Predicting High-income Individuals

Truong Vo Minh Hieu

TABLE OF CONTENTS

Data Information

Dataset Overview

- This dataset, sourced from the UCI Machine Learning Repository, is based on the 1994 U.S. Census.
- Its goal is to classify adults into two income groups:
- <=50K USD per year
- >50K USD per year
- By analyzing factors such as education, gender, race, and occupation, the study aims to understand how these variables influence income levels in the 1990s.
  These insights also help reveal social and economic patterns that can be compared with today's labor market conditions.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

Data Information

Features	Description
Age	Age of the individual in years.
Workclass	Employment type or class of the worker.
fnlwgt	a sampling weight indicating the number of people the record represents in the population.
education	Highest level of education attained.
education-num	Numerical representation of education level
marital-status	Marital status of the individual.
occupation	Type of job or occupation.

Features	Description
relationship	Relationship of the individual to others in the household.
race	Race of the individual.
sex	Gender of the individual.
capital-gain	Capital gains from investments
capital-loss	Capital losses from investments.
hours-per-week	Average number of working hours per week.
native-country	Country of origin of the individual.
salary	Income class of the individual.

02

Initial Data Exploration

Initial Data Exploration

1. Data Inspection

- Checked data types for each column.
- Renamed columns.
- Handled invalid values.
- Treated missing and duplicated records.
- Removed unnecessary columns.
- Examined statistical summary (mean, median, std, min, max) to understand data distribution and detect anomalies.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   age             32561 non-null  int64 
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64 
 3   education       32561 non-null  object
 4   education.num   32561 non-null  int64 
 5   marital.status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital.gain    32561 non-null  int64 
 11  capital.loss    32561 non-null  int64 
 12  hours.per.week  32561 non-null  int64 
 13  native.country  32561 non-null  object
 14  income          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

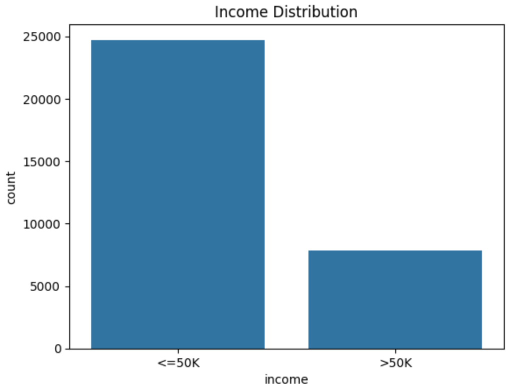
2. Findings:

- Dataset contains 32,561 records and 15 columns.
- Numerical features such as age, education_num, and hours_per_week show reasonable ranges without extreme outliers.
- Mean age is 38.6, with a typical working time of about 40 hours per week.
- Most columns are object (categorical) type, requiring encoding later in preprocessing.

	age	education_num	capital_gain	capital_loss	hours_per_week
count	32537.000000	32537.000000	32537.000000	32537.000000	32537.000000
mean	38.585549	10.081815	1078.443741	87.368227	40.440329
std	13.637984	2.571633	7387.957424	403.101833	12.346889
min	17.000000	1.000000	0.000000	0.000000	1.000000
25%	28.000000	9.000000	0.000000	0.000000	40.000000
50%	37.000000	10.000000	0.000000	0.000000	40.000000
75%	48.000000	12.000000	0.000000	0.000000	45.000000
max	90.000000	16.000000	99999.000000	4356.000000	99.000000

Initial Data Exploration

Target features



Income Distribution:

- Individuals earning <=50K account for approximately **75%** of the total samples.
- Individuals earning >50K make up only about **25%**.

-> Imbalanced classes

Numeric Features

Age Distribution:

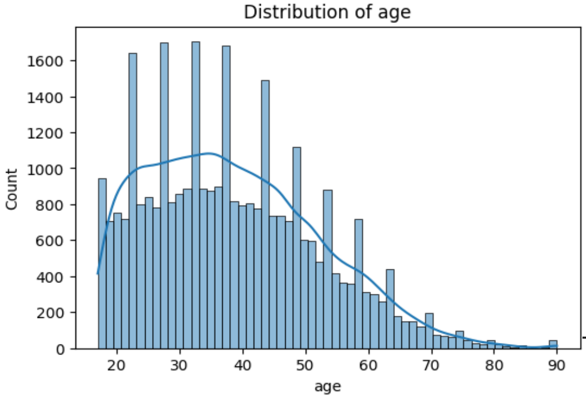
- Right-skewed
- Most adults are in working age.
- Higher age -> more experience -> higher income.

Education Level

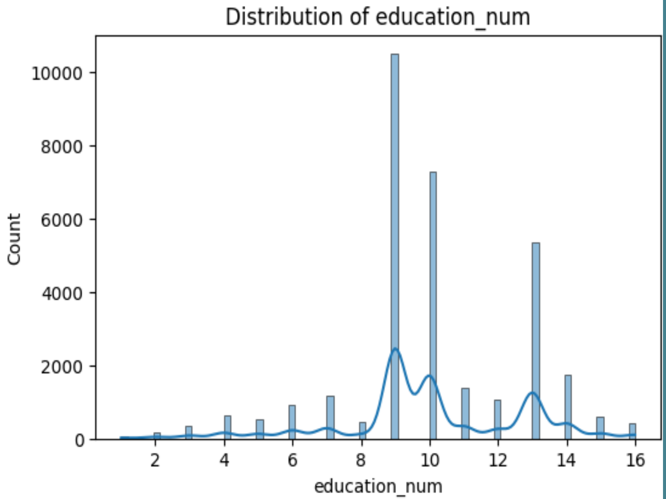
- Strong peaks at 9, 10, 13
- high school or some college education.
- Higher education levels -> earning >50K.

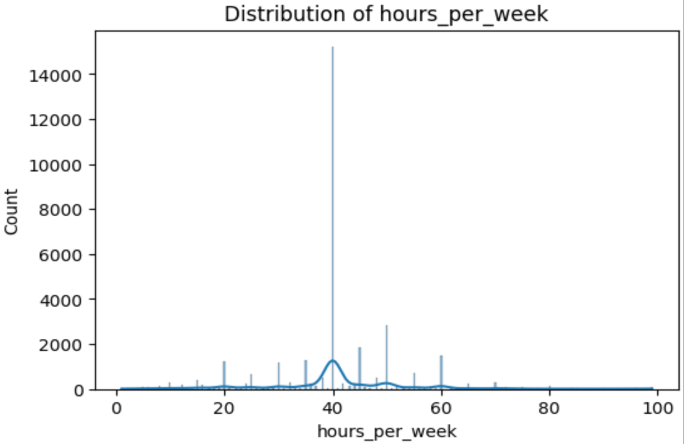
Hours per Week

- Peaked
- Standard workweek
- more than 40 hours/week -> earning >50K



Univariate Analysis



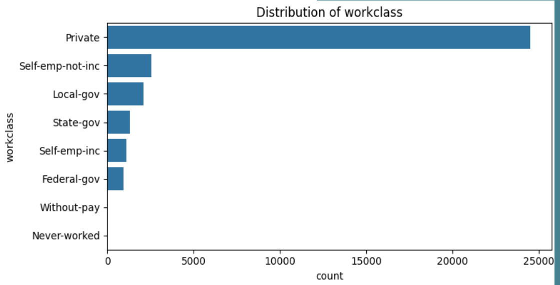


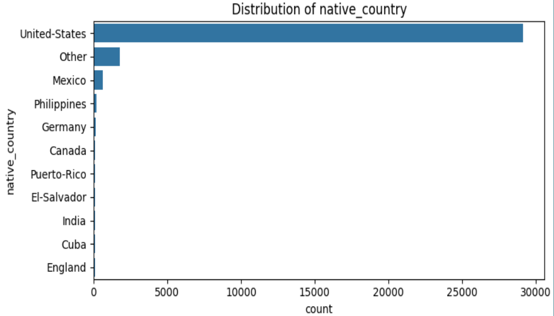
Initial Data Exploration

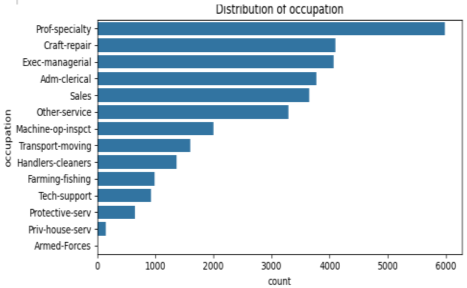
Univariate Analysis

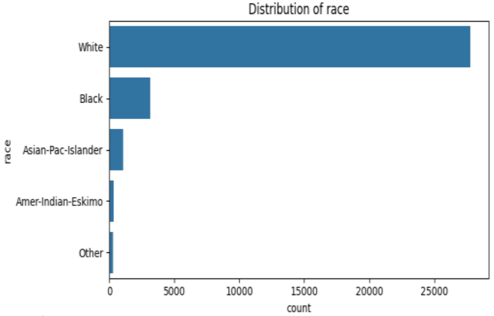
Categorical Features

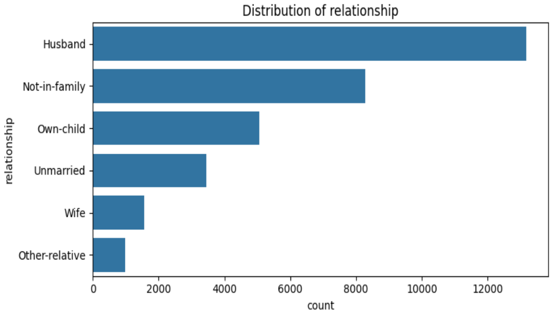
- **Workclass**: Most individuals work in the Private sector
- **Occupation**: The most common occupations are Prof-specialty.
- **Native Country**: Almost all participants are from the United States.
- **Race**: The majority are White
- **Relationship**: Husband is the most frequent category
- **Marital Status**: Married-civ-spouse is the dominant group

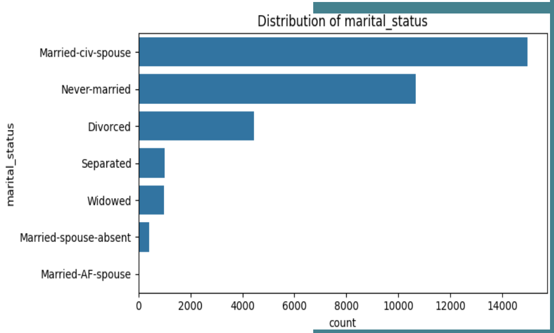






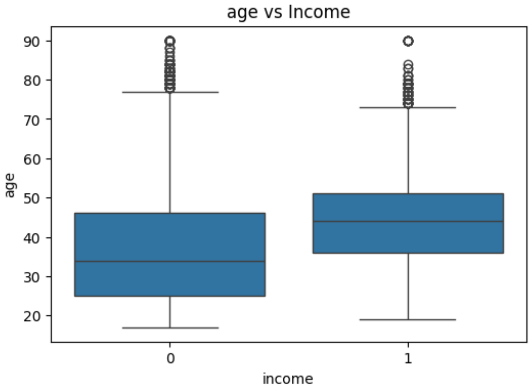


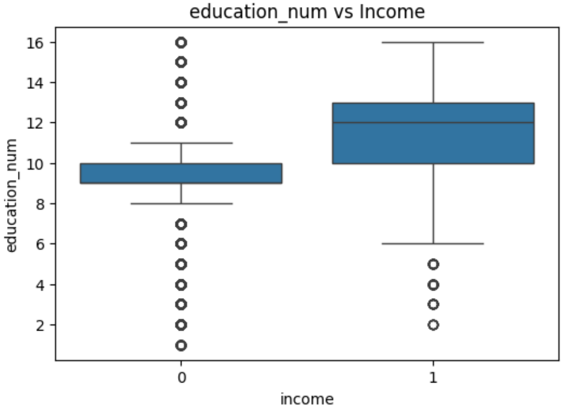




Initial Data Exploration

Bivariate Analysis



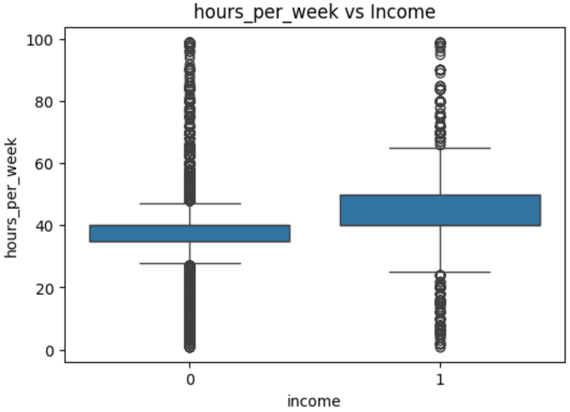


Education Level vs Income

- more years of schooling earn >50K.
- lower-income group is rough high school.
- Education strongly influences income

Age vs Income

- earning >50K tend to be older.
- >50K group has high median age
- outliers in both groups
- Work experience and seniority –> high income



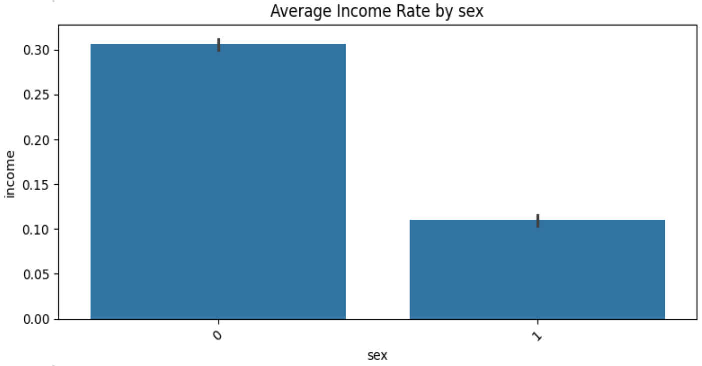
Hours per Week vs Income

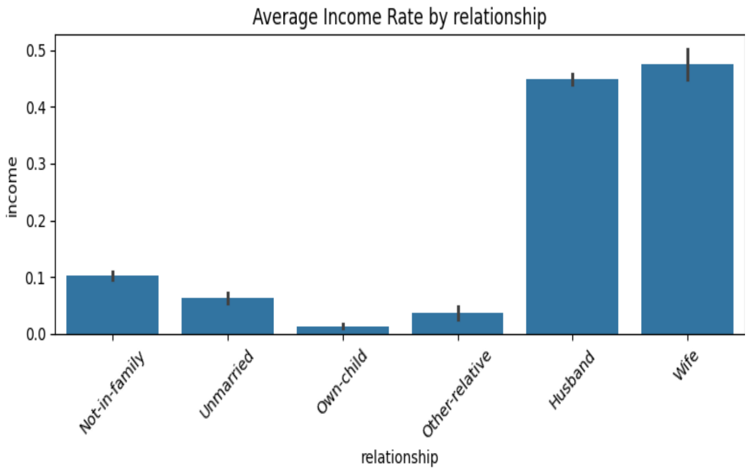
- Work more hours per week -> earning >50K
- Median for >50K group is higher than for <=50K group.
- Outliers in both

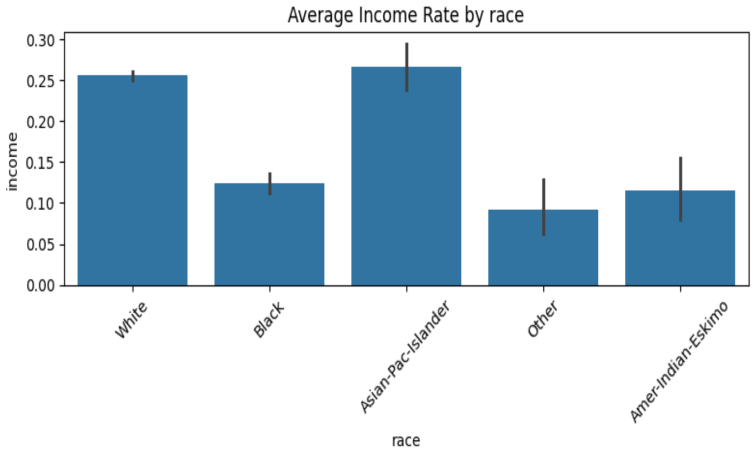
Initial Data Exploration

Bivariate Analysis

Categorical vs Target







Average Income Rate by Sex

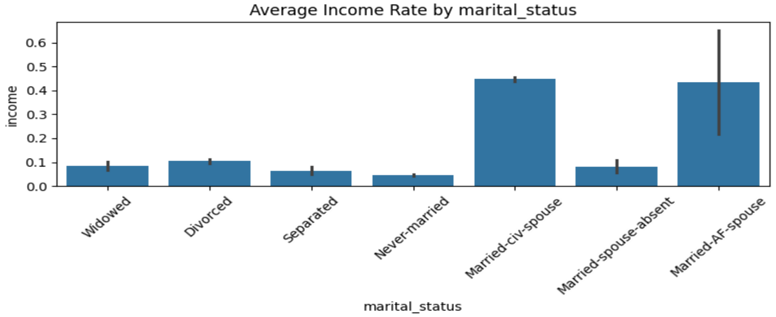
- Males have a significantly higher average income rate than females.

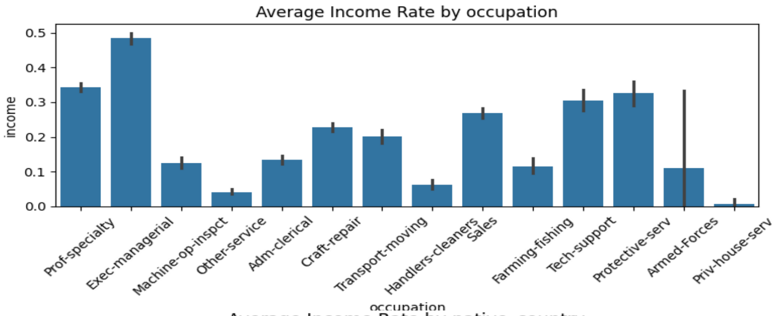
Average Income Rate by Relationship

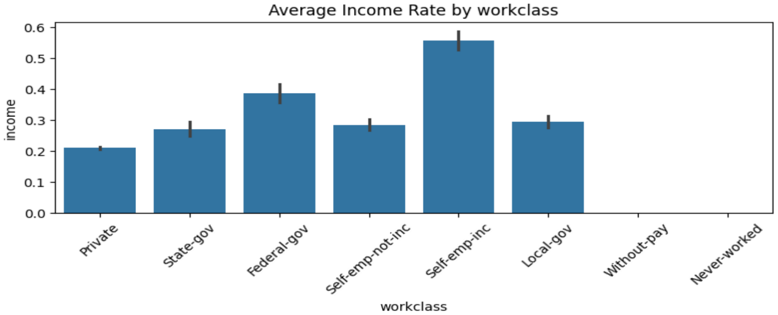
- Husband and Wife categories have much higher income rates

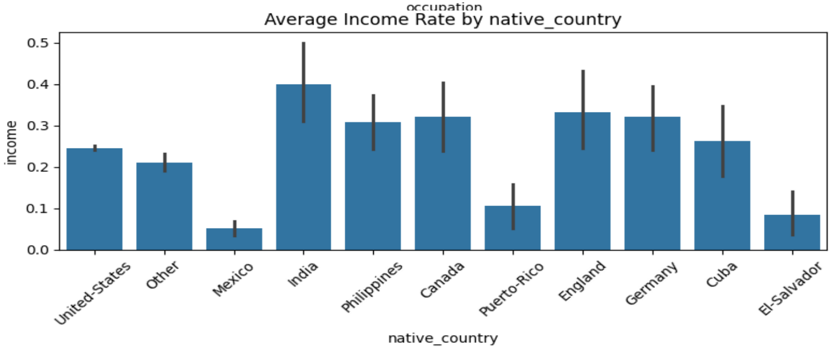
Average Income Rate by Race

- The Asian-Pac-Islander and White groups show higher average income rates.









Average Income Rate by Marital Status

- Individuals who are Married-civ-spouse have the highest income rate

Average Income Rate by Workclass

- Self-emp-inc (self-employed incorporated) shows the highest income rate among all work classes.

Average Income Rate by Occupation

- The Prof-specialty and Exec-managerial occupations have the highest proportion of high-income individuals.

Average Income Rate by Native Country

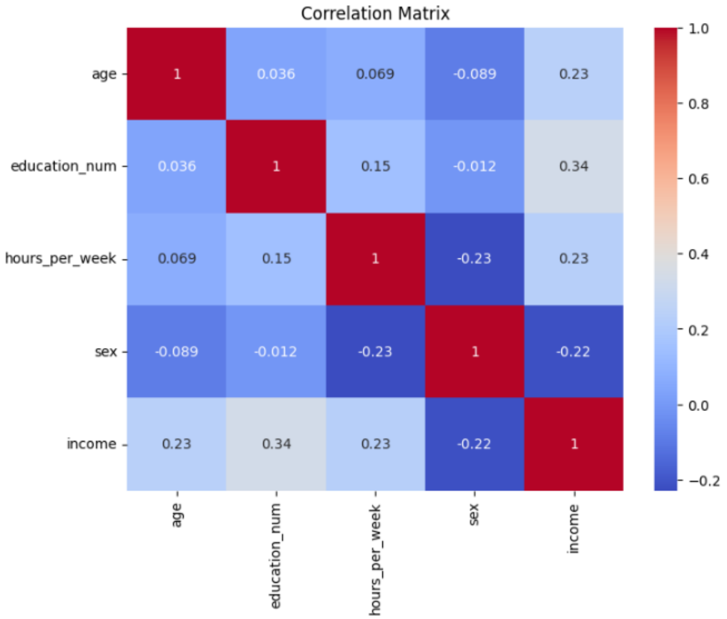
- The majority of individuals from the United States have a moderate average income rate.

Model Selection and Analysis

	age	education_num	sex	hours_per_week	income	workclass_Local-gov	workclass_Never-worked	workclass_Private	workclass_Self-emp-inc	workclass_Self-emp-not-inc	...	native_
0	1.000000	0.533333	1	0.397959	0	False	False	True	False	False	...	
1	0.890411	0.533333	1	0.173469	0	False	False	True	False	False	...	
2	0.671233	0.600000	1	0.397959	0	False	False	True	False	False	...	
3	0.506849	0.200000	1	0.397959	0	False	False	True	False	False	...	
4	0.328767	0.600000	1	0.397959	0	False	False	True	False	False	...	

Feature Engineering
- **One-hot encoding**
- **Feature scaling**
- **Class imbalance** was handled using **SMOTE** to balance income classes

Correlation Analysis
- Education level has the strongest positive correlation with income.
- Age and hours per week also show moderate positive correlations with income.
- Sex has a weak negative correlation



Models:
- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting

Training:
- Split into an 80:20 ratio.
- Train all baseline models (before hyperparameter tuning).
- Tune the hyperparameters of each model.
- Compare the results

Results

Models	Target	Precision	Recall	F1 Score	Accuracy Score	Confusion Matrix
Logistic Regression	<=50k	0.93	0.77	0.84	0.78	[3811 1129] [278 1290]
	> 50k	0.53	0.82	0.65		
	Macro Avg	0.73	0.8	0.75		
	Weighted Avg	0.84	0.78	0.8		
Decision Tree	<=50k	0.86	0.85	0.85	0.78	[4202 738] [710 858]
	> 50k	0.54	0.55	0.54		
	Macro Avg	0.7	0.7	0.7		
	Weighted Avg	0.78	0.78	0.78		
Randon Forest	<=50k	0.88	0.86	0.87	0.8	[4235 705] [603 965]
	> 50k	0.58	0.62	0.6		
	Macro Avg	0.73	0.74	0.73		
	Weighted Avg	0.8	0.8	0.8		
Gradient Boosting	<=50k	0.93	0.8	0.86	0.8	[3958 982] [317 1251]
	> 50k	0.56	0.8	0.66		
	Macro Avg	0.74	0.8	0.76		
	Weighted Avg	0.84	0.8	0.81		

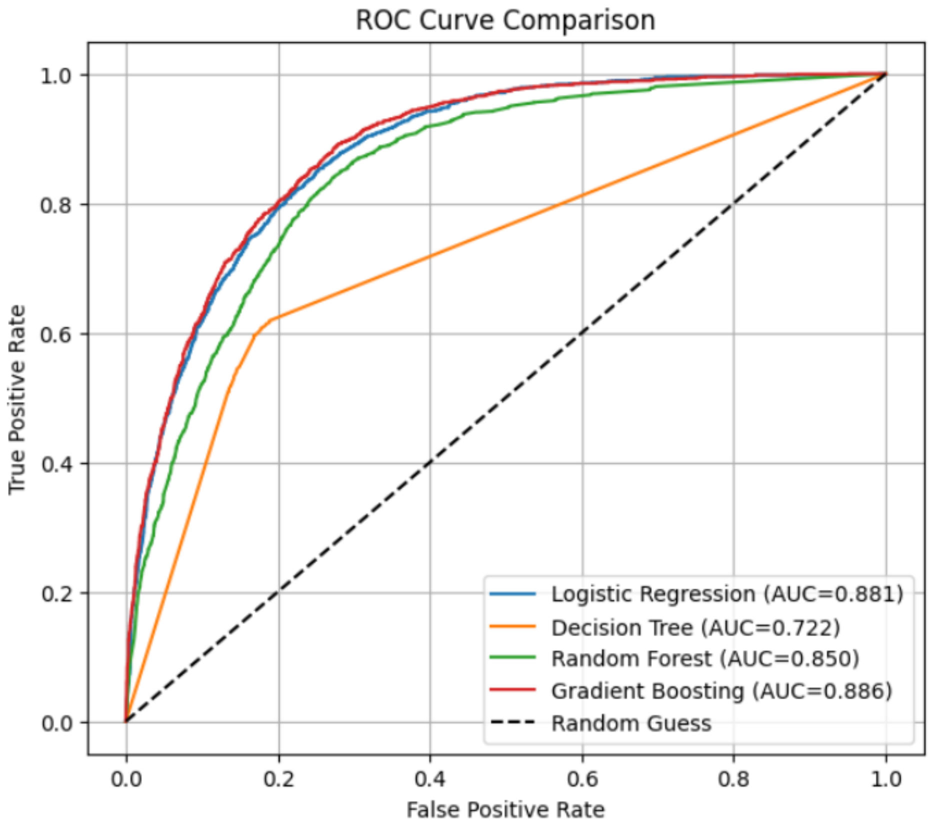
Logistic Regression:
- High precision for <=50K, quite high accuracy.
- F1 (>50K) 0.65 -> poor in detecting high-income individuals.

Decision Tree:
- Balanced results in both class
- Accuracy 0.78, risk of overfitting.

Random Forest:
- High accuracy (0.8), low precision for >50K.

Gradient Boosting:
- Best F1 overall, balanced between precision – recall.

To sum up:
- All models have similar accuracy score (0.78-0.8)
- Gradient Boosting provide the best balance precision, recall, F1 score.
- Logistic Regression performs well for <=50K, but poorly for >50K.
- Decision Tree, Random Forest perform moderately.



- Logistic Regression and Gradient Boosting achieved the highest AUC scores.

➔ Strong discriminative ability between income classes.

- Decision Tree performs the weakest (overfitting).

ROC-AUC Chart

Results Models

After Tunning Hyperparameter

Models	Target	Precision	Recall	F1 Score	Accuracy Score	Confusion Matrix
Logistic Regression	<=50k	0.93	0.77	0.85	0.79	[3819 1121] [277 1291]
	> 50k	0.54	0.82	0.65		
	Macro Avg	0.73	0.8	0.75		
	Weighted Avg	0.84	0.79	0.8		
Decision Tree	<=50k	0.92	0.78	0.85	0.78	[3857 1083] [329 1239]
	> 50k	0.53	0.79	0.64		
	Macro Avg	0.73	0.79	0.74		
	Weighted Avg	0.83	0.78	0.8		
Randon Forest	<=50k	0.89	0.86	0.87	0.81	[4247 693] [550 1018]
	> 50k	0.59	0.65	0.62		
	Macro Avg	0.74	0.75	0.75		
	Weighted Avg	0.82	0.81	0.81		
Gradient Boosting	<=50k	0.93	0.81	0.86	0.81	[3996 944] [322 1246]
	> 50k	0.57	0.79	0.66		
	Macro Avg	0.75	0.8	0.76		
	Weighted Avg	0.84	0.81	0.82		

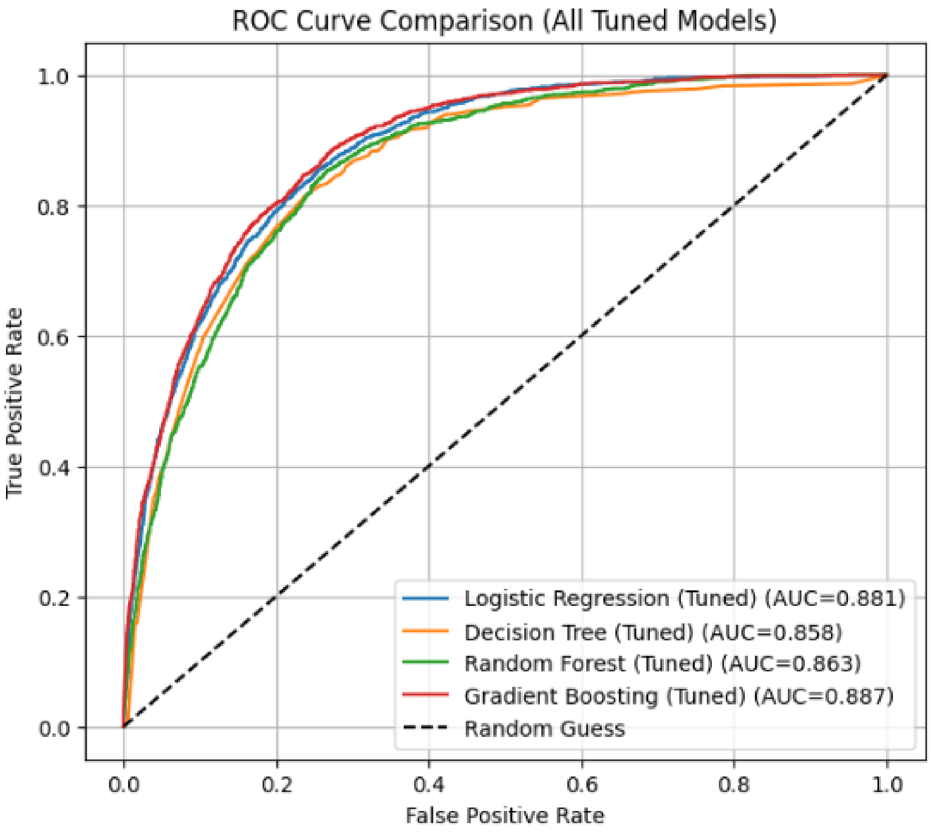
Logistic regression:
- Accuracy score improved slightly.
- Limitation improvement, F1 for > 50K still 0.65.

Decision Tree:
- Recall for >50K increased, accuracy score 0.78.
- F1 score is only 0.74.

Random Forest:
- High accuracy score (0.81), F1 score improved.
- Precision (>50K) is quite low.

Gradient Boosting:
- Best F1 and accuracy score, good balanced between precision – recall.

To sum up:
- All models improved slightly after hyperparameter tuning.
- Random Forest and Gradient Boosting reached the highest accuracy (0.81).
- Gradient Boosting remained the top model overall.
- Decision Tree and Logistic Regression showed moderate gains but remain weaker than others



After Tunning Hyperparameter

- Decision Tree showed the most significant improvement (0.722 – 0.858)
- Gradient Boosting achieved the highest overall performance.

➔ Tuning helps enhance models.
➔ Gradient Boosting was therefore selected as the final model (high AUC and balance predictive power.)

ROC-AUC chart

Conclusion

Model Evaluation Summary:

- ➢ The models perform better when using all available features.
- ➢ Hyperparameter tuning improves the performance of all models.
- ➢ Gradient Boosting achieved the best overall result with:
- Highest Accuracy score: 0.81
- Balanced precision, recall, F1 score
- Highest ROC-AUC score: 0.887

To sum up:
Gradient Boosting demonstrates the strongest discriminative ability.

Identification of Individuals by Income Group:

<= 50K Group(lower income):

- Majority of individuals in the dataset (75%).
- Typically younger and with years of education (high school or some college).
- Often work standard or fewer hours per week (<40 hours).
- Commonly Private employees in services or manual jobs (sales, craft,…).
- More likely to be female, never-married or not in family relationship categories.

>50K Group(higher income) :

- Minority of the population (25%).
- Usually older, indicating more work experience.
- High education levels (Bachelor’s degree and above.
- Work longer hours per week (>40 hours)
- Frequently in professional or managerial occupation (Prof-specialty, Exec-managerial)
- Commonly married and self-employed or in higher work position.
- Higher percentage found among White and Asian-Pac-Islander races.

THANKS!