INCOME PREDICTIONS

Presented By: Maybelline Monge

AGENDA

- Data Description
- Project Objective
- Exploratory Data Analysis
- Machine Learning Model
- Recommendation

DATA DESCRIPTION

- Classification Dataset
- The target is adult income levels based on various demographic attributes
- Each row in the dataset provides statistical record pertaining to income level
- Each column represents a demographic feature of an individual

PROJECT OBJECTIVE

- EXPLORE DATA
- CREATE & RECOMMEND A MACHINE LEARNING MODEL THAT PREDICTS WHETHER AN ADULTS SALARY CROSSES A THRESHOLD OF $50K OR HIGHER
- CLASSIFYING YES OR NO FOR OUR TARGET “ADULT INCOME”

EXPLORATORY DATA ANALYSIS

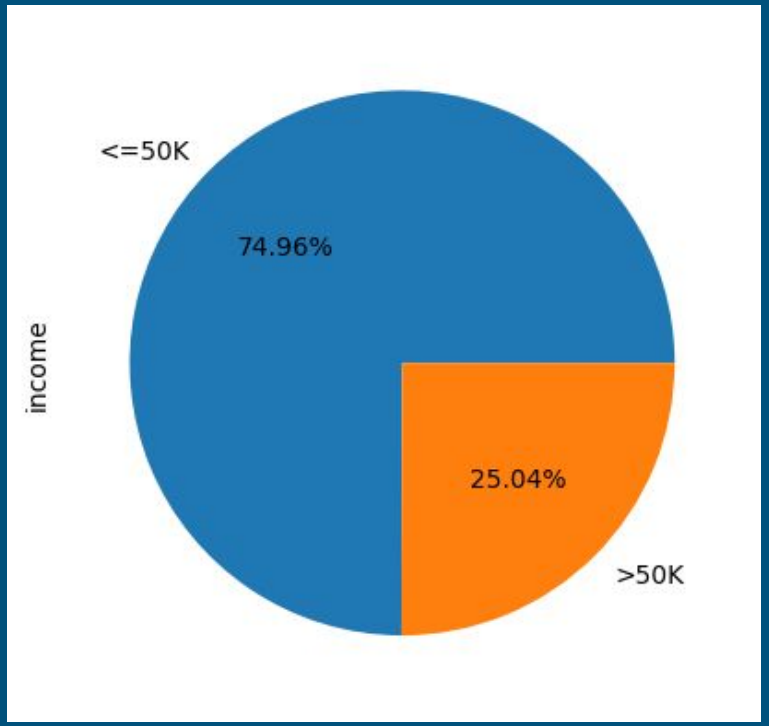
What are some key features correlated to income level?

age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K
38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K
44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K
18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K
34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	Male	0	0	30	United-States	<=50K
29	?	227026	HS-grad	9	Never-married	?	Unmarried	Black	Male	0	0	40	United-States	<=50K
63	Self-emp-not-inc	104626	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	3103	0	32	United-States	>50K

age	workclass	education	marital_status	occupation	race	gender	hours-per-week	continent	income
25	Private	11th	Never-married	Machine-op-inspct	Black	Male	40	N_America	<=50K
38	Private	HS-grad	Married-civ-spouse	Farming-fishing	White	Male	50	N_America	<=50K
28	Government	Assoc-acdm	Married-civ-spouse	Protective-serv	White	Male	40	N_America	>50K
44	Private	Some-college	Married-civ-spouse	Machine-op-inspct	Black	Male	40	N_America	>50K
34	Private	10th	Never-married	Other-service	White	Male	30	N_America	<=50K

Completed some feature engineering to refine features to be relative, concise and non-redundant.

Example: Feature “native-country” listed every country. Instead I’d grouped them into continents which are more feasible when exploring data.



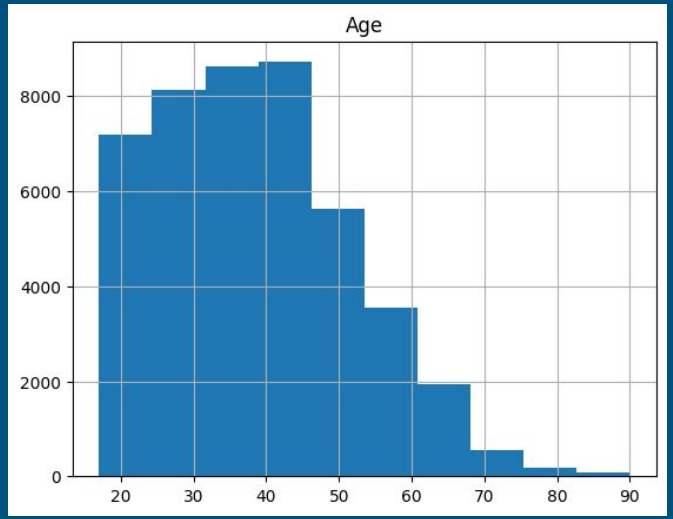
- **ABOUT 75% OF ADULTS IN THE DATASET MAKE A SALARY EQUAL OR LESS THAN $50K**
- **REMAINING 25% MAKE A SALARY GREATER THAN $50K**

HIGHLY IMBALANCED

Hmm... What are some attributes that play a role in classifying the two income classes?

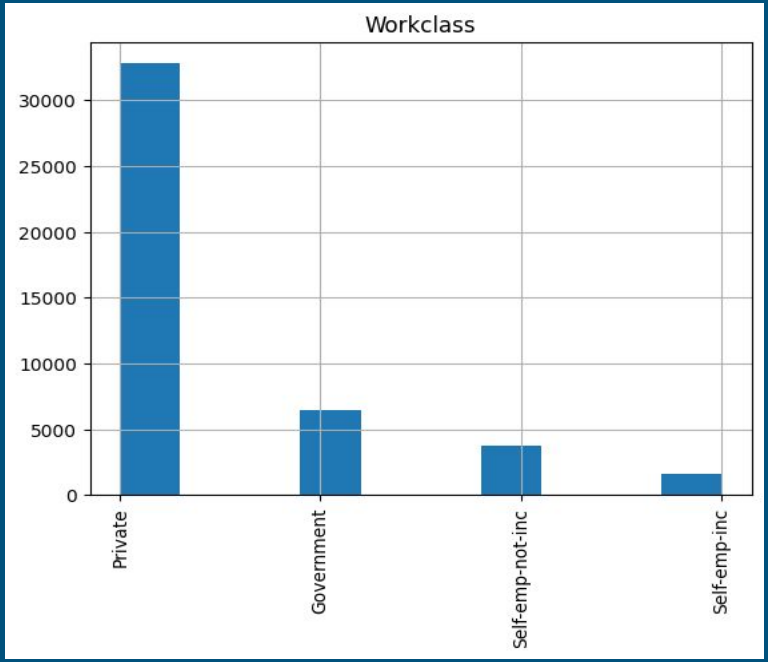
- THE “AGE” FEATURE APPEARS TO HAVE OUTLIERS FROM 47 YEARS OF AGE AND BEYOND 75.

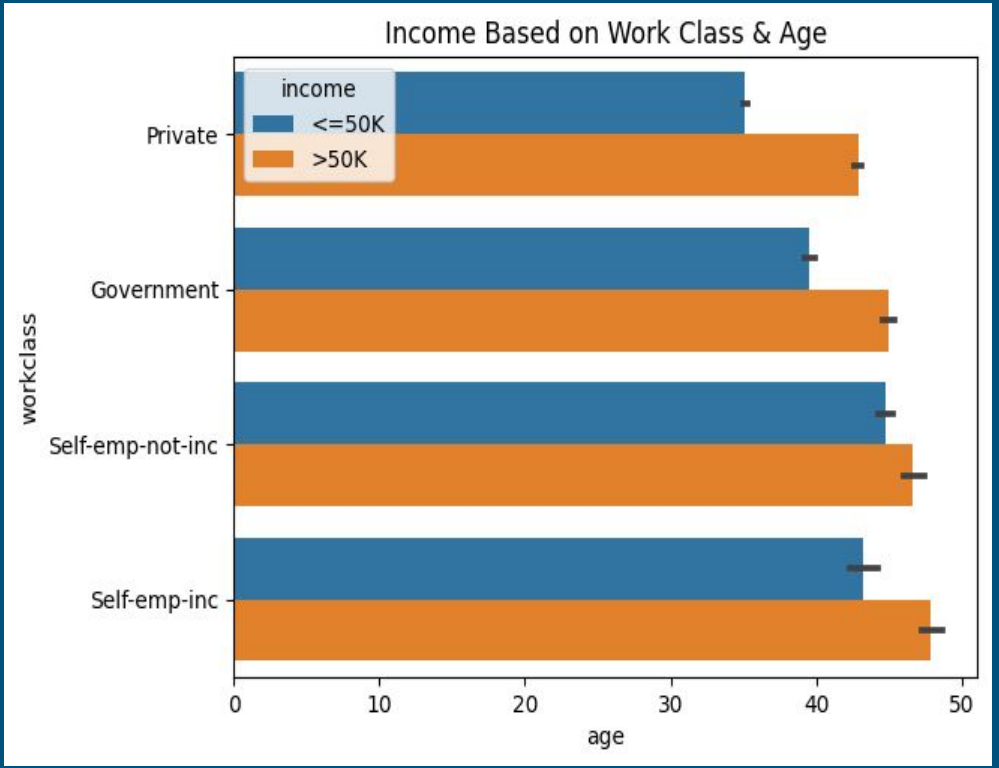
- AGE 37 IS THE MEDIAN IN THE WHOLE DATASET



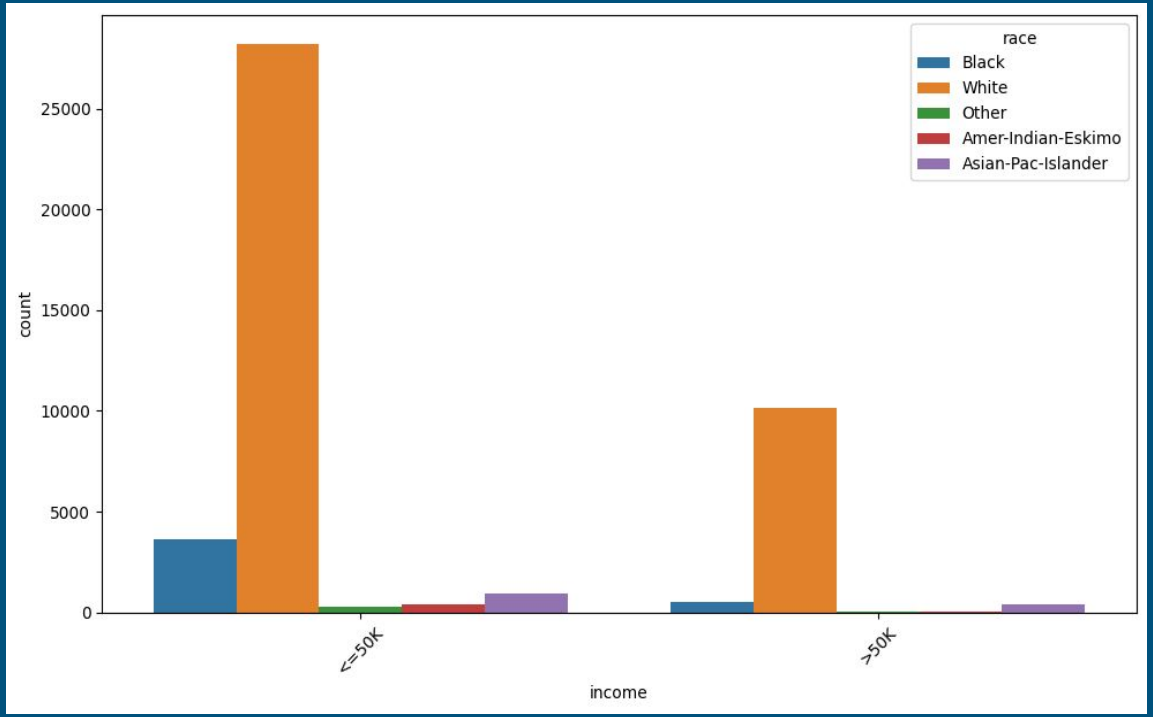
Thanks to feature engineering were able to compare four categories for the "workclass" feature instead of many.

Chart below displays the majority of statistical data obtained in the dataset is primarily from the private sector.





Based on the barplot above, it displays the majority of adults between the ages of 40-50 as self employed making a salary over 50K. Whereas, majority of young adults in the private sector make a salary less than 50k.



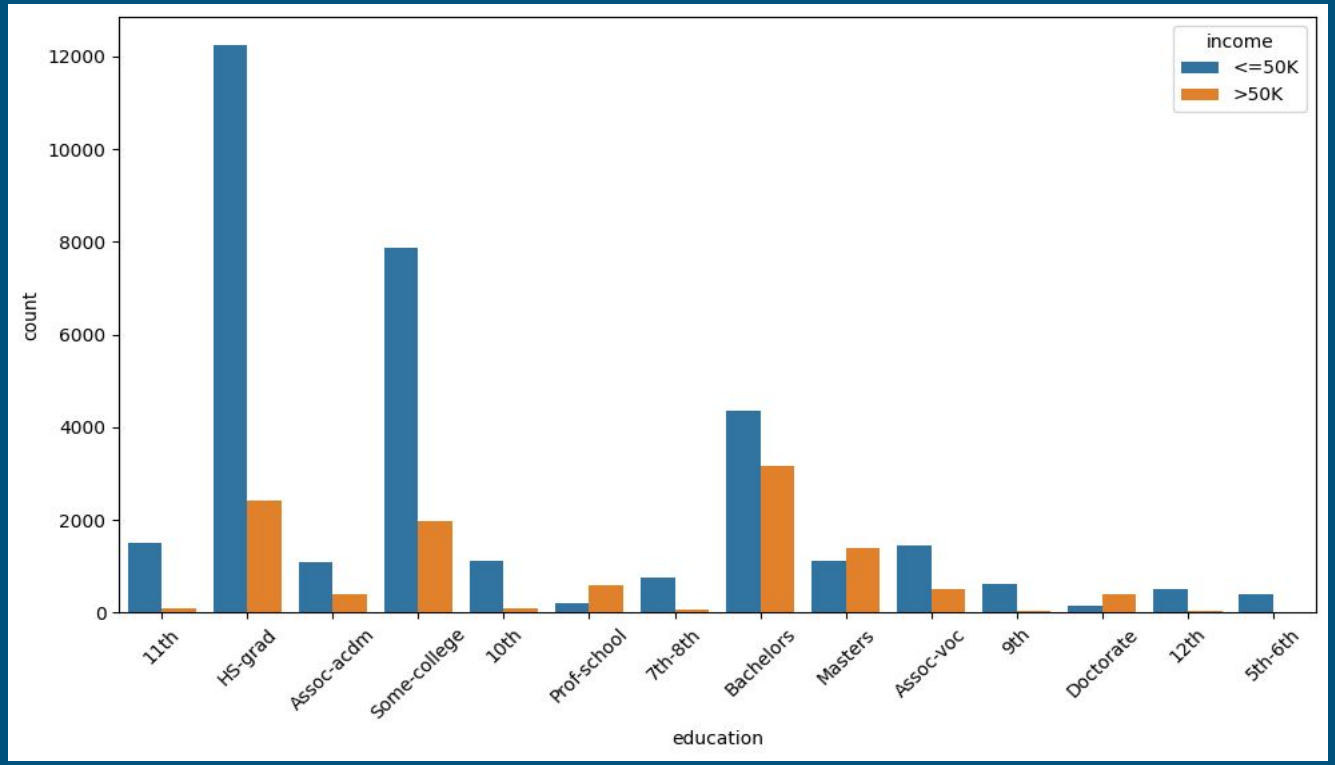
GREAT DISPARITY BETWEEN INCOME & RACE DISTINCTION

NOTABLE INFORMATION

Which race groups can benefit greatly from financial literacy?

CORRELATION OF INCOME RELATIVE TO EDUCATION

YIELDS SOME INTERESTING TRENDS



Based on the data, surprisingly enough individuals who graduated high school came in as second highest in making a salary over $50k. First being individuals with a Bachelor's degree.

What are some influences that minimizes income levels with higher education?

MACHINE LEARNING MODELS

	Train Accuracy	Train Recall	Train Precision	Train F1-Score	Test Accuracy	Test Recall	Test Precision	Test F1-Score
KNN Tuned Model	0.850558	0.661957	0.721496	0.690445	0.818223	0.611354	0.636605	0.623724

	Train Accuracy	Train Recall	Train Precision	Train F1-Score	Test Accuracy	Test Recall	Test Precision	Test F1-Score
Random Forest Tuned Model	0.83277	0.5	0.752771	0.600885	0.827459	0.495997	0.716614	0.586237

	Train Accuracy	Train Recall	Train Precision	Train F1-Score	Test Accuracy	Test Recall	Test Precision	Test F1-Score
RF PCA Model	0.959403	0.906792	0.930207	0.91835	0.805668	0.555677	0.617469	0.584945

	Train Accuracy	Train Recall	Train Precision	Train F1-Score	Test Accuracy	Test Recall	Test Precision	Test F1-Score
LREG PCA Model	0.822606	0.539064	0.688714	0.604769	0.823693	0.549491	0.67471	0.605696

	Train Accuracy	Train Recall	Train Precision	Train F1-Score	Test Accuracy	Test Recall	Test Precision	Test F1-Score
LREG Tuned Model	0.826702	0.552007	0.696688	0.615966	0.831316	0.568049	0.692239	0.624026

RECOMMENDATION

- **The Tuned Logistic Regression Model is the best production model for the business problem at hand.**
- **The model accuracy score being the highest at 83% implies it's the highest predictive model with the most number of samples that display out to be in the positive class.**
- **As this is a classification task of whether the income in the sample crosses a threshold of $50k or higher, yes or no, this metric yields the most value with fewer false positives.**

In other words, the Tuned Logistic Regression Model is able to cypher through the dataset and has the highest probability rate of predicting whether or not an individual's income crosses a threshold of $50k or higher.