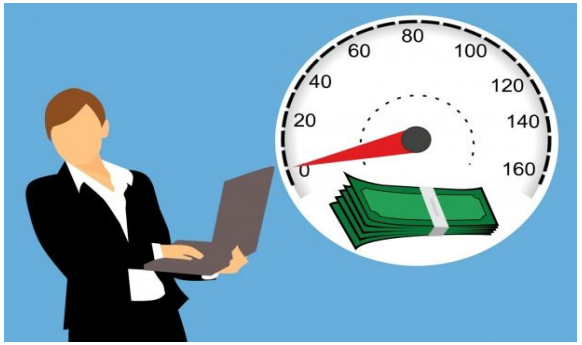
Adult Income Prediction

Anjali Prakash

Project Description



The purpose of this project is to use various features (such as age, race, hours worked etc) to predict the income of an individual.

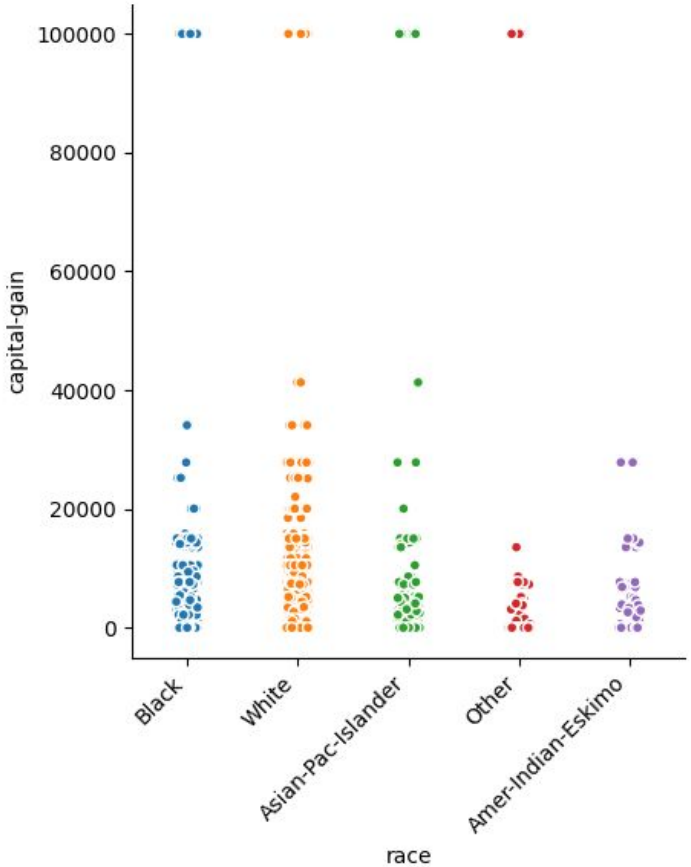
This data set was sourced from Kaggle - Adult Income Data Set. This dataset has 16 columns, utilizing 15 features to predict income of an individual.

Stakeholders

The potential stakeholders for this project is extensive. Ranging from companies trying to market their product to individuals based on income or government programs that are trying to have an understanding of individuals’ incomes based on various demographics.

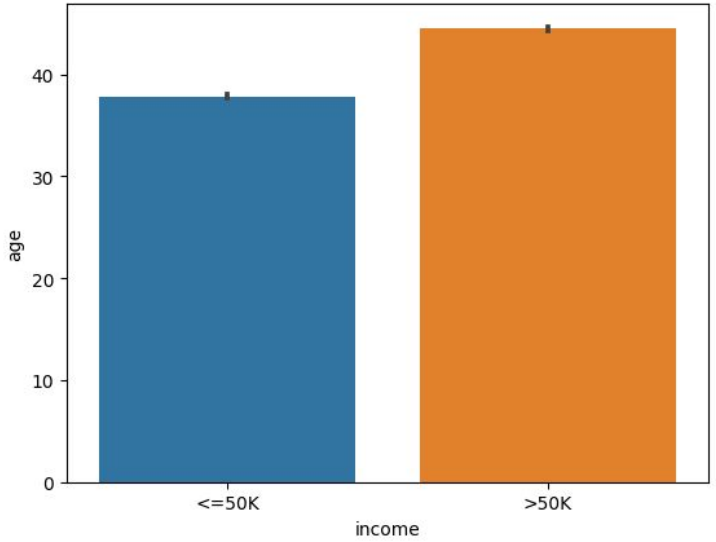
Key Findings

It appears that race has an interesting correlation to capital gain. White individuals appear to have the highest capital gain and individuals labelled as “other” have the least. This could be due to many factors including opportunity, generational wealth etc.

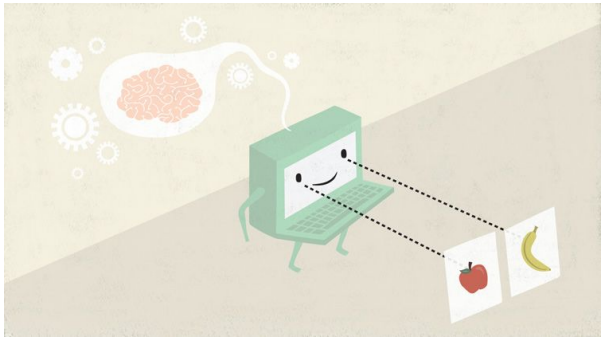


Key Findings

Looking at the following visualization, individuals making >50k appear to be older than individuals making <=50k on average. This can be due to older individuals having more work experience and therefore more income, or they may have more opportunities for job growth or promotions.



Model Evaluation



- After experimenting with several machine learning models to predict an individual's income, the Random Forest classification model was selected due to best performance
- Our false-positive rate, or rate at which our model incorrectly identifies an individual as making <=50k was 5.2%
- Our false-negative rate, or rate at which our model incorrectly identifies an individual as making >50k was 40%
- The high false-negative rate may be due to uneven class balance of 75.6% of individuals making <=50k and 24.4% making >50k.
- To improve/lower our false-negative rate we can resample our training data for our model to even out the class balance

Recommendations

- I would recommend that stakeholders use the Random Forest Classification model as it appears to have the highest recall (rate of correctly identifying true-positives) at 95% and accuracy (rate of correct predictions) at 83% amongst all the classification models used.
- False negative are still not at the ideal accuracy rate, so ideally they go through a manual review by stakeholders.