

A photograph of two women in a room. In the foreground, a woman with short blonde hair is smiling and looking towards the right. In the background, a woman with dark curly hair is also smiling and looking towards the left. The wall behind them is covered with several framed photographs. The image is partially obscured by a purple and pink gradient overlay at the bottom.

20XX

# Adult Income Prediction

---

Machine Learning Project

# Data Exploration and Visualization

Individuals income if they are below or above a certain threshold (50,000) based on several characteristics

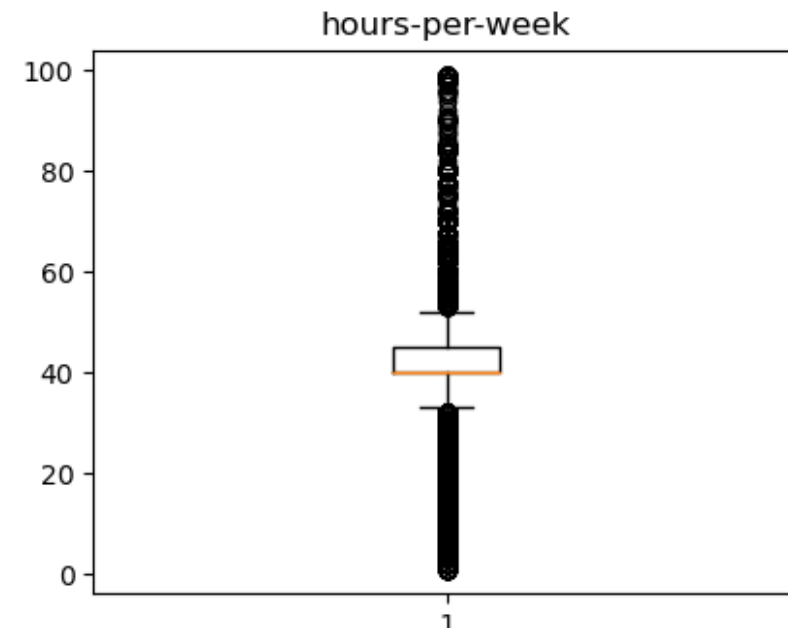
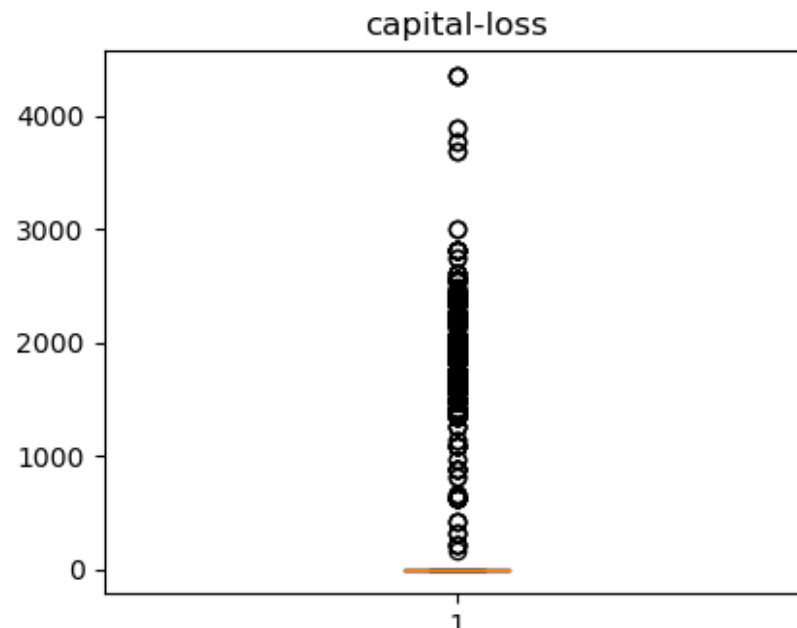
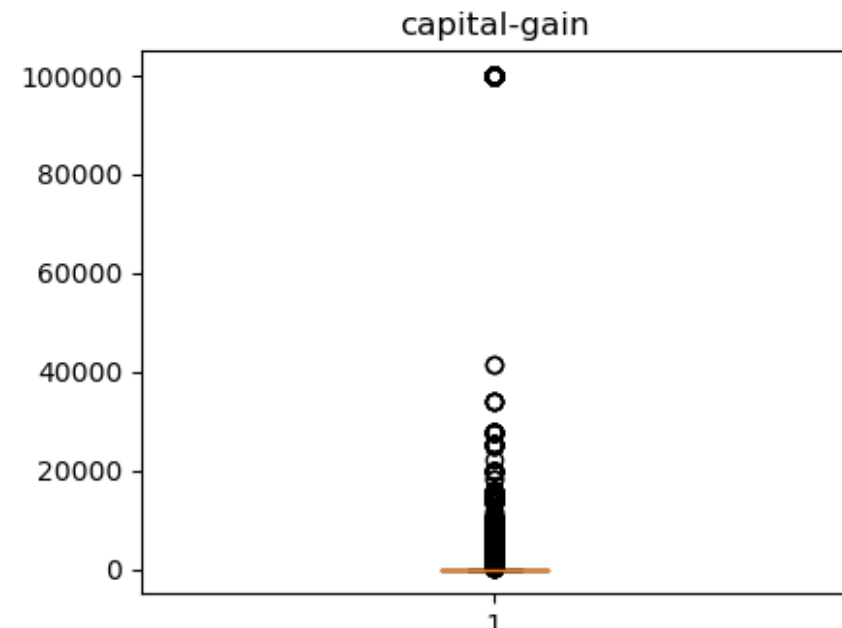
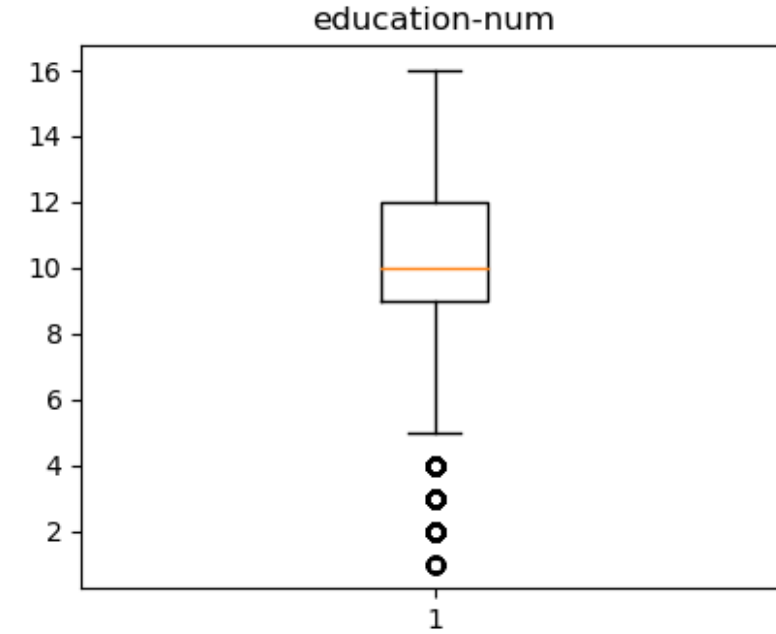
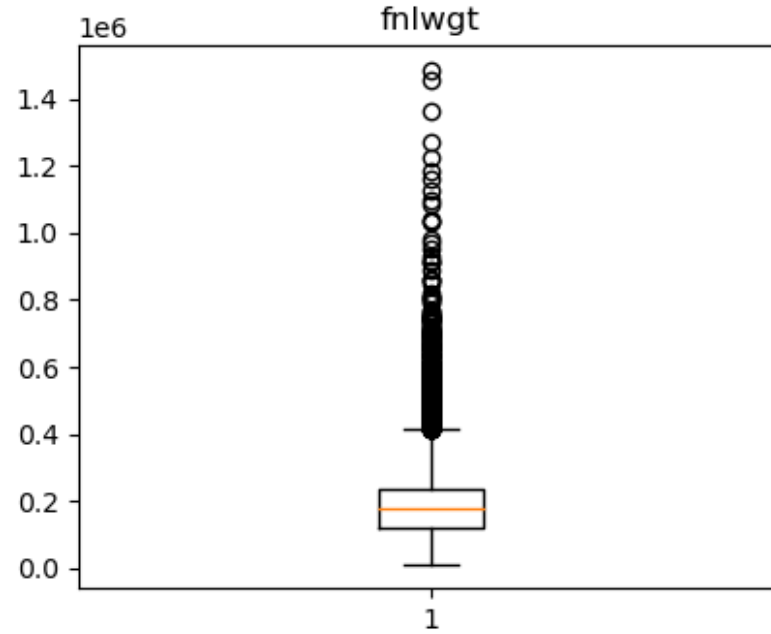
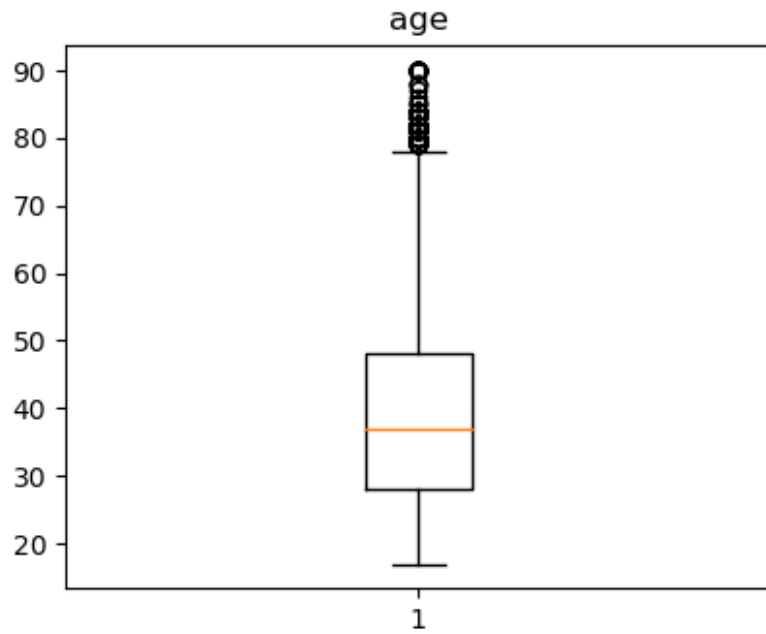
*Adult Income Prediction*

# Adult Dataset

From UCI, about 49000 samples

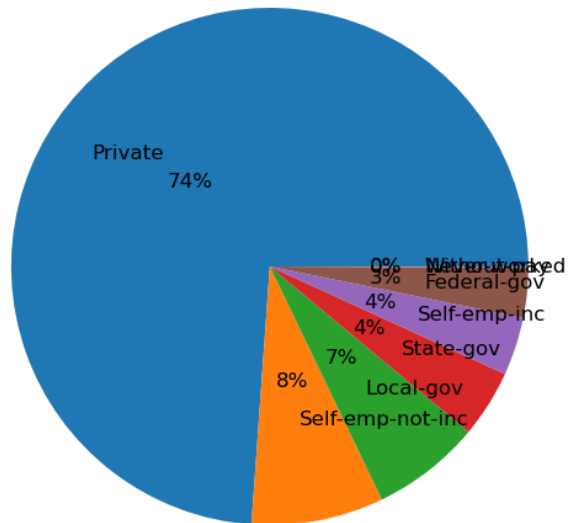
Id	Feature	Type
1	Age	numeric
2	workclass	categorical
3	fnlwgt	numeric
4	education	categorical
5	education_num	numeric
6	marital_status	categorical
7	occupation	categorical
8	relationship	categorical
9	race	categorical
10	sex	categorical
11	capital_gain	numeric
12	capital_loss	numeric
13	hours_per_week	numeric
14	native country	categorical

# Boxplot for Numerical

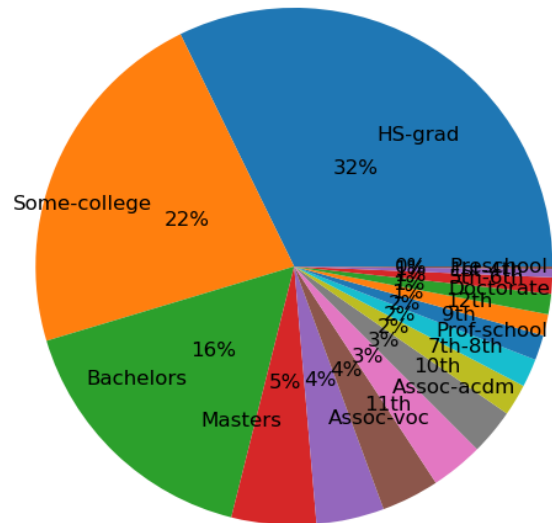


# Pie Chart for Categorical

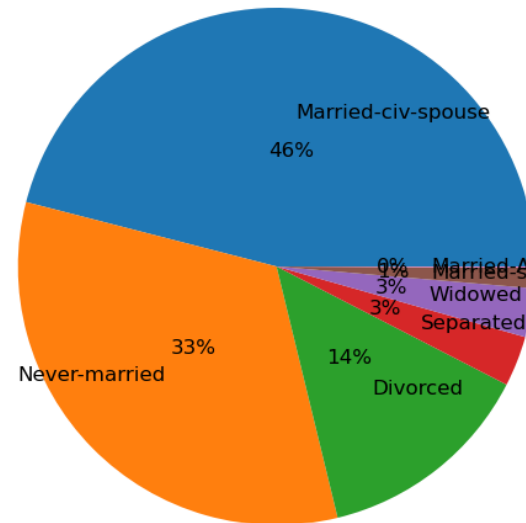
workclass



education



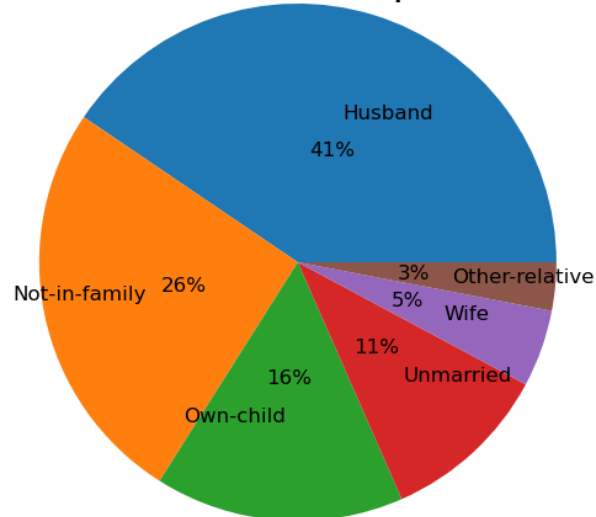
marital-status



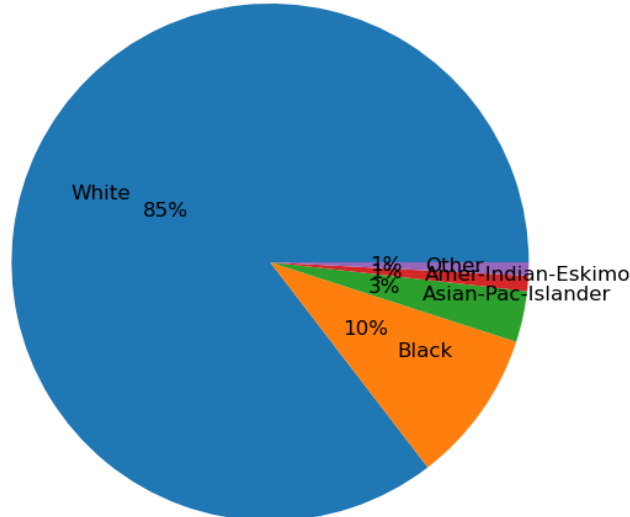
occupation



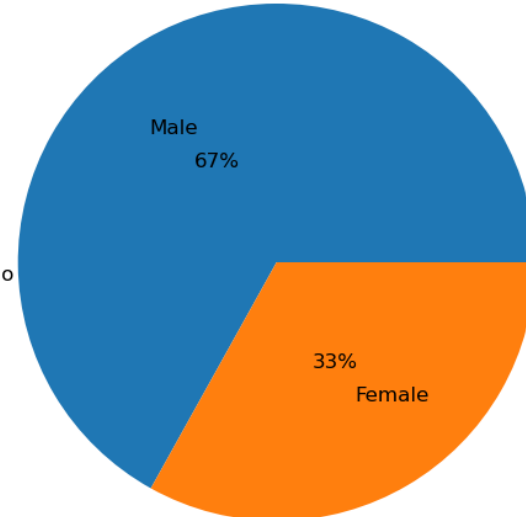
relationship



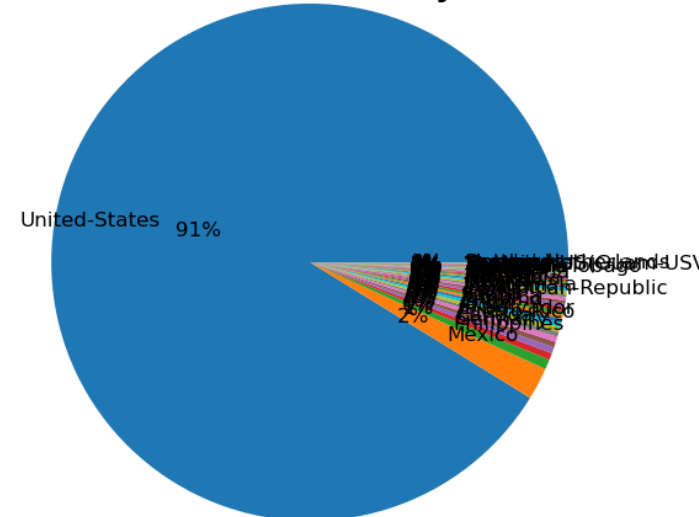
race



sex



native-country



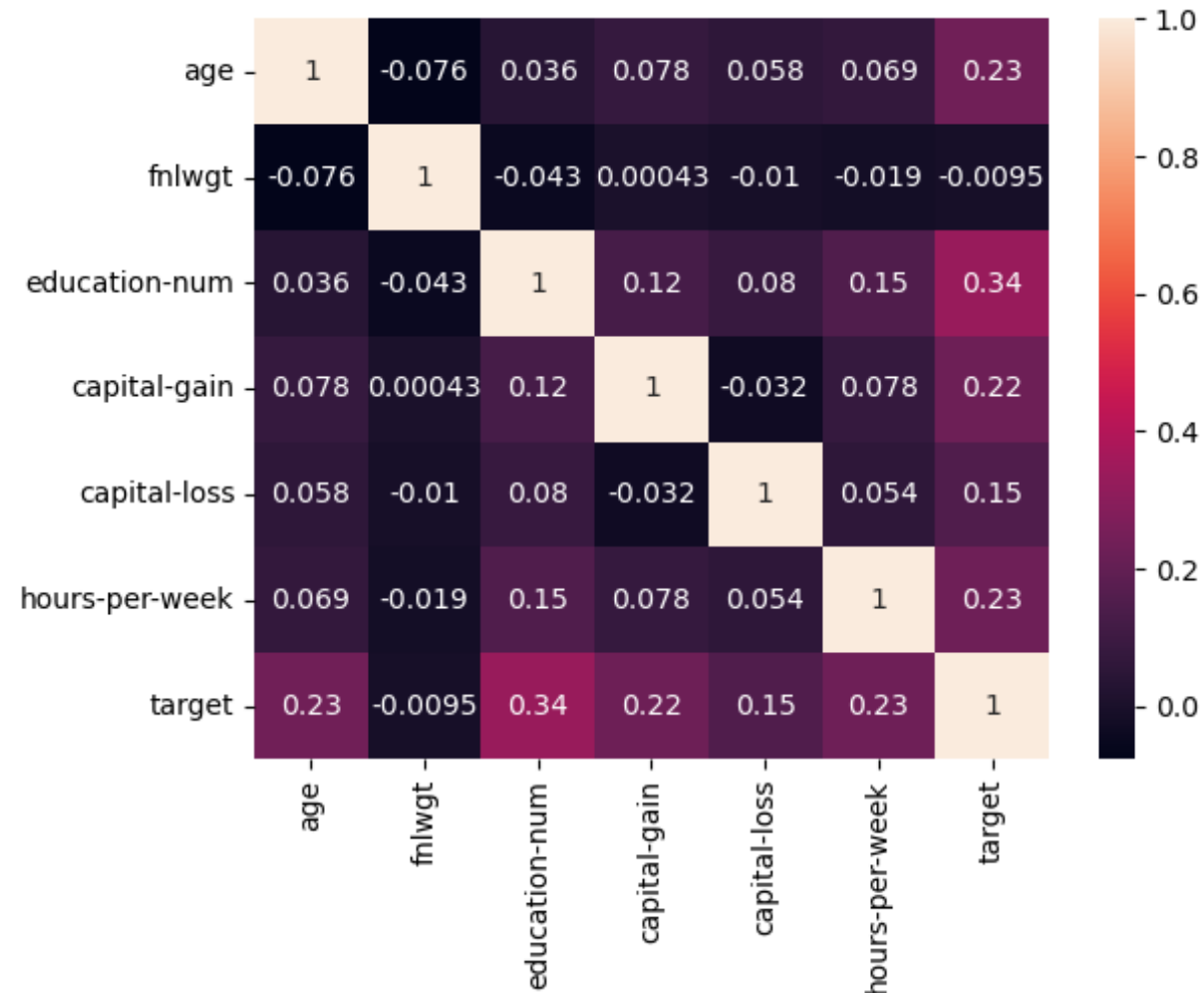


# Data Mining

20XX



# Correlations



## Correlations for Numerical and Target

- numerical features have low correlation with each other.
- the fnlwgt feature has low correlation with the target.

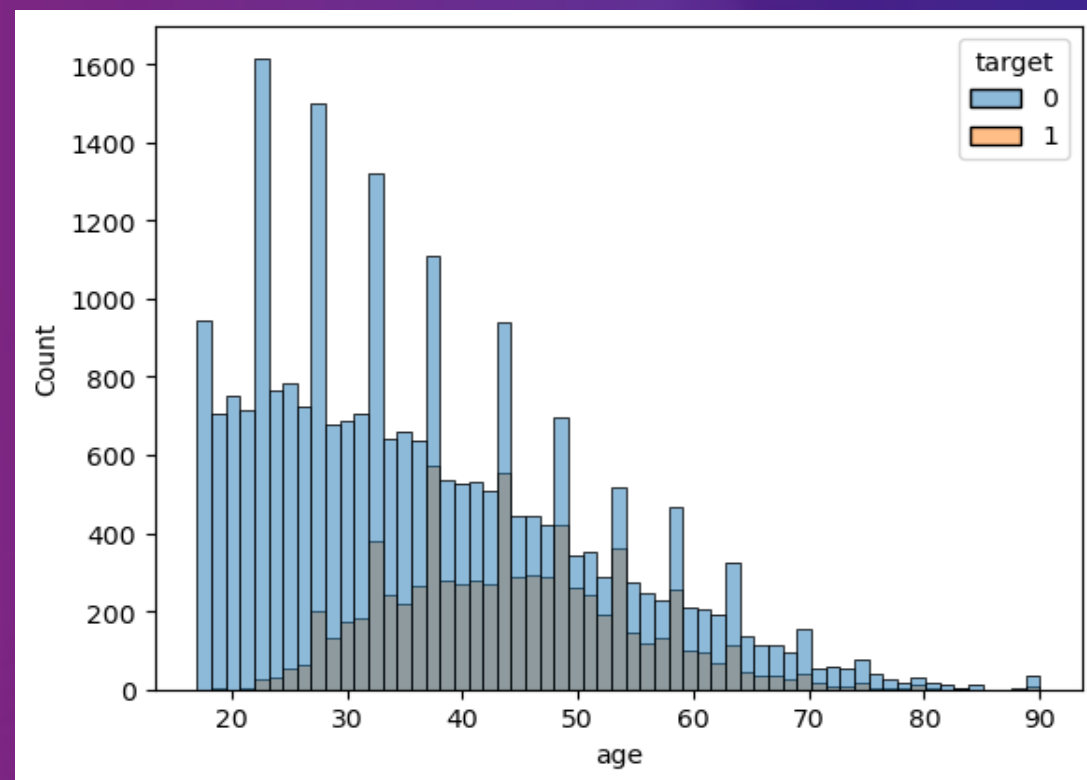
# Age Feature

20XX

## Explanation

when age increase, income does. so there is a positive relation between age and target.

## Histogram



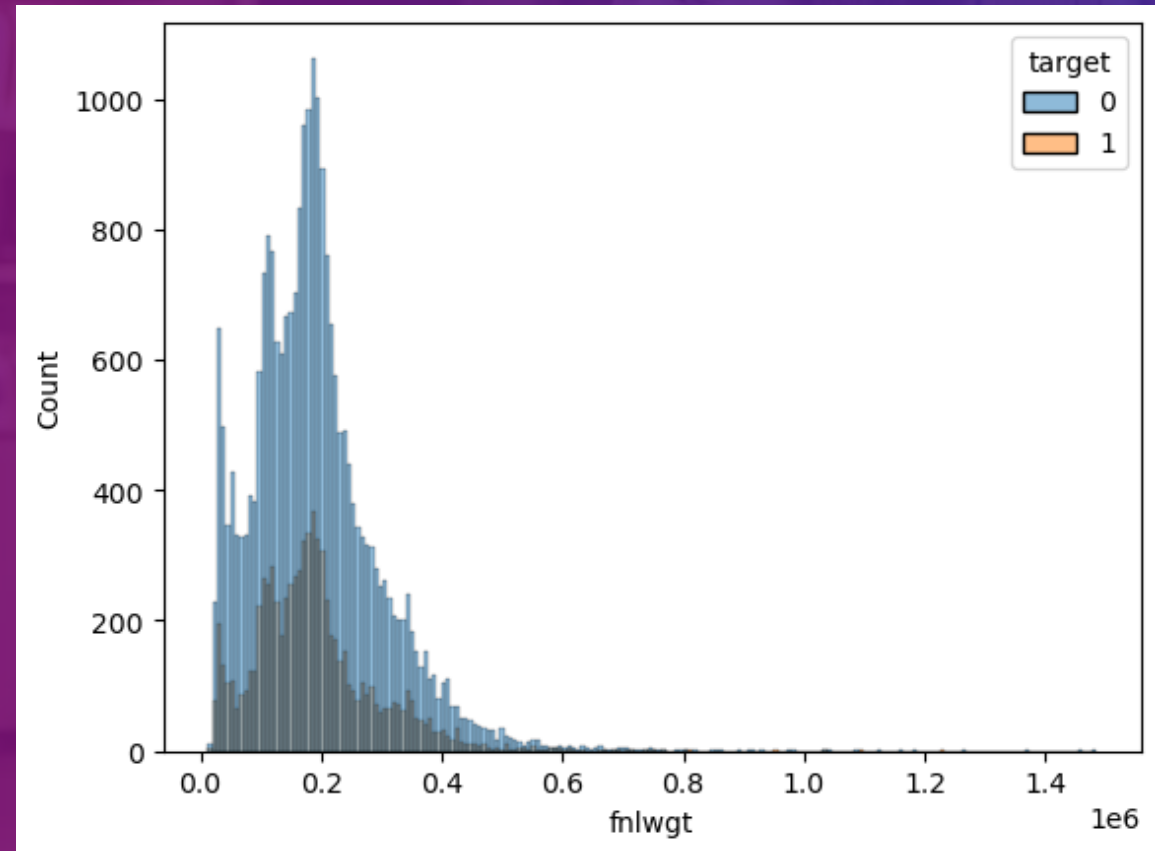


# Final Weight

## Explanation

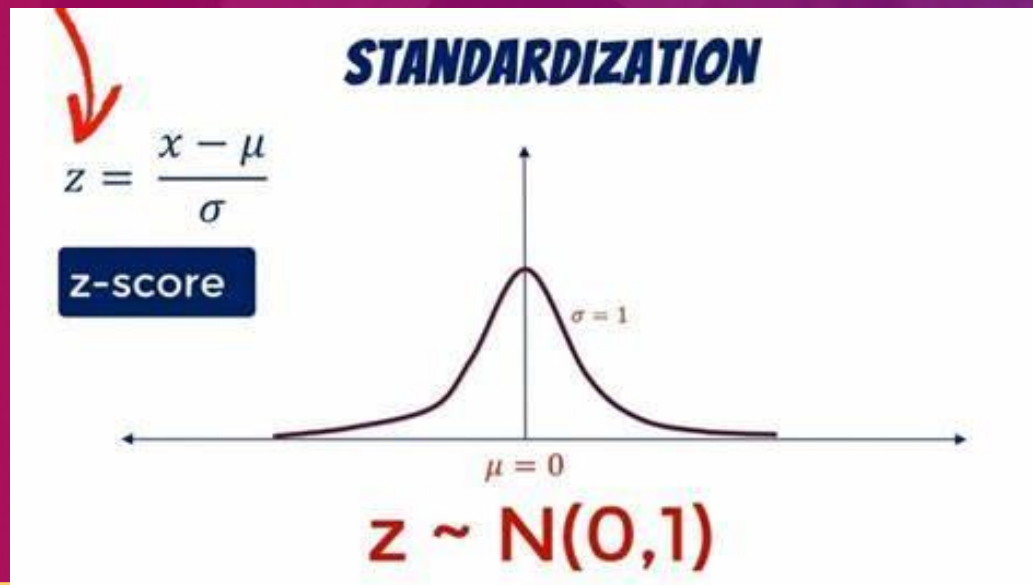
fnlwgt has similar variance for target classes which means it is not useful to make prediction.

Fraction = 1.07817



# Preprocessing

## Z-score scaling

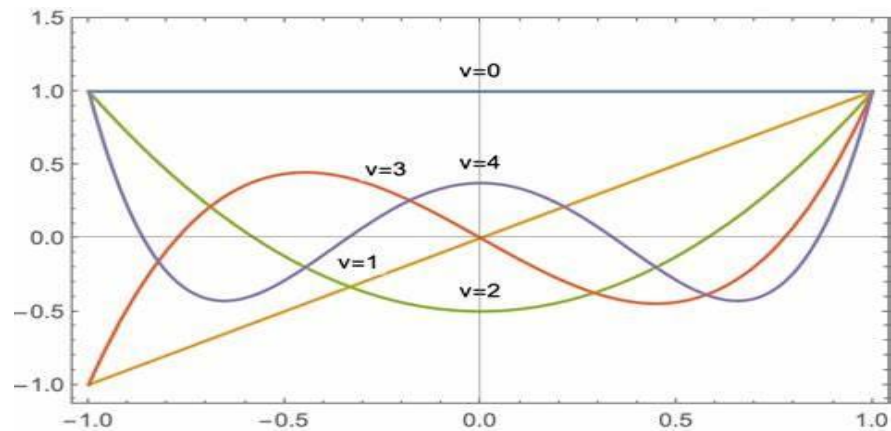


## One Hot Encoding

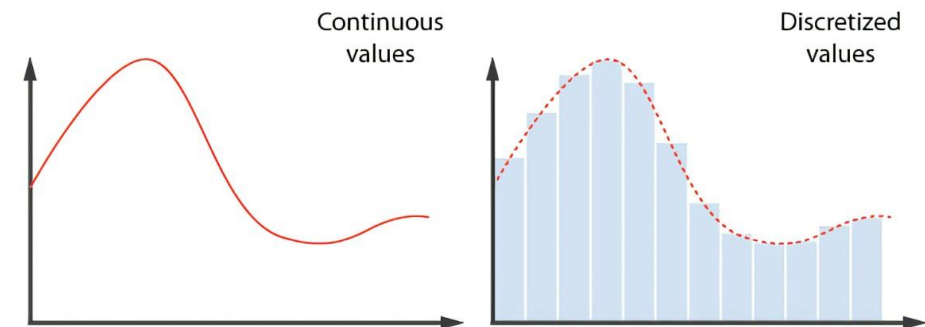
0	→	[ 1 , 0 , 0 , 0 ]
1	→	[ 0 , 1 , 0 , 0 ]
2	→	[ 0 , 0 , 1 , 0 ]
3	→	[ 0 , 0 , 0 , 1 ]

# Feature Extraction

## Polynomial features



## Discretization continuous features



# Training

*Adult Income Prediction*

20XX



# Experiments

Considering AUC for  
Optimization

Model	accuracy	ROC
Dummy classifier	75%	-
Logistic regression	85%	0.911
K-nearest neighbor	85%	0.905
Decision tree	85%	0.896
AdaBoost	86.7%	0.926
Gradient Boosting	86.7%	0.929





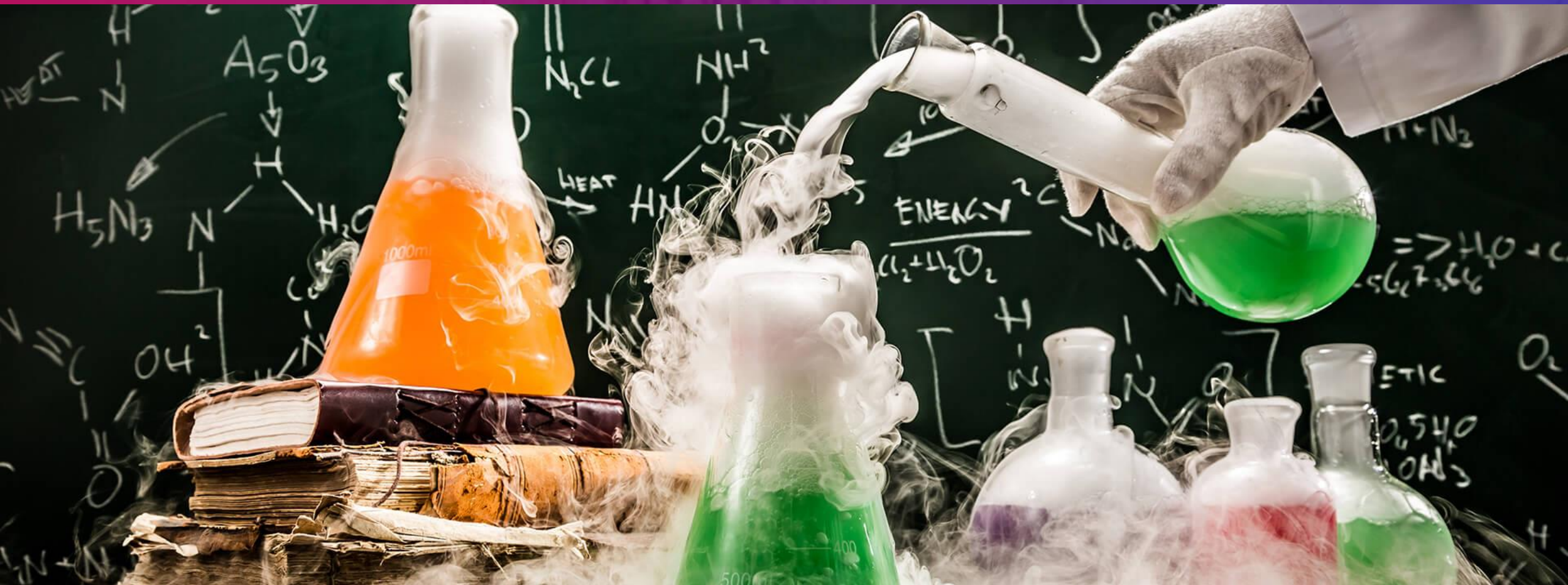
# The Best Model

## Gradient Boost Classifier

The best model with the highest accuracy and generalization is GBoost with a maximum depth of 4 and a number of estimators equal to 250.

AdaBoost model also achieved similar results.

# Testing



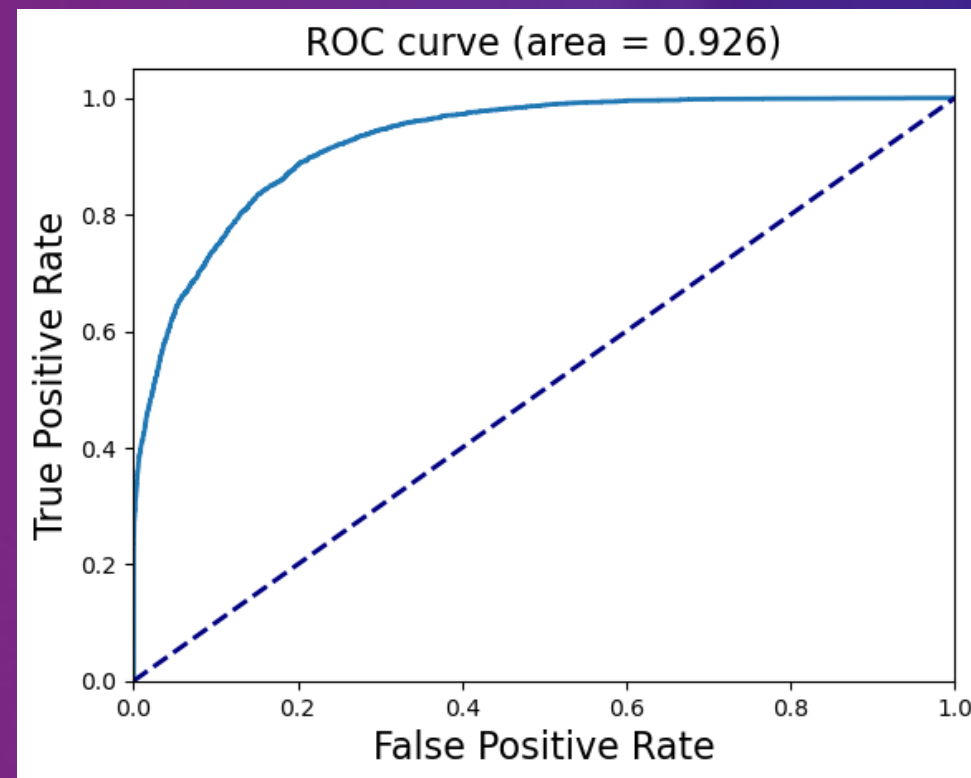
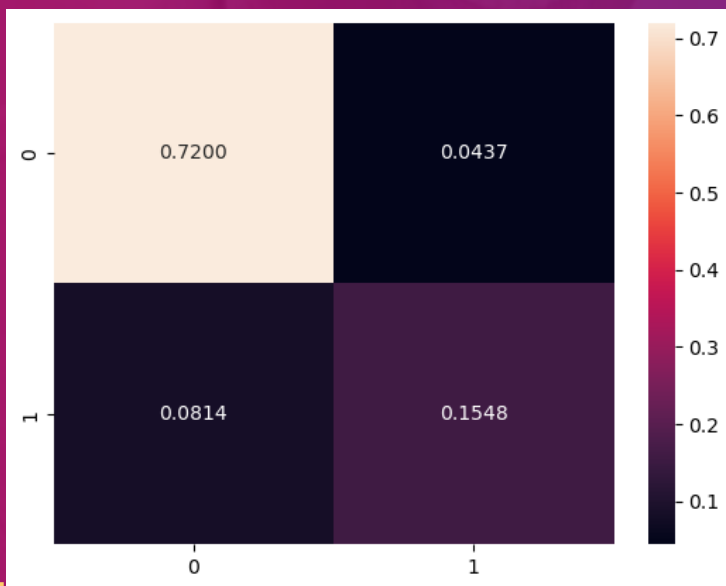
accuracy	recall	precision	auc
0.87488483	0.65548621	0.77977111	<b>0.92619429</b>

# Testing Scores

*Adult Income Prediction*

# ROC Curve

## Confusion Matrix





A woman with dark curly hair and glasses is smiling and looking off to the side. The background is a blurred office or indoor setting. A purple gradient bar is at the bottom of the image.

20XX

**THANK YOU!**

---

ML Godfather