Adult Income Prediction

Nian Vrey

October 2023

Introduction to Project

Project Description

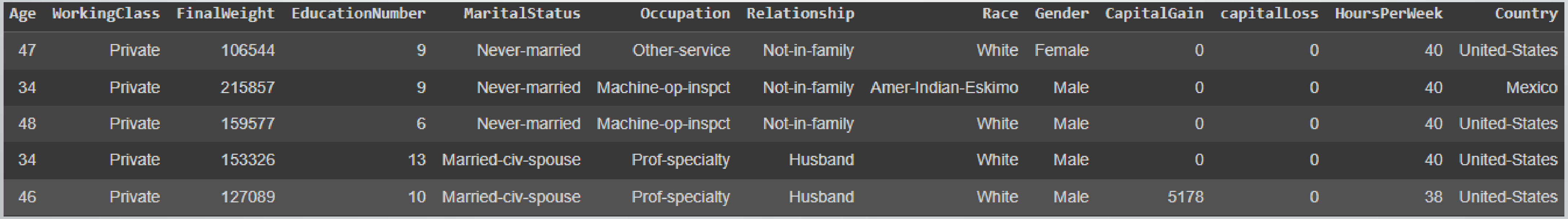
This project strives to analyse the dataset put forth on Kaggle and create a model that will predict which income class an adult belongs to.

The two classes are adults making more than $50,000 p/a (> 50k) and those making $50,000 and less p/a (<= 50k).

Project & Data Source:
https://www.kaggle.com/datasets/wenruliu/adult-income-dataset

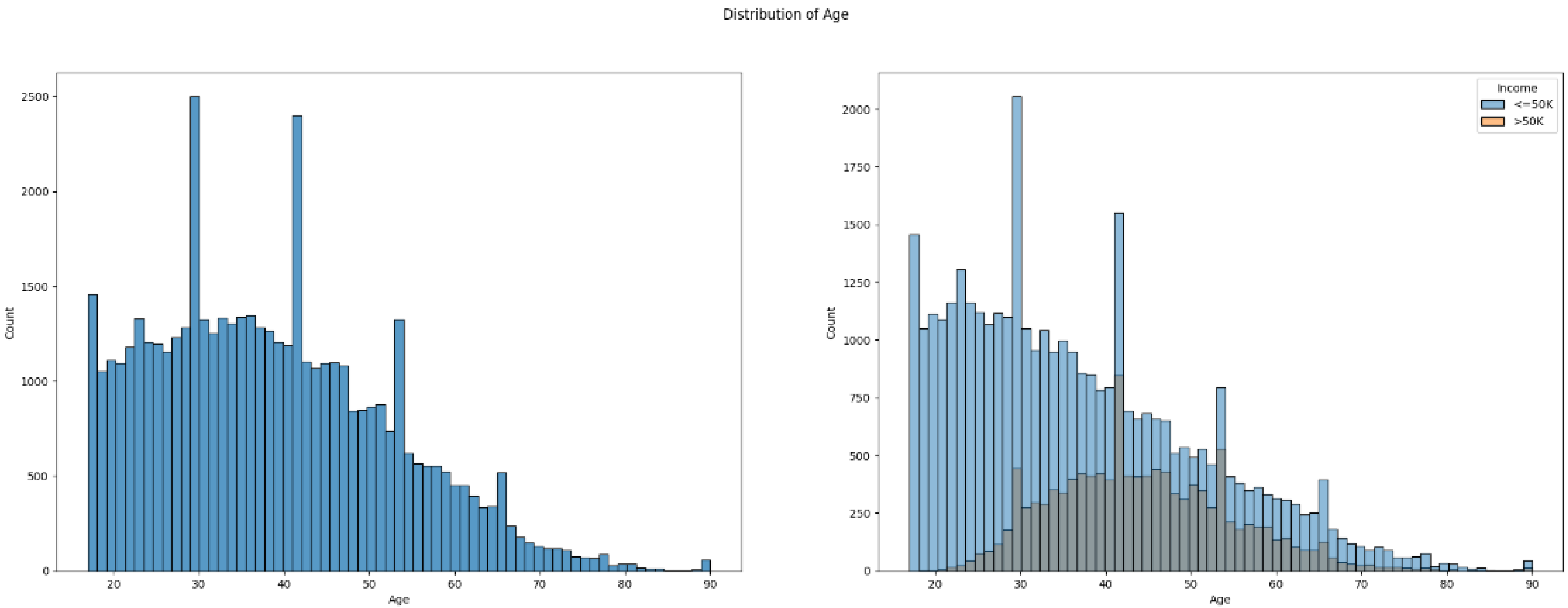
Adult Income Dataset

- The model utilizes 13 attributes used to describe an Adult
- The dataset had almost 49,000 rows

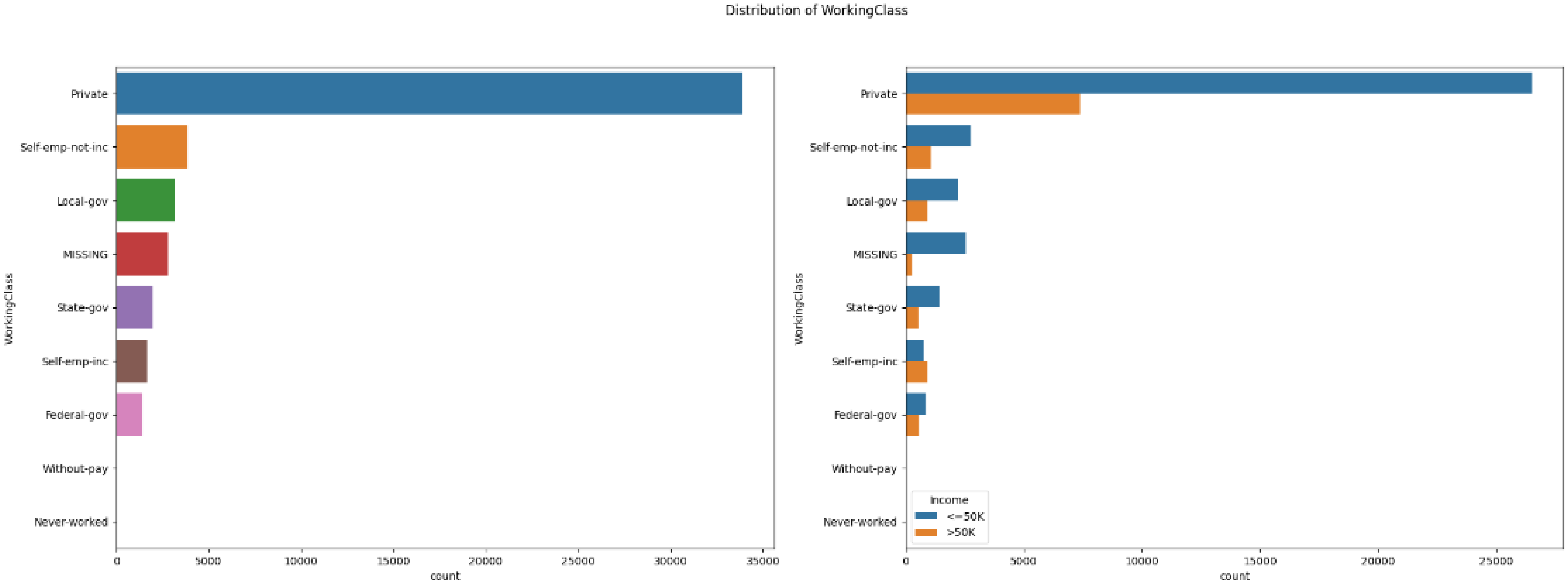


Age	WorkingClass	FinalWeight	EducationNumber	MaritalStatus	Occupation	Relationship	Race	Gender	CapitalGain	capitalLoss	HoursPerWeek	Country
47	Private	106544	9	Never-married	Other-service	Not-in-family	White	Female	0	0	40	United-States
34	Private	215857	9	Never-married	Machine-op-inspct	Not-in-family	Amer-Indian-Eskimo	Male	0	0	40	Mexico
48	Private	159577	6	Never-married	Machine-op-inspct	Not-in-family	White	Male	0	0	40	United-States
34	Private	153326	13	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States
46	Private	127089	10	Married-civ-spouse	Prof-specialty	Husband	White	Male	5178	0	38	United-States

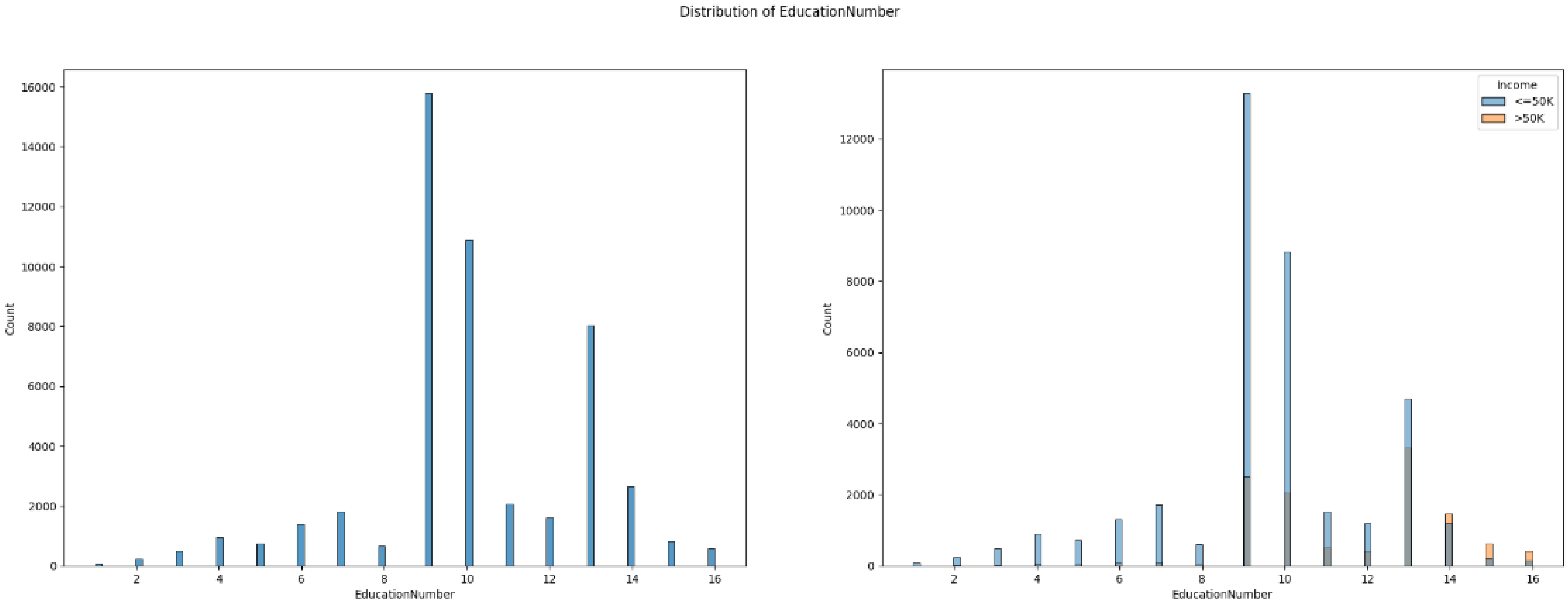
Distribution Charts for Noteworthy Findings



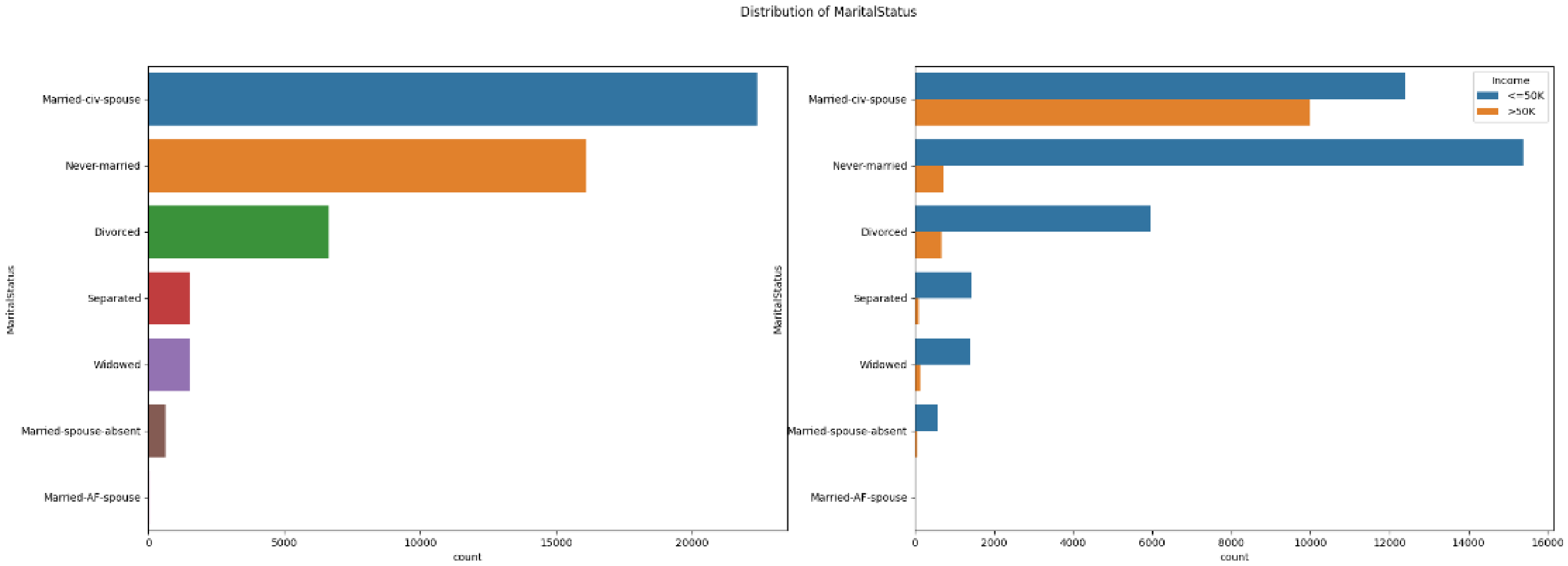
The distribution of >50k in relation to <=50k seems to be highest in age groups of high 30s to high 50s. Before and after that only a small portion reaches >50k



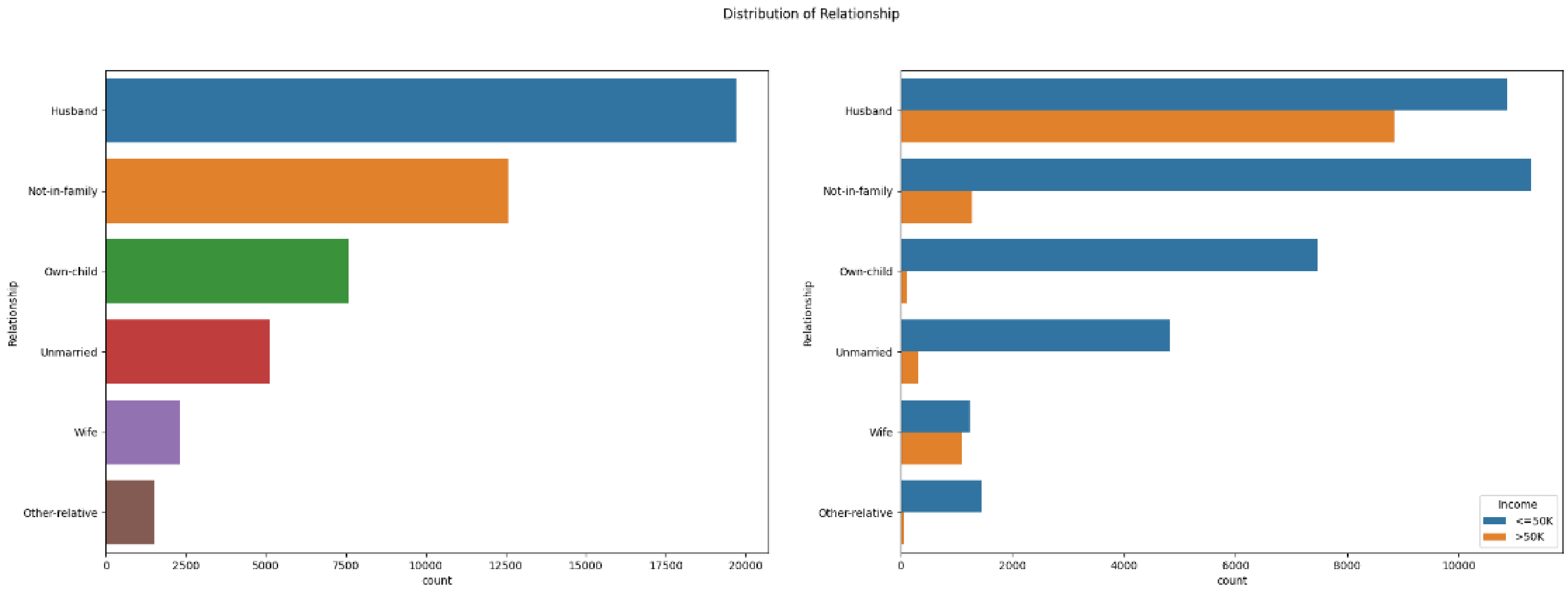
The distribution of >50k in relation to <=50k seems to be highest in Self-Emp-Inc and Federal-Gov.



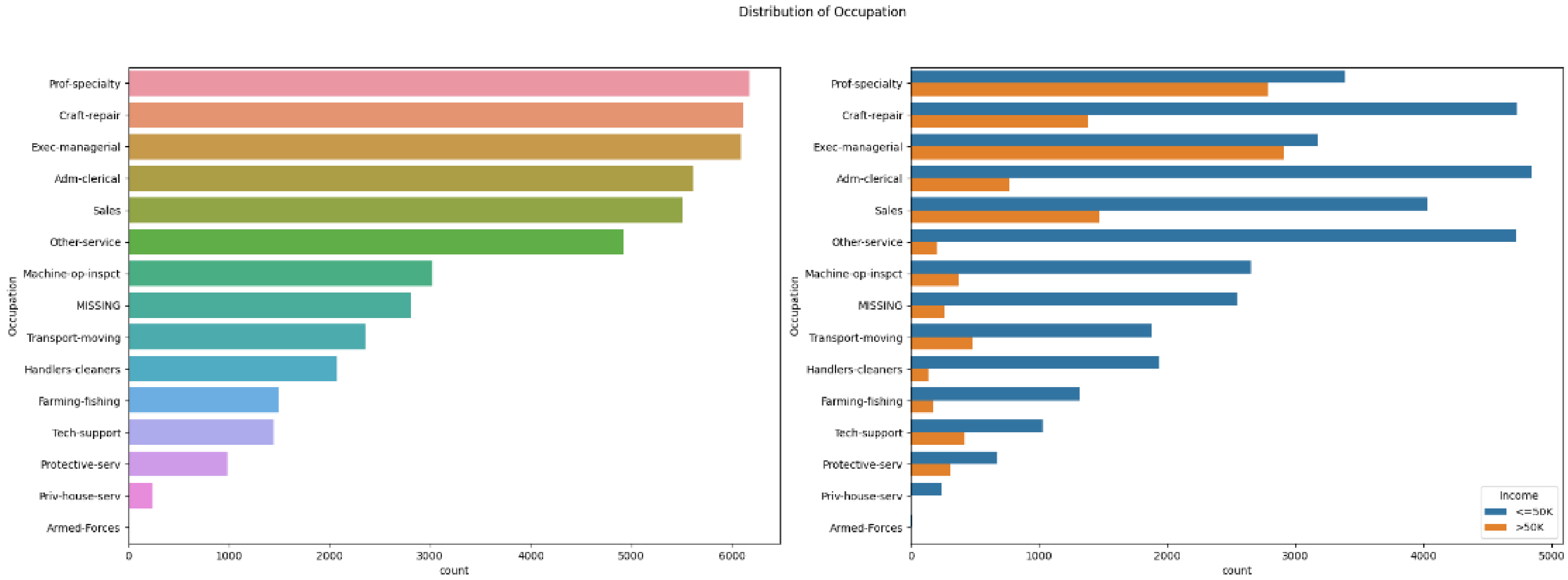
Higher Education results in higher income. Before an education level of 9, it is rare to see >50k income, and after an education level of 14 more than half of adults reaches >50k.



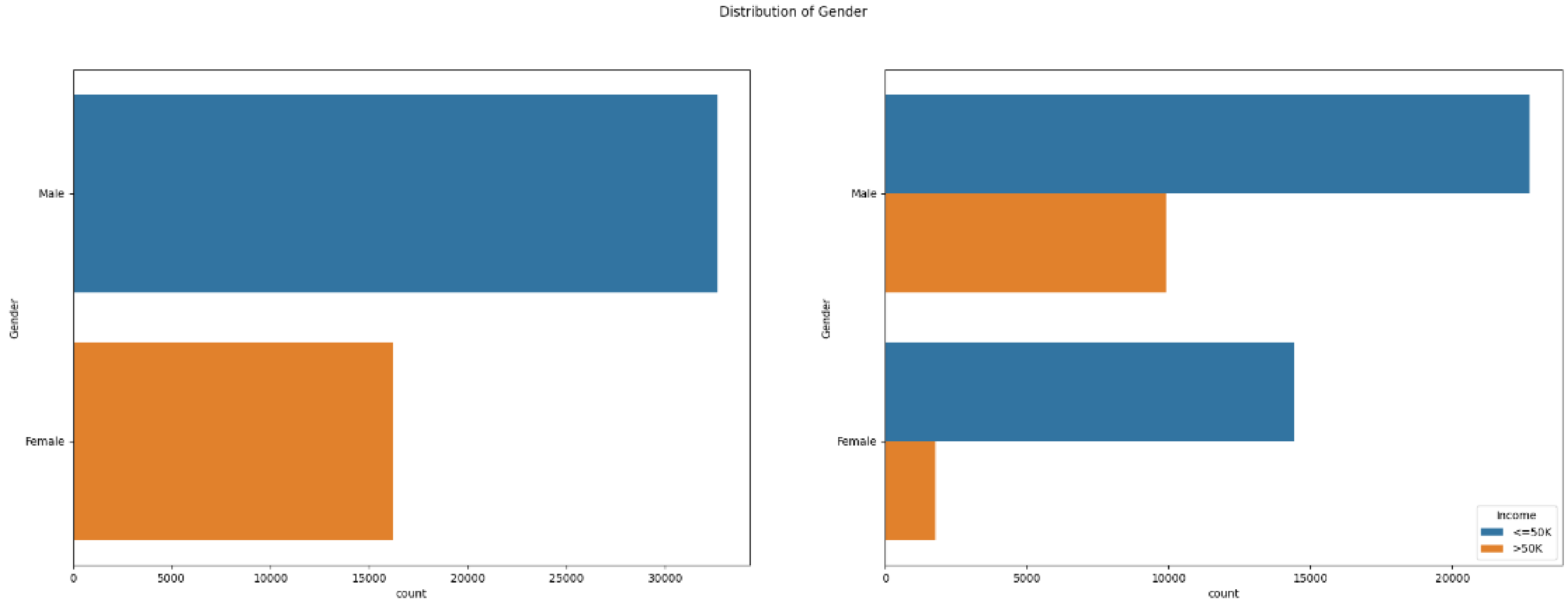
The distribution of >50k in relation to <=50k seems to be highest for Adults with a Marital Status of Married-Civ-Spouse.



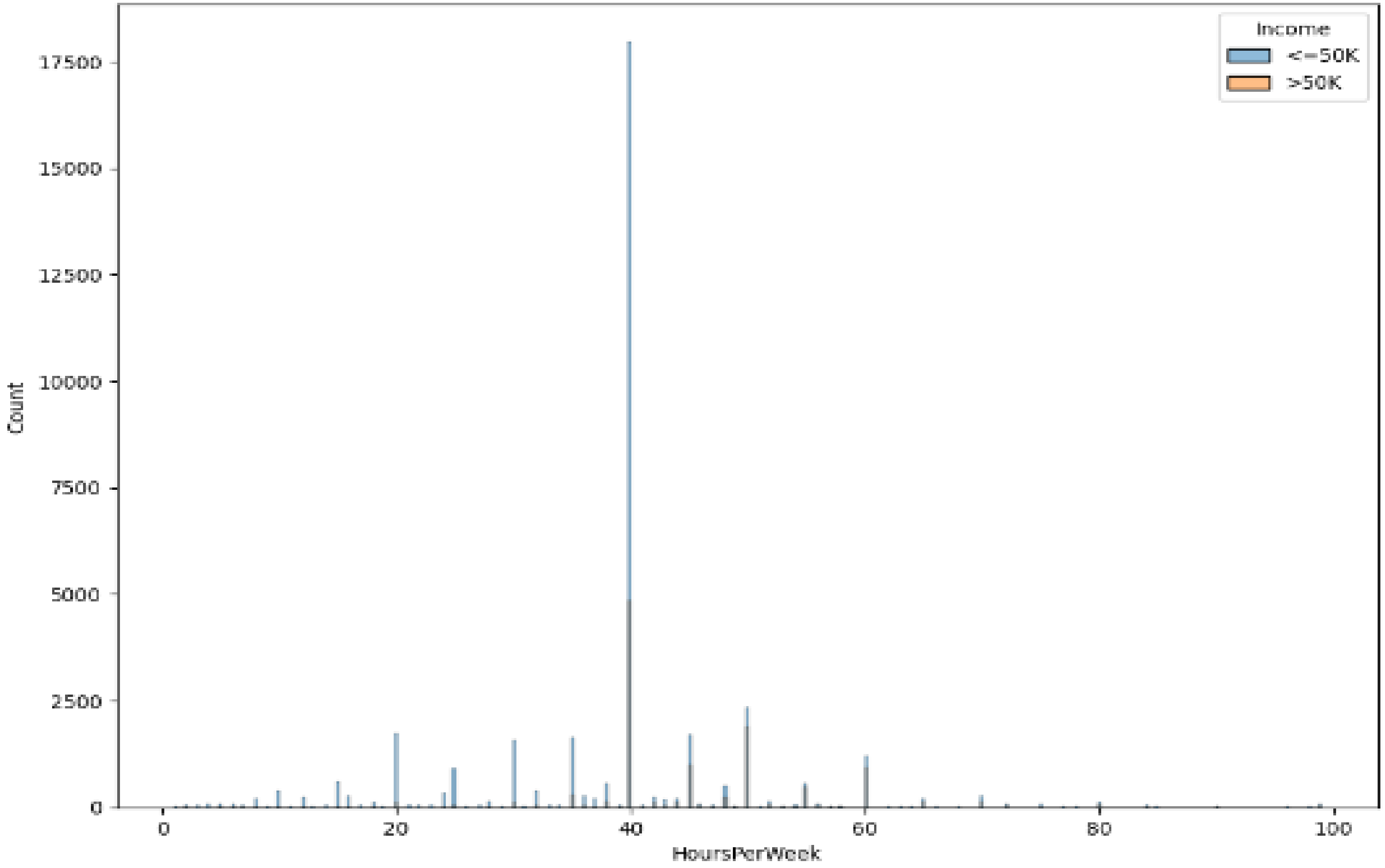
The distribution of >50k in relation to <=50k seems to be highest for Husband and Wife. Correlates to the Marital Status distribution (Married-Civ-Spouse having almost 50/50 spread)



The distribution of >50k in relation to <=50k seems to be highest for Exec-Managerial and Prof-Specialty, as well as a relatively good amount in Craft-Repair and Sales.



The distribution of >50k in relation to <=50k seems to be highest for Male



The distribution of >50k in relation to <=50k seems to increase as HoursPerWeek increases. There's a sizable jump of >50k adults when you reach over 40 hours.

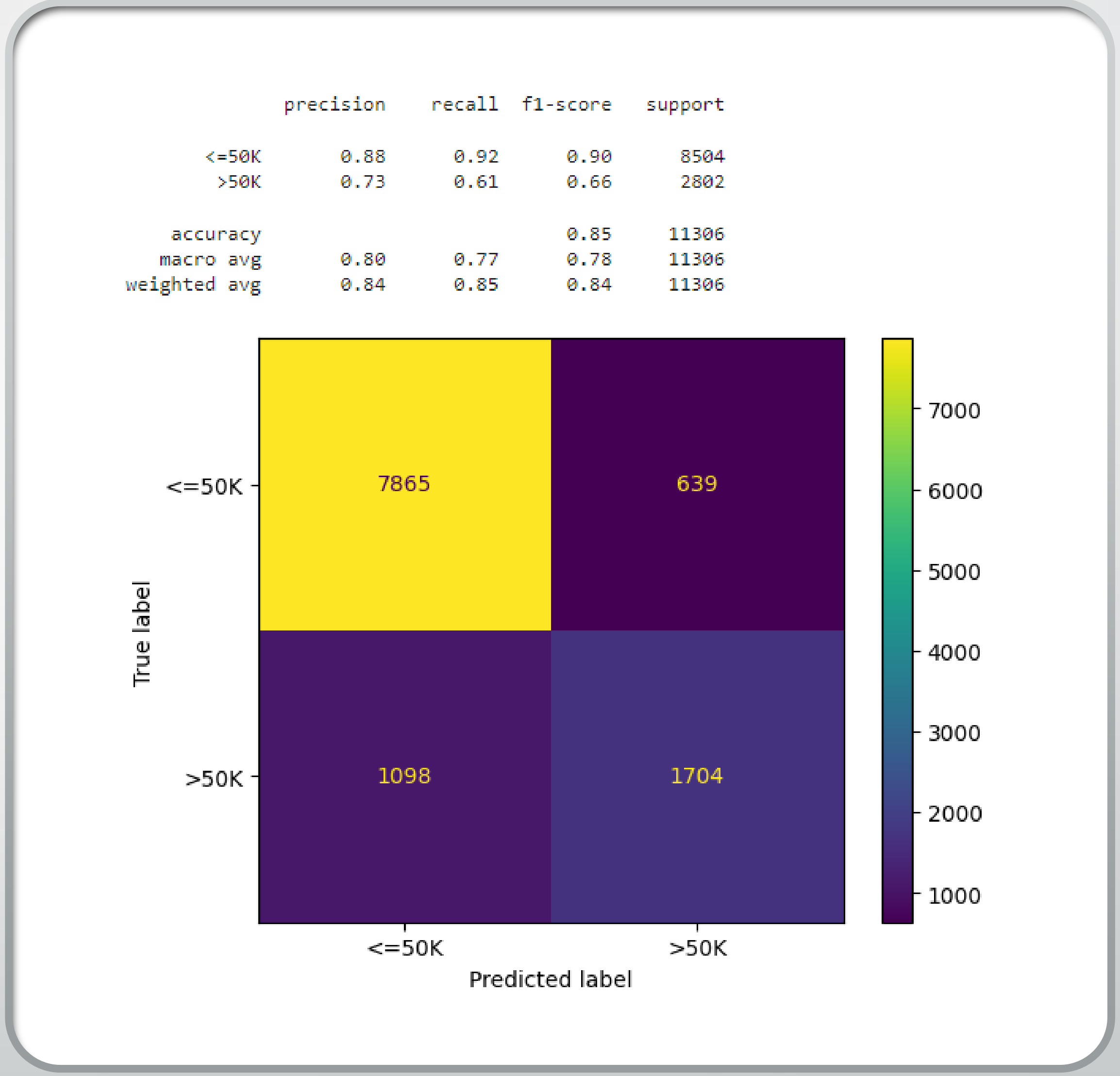
Model Performance

Recall was deemed as the important metric, meaning the number of positives that were correctly identified.

- <= 50k Accuracy: 92%
- > 50k Accuracy: 61%
- Overall Accuracy: 85%

Model seems to be pretty good at identifying <=50k but struggles a bit with >50k classifications.

It is recommended to not only rely on this model to predict adults earning >50k p/a due to the relatively low accuracy.



```
              precision    recall  f1-score   support

       <=50K       0.88      0.92      0.90      8504
        >50K       0.73      0.61      0.66      2802

    accuracy                           0.85     11306
   macro avg       0.80      0.77      0.78     11306
weighted avg       0.84      0.85      0.84     11306
```

Questions