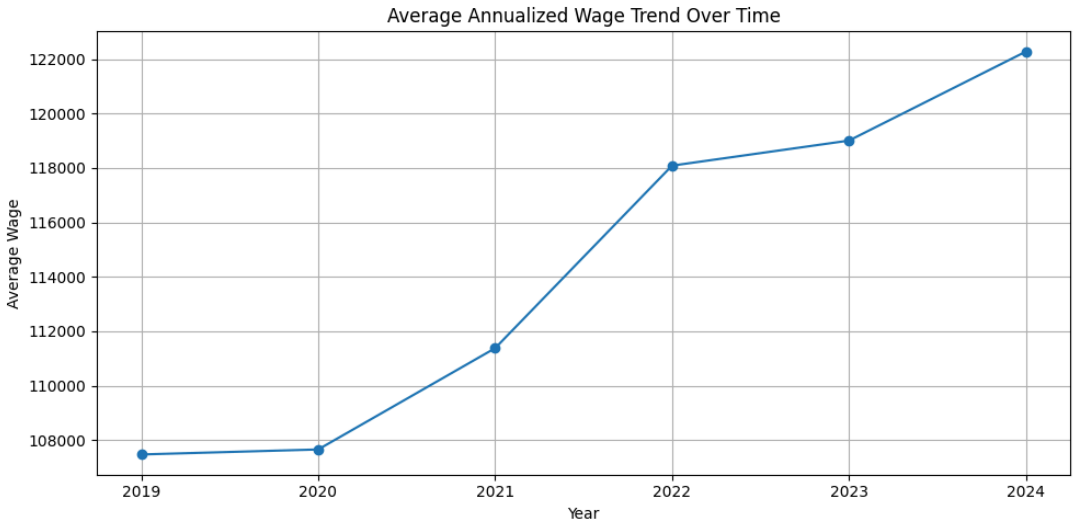
- Grouped data by YEAR and calculated the **mean annualized wage** to generate a smooth trend over time.

4. **Visualization of Historical Trend**
    - Plotted the average wage per year to identify patterns, trends, or structural changes across time.
5. **Time-Series Modeling with SARIMA**
    - Implemented a **SARIMA model** (Seasonal AutoRegressive Integrated Moving Average) with the following parameters:
        - order=(1,1,1): Captures autoregressive, differencing (trend), and moving average components.
        - seasonal_order=(1,1,1,2): Adds seasonal influence assuming a pattern repeating every 2 years.
    - Fitted the model to historical wage data and forecasted wages for the **next 3 years**.
6. **Forecast Visualization**
    - Plotted:
        - Observed wage trend.
        - Forecasted average annualized wages for upcoming years.
        - **95% confidence intervals** (shaded pink) to account for uncertainty in prediction.

Importance of the Model

- **Trend Insight**: Helps track wage growth and compensation expectations in the H1B labor market.

- **Planning Tool**: Employers can use wage forecasts to plan budget allocations for foreign talent.
- **Policy Relevance**: Policymakers and analysts can assess how compensation evolves with labor demand and visa sponsorship volumes.
- **Modeling Practice**: Demonstrates the application of advanced time-series techniques (SARIMA) in real-world labor data forecasting.

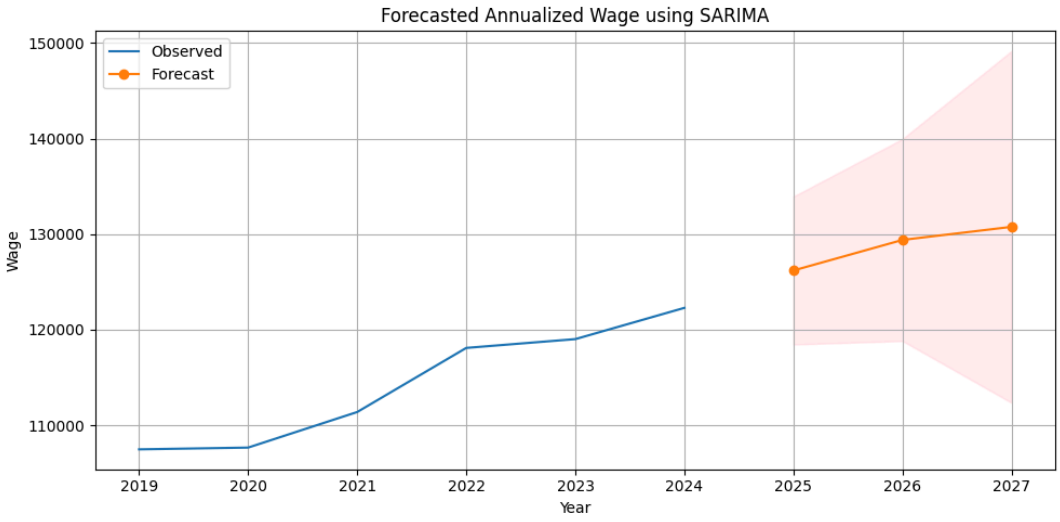


Key Insights:

1. **Overall Upward Trend**
    - There is a **steady increase in wages** across the six-year period.
    - From 2019 ($107,500) to 2024 ($122,200), average wages have increased by roughly **13.7%**.
2. **Post-2021 Acceleration**
    - Wages remained relatively flat between **2019 and 2020**, suggesting stability.

- From **2021 onwards**, there's a clear **acceleration**, likely influenced by:
  - Increased demand for specialized talent
  - Shifts in the labor market post-COVID
  - Inflation and competitive wage adjustments

3. **Sustained Growth in 2023–2024**
   - Wages continue to rise in 2023 and 2024, indicating **consistent growth** rather than short-term spikes.



Key Insights

1. **Continued Upward Wage Trend**
   - The model forecasts a **steady increase** in average wages:
     - ~ $126,000 in 2025
     - ~ $129,000 in 2026
     - ~ $131,000 in 2027
   - This implies ongoing wage growth in the H1B labor market.

2. **Moderate Growth Rate**
    - The slope of the forecast line suggests **moderate but stable growth**, following the steeper rise observed post-2021.
3. **Uncertainty Ranges Widen**
    - The confidence interval (shaded region) widens over time, indicating **growing uncertainty** in longer-term forecasts.
    - By 2027, predicted wages could range between **$113K and $149K**, depending on economic or policy shifts.
4. **Model Confidence is Strong Near-Term**
    - The interval for 2025 is relatively tight, meaning the model is **more confident** in short-term predictions.
5. **Economic and Market Influence**
    - Forecasted growth may reflect:
        - Inflation-adjusted wage increases
        - Continued demand for high-skill foreign labor
        - Policy-driven wage floor increases (e.g., prevailing wage rules)

Conclusion

The SARIMA model successfully captures the wage trend across years and provides **data-driven projections** for future wages under the H1B program. By filtering noise and applying seasonal modeling, the forecast offers **reliable guidance** for stakeholders. The projected growth or stability in wages highlights the evolving dynamics of the U.S. labor market for skilled foreign workers.

Research Question 6: Cost of Living and Crime Impact on H-1B Application Outcomes

Objective

To assess how regional economic conditions—specifically cost of living and crime rates affect H-1B visa application outcomes, and can machine learning models be used to predict high-risk areas and approval likelihoods based on these factors.

1. Data Loading & Cleaning

- Loaded H-1B dataset (Combined_LCA_Disclosure_Data_FY2020_to_FY2024.csv).
- Cleaned column names.
- Dropped columns with more than 50% missing values.
- Filled missing numeric values with median and categorical with mode or "No Data".
- Confirmed zero missing values post-cleaning.

2. Cost of Living and Crime Data Integration

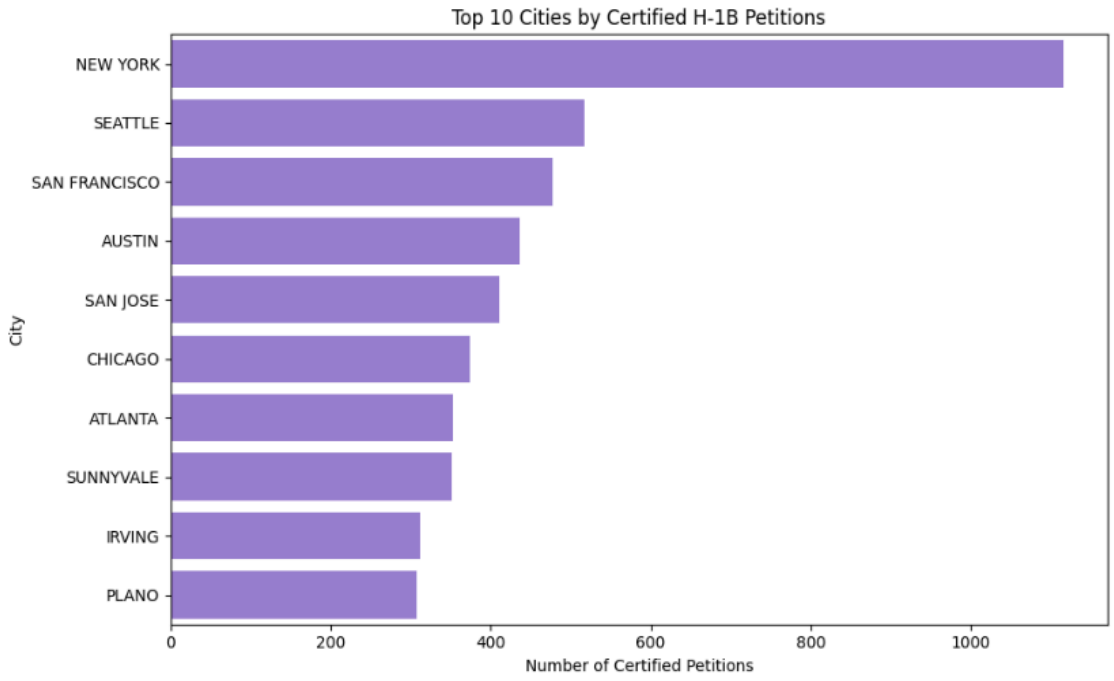
- Loaded and standardized ZIP codes in cost_of_living_by_zip_code.xlsx and crimedata.csv.
- Merged both into the main dataset using worksite_postal_code.
- Final dataset includes cost, wage, and crime metrics.

3. Filter and Derived Feature

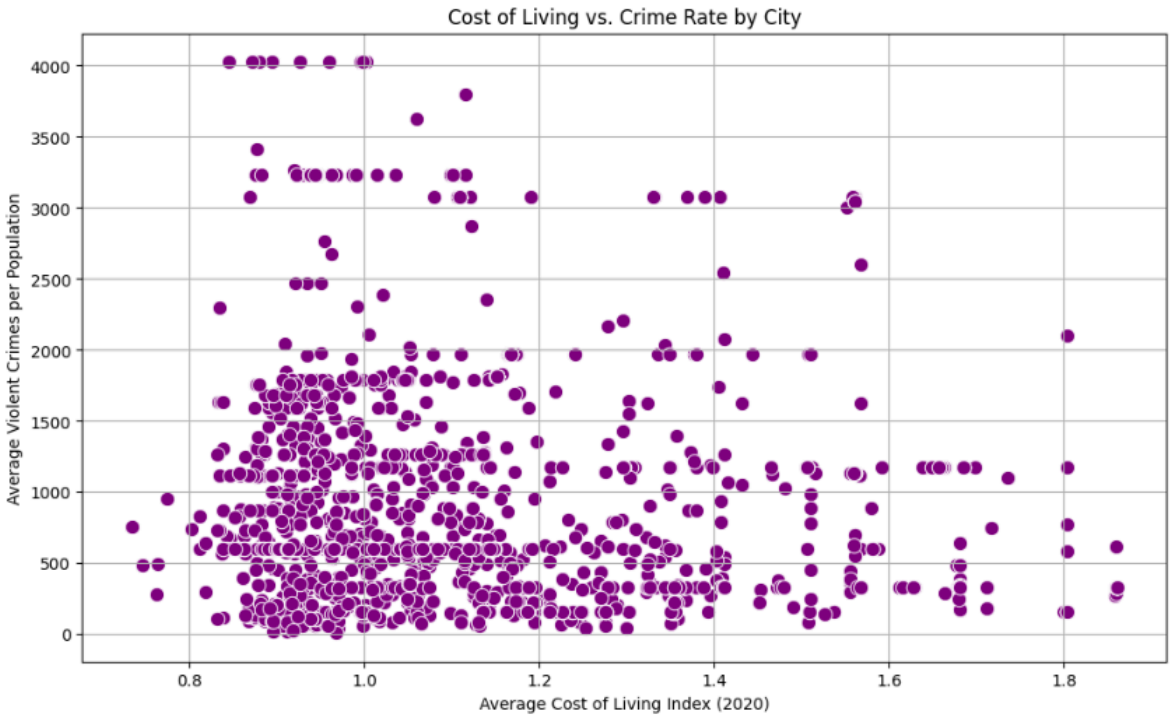
- Filtered certified cases only.
- Dropped rows missing cost, wage, or crime.
- Created wage_to_cost_ratio to analyze job value relative to cost of living.
- Created case_approved as binary variable for approval analysis.

4. Visualizations

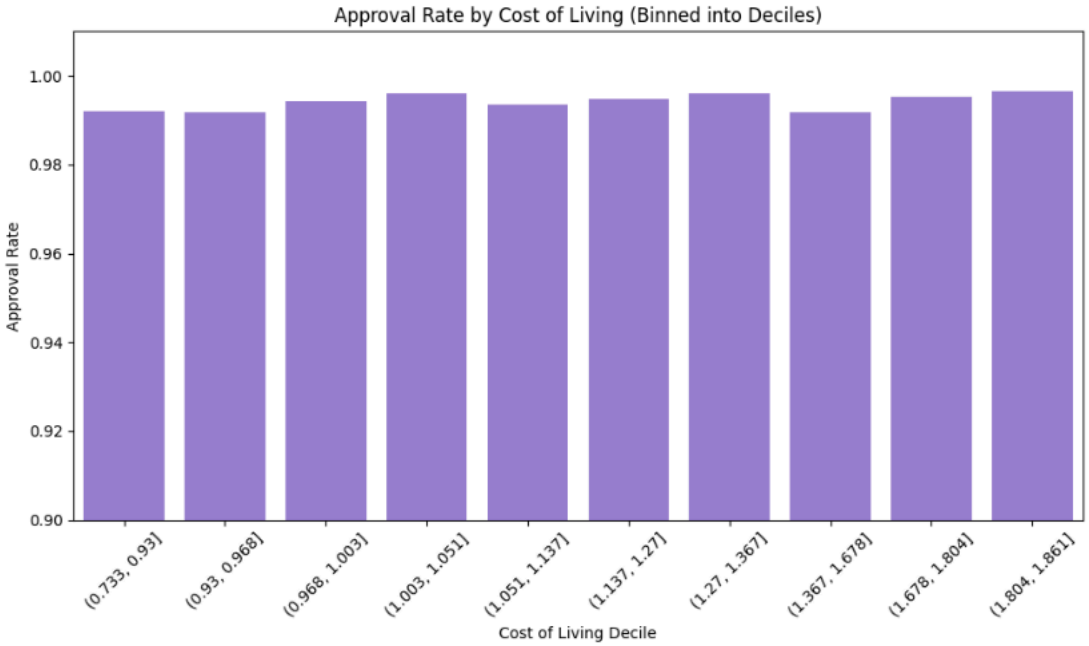
1. Top 10 Cities by Certified H-1B Petitions



2. Cost of Living vs. Violent Crime Rate



3 Approval Rate by Cost of Living Decile



4. Machine Learning Models

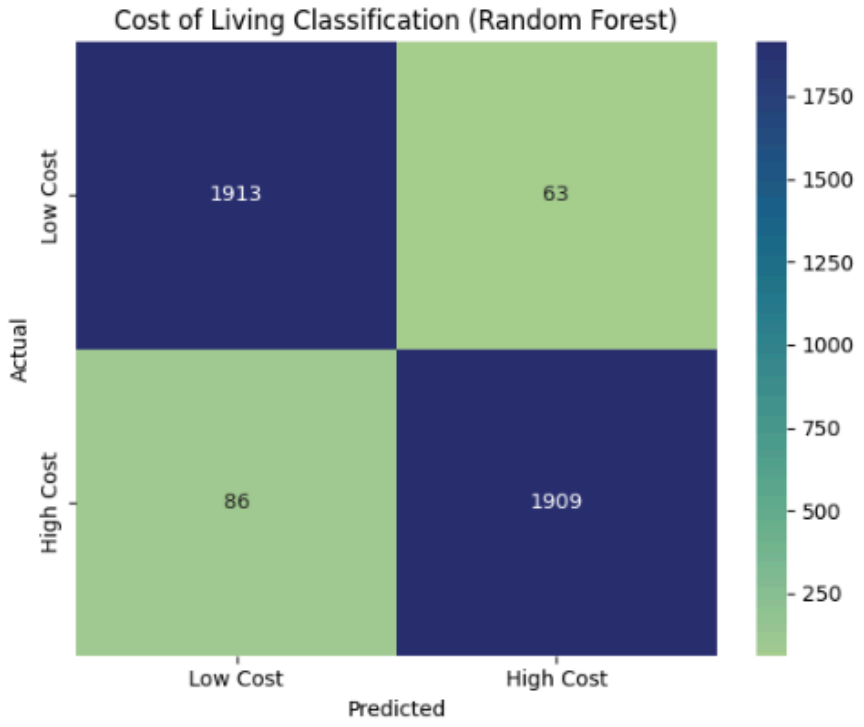
a. High Cost Area Prediction

- **Model:** Random Forest
- **Target:** Top 25% cost of living ZIPs
- **Preprocessing:** SMOTE + Scaling

```
Accuracy: 0.9625
ROC AUC Score: 0.9955
Confusion Matrix:
[[1913   63]
[  86 1909]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.97      0.96      1976
           1       0.97      0.96      0.96      1995

    accuracy                           0.96      3971
   macro avg       0.96      0.96      0.96      3971
weighted avg       0.96      0.96      0.96      3971
```



Key Takeaways:

- Achieved high performance with 96.25% accuracy and ROC AUC score of 0.9955.
- F1-scores were balanced across both classes: Low-Cost (0.96) and High-Cost (0.96).
- The model misclassified only 149 out of 3,971 samples (≈ 3.7% error rate).
- Demonstrated that job location (city/state) and title are strong indicators of a region's cost tier.
- Suggests that employers in high-cost areas tend to follow consistent patterns — making these ZIPs more predictable.

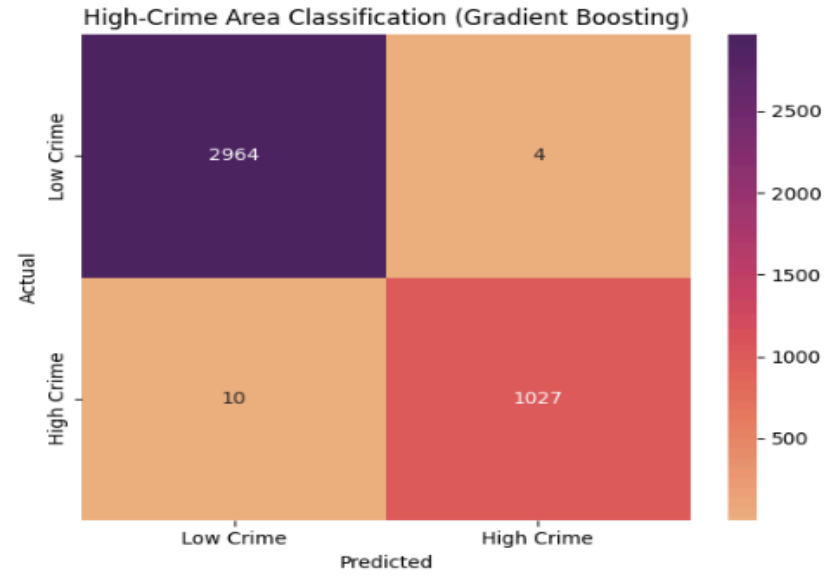
b. High Crime Area Prediction

- **Model:** Gradient Boosting Classifier
- **Target:** Top 25% crime ZIPs
- **Preprocessing:** SMOTE + Scaling

```
Accuracy: 0.9965
ROC AUC Score: 0.9989
Confusion Matrix:
[[2964    4]
 [  10 1027]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2968
           1       1.00      0.99      0.99      1037

    accuracy                           1.00      4005
   macro avg       1.00      0.99      1.00      4005
weighted avg       1.00      1.00      1.00      4005
```

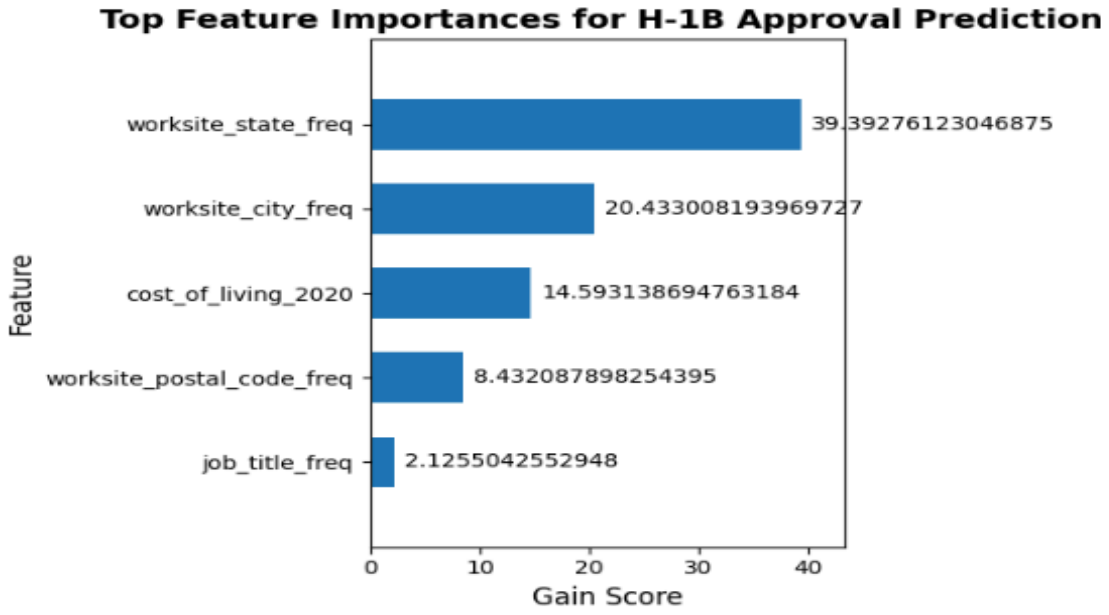


Key Takeaways:

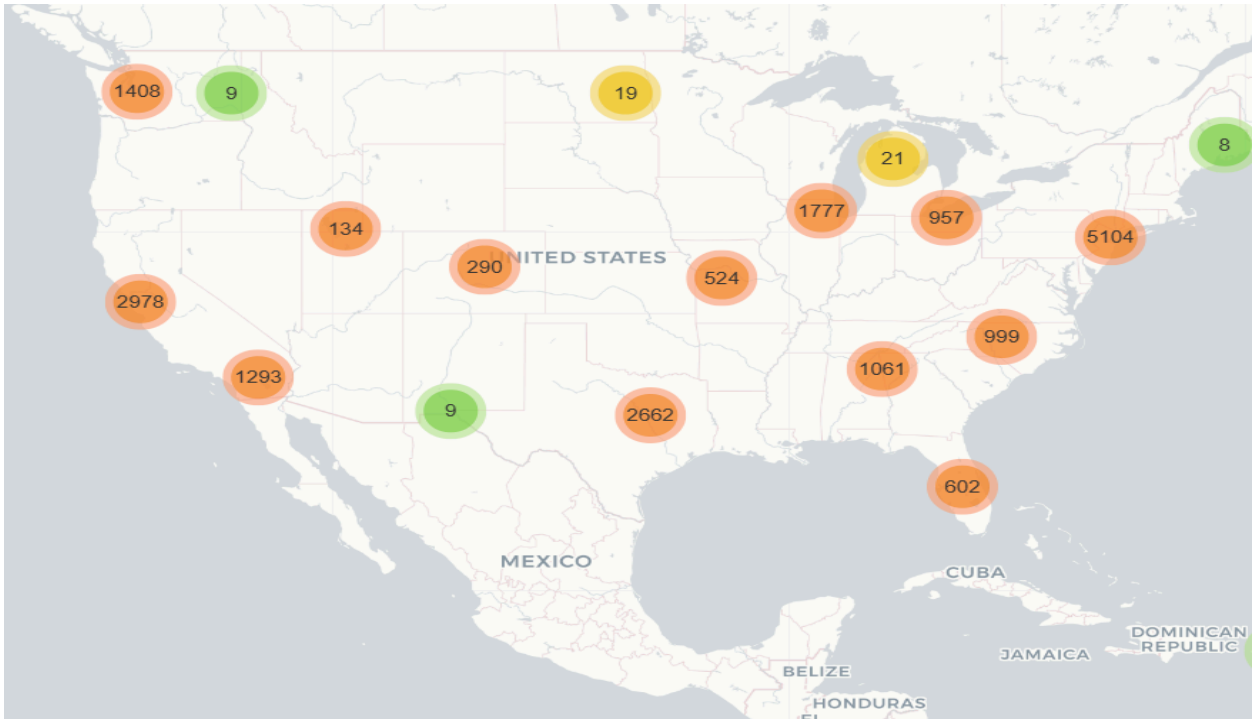
- Extremely high accuracy: 99.65%, with ROC AUC score of 0.9989.
- Misclassified only 14 out of 4,005 samples, showing excellent generalization.
- Precision, recall, and F1-score were nearly perfect across both classes (1.00 for Low Crime, 0.99 for High Crime).
- Indicates that geographic signals (e.g., worksite city/state) strongly correlate with regional crime risk.
- Validates that crime level can be effectively predicted based on H-1B-related features.

c. Feature Impact on H-1B Approval

- **Model:** XGBoost Classifier
- **Goal:** Identify which features most influence H-1B approval outcomes
- **Preprocessing:** StandardScaler + SMOTE + Stratified Train-Test Split
- **Evaluation:** Feature importance ranked by gain (top predictors only)
    - **Visualization:** Bar plot of top 10 feature importances from XGBoost model



5 Interactive Folium Map



Key Takeaways

- Each dot represents a certified H-1B job site, based on ZIP-level data merged from our final dataset.
- The dot color indicates the violent crime level:
    - Green = Low crime
    - Yellow = Moderate crime
    - Red/Orange = High crime
- Popups display detailed location metrics, including the area's violent crime rate and cost of living index .
- Areas with dense clusters of red/orange dots reflect a high volume of certified jobs in potentially higher-risk regions.

- The map highlights a geographic imbalance: many certified jobs are located in expensive or high-crime areas, raising concerns about location-based job accessibility and safety for international workers.

6. Strategic Recommendations for H-1B Job Seekers

1. **Apply in High-Certification Zones**

   Target cities like New York, Seattle, Austin, and San Francisco, where employers have strong H-1B process familiarity and higher historical approval rates.

2. **Maximize Wage-to-Cost Value**

   Prioritize job offers in locations with a strong wage-to-cost-of-living ratio to ensure both financial sustainability and favorable adjudication likelihood.

3. **Avoid Data-Deficient ZIP Codes**

   Refrain from applying to regions lacking crime or wage data, as insufficient transparency may signal elevated visa risk or administrative uncertainty.

4. **Select Low-Risk Job Titles**

   Favor job roles historically aligned with high approval rates (e.g., Software Developer, Data Analyst, Financial Analyst) to reduce scrutiny under specialty occupation criteria.

5. **Leverage High-Filing Regions' Experience**

   States like California, Texas, and New York have established legal and HR infrastructure, boosting procedural compliance and applicant support.

Strategic Recommendations for Policymakers and Immigration Analysts

1. **Address Geographic Disparities**

   Expand employer outreach in underrepresented regions to reduce

overconcentration of approvals in coastal metro areas and promote equitable access.

2. **Mandate Reporting Compliance for Local Data**

   Enforce municipal-level wage and crime data reporting to eliminate bias from low-quality ZIP-code records flagged by predictive models.

3. **Encourage Growth in Moderate-Cost Hubs**

   Support scalable visa pipelines in cities with mid-range living costs, where approval outcomes and affordability align most effectively.

4. **Integrate ML Forecasting into Policy Design**

   Utilize machine learning models (e.g., XGBoost, SARIMA) to anticipate approval bottlenecks, employer risk patterns, and regional labor demand trends.

5. **Balance Policy for Low-Filing Employers**

   Avoid saturation bias in policy frameworks by ensuring compliance-focused small and mid-sized firms receive equal access to visa allocations.

References

1. https://www.dol.gov/agencies/eta/foreign-labor/performance
2. https://www.researchgate.net/publication/340109729_A_Deep_Learning-Based_Approach_for_Predicting_the_Outcome_of_H-1B_Visa_Application/link/60f86b17169a1a0103ab192c/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19
3. RedBus2US. (n.d.). Requirements for H1B, https://redbus2us.com/h1b-visabasics-requirements-filing-dates-cap-fee-faqs
4. Neil G. Ruiz (2017, April 27), Key facts about U.S. H1B Visa program,https://www.pewresearch.org/fact-tank/2017/04/27/key-facts-about-the-u-sh-1b-visa-program/
5. https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/
6. Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: Principles and techniques for data scientists. O'Reilly Media.
7. National Bureau of Economic Research. (n.d.). The role of the H-1B visa in the U.S. economy. Retrieved from https://www.nber.org/
8. Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). Retrieved from https://otexts.com/fpp3/
9. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.
10. Zongaobian. (2024). H-1B LCA disclosure dataset (2020-2024) [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/zongaobian/h1b-lca-disclosure-data-2020-2024/data