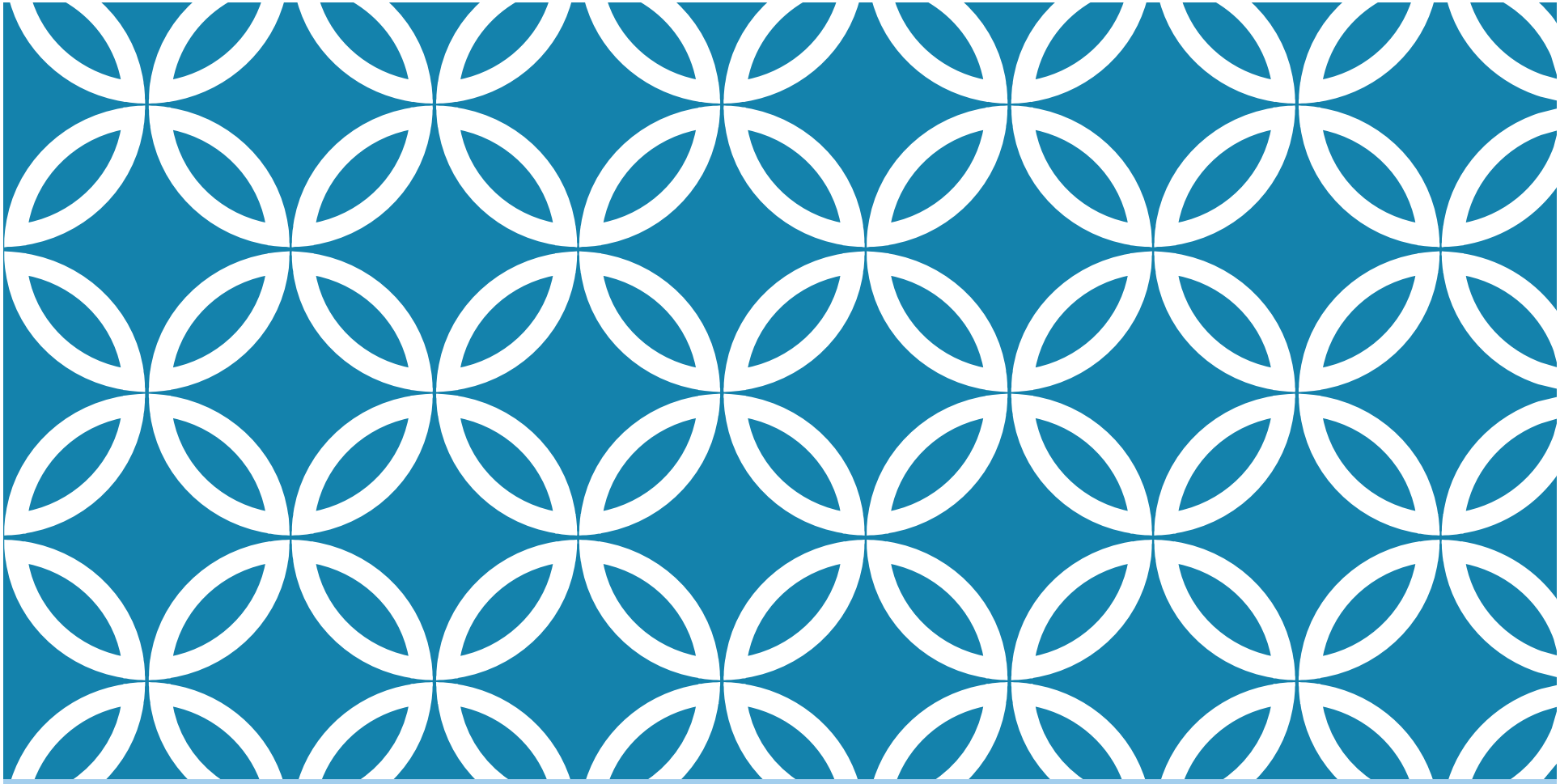


INCOME PREDICTION PROJECT

By Rolddy Surpris

INTRODUCTION

Objective: Predict whether a person earns more than 50K per year.

Dataset: UCI Adult Census dataset.

Source: Kaggle

DATASET OVERVIEW

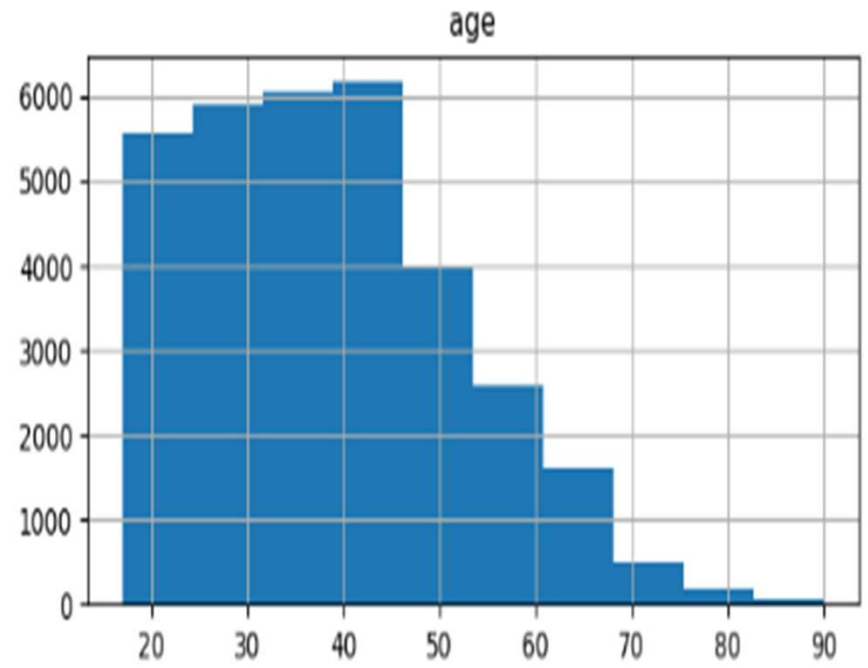
- Rows: Roughly 32,000
- Features: Age, education, occupation, workclass, hours per week, etc.
- Target: Income (<=50K, >50K).

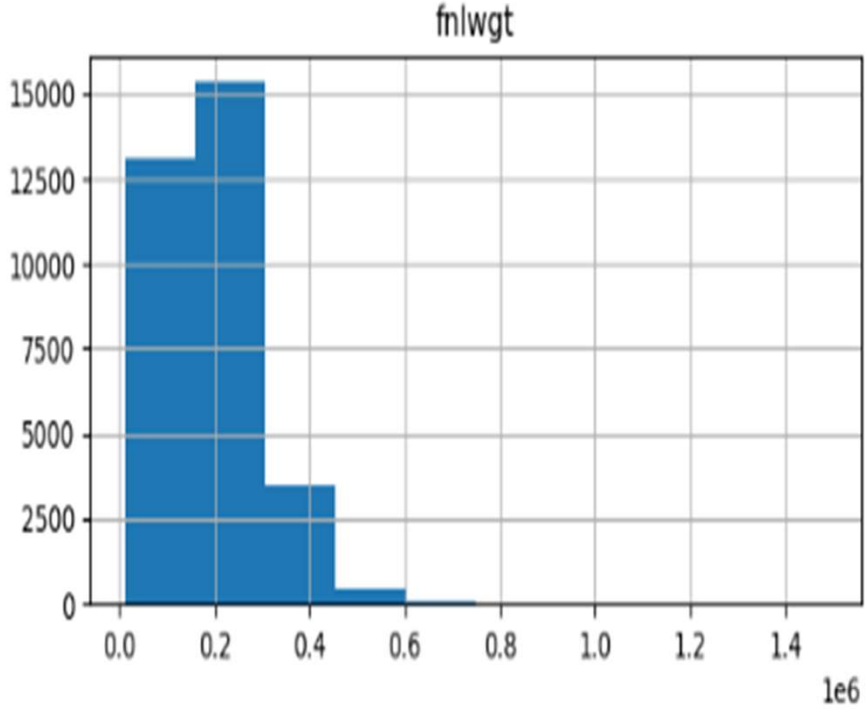
DATA PREPROCESSING

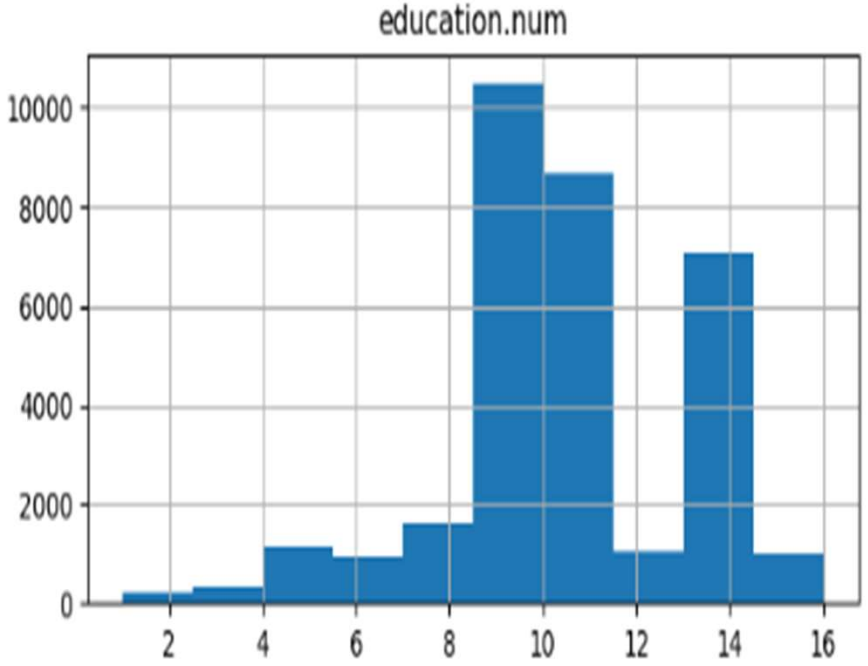
- Handle missing values
- Encode categorical variables (One-Hot, Label Encoding).
- Feature scaling for numerical values.
- Train-test split.

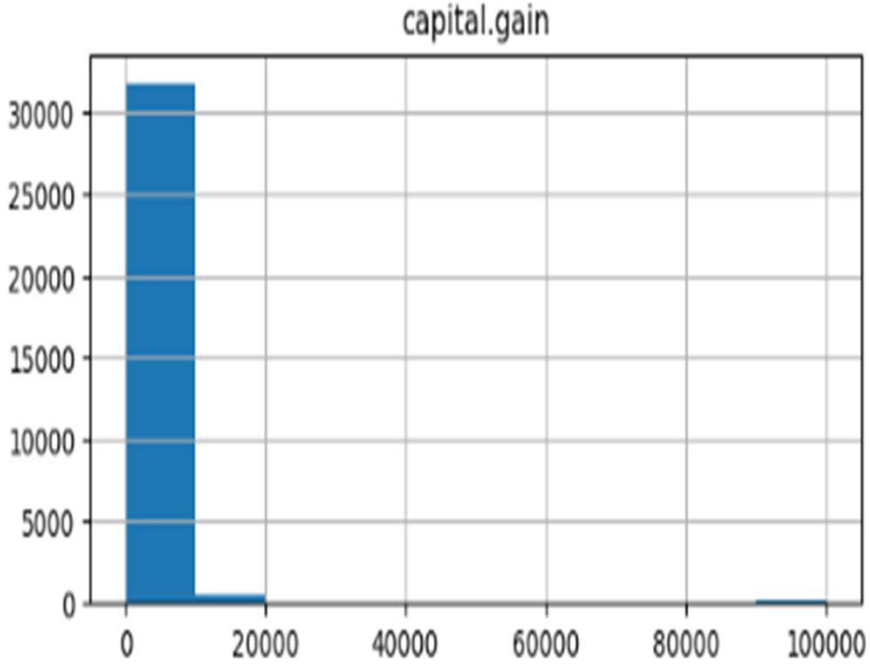
EXPLORATORY DATA ANALYSIS

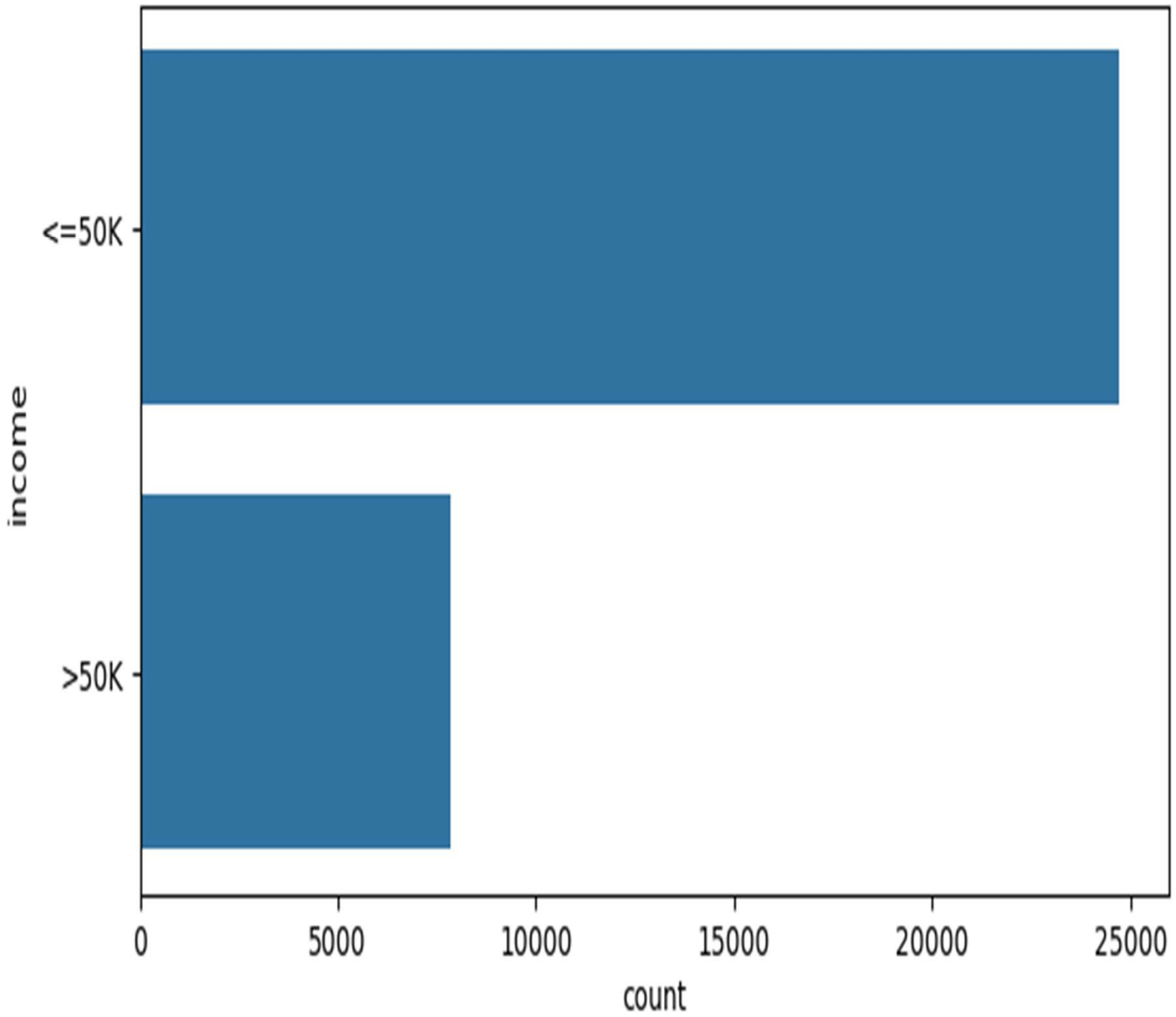
- Distribution of income levels.
- Correlation between age, education, hours worked, and income.
- Class imbalance checked.


age
6000
5000
4000
3000
2000
1000
0
20
30
40
50
60
70
80
90


fnlwgt
15000
12500
10000
7500
5000
2500
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e6


education.num
10000
8000
6000
4000
2000
0
2
4
6
8
10
12
14
16


capital.gain
30000
25000
20000
15000
10000
5000
0
0
20000
40000
60000
80000
100000


<=50K
>50K
income
0
5000
10000
15000
20000
25000
count

MODELING APPROACH

- Algorithms tested:
  - Logistic Regression
  - Decision Tree
  - Random Forest

EVALUATION METRICS

- Metrics used:

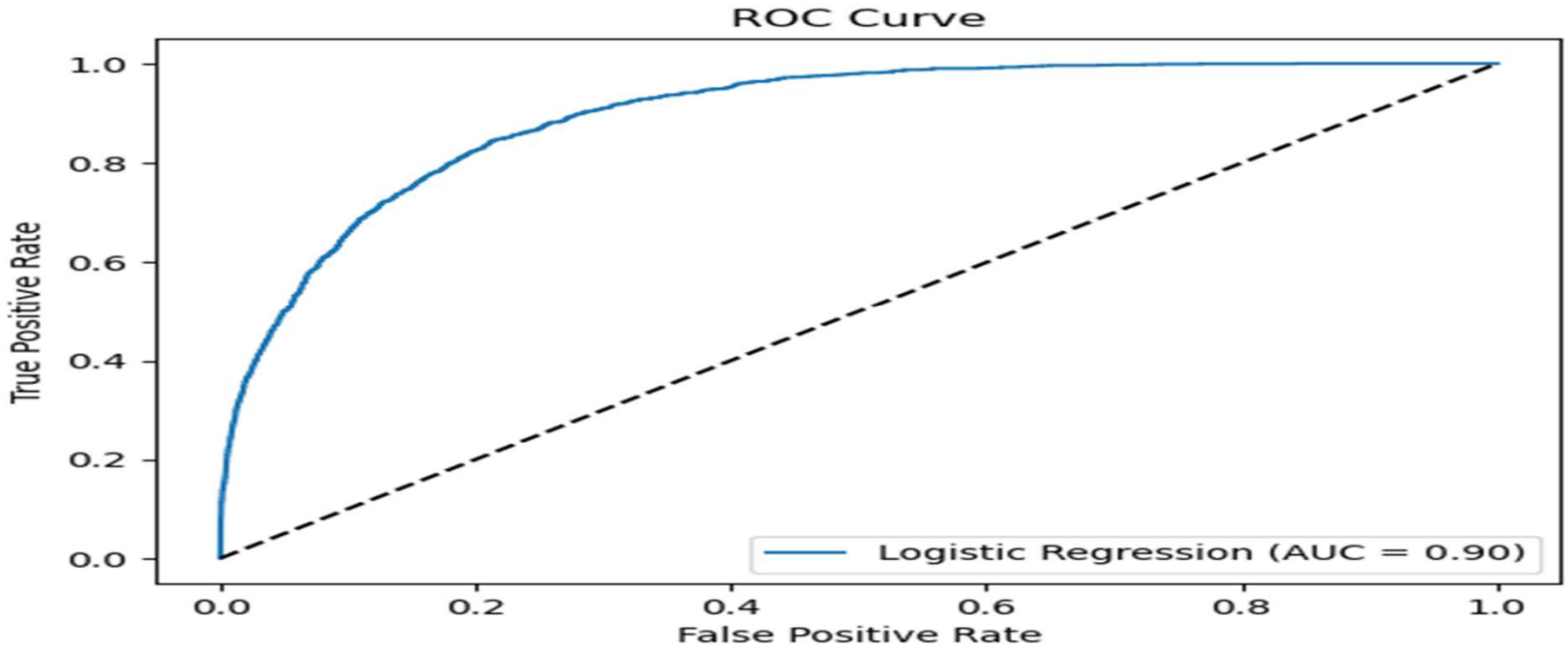
- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

RESULTS

- Best performing model: Random Forest
- Achieved accuracy :84%.
- Precision and recall balanced.
- Confusion matrix indicates good separation.

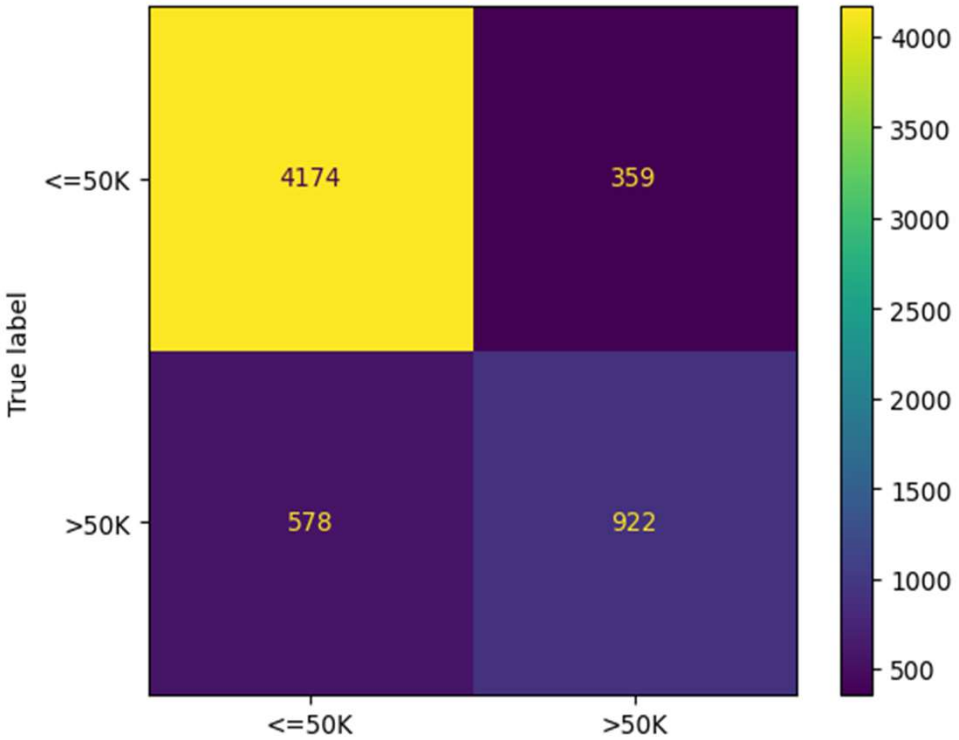
Logistic Regression Results:

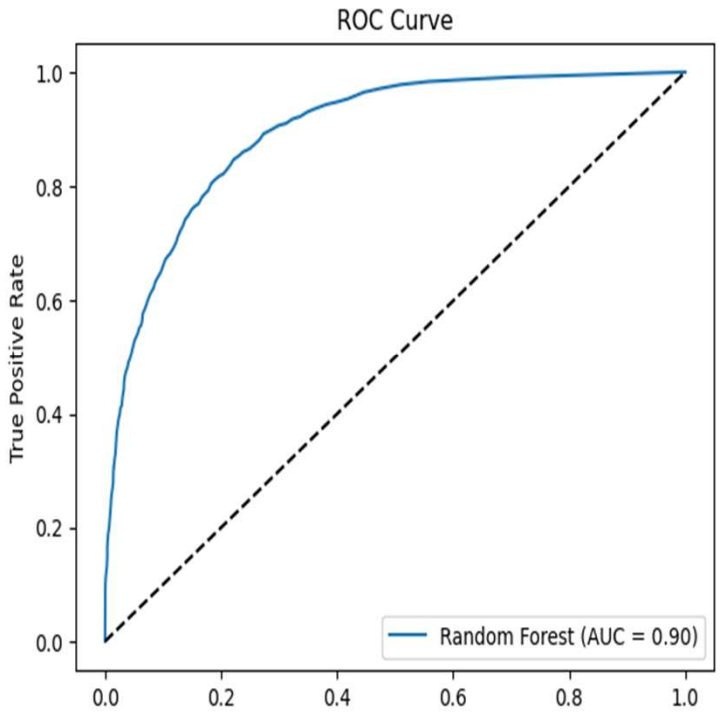
	precision	recall	f1-score	support
<=50K	0.87	0.92	0.90	4533
>50K	0.72	0.60	0.65	1500
accuracy			0.84	6033
macro avg	0.80	0.76	0.78	6033
weighted avg	0.84	0.84	0.84	6033



Random Forest Results:

	precision	recall	f1-score	support
<=50K	0.88	0.92	0.90	4533
>50K	0.72	0.61	0.66	1500
accuracy			0.84	6033
macro avg	0.80	0.77	0.78	6033
weighted avg	0.84	0.84	0.84	6033





CONCLUSION

- Key insights:

  • Education, hours per week, and occupation are strong predictors.

  • Model performs well on unseen data.

- Applications: HR screening, policy making, financial risk analysis.

NEXT STEPS

- Improve feature engineering.
- Explore deep learning models.
- Monitor performance in real-world use.