Adult Income

By: Rosa R

Adult Income Project Description

The Adult dataset is from the Census Bureau and the task is to predict whether a given adult makes more than $50,000 a year based attributes such as education, hours of work per week, etc..

Income Dataset

The dataset was sourced from Kaggle-Adult Income

The dataset is credited to Ronny Kohavi and Barry Becker and was drawn from the 1994 United States Census Bureau data and involves using personal details such as education level to predict whether an individual will earn more or less than $50,000 per year.

This dataset utilizes 14 features for predicting income.

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K

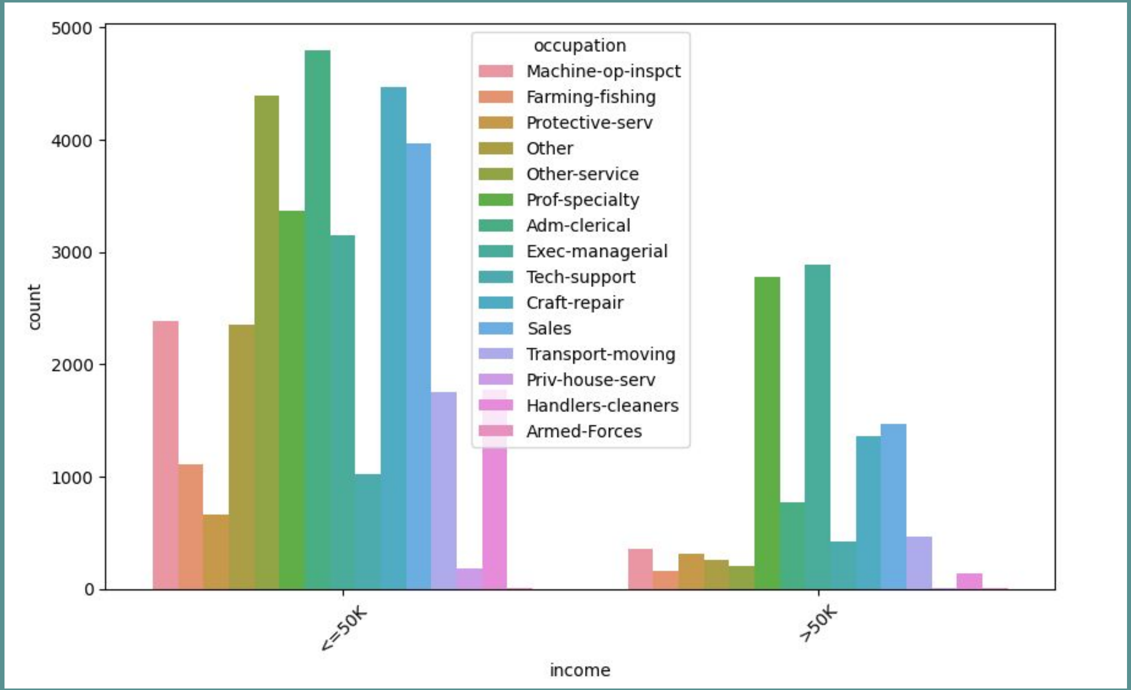
Stakeholders

This analysis will be utilized by the US Department of Labor and Education in order to assist with predicting income to push for higher education and better career opportunities.

This will help decrease poverty and low income communities.

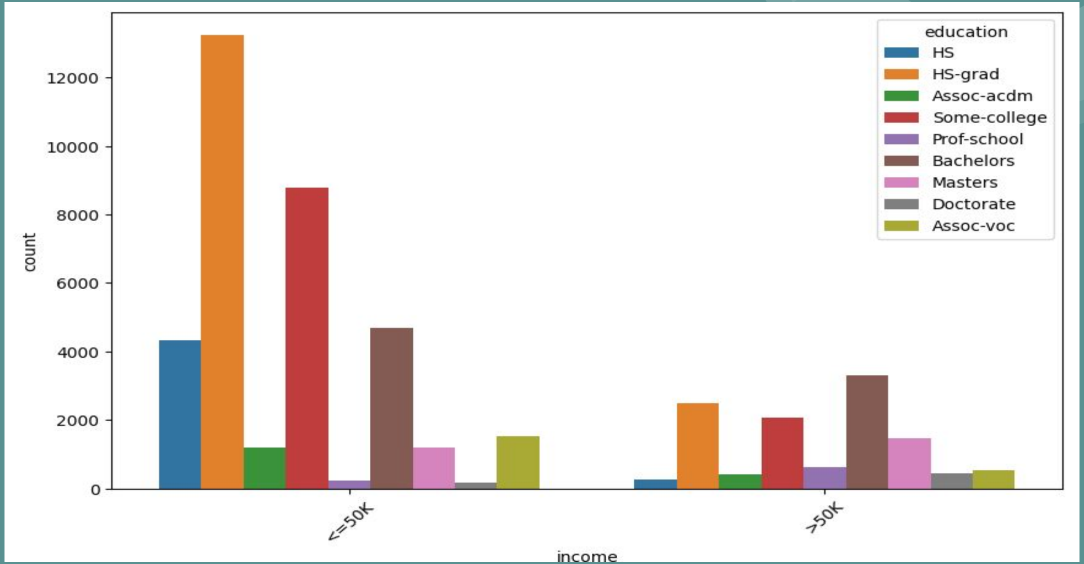
Key Findings:

This graph shows that a person can have the same occupation but be on different pay scales.



Key Findings

Here you can see a clear correlation between income and eduation. Very few people with a high school education (not grad) can still make over 50k.



Models Evaluated & Results

The Final Model Chosen was a Logistic Regression PCA Tuned Model. As you can see here this model performed better in F1 Score and ROC.

	Model	Accuracy Score	Precision Score	Recall Score	F1 Score	ROC	Best Parameters	Execution Time
0	Random Forest	0.815565	0.647258	0.584068	0.614041	0.738645	N/A	16.70
1	Logistic Regression	0.794108	0.562208	0.814915	0.665375	0.801021	N/A	5.08
2	KNeighbors	0.689629	0.438200	0.835254	0.574828	0.738016	N/A	13.83
3	Logistic Regression Tuned	0.795300	0.563964	0.815932	0.666944	0.802155	{'logisticregression__C': 0.5, 'logisticregres...	633.42
4	Logistic Regression PCA Tuned	0.793767	0.560998	0.823051	0.667216	0.803497	{'logisticregression__C': 0.5, 'logisticregres...	531.10
5	Random Forest PCA Tuned	0.790446	0.556794	0.812542	0.660786	0.797788	{'randomforestclassifier__max_depth': 10, 'ran...	3511.21
6	KN Tuned	0.743784	0.493762	0.791525	0.608152	0.759647	{'kneighborsclassifier__n_neighbors': 10}	121.06

Recommendations to the Stakeholders

- Given the results of the models after running with PCA I think Logistic Regression PCA Tuned performed slightly better than Random Forest therefore I would use this as my prediction model and suggest to my stakeholder. KNeighbors actually scored lower before using PCA but still lower than both other models.
- I think more information regarding the school district is needed to get a better understanding of where and why the lack of education stands.


Thank You