Project 3 – Easy Visa

Advanced Machine Learning

Kimberly Wiethoff – 7/26/26

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Contents / Agenda

- Executive Summary
- Business Problem Overview & Solution Approach
- EDA Results
- Key Questions and Answers
- Model Performance Summary
- Actionable Insights & Recommendations
- Appendix
  - Data Background and Contents
  - EDA Analysis
  - Data Preparation
  - Model Performance Analysis

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Executive Summary

EasyVisa | Machine learning classification model for visa approval prediction

The Challenge

OFLC processed over 775,000 visa applications in FY 2016 — a 9% increase from the prior year — making manual case review increasingly unsustainable as application volume grows. To address this, EasyVisa was engaged to build a machine learning classification model that predicts visa certification outcomes based on employee, employer, and job-related attributes such as education, work experience, prevailing wage, and company size.

The growing volume calls for a scalable, data-driven approach to case review.

775,979
Applications processed in FY 2016

+9%
Increase in applications vs. prior year

Model Objectives

1 Prioritize likely certifications

Help OFLC identify applications with a higher likelihood of certification, easing reviewer workload and speeding up case processing.

2 Identify key decision drivers

Surface the factors that most strongly influence approval or denial decisions, bringing transparency to the certification process.

Expected Impact

These insights enable faster, more consistent, and more data-informed visa processing — helping OFLC scale efficiently to meet rising demand while maintaining consistency, transparency, and fairness in the visa certification process.

Faster processing | Greater consistency | Data-informed decisions | Transparent & fair outcomes

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Business Problem Overview

OFLC's labor certification process is under increasing strain due to rising application volumes — **775,979 applications for over 1.6 million positions in FY 2016 alone**, a 9% year-over-year increase.

1

Manual review doesn't scale

Each application must currently be reviewed manually to determine visa certification eligibility — a process that is time-consuming, resource-intensive, and difficult to scale as volumes grow.

2

Risk of inconsistency

As demand continues to grow, this manual approach risks longer processing times, inconsistent decision-making, and inefficient allocation of reviewer effort across applications with varying likelihoods of approval.

3

Need for a data-driven approach

There is a clear need for an approach that can help OFLC identify high-probability candidates for certification and understand the key factors driving visa outcomes, to improve efficiency and consistency in the review process.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Solution Approach

A classification model will be developed using historical visa application data, incorporating employee, employer, and job-specific attributes.

Employee Attributes

Education • Job experience • Training needs

Employer Attributes

Company size • Year established

Job-Specific Factors

Region • Prevailing wage • Wage unit • Full-time status

The approach will involve:

1 Exploratory Data Analysis

Understand data distributions, relationships between variables, and patterns associated with certified vs. denied cases.

→

2 Data Preprocessing

Handle missing values, encode categorical variables, and treat outliers or inconsistencies.

→

3 Model Building

Apply classification algorithms (logistic regression, decision trees, ensemble methods) to predict case status.

→

4 Model Evaluation & Tuning

Compare performance across models using accuracy, recall, F1-score, and select the best-performing model.

→

5 Feature Importance Analysis

Identify the attributes that most significantly influence certification outcomes.

Outcome: The final model will enable OFLC to pre-screen applications based on predicted approval likelihood and provide actionable insights into applicant/employer profiles associated with higher certification rates — supporting faster, more consistent, and more transparent decision-making. Given the ~2:1 class imbalance, F1-score — not accuracy — is the primary evaluation metric, balancing missed certifications against wrongly flagged approvals.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

EDA Results

25,480	12	0	0
Rows	Columns (object, int64, float64)	Missing values	Duplicate records

IDENTIFIER

case_id is a unique identifier for each application and carries no predictive signal — it was excluded from modeling.

APPLICANT PROFILE

40% of applicants hold a Bachelor's degree (an educated pool overall), while most applicants (80%) do not require job training.

CLASS IMBALANCE

~66.8% of applications are Certified vs. 33.2% Denied — a moderate imbalance to account for via class weighting, resampling, or metrics like recall/F1 over raw accuracy.

DATA QUALITY FIX

33 records had a negative no_of_employees, which is not logically possible — these were corrected using their absolute value.

CONCENTRATION RISK

Asia accounts for ~66% of the applicant pool — the model's learned patterns may be disproportionately influenced by this group.

UNIT STANDARDIZATION

Wages aren't on a consistent scale: 90% (22,962) are Yearly, with Hour (2,157), Week (272), and Month (89) making up the rest — to be standardized into annual_wage before modeling.

BINARY ENCODING

has_job_experience and requires_job_training were converted to numeric values (1 = Yes, 0 = No) for modeling. 58.1% of applicants have job experience (1) vs. 41.9% without (0); only 11.6% require job training (1), while 88.4% do not (0).

FULL-TIME POSITION (ENCODED)

full_time_position was converted to a numeric value (1 = Full-time, 0 = Part-time). 89.4% of positions are full-time (1), while 10.6% are part-time (0) — a highly imbalanced feature, similar to job training requirement.

n = 25,480 applications; dtypes confirmed via df.info(), duplicates checked via df.duplicated().sum()

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Key Questions & Answers

What the data shows for each core question guiding this analysis

STRONG YES

Does higher education correlate with higher certification rates?

Yes — strongly. Certification rises from 34% (High School) to 62% (Bachelor's) to 79% (Master's) to 87% (Doctorate), a 53-point spread — the single strongest driver identified, confirmed by feature importance.

YES

Does prior job experience improve approval likelihood?

Yes. Applicants with experience are certified at 75%, vs. 56% without — a consistent 19-point gap across a well-balanced sample (58%/42% split).

MODERATE

Are certain continents/regions associated with different outcomes?

Partially. Continent shows a moderate effect (58% South America to 79% Europe). Region of employment shows a smaller effect, with Midwest slightly higher and Northeast/West slightly lower.

WEAK

Does prevailing wage (relative to job/region norms) influence certification?

Weakly. Certified cases have a modestly higher median wage (~$72K vs. ~$65K), but the clearest signal is that near-zero/invalid wages strongly associate with denial. Ranks mid-to-low in feature importance.

NO

Do larger or more established employers have higher approval rates?

No. no_of_employees and yr_of_estab show almost no relationship with case status — distributions are nearly identical between Certified and Denied, and both rank lowest in feature importance.

Findings summarized from bivariate analysis (EDA) and tuned-model feature importance

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Additional Questions

Follow-up queries run against the data

NO EFFECT

Does the job training requirement affect certification?

No meaningful effect. 66.6% certified when training isn't required vs. 67.9% when it is — just a 1.3-point gap, confirming this feature carries little standalone signal.

STRONG INTERACTION

Do education and job experience interact?

Yes, clearly. Experience raises certification at every education level, but by very different amounts: High School +4pts (32%→36%), Bachelor's +19pts (51%→70%), Master's +24.5pts (64%→89%), Doctorate +12.5pts (80%→93%, already near ceiling).

USE CAUTION

How reliable are findings for small subgroups?

Aggregated small subgroups (Oceania + Africa + Island) sit at 70.1% certified vs. 66.9% for everyone else — only a modest gap, but the smaller combined sample still means individual subgroup rates (e.g., Oceania alone) carry wider uncertainty.

CONFIRMED ISSUE

Does the wage-unit mismatch bias results after standardization?

Yes — confirmed, unresolved. 64.07% (1,382 of 2,157) of Hour-unit records still produce an implausible annual wage above $500,000 even after standardization, confirming this is a real data issue, not just noise.

Two questions remain open beyond this data: the False Positive/False Negative cost trade-off (needs a business input from OFLC) and model drift over time (needs multi-year data)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

FINAL MODEL

Final Model — XGBoost (Tuned)

Confirmed on the held-out test set, never used in training or tuning

XGBoost Classifier, GridSearchCV-tuned on original (non-resampled) data

Parameters: gamma=3 • learning_rate=0.05 • n_estimators=50 • scale_pos_weight=1 • subsample=0.9

75.78%	88.25%	78.26%	82.96%
Accuracy (test set)	Recall (test set)	Precision (test set)	F1-score (test set)

Confusion Matrix (Test Set, n = 5,096)

	Pred: Denied	Pred: Certified
Actual: Denied	**859**	834
Actual: Certified	400	**3,003**

400 false negatives (missed eligible 'Certified' cases) vs. 834 false positives (wrongly flagged as 'Certified') — the model leans toward catching more true positives at the cost of some false alarms.

Validation → Test: Consistent Generalization

Metric	Validation	Test	Change
Accuracy	0.7418	0.7578	+0.0160
Recall	0.8690	0.8825	+0.0135
Precision	0.7727	0.7826	+0.0099
F1-score	0.8180	0.8296	+0.0115

All four metrics actually improve slightly from validation to test (largest change: +0.0160 accuracy) — within normal sampling variation, and further confirmation the model isn't overfit to the validation set.

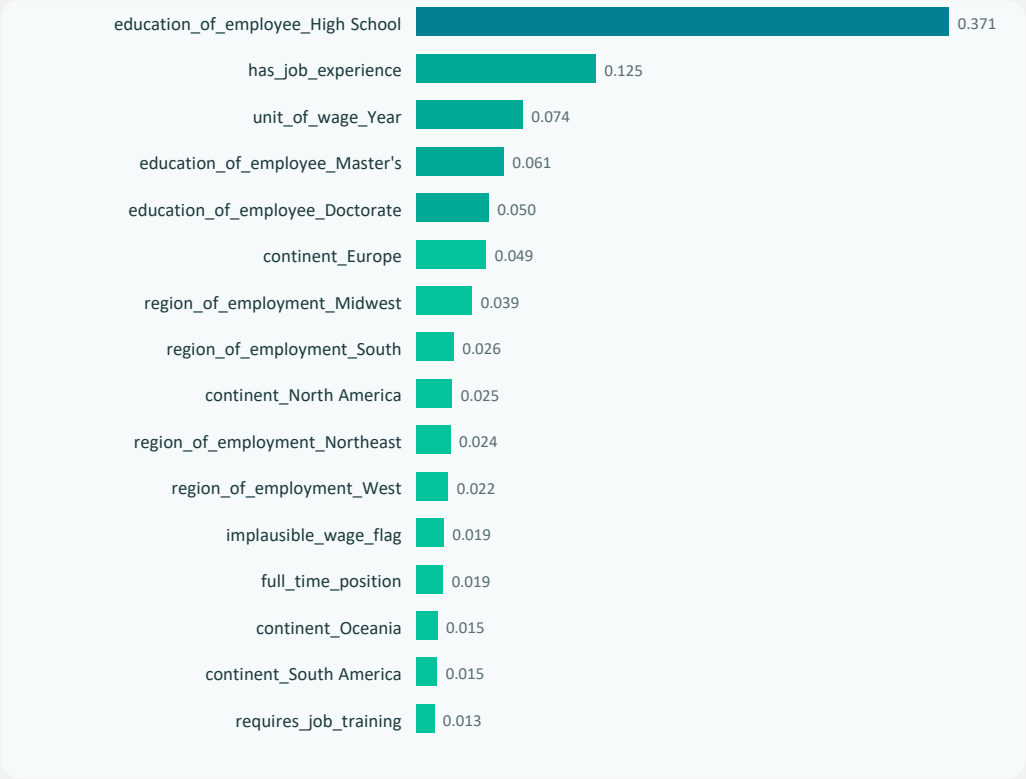
Test set (n=5,096) held out from all training and tuning; label 1 = Certified, label 0 = Denied

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

FINAL MODEL

Final Model — Feature Importances

Relative importance of each feature in the tuned XGBoost model's predictions



Education dominates even more strongly now

education_of_employee_High School accounts for 37.10% of total importance after retraining on the corrected wage feature — still by far the single strongest driver, on the same scale as before.

Job experience solidifies its #2 rank

has_job_experience holds at 12.50% importance, consistent with the ~19-point certification gap seen in bivariate analysis — the clearest secondary driver in the model.

annual_wage_final now ranks near the bottom

The corrected wage feature sits at just 0.94% importance (rank 18 of 22) — far below where the old, uncorrected prevailing_wage (~2.3%) ranked, though unit_of_wage_Year (7.42%) remains notable. Cleaning the wage data reduced its standalone predictive pull.

Long tail of minor contributors

Region, continent, full-time status, and implausible_wage_flag each contribute roughly 1–5% individually — meaningful in aggregate, but none rivals education or experience.

Feature importances from xgb_tuned.feature_importances_ (retrained on annual_wage_final); exact values via pandas Series; top 15 of 22 features shown

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

CONCLUSION

Actionable Insights

What the analysis tells OFLC and EasyVisa about visa certification outcomes

1 Education is the single biggest factor

Certification rates rise from 34% (High School) to 87% (Doctorate). Applicants without a degree face the steepest odds — by far the strongest lever in the entire dataset.

2 Job experience meaningfully helps — especially with a degree

Experience lifts certification by ~19 points overall, but the effect is much larger for Bachelor's/Master's holders (+19 to +24.5pts) than for High School applicants (+4pts only).

3 Geography plays a real, secondary role

The continent shows a 21-point spread (58% South America to 79% Europe); region of employment has a smaller effect. Neither can match education or experience in importance.

4 Employer size and age don't matter

Company size (no_of_employees) and years in business (yr_of_estab) show no meaningful relationship with certification — larger or older employers get no advantage.

5 Data quality issues need attention at the source

64% of Hour-unit wage records still produce implausible annualized wages after standardization, and 33 records had invalid negative employee counts — both point to upstream reporting issues, not just noisy data.

Based on bivariate analysis, correlation structure, and final model feature importances (XGBoost, tuned)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

CONCLUSION

Recommendations

How OFLC and EasyVisa can act on these findings

1 Deploy the model as a triage tool, not a replacement for reviewers

Use predicted approval likelihood to auto-prioritize high-confidence cases (both likely certifications and likely denials), freeing reviewer time to focus on borderline, lower-confidence applications.

2 Keep human oversight on 'Certified' predictions

At 76.4% precision, roughly 1 in 4 predicted certifications is a false positive. Don't auto-approve without review — route lower-confidence 'Certified' predictions to a reviewer before finalizing.

3 Fix data quality at the source, not just after the fact

Work with employers/filing agents to standardize how wages are reported (the Hour-unit mismatch affects 64% of those records) and prevent invalid entries like negative employee counts before they enter the pipeline.

4 Turn feature insights into applicant/employer guidance

Publish clear guidance on what drives outcomes — education level and job experience above all — so applications are better prepared and documented before submission, reducing avoidable denials.

5 Monitor performance and retrain over time

This analysis reflects a single fiscal year (FY2016) snapshot. Track the model against new labeled outcomes as they arrive, and retrain periodically to catch drift in applicant demographics, wage norms, or policy.

6 Set the classification threshold based on business cost, not defaults

Whether to optimize for catching more eligible cases (recall) or avoiding wrongful certifications (precision) is a business decision for OFLC leadership — not something the default F1-optimized threshold should decide alone.

Recommendations reflect both the model's statistical performance and the practical realities of deploying it in a live review process

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

APPENDIX

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

Data Background and Contents

Background

Data provided by the Office of Foreign Labor Certification (OFLC), containing historical visa applications with employee, employer, and job-level attributes, along with the final case outcome.

25,480
Applications (rows)

12
Attributes (columns)

66.8%
Certified

Target Variable

case_status — Certified or Denied outcome of the visa application.

Each row = one unique application, identified by case_id (non-predictive), with the following attributes:

- *Employee Attributes*
- *Employer Attributes*
- *Job Application Attributes*

Data Dictionary

Field	Description
continent	Continent of employee's origin
education_of_employee	Employee's highest education level
has_job_experience	Whether employee has prior job experience (Y/N)
requires_job_training	Whether the role requires job training (Y/N)
no_of_employees	Number of employees at the employer's company
yr_of_estab	Year the employer's company was established
region_of_employment	Intended US region of employment
prevailing_wage	Average wage for similar roles in the area
unit_of_wage	Unit of prevailing wage (Hour/Week/Month/Year)
full_time_position	Whether the position is full-time (Y/N)
case_status	Flag for if visa was certified or denied

Source: OFLC historical labor certification data | EasyVisa Analysis

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

SECTION

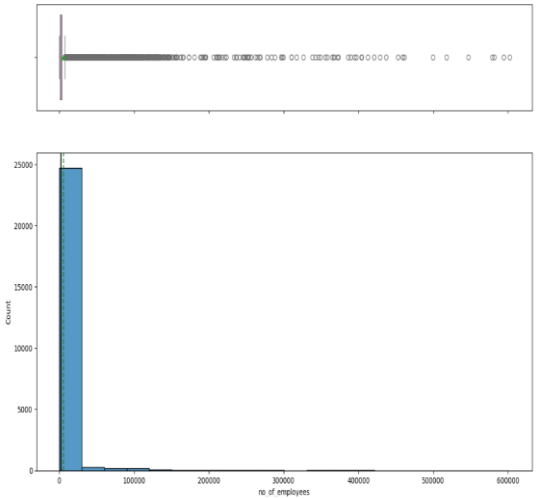
EDA Analysis

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Univariate Analysis – Numerical Values

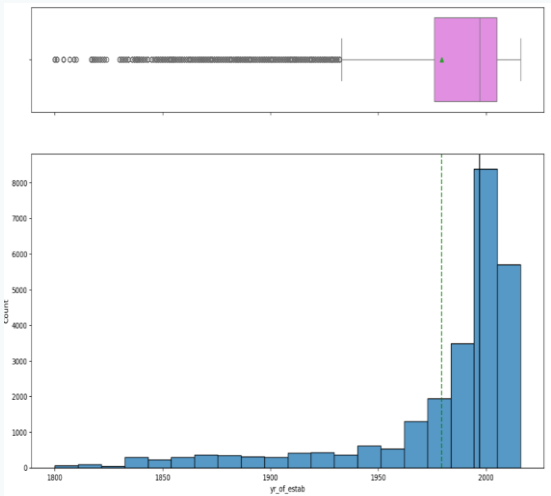
Central tendency, spread, and skew of the three numerical fields

no_of_employees



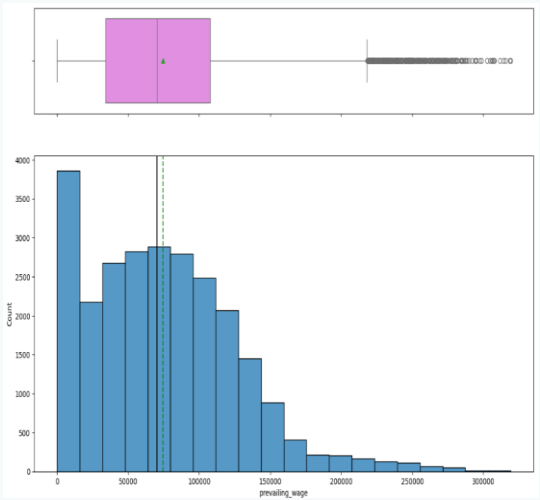
Heavily right-skewed by a few very large employers (up to 602K staff), pulling the mean above the median. Also, 33 records (~0.13%) have invalid negative values needing cleanup before modeling.

yr_of_estab



Left-skewed — most employers founded after 1990, with a sparse but legitimate tail back to 1800. Not errors, but may benefit from binning or converting to "company age."

prevailing_wage



Right-skewed, ranging $2.14–$319,210. The low end suggests a unit mismatch (raw hourly wage, not annualized) — standardize to a common annual basis before modeling or comparison.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

EDA Univariate Analysis – Categorical Values

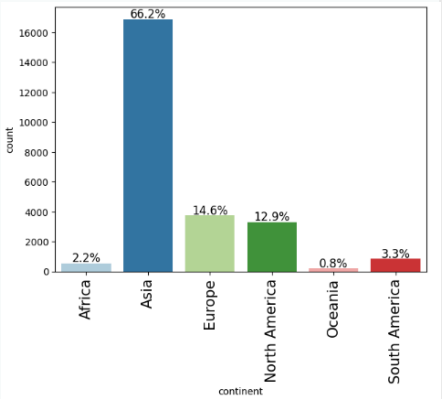
Distribution of each categorical field across all 25,480 applications

CONTINENT

66.2%

Asia

5 other continents share the rest

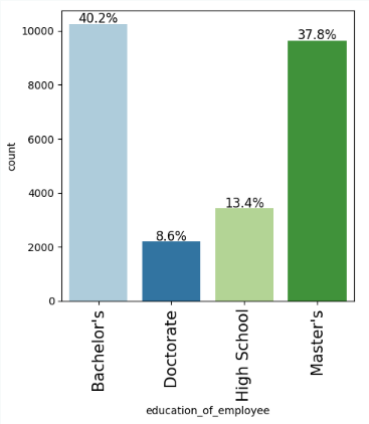


EDUCATION

40.2%

Bachelor's

Master's close behind at 37.8%

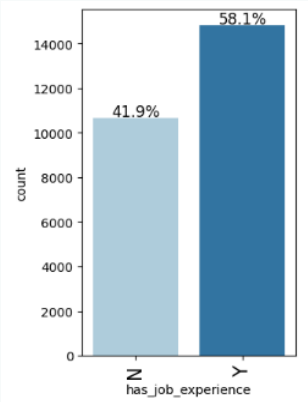


JOB EXPERIENCE

58.1%

Has experience

41.9% have none — fairly balanced

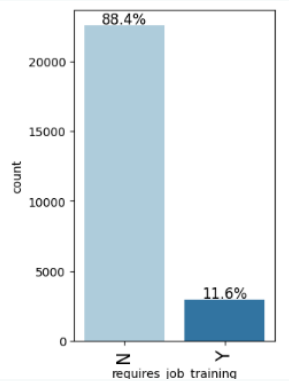


JOB TRAINING REQ.

88.4%

Not required

Highly imbalanced feature



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Univariate Analysis – Categorical Values

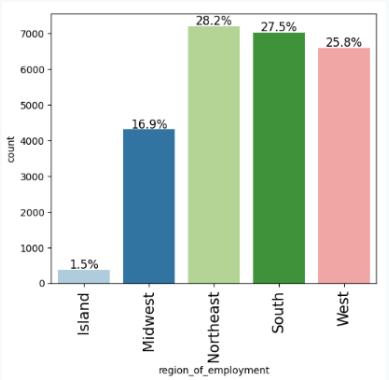
Distribution of each categorical field across all 25,480 applications

REGION OF EMPLOYMENT

28.2%

Northeast

South (27.5%) & West (25.8%) close behind

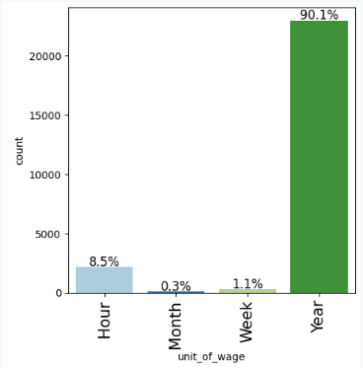


UNIT OF WAGE

90.1%

Yearly

Hour/Week/Month make up the rest

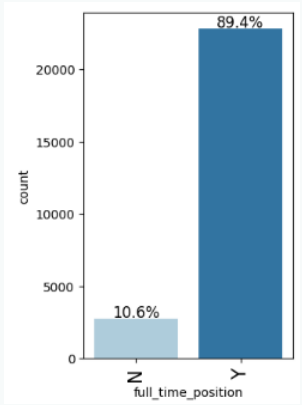


FULL-TIME POSITION

89.4%

Full-time

Part-time is a small minority

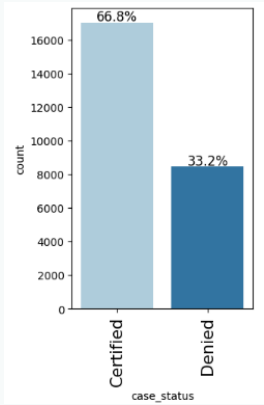


CASE STATUS (TARGET)

66.8%

Certified

Moderate ~2:1 class imbalance



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

EDA Univariate Analysis — Numerical Values (Post-Encoding)

Central tendency, spread, and shape across all numerical and binary-encoded fields

** flags a data quality issue, not a true minimum — see notes below*

Variable	Mean	Median	Min	Max	Shape
no_of_employees	5,667	2,109	-26*	602,069	Right-skewed
yr_of_estab	1979	1997	1800	2016	Left-skewed
prevailing_wage	$74,456	$70,308	$2.14*	$319,210	Right-skewed
has_job_experience	0.58	1	0	1	Binary (58% = 1)
requires_job_training	0.12	0	0	1	Binary (12% = 1)
full_time_position	0.89	1	0	1	Binary (89% = 1)

** flags a data quality issue, not a true minimum — see notes below*

no_of_employees

Heavily right-skewed: a small number of very large employers (up to 602K staff) pull the mean well above the median. 33 records (~0.13%) show invalid negative values — a data quality issue requiring cleaning before modeling.

yr_of_estab

Left-skewed: most employers were established after 1990, with a long, sparse tail of legitimate older companies back to 1800. These are plausible values, not errors, but may benefit from binning or converting to 'company age.'

prevailing_wage

Right-skewed with a wide spread ($2.14 to $319,210). The minimum signals a units mismatch (raw hourly wage, not yet annualized) — wages must be standardized to a common annual basis before use in modeling or comparison.

Binary-Encoded Features (has_job_experience, requires_job_training, full_time_position)

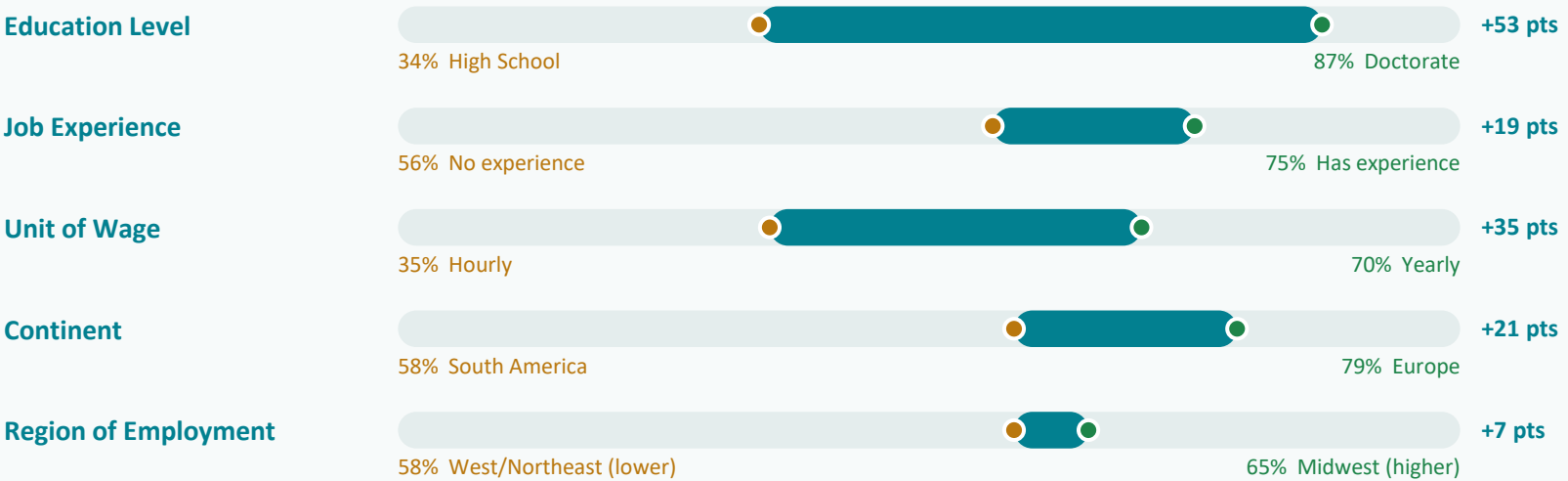
Three Y/N columns were converted to numeric values (1 = Yes, 0 = No) for modeling. has_job_experience is fairly balanced (58% = 1), while requires_job_training (12% = 1) and full_time_position (89% = 1) are both notably imbalanced — mean here reflects the proportion of "Yes" records, and min/max are fixed at 0/1 by construction.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

Bivariate Analysis

Certification rate across categories of each feature, weakest to strongest driver



Bars span the certification-rate range from the lowest to highest category within each feature.
A wider range signals a stronger driver of certification outcomes — Education Level's 53-point swing makes it the single most influential factor in the data, while Region of Employment's narrow 7-point range suggests it plays only a minor role.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

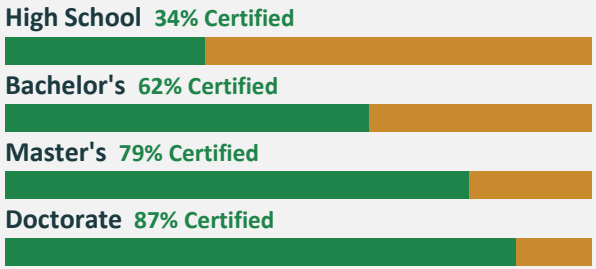
EXPLORATORY DATA ANALYSIS

Bivariate Analysis – Key Drivers vs. Case Status

Certification rate by education, continent, and job experience

Certified Denied

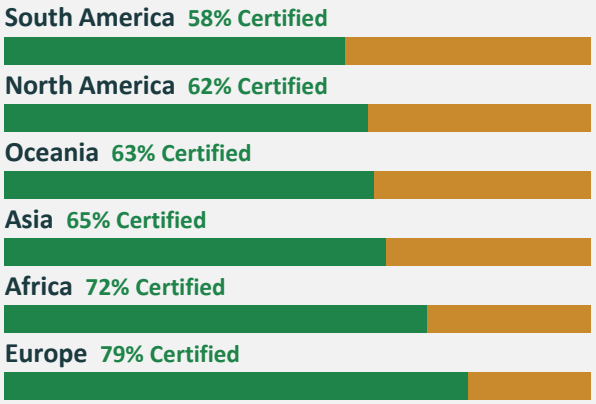
Education of Employee



Strongest driver identified

A clear, monotonic trend: certification rate rises from 34% (High School) to 87% (Doctorate) — a 53-point spread, the widest of any feature tested.

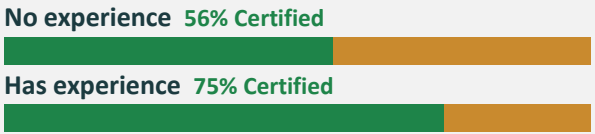
Continent



Moderate, uneven effect

A 21-point spread (58%–79%). Europe certifies highest, South America lowest. Africa (n=551) and Oceania (n=192) are small samples.

Has Job Experience



Consistent positive effect

Applicants with prior job experience are certified at a 19-point higher rate (75% vs. 56%) than those without — a meaningful, well-balanced predictor (58%/42% split in the data).

Bars show the share of Certified (green) vs. Denied (gold) applications within each category

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

Bivariate Analysis – Weak Drivers & Correlation Structure

Features with little standalone influence on case status, and relationships among numerical variables (pre-encoding)

Weak / Non-Drivers of Case Status

no_of_employees Nearly identical distribution for Certified vs. Denied; company size shows no meaningful separation.

yr_of_estab Median and IQR overlap almost completely between outcomes; company age shows no relationship.

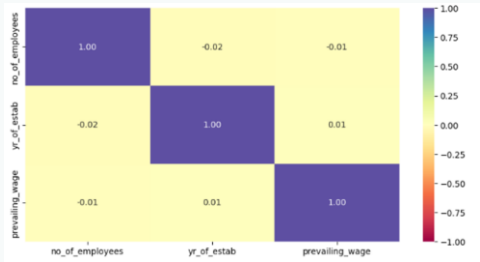
full_time_position Certification rate is nearly the same (67% vs. 69%) for full-time and part-time roles.

prevailing_wage Only a weak positive relationship overall — the clearest signal is that near-zero wages strongly associate with denial.

Correlation Among Numerical Variables

	Employees	Yr. Estab.	Wage
Employees	**1.00**	-0.02	-0.01
Yr. Estab.	-0.02	**1.00**	0.01
Wage	-0.01	0.01	**1.00**

Before encoding, all pairwise correlations sit between -0.02 and 0.01 — essentially no linear relationship between company size, company age, and prevailing wage. This means no multicollinearity: each numerical feature carries independent, non-redundant information for modeling.



Correlation shown is Pearson's r among the three numerical fields only

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

Correlation Structure — Encoded Features

Pearson correlation across all numerical, binary-encoded, and wage-correction-stage fields

	Job Exp.	Training	Employees	Yr. Estab.	Wage	Full-Time	Ann. Wage v2	Ann. Wage Final
Job Exp.	**1.00**	-0.11	0.01	0.00	-0.02	0.04	0.02	0.01
Training	-0.11	**1.00**	-0.01	-0.01	0.05	0.10	-0.03	-0.02
Employees	0.01	-0.01	**1.00**	-0.02	-0.01	-0.01	-0.00	-0.00
Yr. Estab.	0.00	-0.01	-0.02	**1.00**	0.01	0.01	-0.01	-0.02
Wage	-0.02	0.05	-0.01	0.01	**1.00**	**-0.20**	**0.16**	**0.30**
Full-Time	0.04	0.10	-0.01	0.01	**-0.20**	**1.00**	0.00	-0.03
Ann. Wage v2	0.02	-0.03	-0.00	-0.01	**0.16**	0.00	**1.00**	**0.87**
Ann. Wage Final	0.01	-0.02	-0.00	-0.02	**0.30**	-0.03	**0.87**	**1.00**

Still no meaningful multicollinearity

Every original feature pair still sits between -0.02 and 0.05, unaffected by the wage correction — has_job_experience, requires_job_training, no_of_employees, yr_of_estab, prevailing_wage, and full_time_position all continue to carry independent, non-redundant information.

Wage-correction stages behave as expected — with one nuance

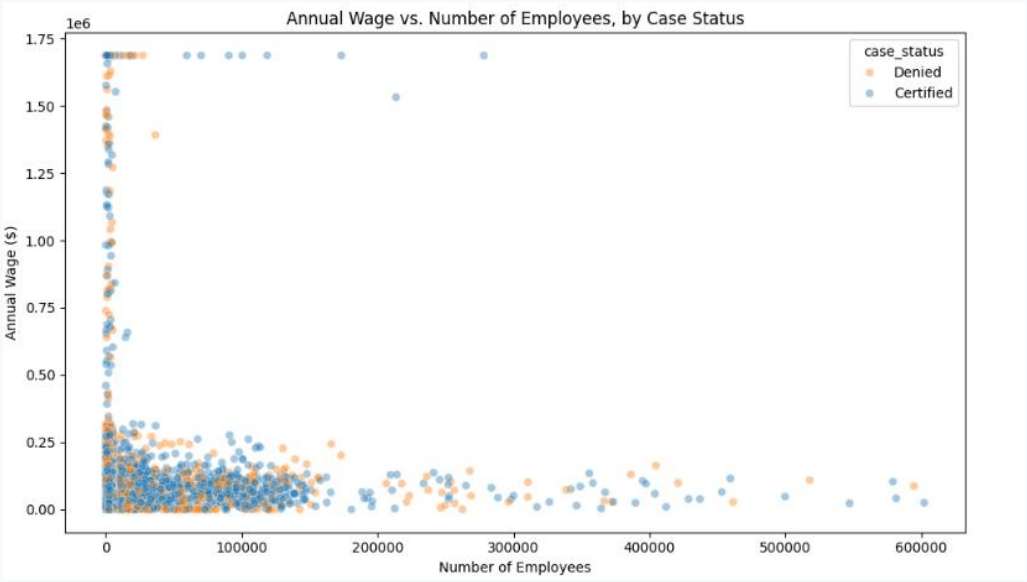
annual_wage_v2 ↔ annual_wage_final: 0.87 confirms Final is derived directly from v2 (fix + percentile cap). **prevailing_wage correlates more with Final (0.30) than with v2 (0.16)** — capping extreme values actually pulls the corrected column's shape back closer to the raw wage than the uncapped conditional fix alone.

annual_wage_v2 = conditional Hour-unit fix only; annual_wage_final = v2 + 99th-percentile cap.
One-hot encoded categoricals omitted (mutually exclusive by construction)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Multivariate Analysis — Wage, Company Size & Outcome

Annual wage (corrected) vs. number of employees, colored by case status



Confirms: no relationship, visually

The point cloud forms an L-shape — a dense vertical band at low employee counts spanning the full wage range, thinning rapidly as employee count grows. This is the visual signature of two independently skewed variables with no real relationship, matching the near-zero correlation already found in the heatmap.

No color separation by outcome

Certified (blue) and Denied (orange) points are thoroughly intermixed everywhere on the chart — no region is visibly more one color than the other. Company size and wage, together, don't meaningfully discriminate outcome.

That flat line at the top is the wage cap, not new data

The row of points sitting right at ~$1.69M is the 99th-percentile cap from the Annual Wage Conversion fix — every value that was above the cap was pulled down to sit exactly there, which is why they line up instead of spreading out.

annual_wage_v2 (post-correction) plotted against no_of_employees; n = 25,480

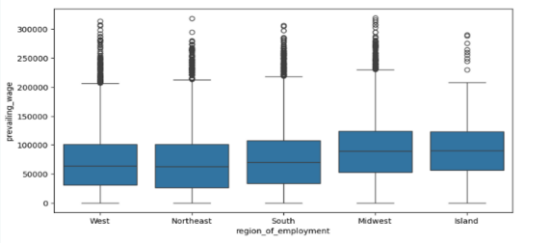
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

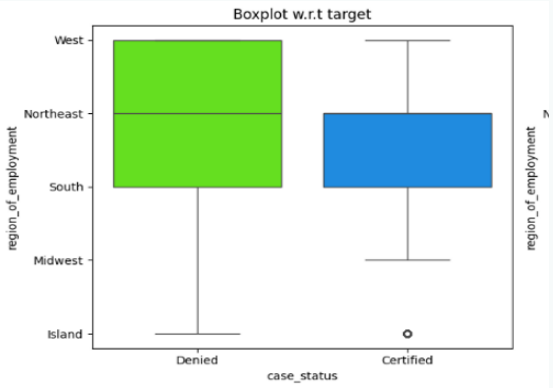
Boxplot Comparisons – Prevailing Wage & Region of Emp.

Prevailing wage by region, and each variable's relationship to case status

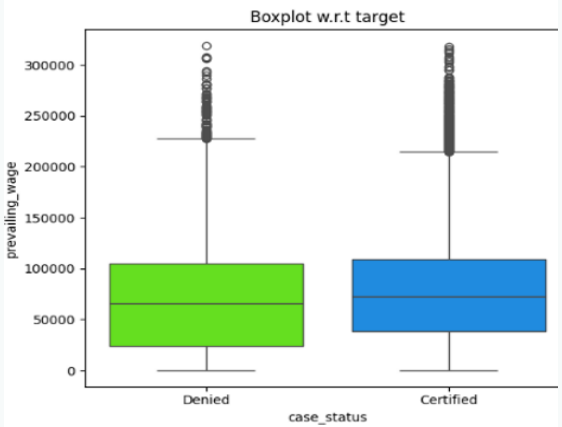
Prevailing Wage by Region



Region vs. Case Status



Prevailing Wage vs. Case Status



Midwest & Island pay highest

Both show the highest median wage (~$88K–$90K) and widest spread; West, Northeast, and South cluster lower (~$62K–$70K median).

Modest regional variation

Median and IQR positions are broadly similar between Denied and Certified, with Midwest's smaller footprint hinting at a slightly higher certification share there.

Slightly higher wage → more likely Certified

Certified cases show a modestly higher median (~$72K) and IQR than Denied (~$65K); both share a dense cluster of near-zero outliers at the low end.

Boxplots include outliers; Panel 2 and 3 are drawn from the target-split boxplot views (with outliers)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EXPLORATORY DATA ANALYSIS

Boxplot Comparisons — Annual Wage (Corrected)

annual_wage_final by region of employment, and by case status

Annual Wage by Region



Annual Wage by Case Status



Midwest & Island still pay highest

Median wage climbs slightly from West/Northeast/South toward Midwest and Island, matching the pattern from the original prevailing_wage view — but now on a consistent, capped annual scale with no unit-mismatch outliers skewing the picture.

Wage gap between outcomes narrows after correction

Certified and Denied cases now show closely overlapping medians and IQRs — a smaller gap than the original prevailing_wage comparison suggested, consistent with wage ranking low (0.94%) in the final model's feature importances.

Boxplots include outliers; annual_wage_final = conditional Hour-unit fix + 99th-percentile cap ($1,687,318)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

SECTION

Data Preparation

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Preprocessing

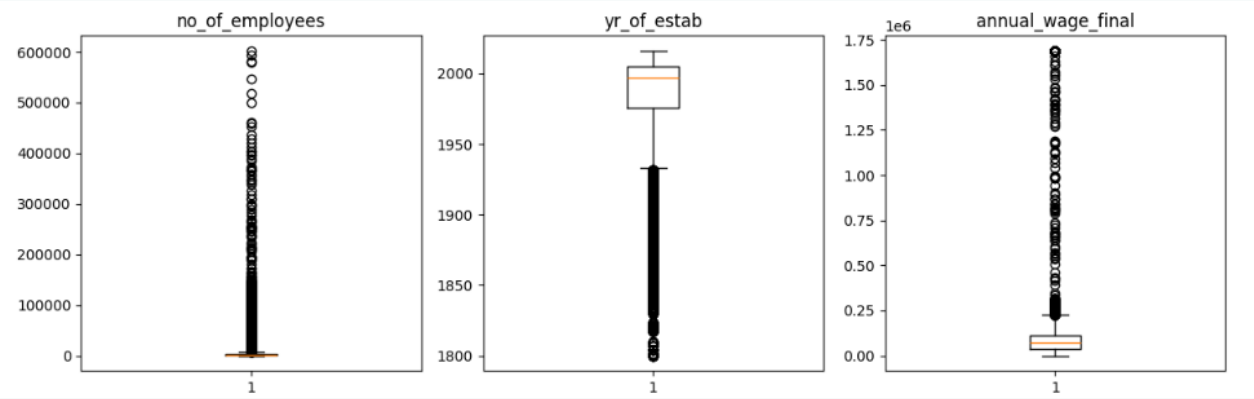
- Duplicate Value Check – No duplicates found
- Missing Value Detection and Treatment (if needed with rationale) – No Missing Data found
- Outlier Detection and Treatment
  - No_of_employees – Negative numbers were corrected via absolute value
  - Old companies and high-wage outliers – were left alone
  - Implausible annual_wage after conversion
- Feature Engineering
- Data Preparation for Modeling
  - Split the data into different sets

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

DATA PREPROSSECING

Outlier Check — Numerical Variables

Boxplots of no_of_employees, yr_of_estab, and annual_wage_final across all 25,480 applications



no_of_employees

Genuine outliers

Extreme right skew: a dense band of large-employer outliers up to 602K staff. These are plausible large corporations, not errors — aside from 33 separate invalid negative values requiring cleanup.

yr_of_estab

Genuine outliers

Left-skewed: most companies founded after 1990, with a long tail of legitimate older companies flagged as outliers back to 1800. No cleaning needed, but binning may help models.

annual_wage_final

Genuine outliers, now clean

Right-skewed with high-wage outliers up to the 99th-percentile cap ($1,687,318). Unlike the original prevailing_wage, this distribution is now on a consistent annual scale — no more units mismatch inflating the outliers.

Recommendation: don't blindly remove outliers — tree-based models are largely robust to them; annual_wage_final's remaining high-end outliers are legitimate, capped salaries, not data errors

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

DATA PREPROCESSING

Feature Engineering

New and transformed features created to prepare the data for modeling

✓ Implemented

annual_wage_final

Standardized prevailing_wage to an annual basis with a conditional fix for mislabeled Hour-unit records (raw wage > $150/hr treated as already annual), then capped remaining extreme values at the 99th percentile ($1,687,318). This is the wage feature actually used in the final model.

implausible_wage_flag

Recomputed against annual_wage_final (>$500K threshold): 335 of 25,480 records flagged (down from 1,717 pre-correction) — the remaining flagged records are legitimately high, capped salaries, not mislabeled data.

Binary encoding

has_job_experience, requires_job_training, and full_time_position converted from Y/N to 1/0.

One-hot encoding

continent, education_of_employee, region_of_employment, and unit_of_wage expanded into individual binary columns (drop_first=True).

○ Recommended, Not Yet Implemented

Company age

Convert yr_of_estab into “years since establishment” rather than a raw year, to reduce sensitivity to very old outliers.

Log transformation

Apply to no_of_employees and annual_wage_final to reduce right-skew — more useful for distance-based/linear models than tree-based ones.

Era binning

Group yr_of_estab into bands (e.g., pre-1950, 1950–1999, 2000+) as an alternative to log-transforming or leaving it continuous.

Explicit interaction term

Bivariate analysis found a real education × experience interaction; this was never engineered as its own feature (the model may pick it up implicitly).

Note: the wage-unit mismatch is now root-cause corrected (annual_wage_final = conditional fix + 99th-percentile cap) — see Annual Wage Conversion slide for the full investigate/fix/cap process

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

DATA PREPROCESSING

Annual Wage Conversion — Found & Fixed

Standardizing prevailing_wage to an annual basis, then correcting a mislabeling issue it revealed

BEFORE

1,717

records flagged (6.74% of dataset)

AFTER CONDITIONAL FIX

0

Hour-unit records still implausible (all 1,382 resolved)

AFTER PERCENTILE CAP

80

records remain above $500K, retained as plausible

Hour (≤ $150/hr)	Hour (> $150/hr)	Week	Month	Year
× 2,080	unchanged*	× 52	× 12	unchanged

**if raw value is already in a plausible annual-salary range — treated as mislabeled rather than multiplied*

1 Investigate

Compared raw prevailing_wage values for flagged vs. non-flagged Hour-unit records. Flagged records clustered in normal annual-salary range ($40K–$150K+), not a plausible hourly rate — strong evidence of unit mislabeling, not random noise.

2 Conditional Conversion

Rebuilt the conversion logic: if an Hour-unit record's raw wage is ≤ $150, multiply by 2,080 as normal. If it's above $150, treat it as already annual (likely mislabeled) instead of multiplying. This resolved all 1,382 flagged Hour-unit records.

3 Percentile Cap

335 Week/Month/Year records remained (untouched by the Hour-specific fix). Capped values above the 99th percentile ($1,687,318) — 255 records capped, leaving 80 high-but-plausible salaries untouched.

Root cause resolved, not just flagged: 80.5% of the original 1,717 implausible records were mislabeled Hour-unit wages, and all of them are now corrected. The remaining 80 records are plausible high salaries, kept as-is; 255 extreme outliers were capped at the 99th percentile ($1,687,318) rather than dropped.

annual_wage_v2 = conditional conversion (hourly_ceiling=$150) applied, then clipped at the 99th percentile in place

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

DATA PREPROCESSING

Data Preparation

Train/validation/test split, class balance, and what it means for model building

TRAINING SET
15,288
rows • 60% of total • 24 columns

VALIDATION SET
5,096
rows • 20% of total • 24 columns

TEST SET
5,096
rows • 20% of total • 24 columns

Class Balance (case_status) — Consistent Across All Sets

Set	Certified	Denied	Ratio
Training (n=15,288)	66.78%	33.22%	~2:1
Validation (n=5,096)	66.80%	33.20%	~2:1
Test (n=5,096)	66.80%	33.20%	~2:1

Near-identical proportions across all three sets strongly indicate a stratified split (train_test_split(..., stratify=y)), which preserves class distribution in every subset.

Consistency Across Sets

Validation and test distributions mirror the training set almost exactly, so there's no distribution shift introduced by the split itself — performance measured on these sets should be a fair, reliable estimate of true generalization.

Implications for Model Building

The ~2:1 imbalance is moderate but not severe — a naive model predicting 'Certified' for everyone would already score ~66.8% accuracy. This means: prioritize F1-score, precision/recall, or ROC-AUC over raw accuracy, and consider class weighting or resampling (oversampling/undersampling) to strengthen minority-class ('Denied') detection.

Categorical features one-hot encoded prior to split; split performed with stratify=y to preserve class ratio

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

SECTION

Model Performance Analysis

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

MODEL BUILDING

Model Performance on Original Data

Six classifiers trained on the unmodified (non-resampled) training set, using annual_wage_final

Primary metric: F1-score, not accuracy. A ~2:1 class imbalance means a naive "always certify" model would already score ~67% accuracy — F1 balances missed certifications against wrongly flagged approvals.

Model	Train F1	Validation F1	Difference	Fit Assessment
Bagging Classifier	0.9886	0.7731	0.2155	Overfitting
Random Forest	1.0000	0.8033	0.1967	Overfitting
Gradient Boosting (GBM)	0.8281	0.8173	0.0107	Best generalization
XGBoost	0.8959	0.8035	0.0924	Moderate overfitting
AdaBoost	0.8194	0.8128	0.0065	Best generalization
Decision Tree	1.0000	0.7532	0.2468	Worst overfitting

GBM & AdaBoost generalize best

AdaBoost shows the smallest gap (0.0065), with GBM close behind (0.0107) — both stay within ~0.01 of their validation F1, indicating strong, reliable performance on unseen data rather than memorization of the training set.

XGBoost sits in the middle

XGBoost's gap (0.0924) is far smaller than the heavy overfitters, but still roughly 9–14x larger than GBM/AdaBoost — some overfitting is present, though validation F1 (0.8035) remains solid.

Tree-based ensembles overfit

Random Forest and Decision Tree reach a near-perfect 1.0000 training F1 but drop sharply on validation (0.803 and 0.753). Bagging shows the same pattern. Decision Tree is the weakest generalizer overall (0.2468 gap).

Metric: weighted F1-score | Evaluation: consistent train/validation split, stratified by case_status

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

MODEL BUILDING

Model Performance on Oversampled Data

Same six classifiers retrained after oversampling the minority (Denied) class, using annual_wage_final

BEFORE — Certified	BEFORE — Denied	AFTER — Certified	AFTER — Denied
10,211	5,077	10,211	10,211

Resulting training set after oversampling: 20,420 rows × 20 columns (classes balanced 1:1).
SMOTE generates synthetic minority-class (Denied) samples via interpolation, rather than duplicating existing rows.

Model	Train F1	Validation F1	Difference	Fit Assessment
Bagging Classifier	0.9861	0.7660	0.2201	Overfitting
Random Forest	1.0000	0.7917	0.2083	Overfitting
Gradient Boosting (GBM)	0.8103	0.8074	0.0030	Best generalization
XGBoost	0.8761	0.7993	0.0768	Moderate overfitting
AdaBoost	0.7839	0.7913	-0.0075	Best generalization
Decision Tree	1.0000	0.7347	0.2653	Worst overfitting

GBM & AdaBoost remain most reliable

GBM shows a very small 0.0030 gap and AdaBoost turns slightly negative (-0.0075) — both continue to generalize at least as well on unseen data as on the oversampled training set.

XGBoost still moderate

XGBoost's gap narrows slightly to 0.0768 (from 0.0924 on original data) — still a step above GBM/AdaBoost, but validation F1 (0.7993) stays solid.

Tree-based ensembles still overfit

Random Forest and Decision Tree still hit a perfect 1.0000 training F1 with validation well below (0.792 and 0.735). Bagging shows the same pattern, similar to the original data results.

Metric: weighted F1-score | Difference = Train F1 − Validation F1 (negative = validation outperforms train).
*Oversampling adds synthetic Denied cases to match Certified — **Certified is unchanged by design.***

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

MODEL BUILDING

Model Performance on Undersampled Data

Same six classifiers retrained after undersampling the majority (Certified) class, using annual_wage_final

BEFORE — Certified	BEFORE — Denied	AFTER — Certified	AFTER — Denied
10,211	5,077	5,077	5,077

Resulting training set after undersampling: 10,154 rows × 22 columns (classes balanced 1:1)

Model	Train F1	Validation F1	Difference	Fit Assessment
Bagging Classifier	0.9789	0.6950	0.2839	Overfitting
Random Forest	1.0000	0.7348	0.2652	Overfitting
Gradient Boosting (GBM)	0.7237	0.7657	-0.0420	Best generalization
XGBoost	0.8727	0.7429	0.1299	Moderate overfitting
AdaBoost	0.7027	0.7628	-0.0601	Best generalization
Decision Tree	1.0000	0.6894	0.3106	Worst overfitting

GBM & AdaBoost still most reliable

Both post their largest negative gaps yet (GBM -0.0420, AdaBoost -0.0601) — validation F1 exceeds train F1, consistent with a smaller, harder undersampled training set giving an easier read on validation.

XGBoost's gap widens

XGBoost's gap grows to 0.1299 — its largest yet across all three sampling strategies (vs. 0.0924 original, 0.0768 oversampled), though validation F1 (0.7429) is still reasonable.

Tree-based ensembles overfit most here

Random Forest and Decision Tree still hit a perfect 1.0000 training F1, with the widest train–validation gaps of any sampling strategy. Decision Tree remains the weakest generalizer overall (0.3106 gap).

Metric: weighted F1-score | Difference = Train F1 − Validation F1 (negative = validation outperforms train)

*Under-sampling removes Certified cases to match Denied — **Denied is unchanged by design.***

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

MODEL IMPROVEMENT

Model Selection Rationale

Which models and which sampling strategy carried forward into tuning

1. **Original data selected over resampling**

   Across all three sampling strategies, models trained on the original (non-resampled) data produced the highest validation F1-scores overall (up to 0.8173), outperforming oversampled (up to 0.8074) and undersampled (up to 0.7657) versions. Original data was carried forward for tuning.

2. **Gradient Boosting (GBM) — best generalizer**

   Highest validation F1 (0.8173) with the smallest train–validation gap (0.0107) of any model on original data — the most reliable generalization observed.

3. **AdaBoost — most consistent performer**

   Near-identical F1 (0.8128) with an even tighter train–validation gap (0.0065) — the most consistent performer between train and validation of any model tested.

4. **XGBoost — best tuning potential**

   Slightly lower F1 (0.8035) but flexible regularization controls (gamma, subsample, scale_pos_weight) made it the strongest candidate for improvement through tuning, despite moderate pre-tuning overfitting (0.0924 gap).

5. **Remaining models tuned for comparison, not as final candidates**

   Random Forest, Bagging, and Decision Tree were also tuned for completeness, but deprioritized as final candidates due to severe overfitting (train F1 up to 1.0000 vs. validation as low as 0.7532).

Selection based on validation F1-score and train–validation generalization gap across all six models on original (non-resampled) data

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

MODEL IMPROVEMENT

Hyperparameter Tuning — Best Performing Models

Validation-set performance after GridSearchCV tuning on annual_wage_final, ranked by F1-score

Rank	Model	Accuracy	Recall	Precision	F1-score
1	XGBoost	0.7418	0.8690	0.7727	**0.8180**
2	AdaBoost	0.7370	0.8702	0.7674	**0.8155**
3	Gradient Boosting (GBM)	0.7321	0.8860	0.7553	**0.8155**
4	Decision Tree	0.7382	0.8561	0.7754	0.8137
5	Bagging Classifier	0.7251	0.8760	0.7528	0.8098
6	Random Forest	0.7306	0.8205	0.7857	0.8027

1 XGBoost — Highest F1

Leads all six models with F1 0.8180 (Recall 0.8690, Precision 0.7727) — a well-balanced pair rather than an extreme trade-off. Train–validation gap of 0.0129 is small, confirming reliable generalization after retraining on the corrected annual_wage_final feature.

2 AdaBoost — Tightest generalization

F1 of 0.8155, essentially tied with GBM, but with the smallest train–val gap of any model (0.0067) — train and validation performance are nearly identical, the most dependable generalizer in this round.

3 GBM — Virtually tied with AdaBoost

F1 of 0.8155 (fractionally behind AdaBoost at full precision) with the highest recall of the top 3 (0.8860) — best at catching true 'Certified' cases among the leading models, with a still-small 0.0082 gap.

Metric basis: validation-set performance after GridSearchCV tuning across all six models

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

MODEL IMPROVEMENT

Tuning Method & Parameter Grids

GridSearchCV with 5-fold cross-validation, optimized for F1-score — retrained on annual_wage_final

Model	Best Parameters Found (GridSearchCV, cv=5, scoring=f1)
Random Forest	Best found: max_depth=10 • max_features='sqrt' • min_samples_split=5 • n_estimators=10
AdaBoost	Best found: base estimator Decision Tree (max_depth=3) • learning_rate=0.06 • n_estimators=100
Gradient Boosting (GBM)	Best found: learning_rate=0.05 • max_features=0.8 • n_estimators=75 • subsample=0.9
XGBoost	Best found: gamma=3 • learning_rate=0.05 • n_estimators=50 • scale_pos_weight=1 • subsample=0.9
Decision Tree	Best found: criterion='gini' • max_depth=8 • max_leaf_nodes=20 • min_samples_leaf=1
Bagging Classifier	Best found: max_features=0.7 • max_samples=0.7 • n_estimators=100

Performance Commentary

After retraining on the corrected annual_wage_final feature, the six-model ranking shifted: **XGBoost now leads on F1 (0.8180)** with a well-balanced recall/precision split (0.869/0.773) rather than an extreme trade-off. **AdaBoost** posts the tightest train–validation gap of any model (0.0067), with GBM essentially tied on F1 (0.8155) just behind it. **Random Forest now ranks last on F1 (0.8027)**, a shift from earlier rounds. **Bagging** remains the clear outlier: train F1 of 0.9944 vs. validation F1 of 0.8098 — a 0.1846 gap, confirming tuning still hasn't meaningfully resolved its overfitting. Decision Tree, by contrast, continues to generalize well (0.0071 gap), even though its raw F1 (0.8137) sits just outside the top 3.

All six models' best parameters confirmed via GridSearchCV.best_params_; performance metrics reflect the retrain on annual_wage_final (scoring = F1-score, cv = 5)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Great Learning

Happy Learning !

