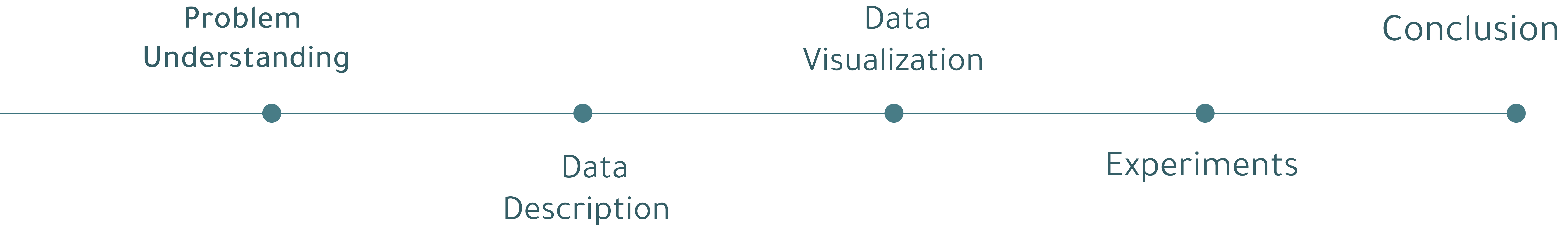
Income Classification

Methodology



Problem Understanding

Data Description

Data Visualization

Experiments

Conclusion

Problem Understanding

Overview

Income is money that a person or a business receives in return for working, providing a product or service, or investing capital. A person's income may also derive from a pension, a government benefit.

Problem Statment

building a model to predict whether an individual's income will be greater than $50,000 per year based on several attributes .Helping Governments for income tax or any finance company .

Data Description

They Data was token from kaggle. Dataset contains information about people and their income.

15 Columns and 32,561 Rows.

Data Description

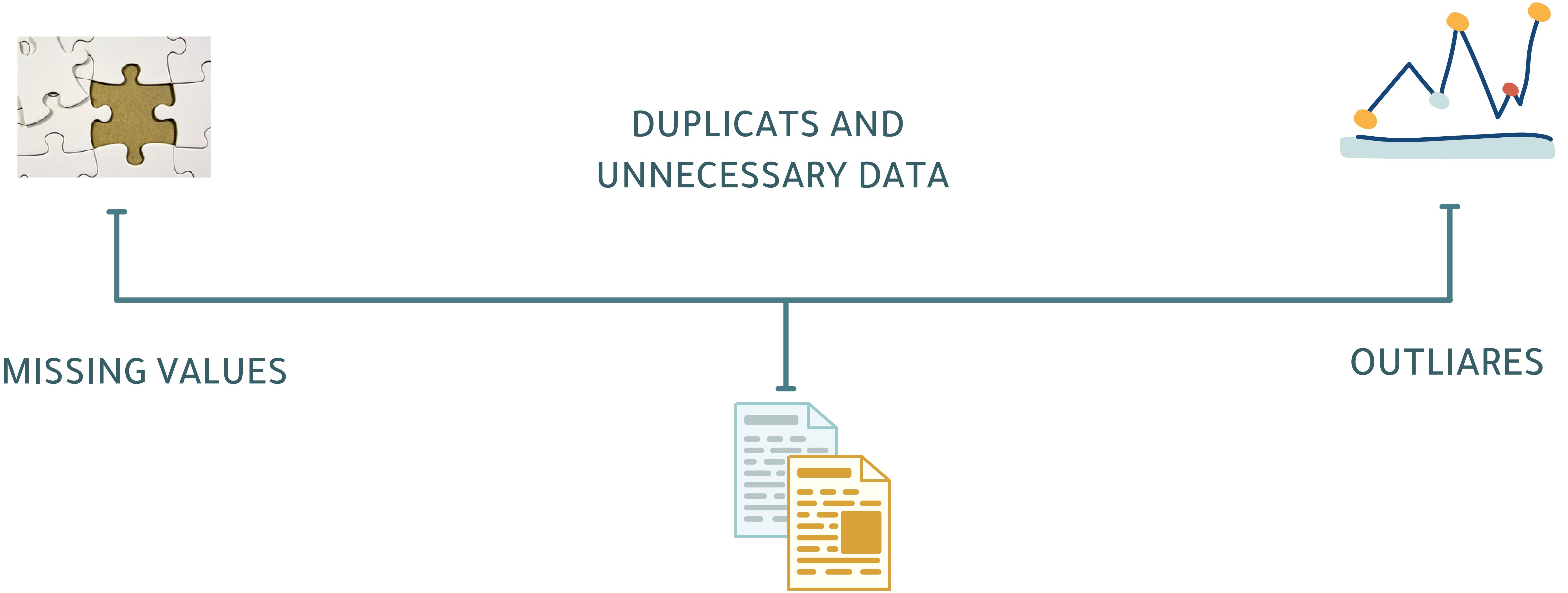
Features

age
workclass
Fnlwgt
education
education-num
marital-status
occupation
Relationship
race
sex
capital-gain
capital-loss
hours-per-week
native-country

Target

Income

Data Validation



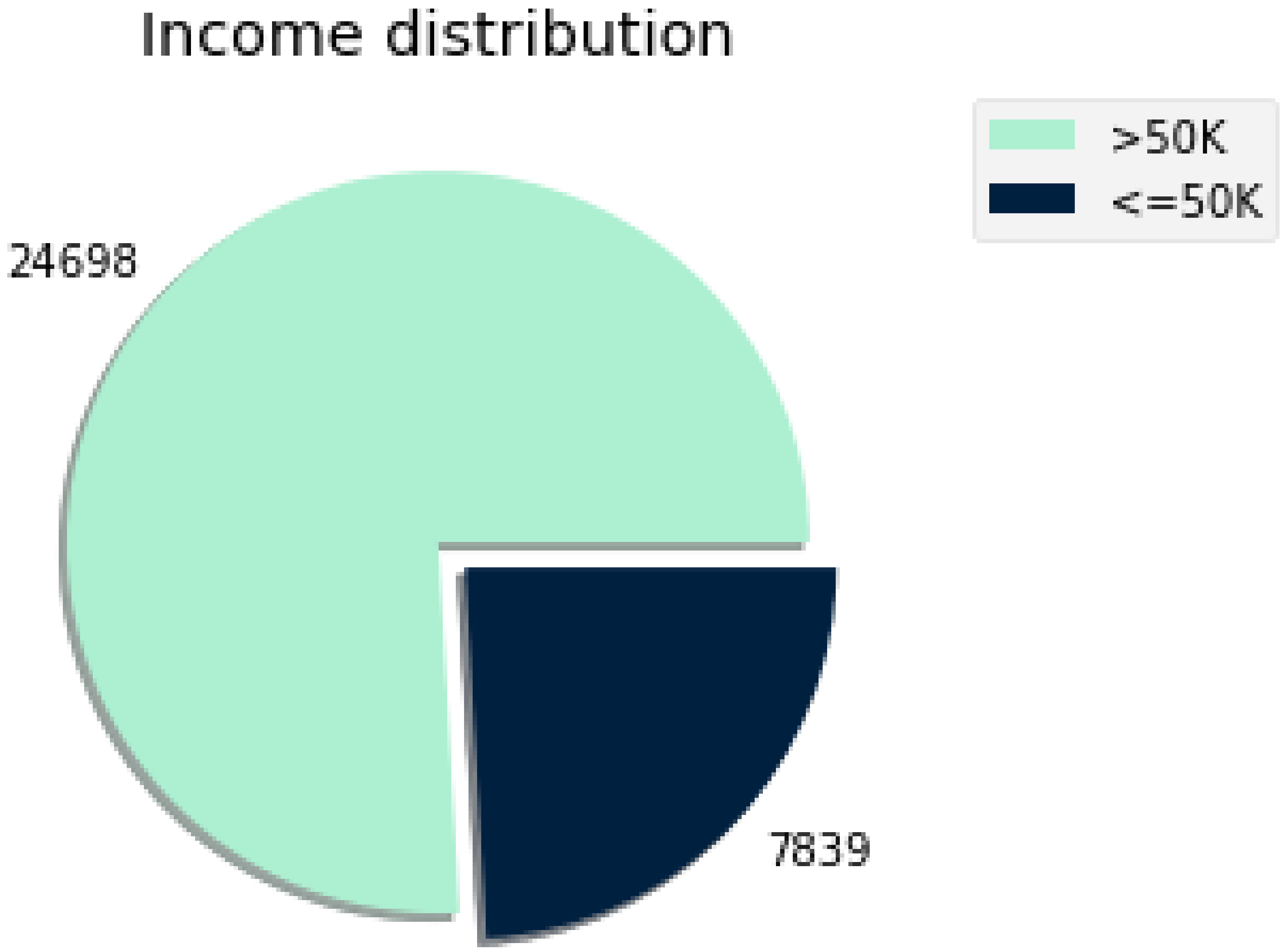
MISSING VALUES

DUPLICATS AND UNNECESSARY DATA

OUTLIARES

Pie Chart

Data Visualization



Observation

This plot shows that there is Imbalance in the data

Baseline Model

Train

F1 score **0.38**

Test

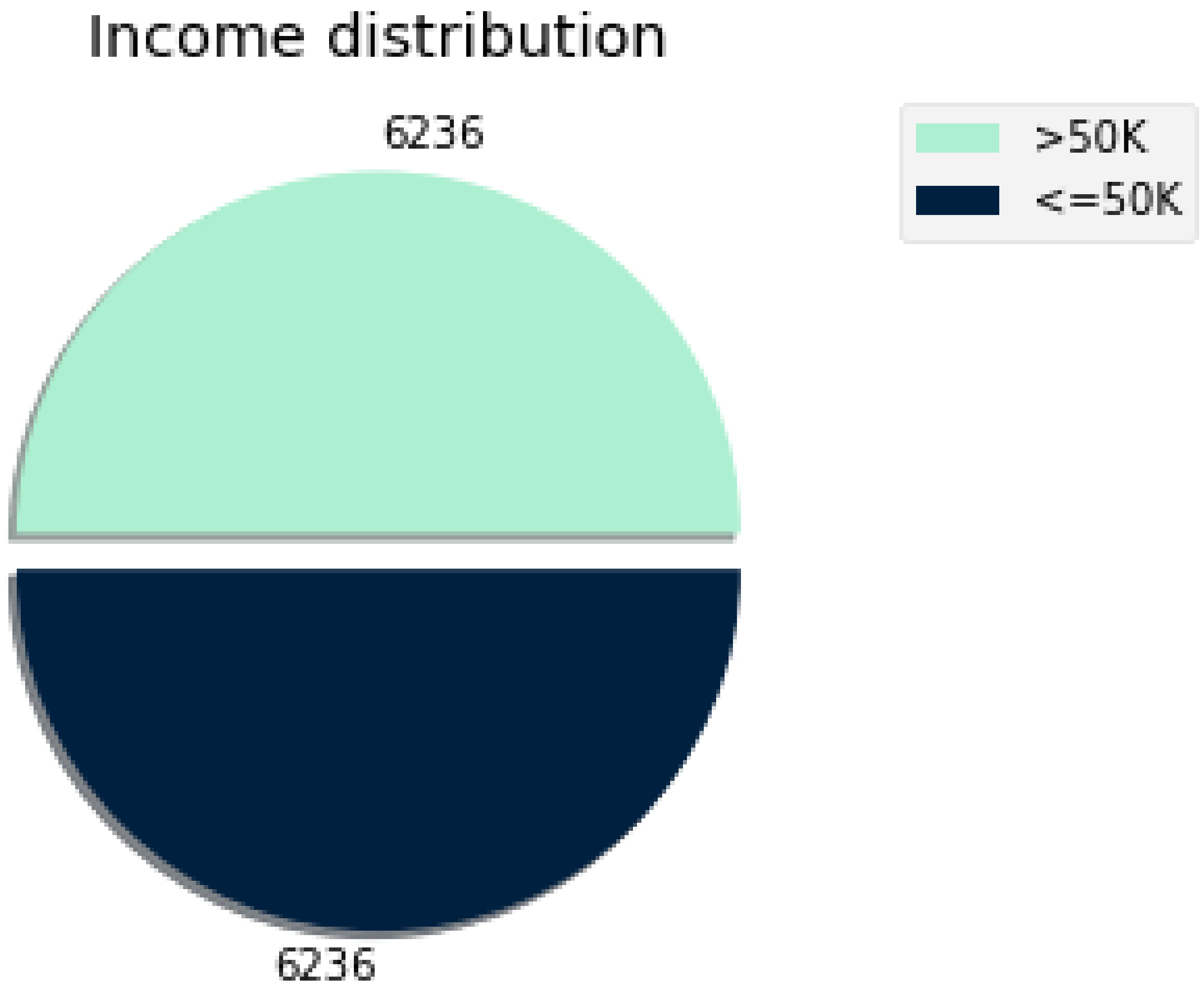
F1 score **0.40**

Dummy
with Resampling

	SMOTE	RandomUnder Sampler	RandomOver Sampler
Test	F1 score 0.420	F1 score 0.433	F1 score 0.428

Pie Chart

Data Visualization



Observation

Data distribution after using RandomUnderSampler

Decision Tree

	Befor Tuning	After Tuning
Train	F1 score 0.999	F1 score 0.674
Test	F1 score 0.620	F1 score 0.657

Random forest

	Befor Tuning	After Tuning
Train	F1 score 0.999	F1 score 0.72
Test	F1 score 0.677	F1 score 0.684

Knn

	Befor Tuning	After Tuning
Train	F1 score 0.60	F1 score 0.43
Test	F1 score 0.41	F1 score 0.39

Xgb

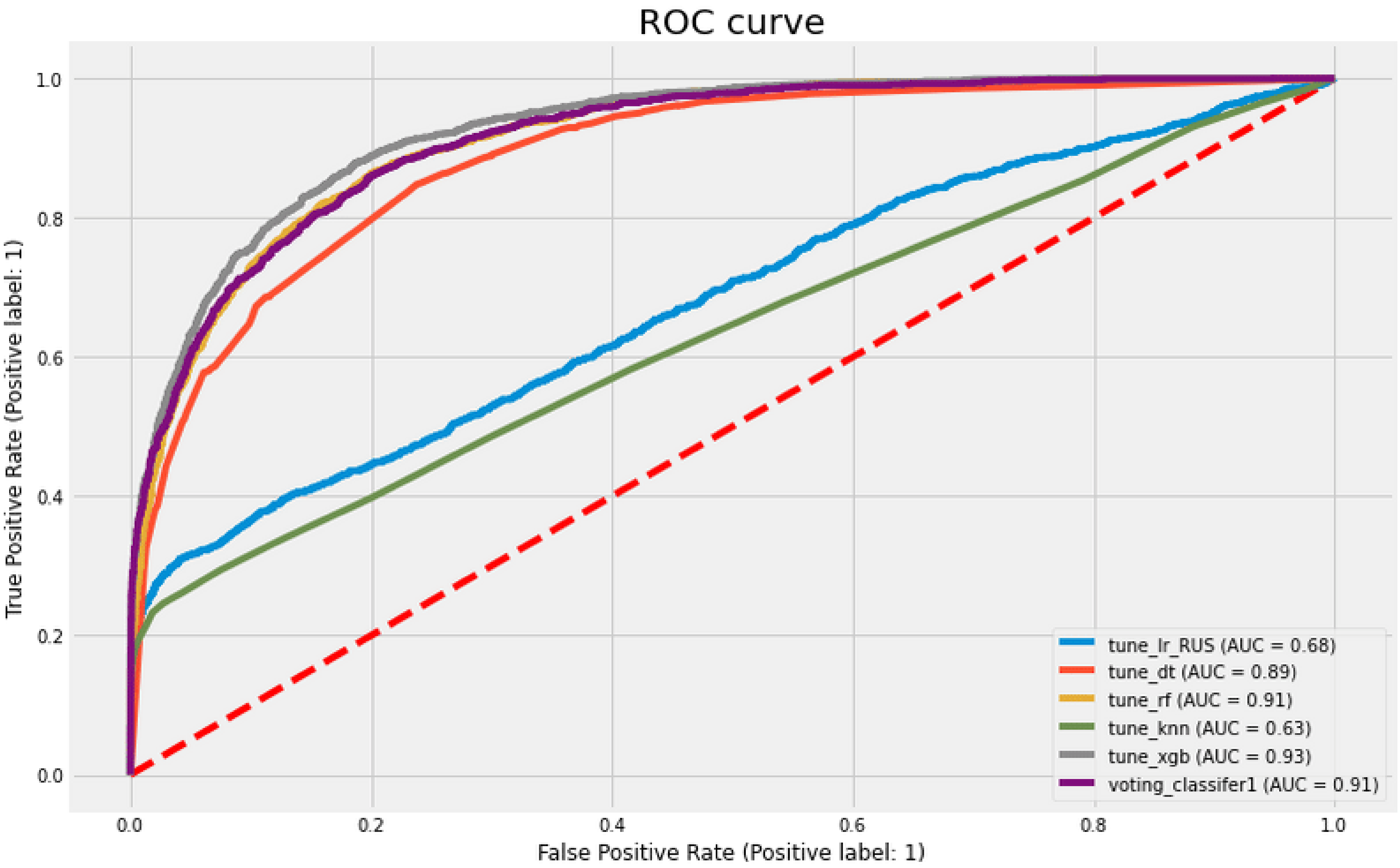
	Befor Tuning	After Tuning
Train	F1 score 0.872	F1 score 0.719
Test	F1 score 0.713	F1 score 0.717

Voting & Stacking Classifier

	Hard Voting	Soft Voting	Stacking
Train	F1 score 0.870	F1 score 0.876	F1 score 0.895
Test	F1 score 0.685	F1 score 0.693	F1 score 0.677

Roc Curve

Data Visualization

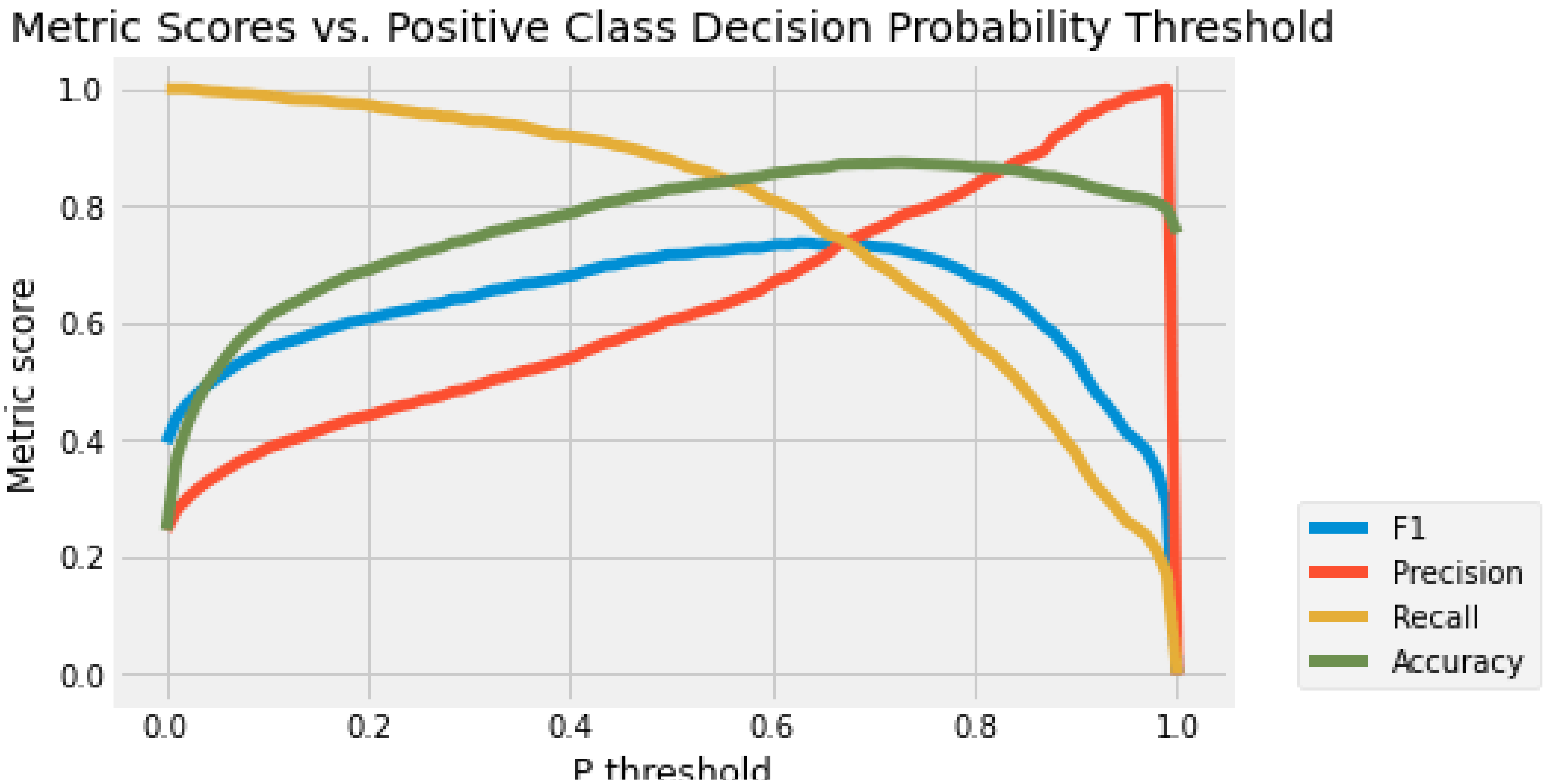


Observation

This plot shows the roc curve for all models.

Xgb Roc Curve

Data Visualization

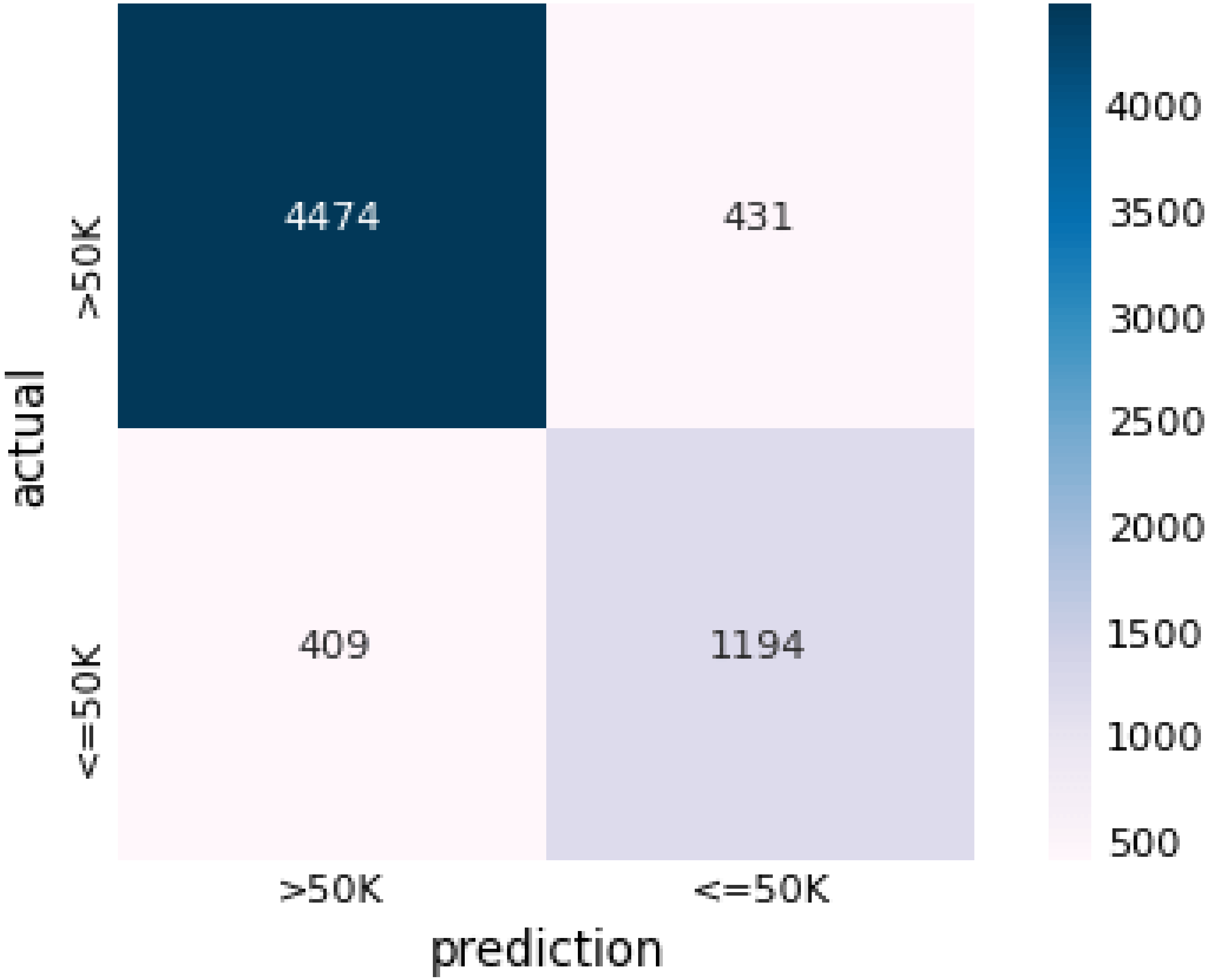


Observation

Roc curve for Xgb
F1 score 0.741
Threshold 0.667

Confusion Matrix

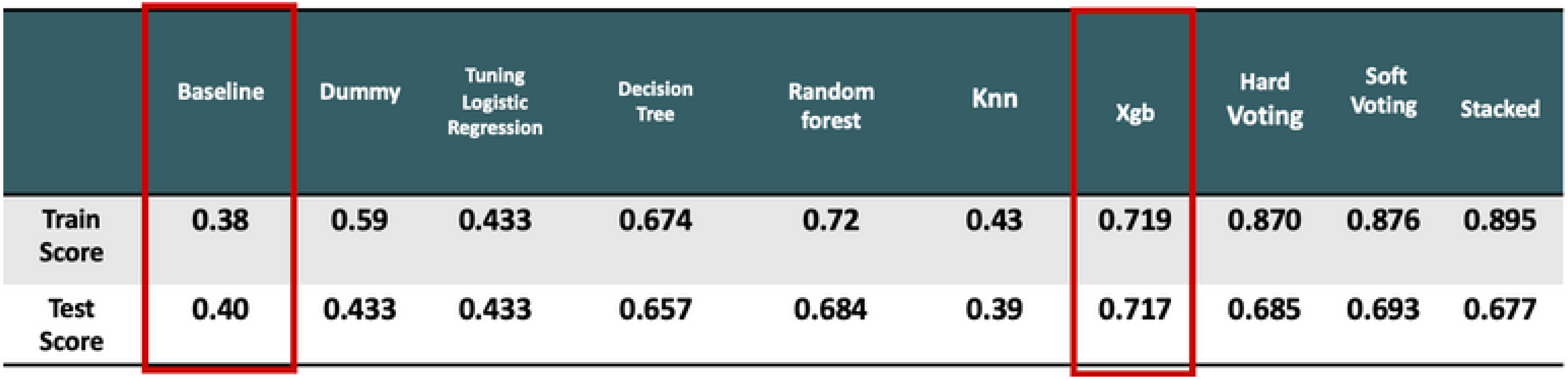
Data Visualization



Observation

This plot shows the Confusion Matrix for best Threshold

Conclusion



	Baseline	Dummy	Tuning Logistic Regression	Decision Tree	Random forest	Knn	Xgb	Hard Voting	Soft Voting	Stacked
Train Score	0.38	0.59	0.433	0.674	0.72	0.43	0.719	0.870	0.876	0.895
Test Score	0.40	0.433	0.433	0.657	0.684	0.39	0.717	0.685	0.693	0.677

Thank You..

Raghad Albarrak

Maryam Aljasham