SHIVAM KUMAR

HR ANALYTICS MACHINE LEARNING PROJECT

PRESENTATION

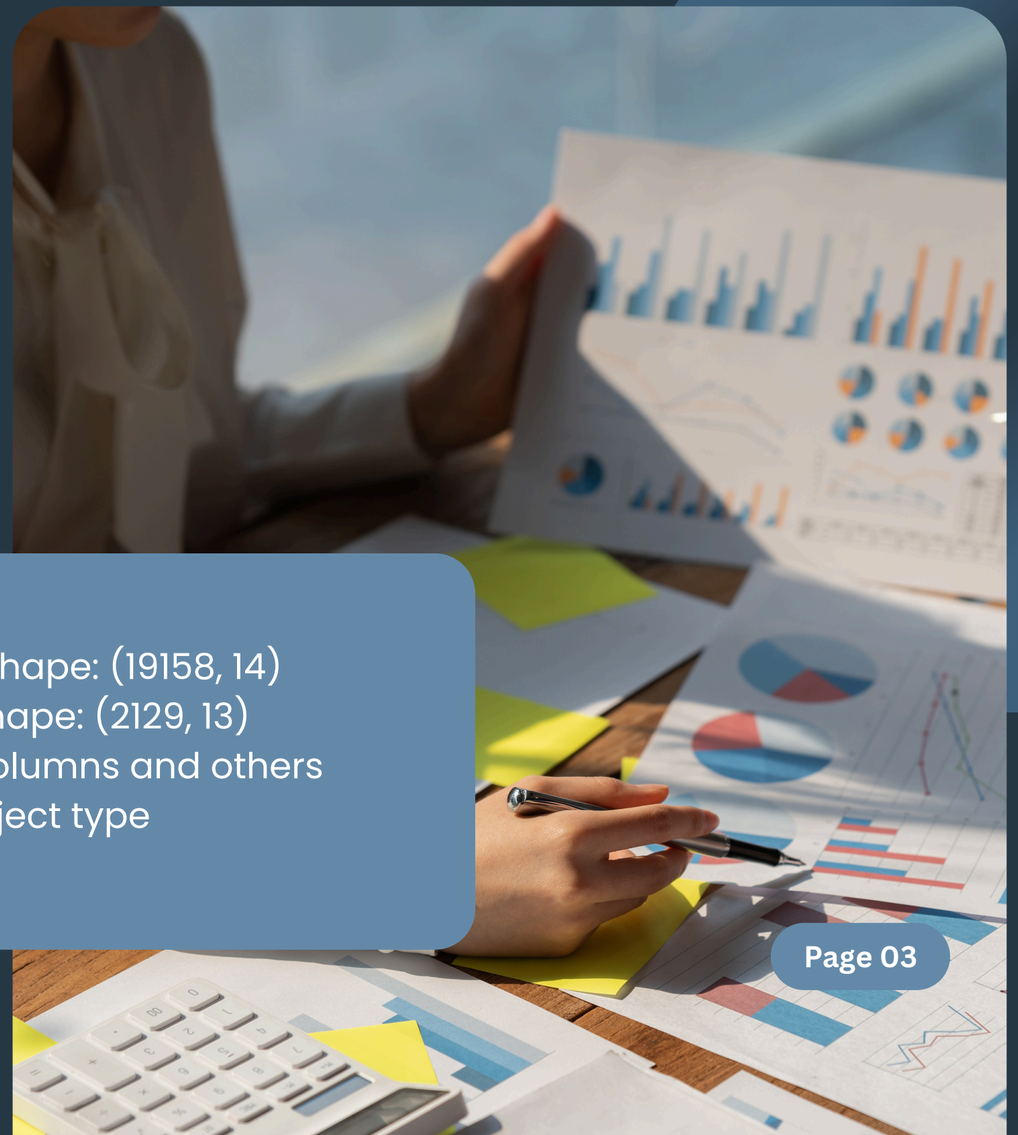
PROBLEM STATEMENT

A data science company wants to optimize hiring by identifying candidates genuinely interested in working for them versus those seeking other opportunities.

Predict whether data science training candidates will stay with the company or seek new employment

DATASET DETAILS

Contains demographics, education, and experience information from candidate signups. Dataset is split into train/test sets with imbalanced target variable.



01.

Key Characteristics

- Imbalanced dataset
- Mostly categorical features (nominal, ordinal, binary)
- High cardinality features
- Missing values present

02.

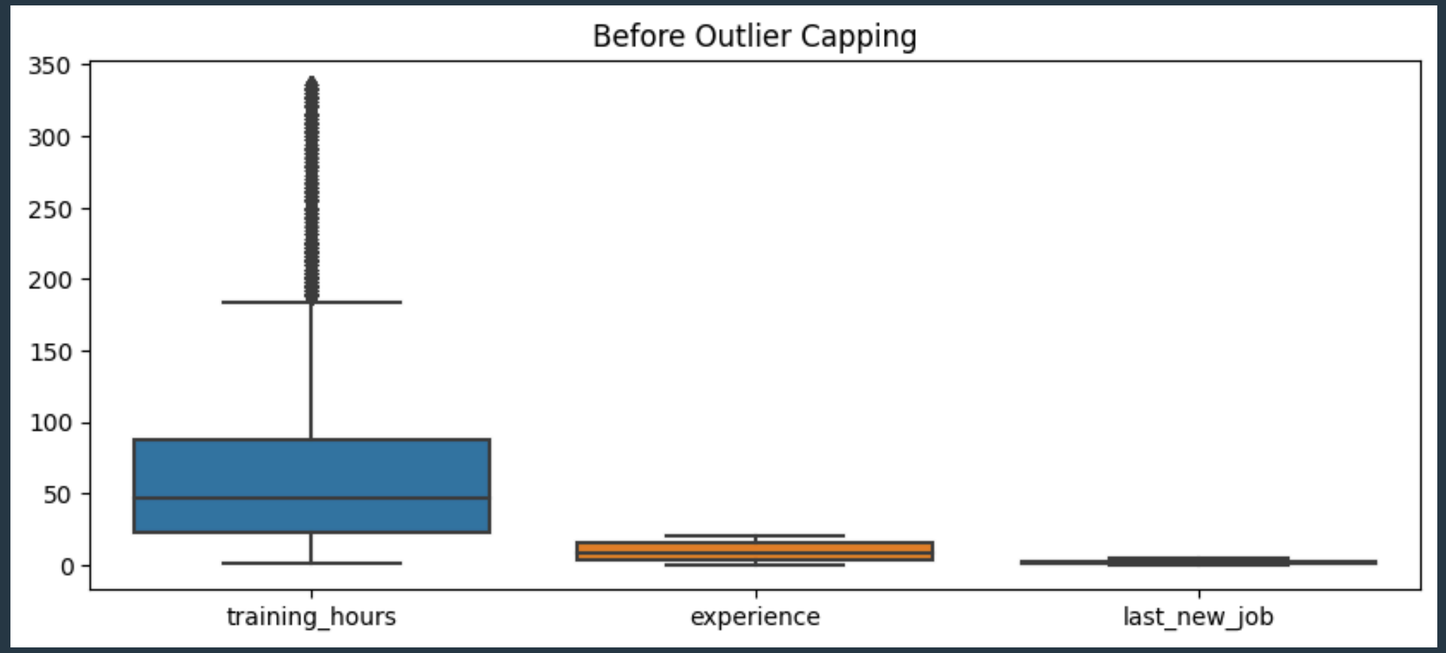
Details

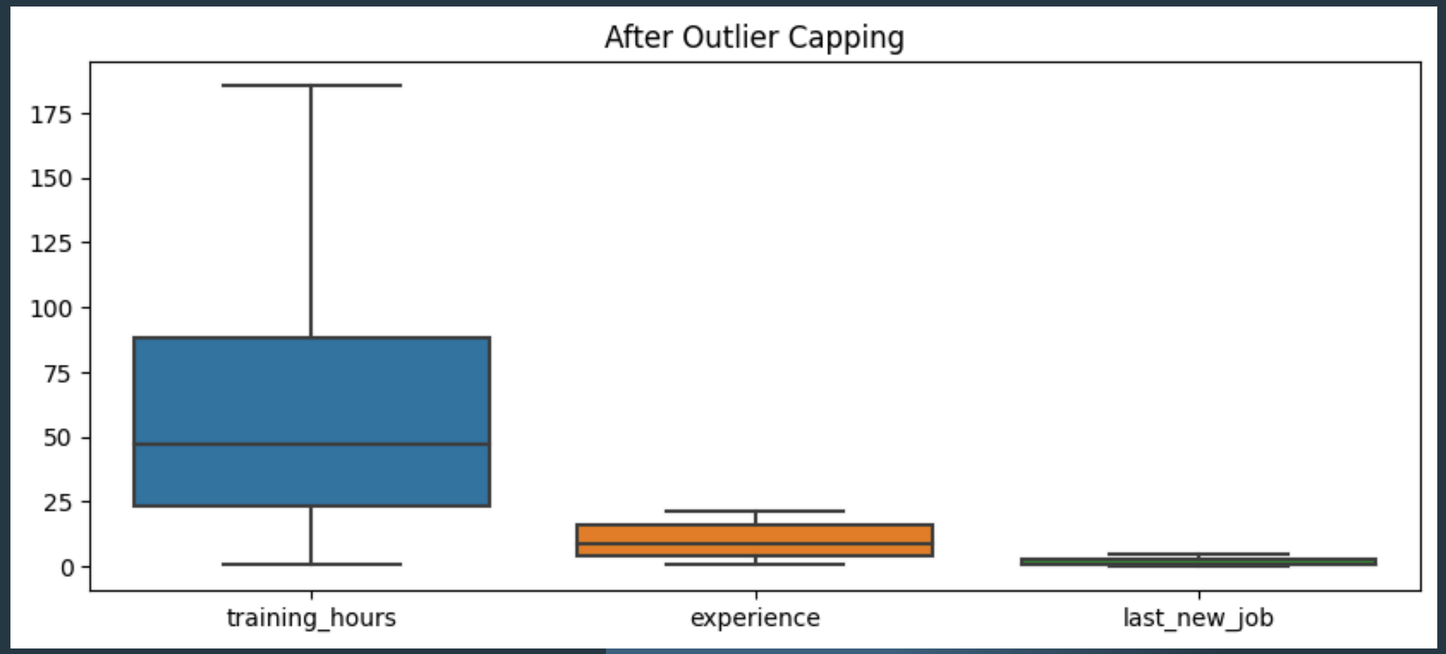
- Train Shape: (19158, 14)
- Test Shape: (2129, 13)
- 4 int columns and others are object type

EXPOLATORY DATA ANALYSIS

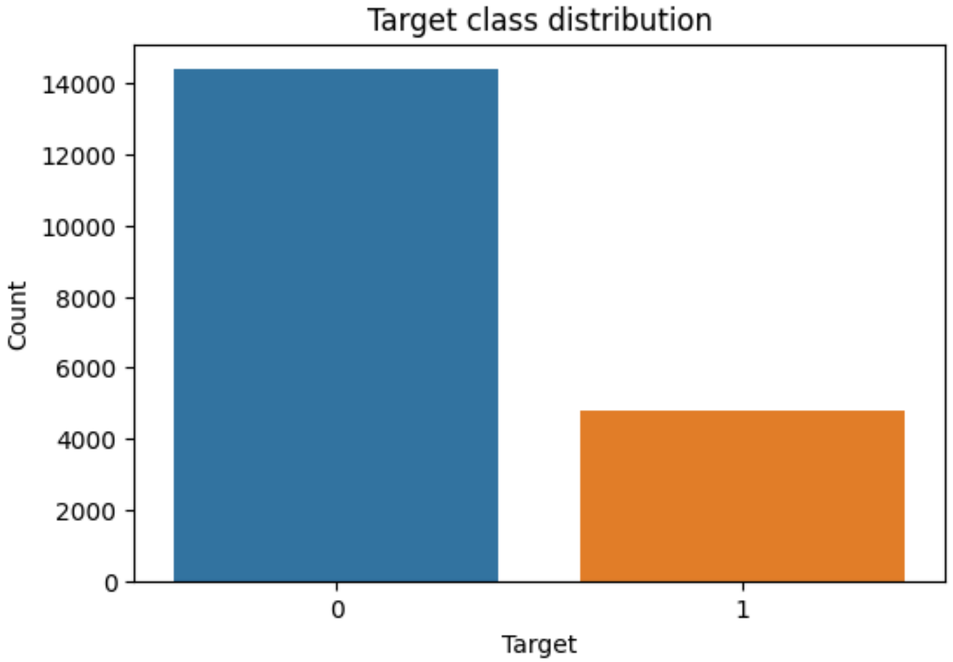
```
No of Columns having null values: 8
Columns:
 gender                  23.530640
enrolled_university       2.014824
education_level           2.401086
major_discipline         14.683161
experience                0.339284
company_size             30.994885
company_type             32.049274
last_new_job              2.207955
dtype: float64
```

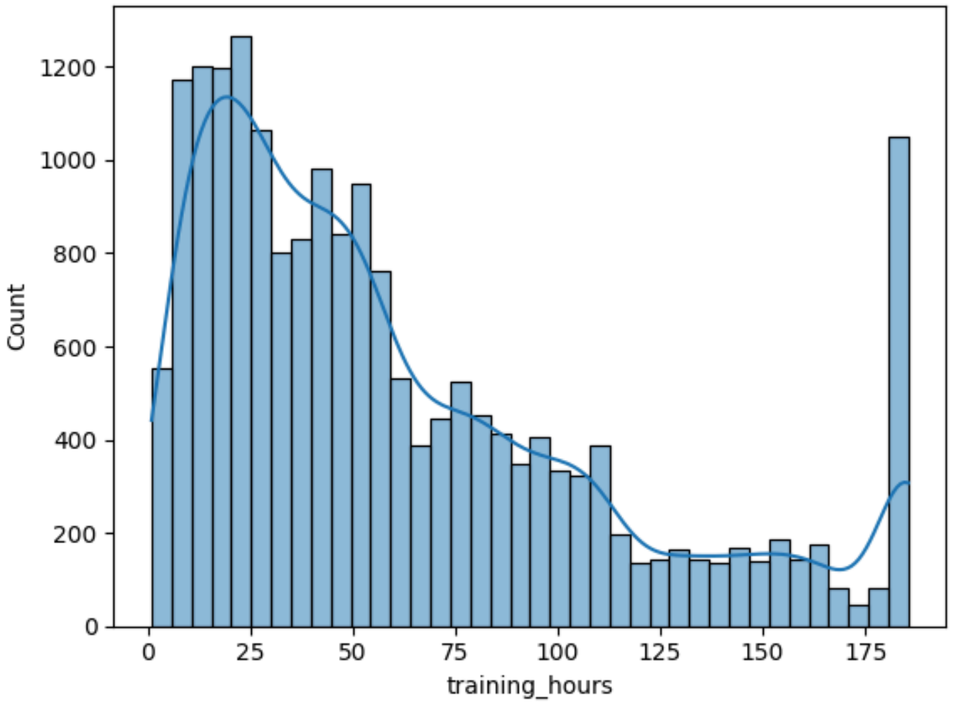
- 8 Columns Contain Null Values
- Company Type have 32% of values as NAN
- NAN valus are replaced with most frequent value
- 0 Duplicates Value





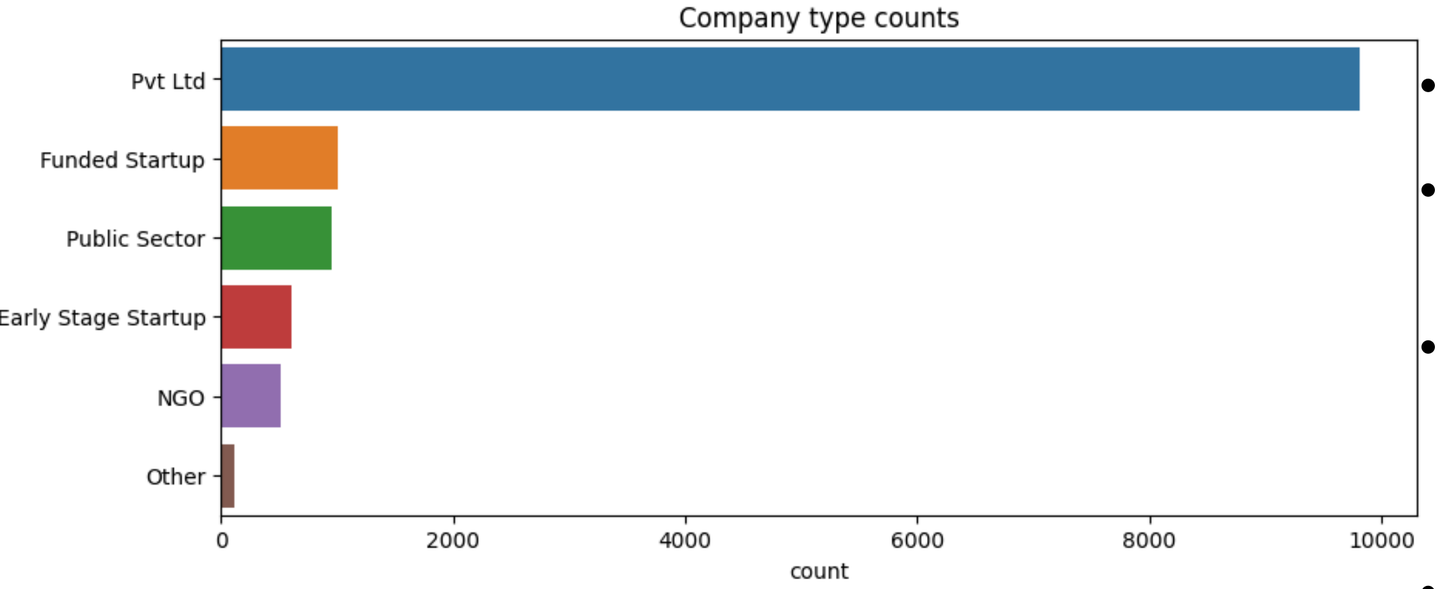
- Training hours had extreme outliers above 300, distorting the scale and affecting distribution.
- After capping, training hours became more balanced and realistic, reducing noise for the model.
- Experience also showed a few high outliers, which were corrected to create a more stable numeric range.
- Last new job had minimal outliers, but capping ensured consistent preprocessing across all numeric features.
- Overall, outlier capping reduced skewness and improved model stability, leading to cleaner inputs and better learning performance.



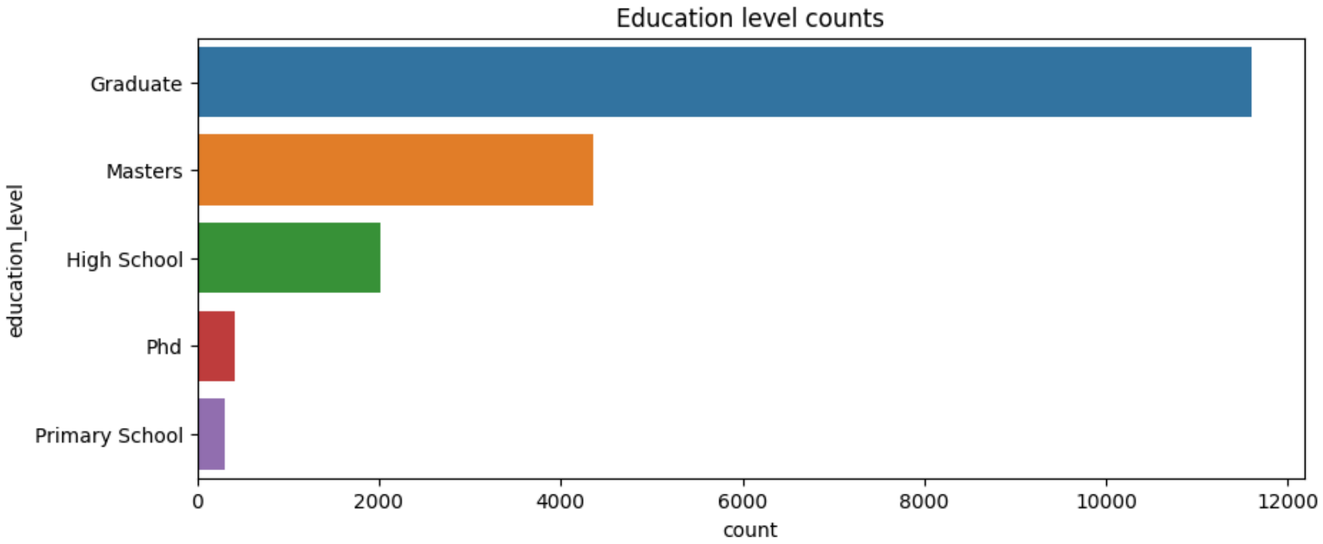


- The dataset is imbalanced, with class 0 significantly higher than class 1.
- A majority of candidates are not looking to change jobs, while a smaller portion belongs to class 1 (job change).

- Training hours are right-skewed, with most candidates completing fewer than 50 hours.
- A long tail exists, showing some candidates with very high training hours, indicating variation in upskilling behavior.
- A noticeable spike near 180 hours suggests a cut-off or maximum training limit for many participants.

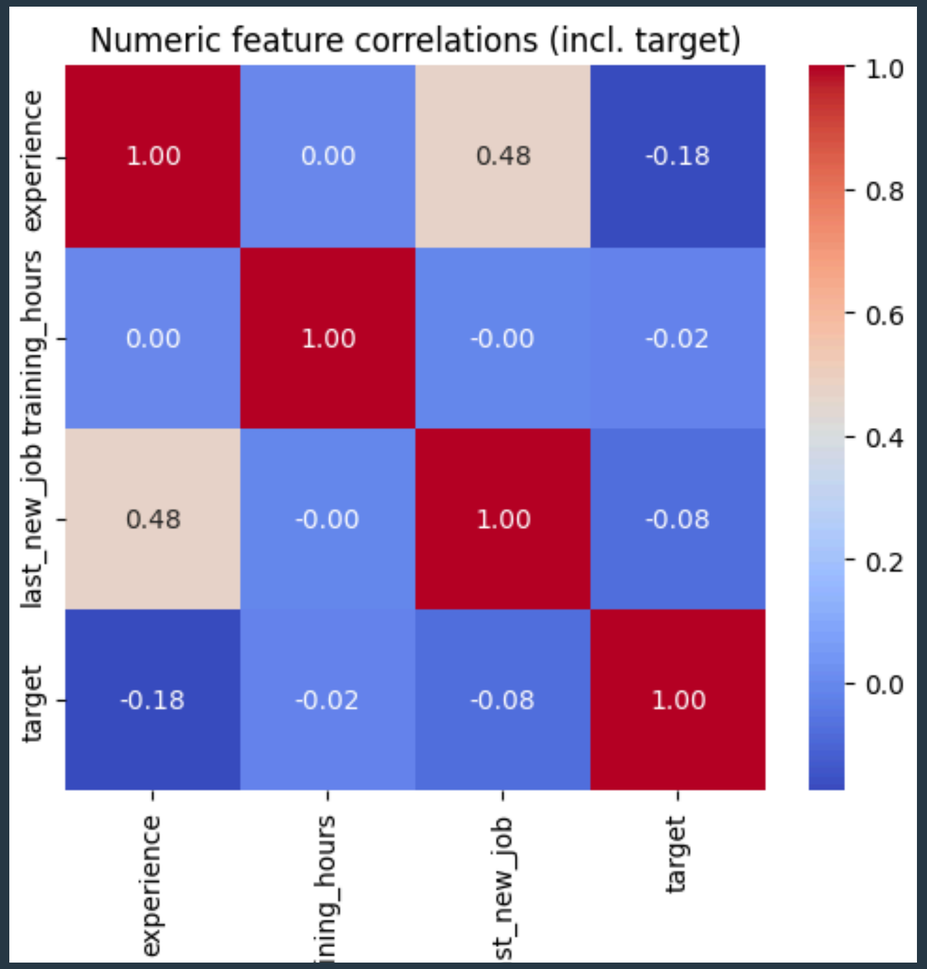


- Pvt Ltd companies dominate the dataset, indicating most candidates work in private-sector organizations.
- Funded startups and public-sector employees form moderate proportions, showing diverse organizational backgrounds.
- Early-stage startups, NGOs, and other small categories appear far less frequently, suggesting limited representation.

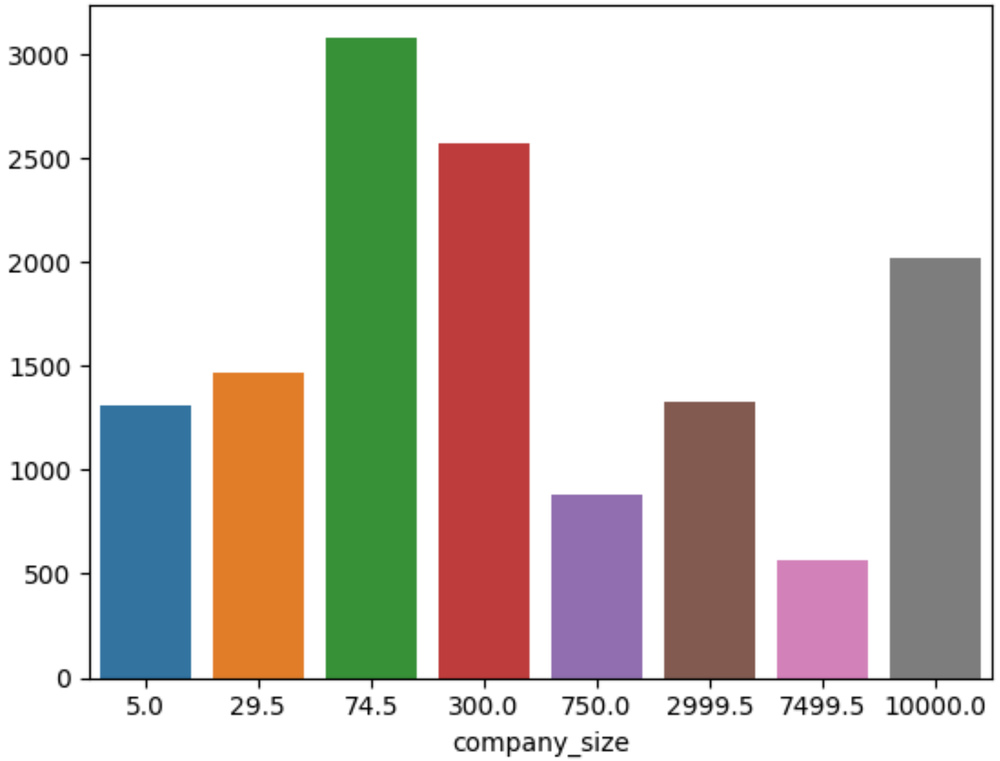


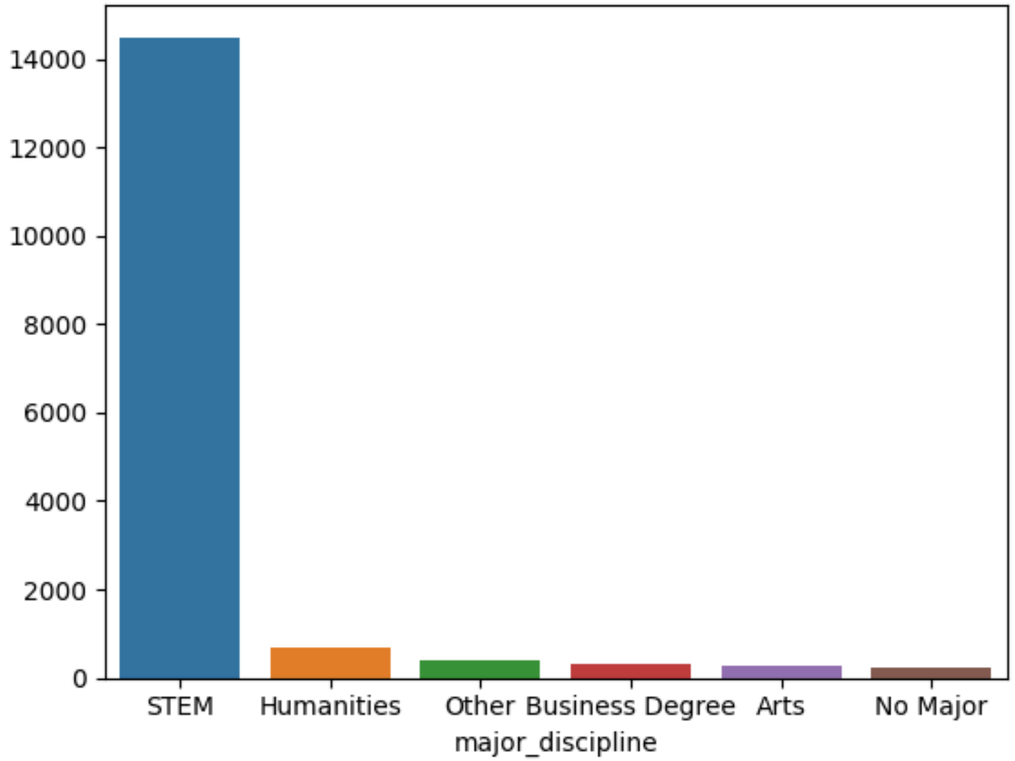
- A majority of individuals are Graduates, forming the largest educational group in the dataset.
- Master’s degree holders are the next largest group, showing many candidates pursue higher education.
- High School, PhD, and Primary School categories have much smaller counts, indicating fewer candidates with these qualifications.
- The distribution shows a workforce primarily composed of degree holders, which may impact job change trends and skill requirements.

CORREALTION MATRIX



- Experience and last_new_job show a moderate positive correlation (0.48), suggesting individuals with more experience tend to have longer gaps since their last job change.
- Training_hours shows near-zero correlation with experience and last_new_job, indicating training time is independent of work experience or recent job changes.
- All numeric features have weak correlations with the target (ranging from -0.18 to -0.02), meaning no single numeric variable strongly predicts job change alone.
- Experience has the strongest negative correlation with the target (-0.18), suggesting slightly lower job-change likelihood among more experienced individuals.
- Overall, numeric variables contribute information but are not dominant predictors, implying categorical features likely play a larger role in model performance.





- The dataset shows a balanced spread of company sizes, with employees from small, medium, and large organizations.
- The largest representation is from companies with 50–99 employees (≈3100 samples), indicating many candidates work in moderately sized firms.
- STEM dominates overwhelmingly, with more than 14,000 candidates, making it the primary educational background in the dataset.
- Humanities, Business, Arts, and Other degrees show much smaller representation, each under 1,000 samples.
- The large STEM dominance indicates the dataset is skewed toward technical fields, likely reflecting the nature of jobs in this domain.

DATA PREPROCESSING

- Handled missing values in categorical features using mode/constant imputation and numerical values using median imputation to ensure completeness of the dataset.
- Performed outlier detection and capping using the IQR method, reducing extreme values in training_hours, experience, and last_new_job to make numeric features more stable.
- Standardized numerical features (e.g., training_hours, experience_years) to ensure uniform scaling for algorithms sensitive to magnitude differences.
- Encoded categorical variables using One-Hot Encoding to transform non-numeric data into model-compatible numerical formats.
- Converted experience into numeric form, including parsing special values like “<1”, “10+”, and transforming them into meaningful numeric ranges.

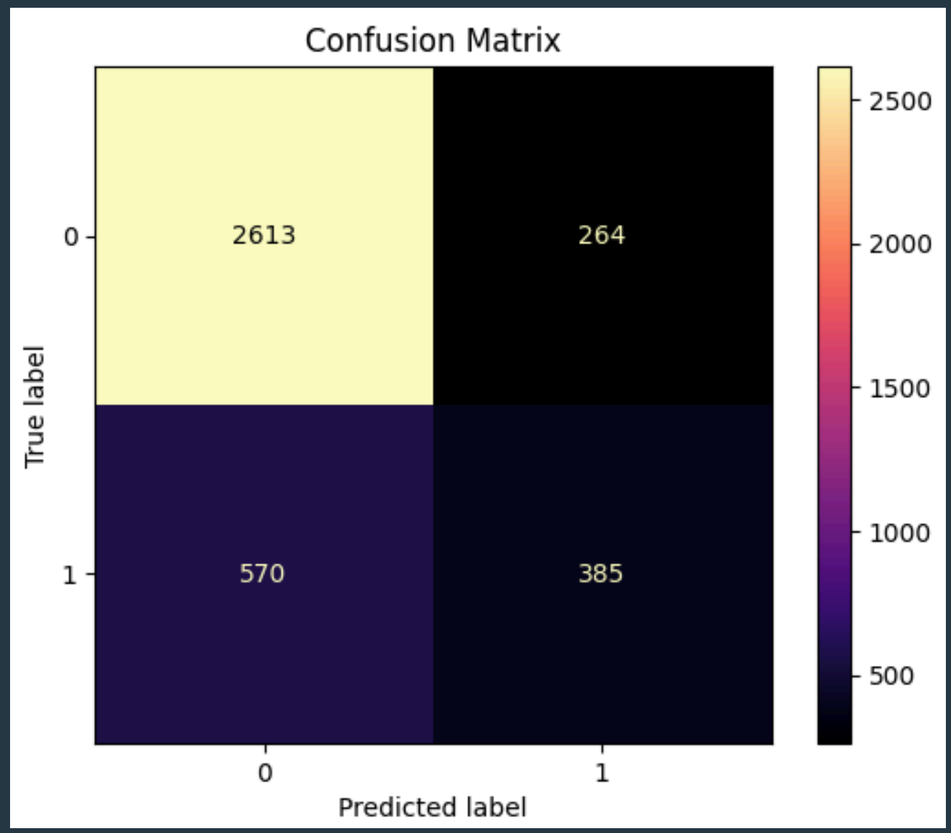
MODEL TRAINING

Linear Regression Results

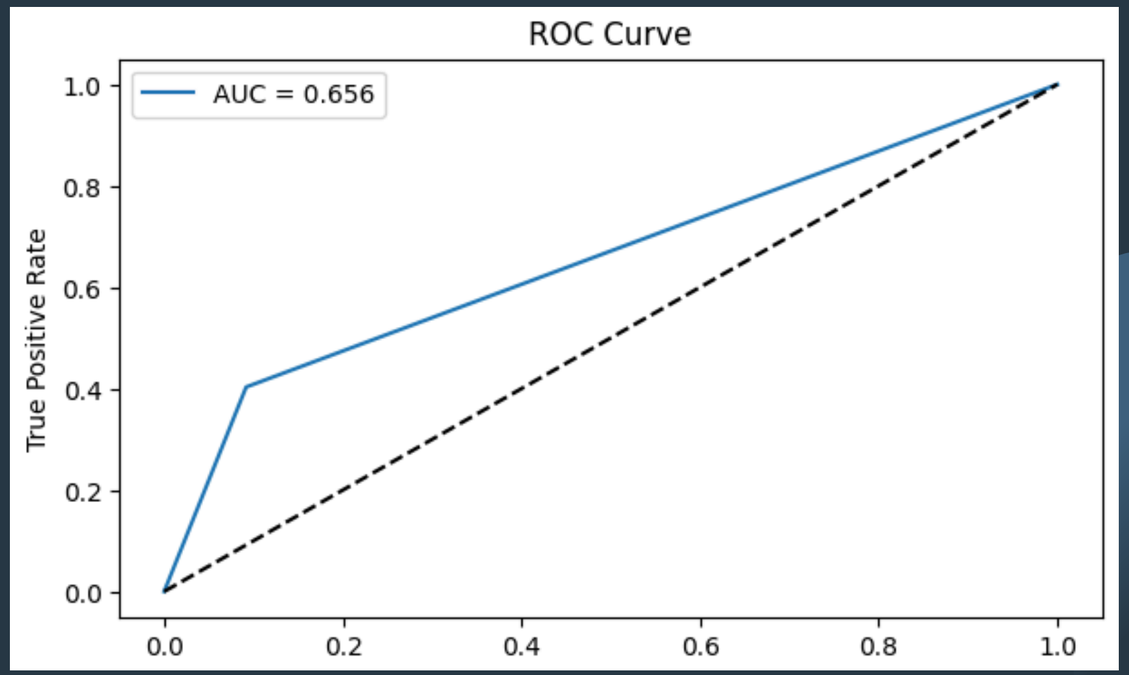
- The model achieves 78% accuracy, correctly predicting most samples overall.
- Class 0 (negative) is classified accurately (high precision and recall).
- Class 1 (positive) is often missed (lower precision and recall, high false negative rate).
- Most negatives are well detected, while a significant portion of actual positives are incorrectly labeled as negatives.
- The model is biased toward class 0

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.91	0.86	2877
1	0.59	0.40	0.48	955
accuracy			0.78	3832
macro avg	0.71	0.66	0.67	3832
weighted avg	0.76	0.78	0.77	3832



- AUC Value (0.656): The classifier has moderate ability to distinguish between classes, performing better than random guessing but far from perfect.
- Model Reliability: At most thresholds, it identifies positives better than chance, but is not highly accurate—many cases are still misclassified.



SVC And Random Forest Results

SVC

Accuracy: 78.2%

- The model correctly predicts class labels for ~78% of cases overall.

ROC AUC: 0.661

- Moderate ability to distinguish between positive and negative classes, better than random but not strong.

Confusion Matrix Insights:

- Class 0 (Negative): Well-predicted (2597 correct, 280 false positives).
- Class 1 (Positive): Poorly detected (401 correct, 554 false negatives).

Random Forest

Accuracy: 73.6%

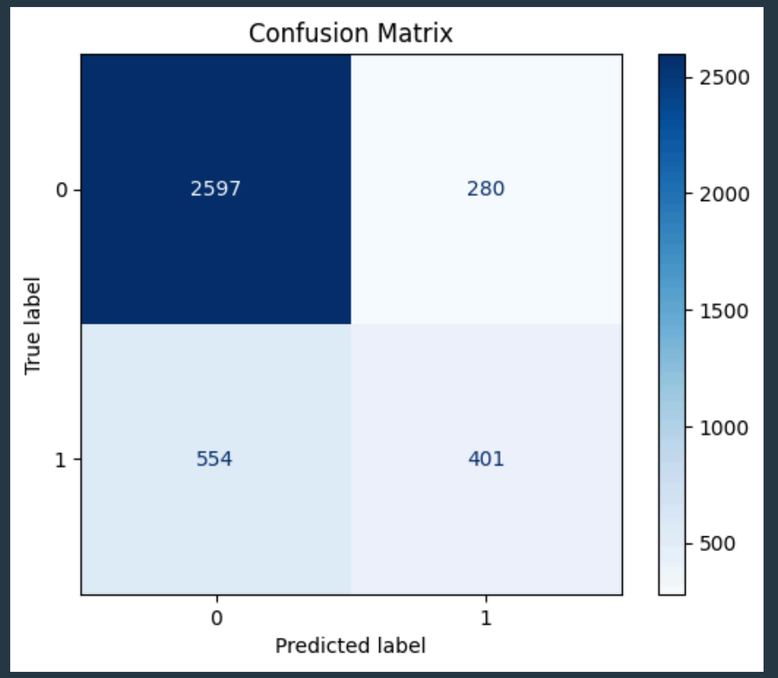
- The model correctly predicts ~74% of all cases.

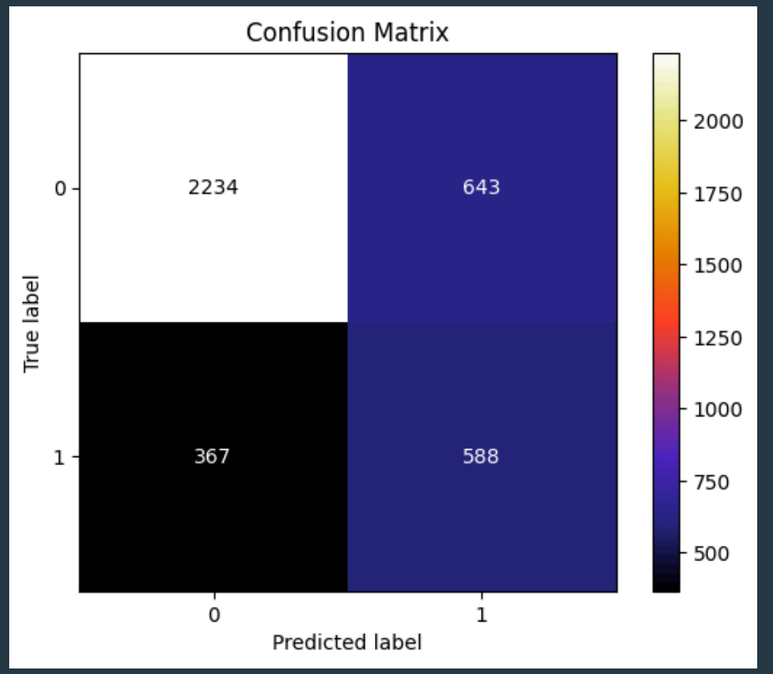
ROC AUC: 0.696

- Moderate discrimination ability between classes, slightly better than the previous model (0.661 vs 0.696).

Confusion Matrix Insights:

- Class 0 (Negative): 2234 correctly predicted, 643 false positives.
- Class 1 (Positive): 588 correctly predicted, 367 false negatives.





KEY TAKEWAYS

- Logistic Regression achieved best accuracy (≈78%).
- Random Forest & KNN performed slightly lower (≈73%).
- Data is partially linearly separable → LR works best.
- Outlier capping improved distribution stability.
- Training hours, city development index, experience were strong predictors.

THANK YOU

FOR YOUR NICE ATTENTION