Factors that influence making money

By Mushtaq Ahmed

Introduction

This project presents an in depth analysis of the common factors that influence an individual making more money and recommends a predictive model combining a set of factors that influence making more money.

Issues

- Income inequality
    1. Woman organization
    2. Racial groups
    3. Political parties
- Career Choice
    1. New Career
    2. Career Transition

Motivations: The New York Public Library is interested to include a machine learning predictive model to enrich its career advice service.

- Identifying Key Factors
- Developing Predictive Machine Learning Model
- Disseminating the Information to interested parties

Dataset information

Dataset: UCI Repository: https://archive.ics.uci.edu/ml/datasets/adult

File type: Labelled

Tools: Pandas in Python.

Total Instances: dataset has 32,560 instances

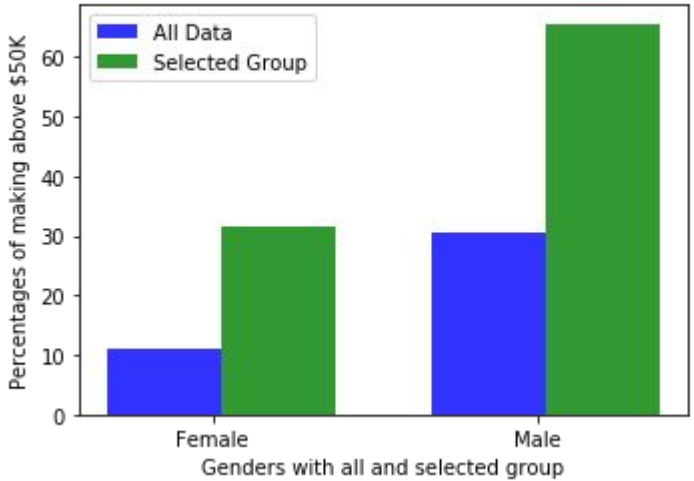
Total Features: 13 features – 5 continuous and 8 categorical

Target variable – with a binary categorical target variable - making more than $50K or less than $50K.

Exploratory Data Analysis

Do males make more money than their female counterparts? Yes, they do.

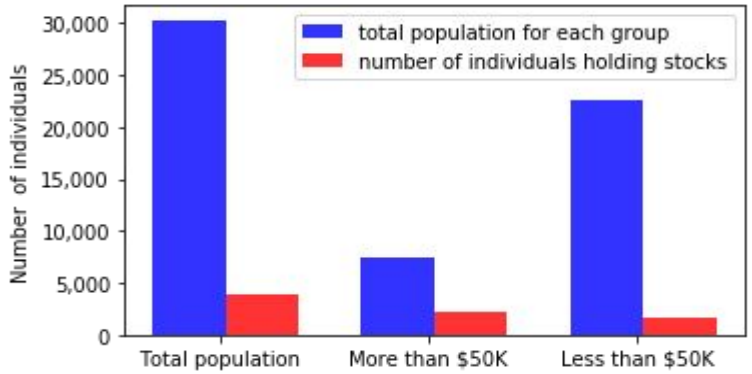
Table 2: Pecentage comparison of males and female making above $50K



Exploratory Data Analysis – continues

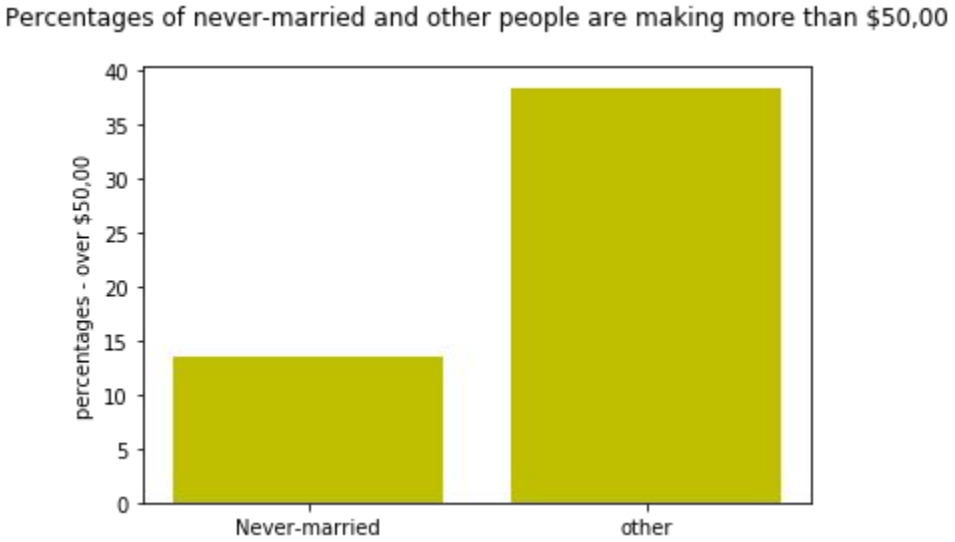
Do rich invests more on stocks ? Yes. They do

Table 4: Number of individuals own stocks for each group of population

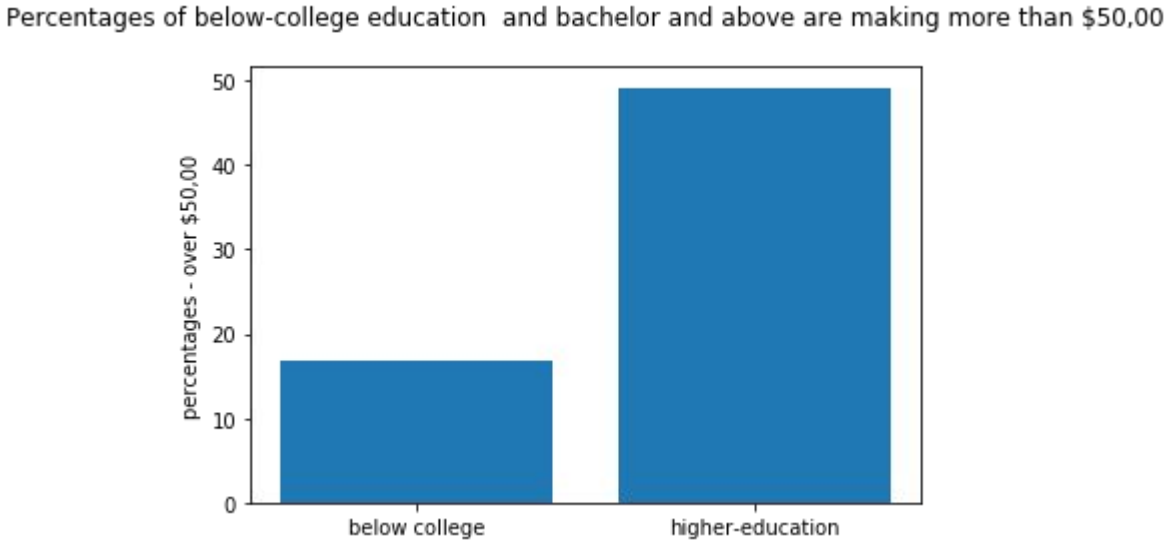


Blue indicates total population for each group and
red indicates the number of individuals holding stocks from that group

Does marriage help to become rich? Yes, it does.



Does education help to make more money? Yes. It does



Statistical Test of Hypothesis

1. Do males make more than their female counterparts? Yes.

   Performed Z test on proportions

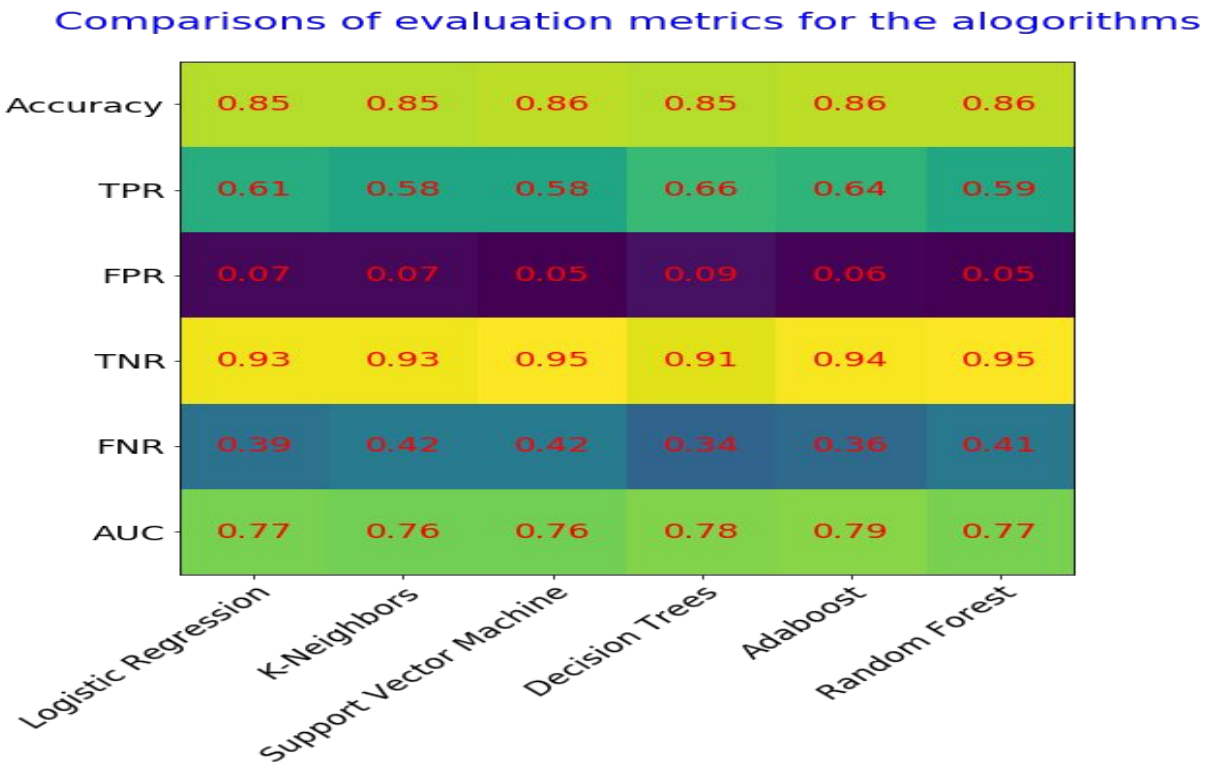
2. Does rich people invests more on stocks? Yes.

   Performed Chi-square test for independence

Machine Learning Algorithms Used

1. Logistic Regression
2. K-Neighbors
3. Support Vector Machine
4. Decision Trees
5. Adaboost
6. Random Forest

Algorithm Evaluations



Analysis

If we look at the previous slide:

- we can see that there is no clear winner.
- The mean of the accuracy rates is 0.855 with a standard deviation of 0.005
- FPR (false positive rate) is also fairly close for all algorithms.
- AUC (Area Under Curve) is also very close for all algorithms.

Interesting Observations

Accuracy Rates:

- Correctly classified a high proportion above 90%
- Correctly classified a low proportion around 60%

Possible Reasons:

- The dataset is more balanced towards negatives
- Making less than $50K have more stability in real life.

Ten most important feature for Decision Tree:

marital_stat_Married-civ-spouse
yrs_edu
capital_gain
age
capital_loss
hr_week

Recommendations: Decision Trees

- Performed very close to all other models in all metrics
- Offers an excellent interpretability
- Computational performance

i

E N D

THANK YOU