PREDICTING

INCOME

A Classification Project - By Kyle Turner

$50,000

THE CLIENT

- Non-profit
- Future Planning
- US Roll out

DATA SCIENCE SOLUTION

- Classification Model
- Use Census Data
- >50K / <=50K

THE DATA

INCOME CLASSIFICATION

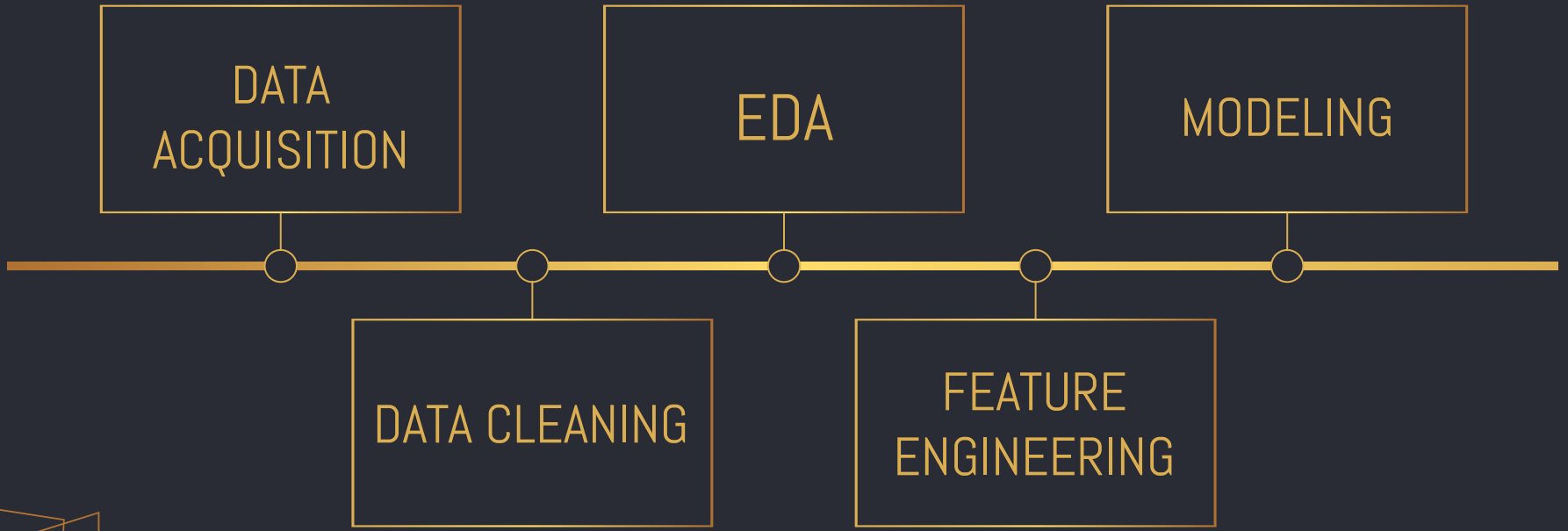
- Census data from Kaggle
- 32,561 rows/ 15 columns

TARGET

Income:

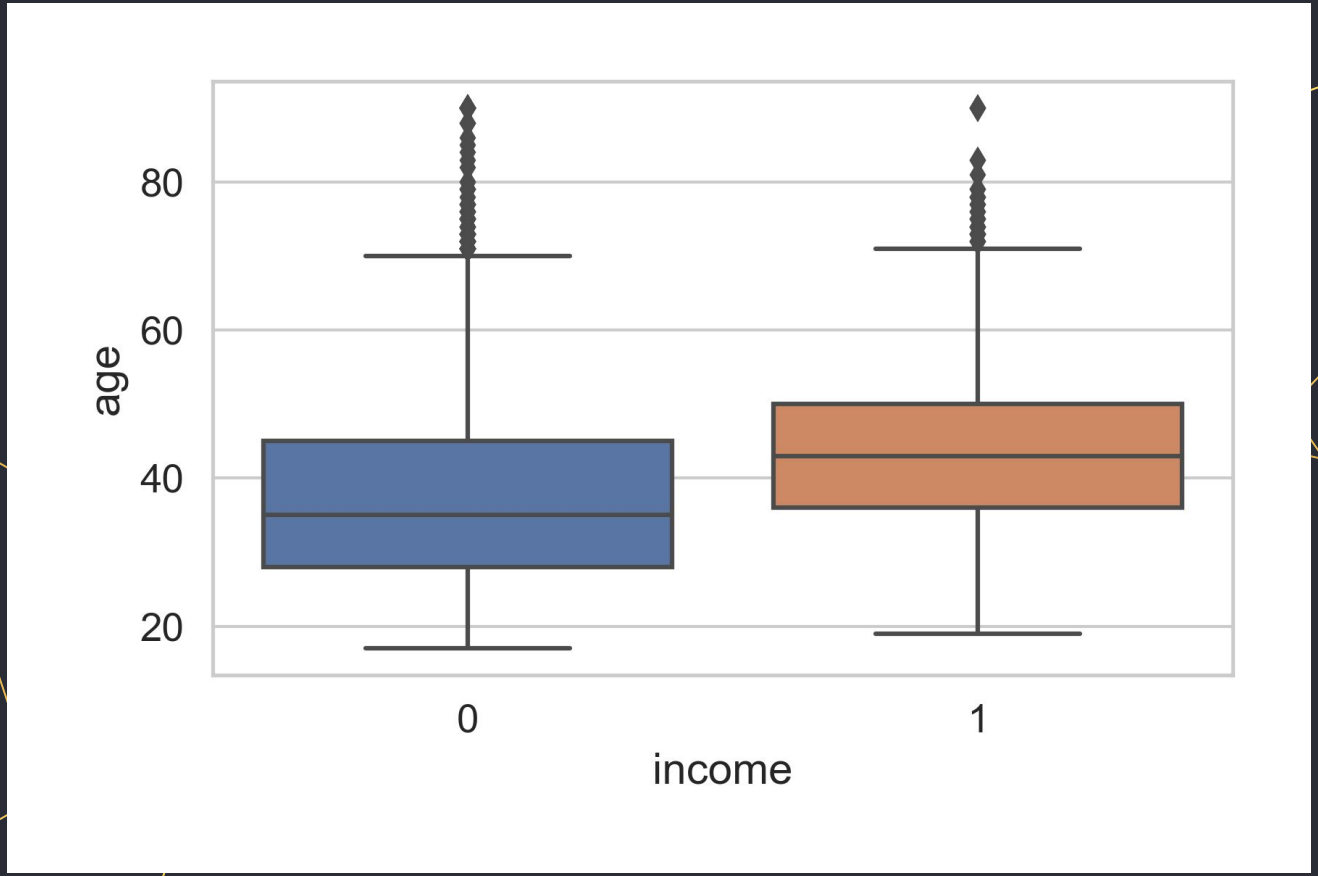
0= (<=50K)
1= (>50K)

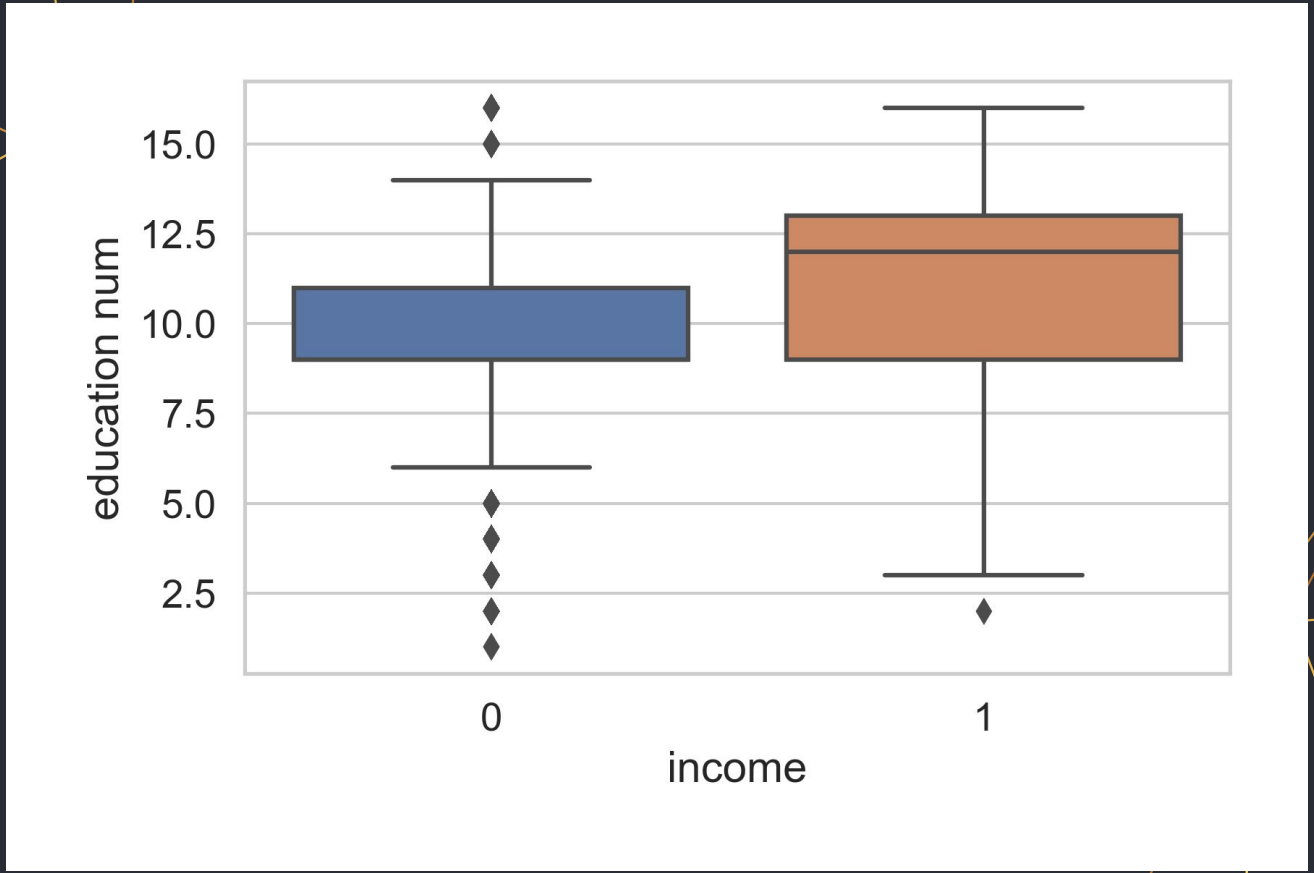
METHODOLOGY

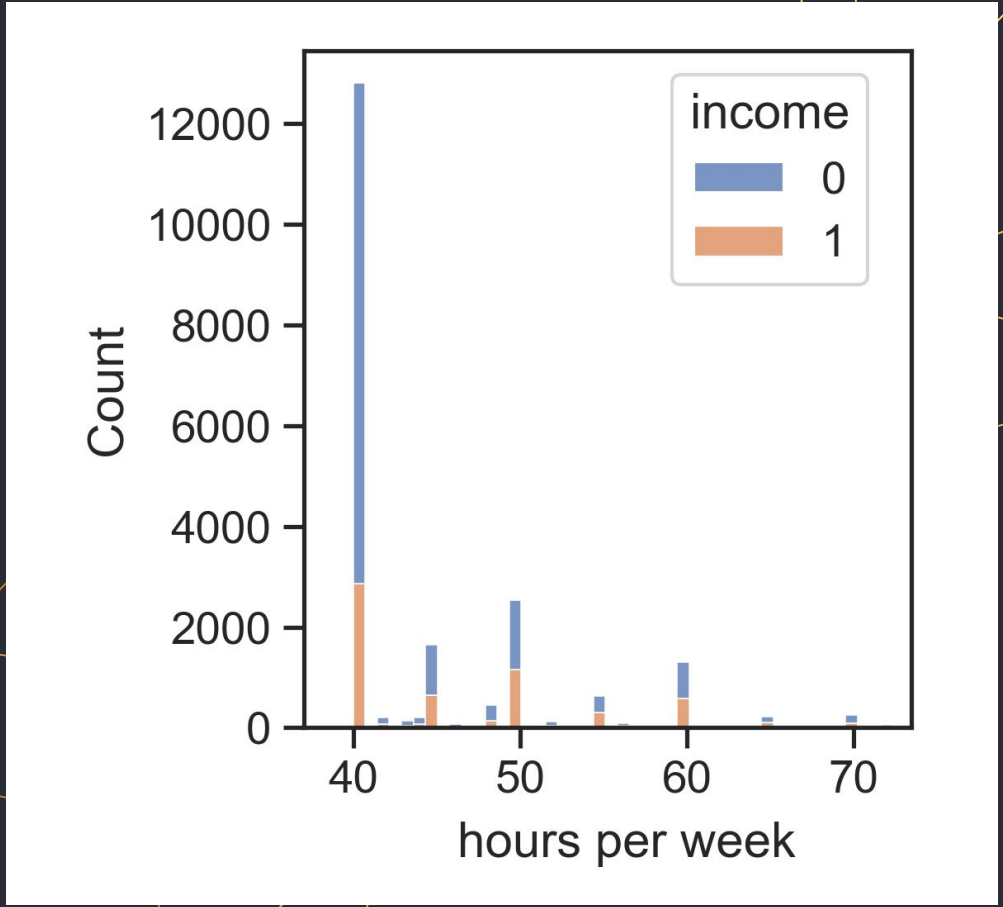


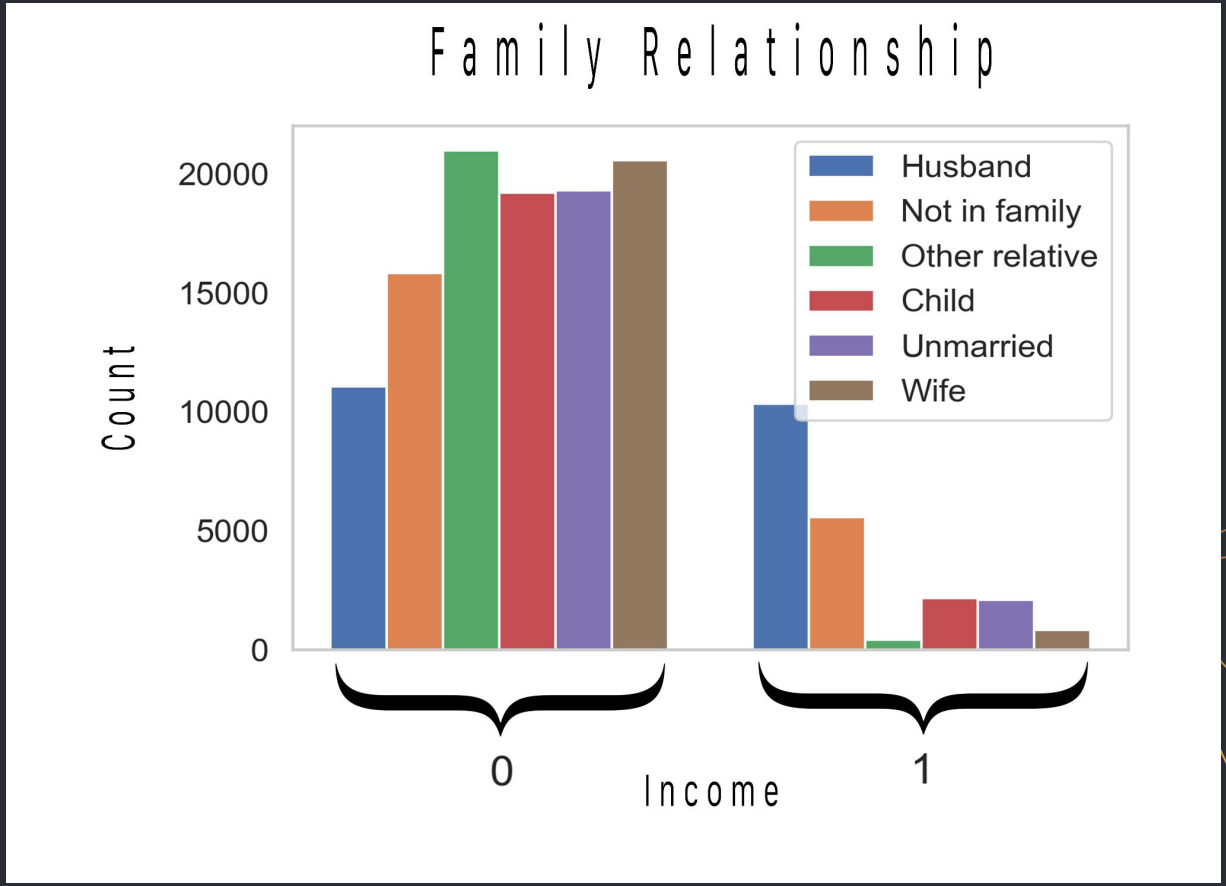
FEATURES

- Age
- Sex
- Race
- Education
- Occupation
- Marital Status
- Hours Worked

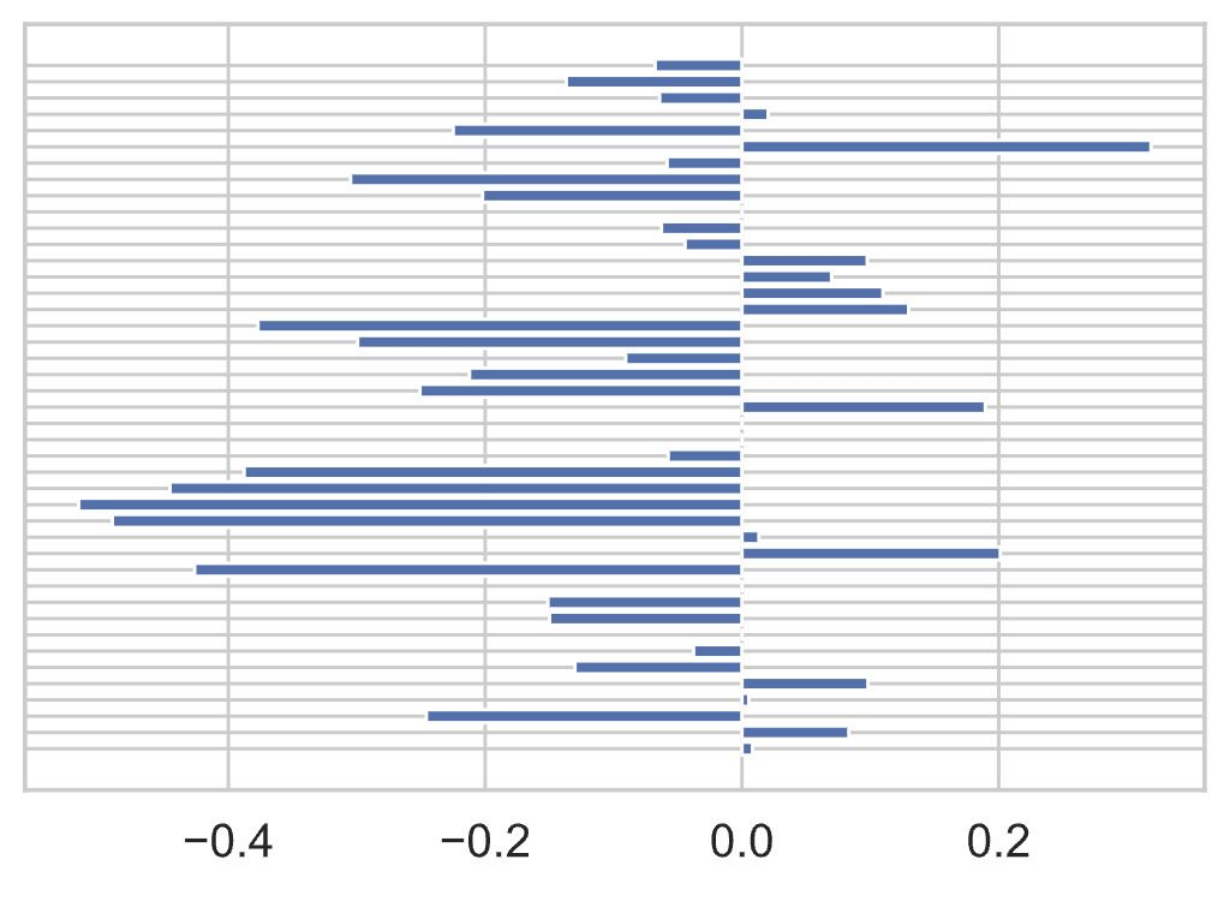

80
60
40
20
age
0
1
income


15.0
12.5
10.0
7.5
5.0
2.5
education num
0
1
income

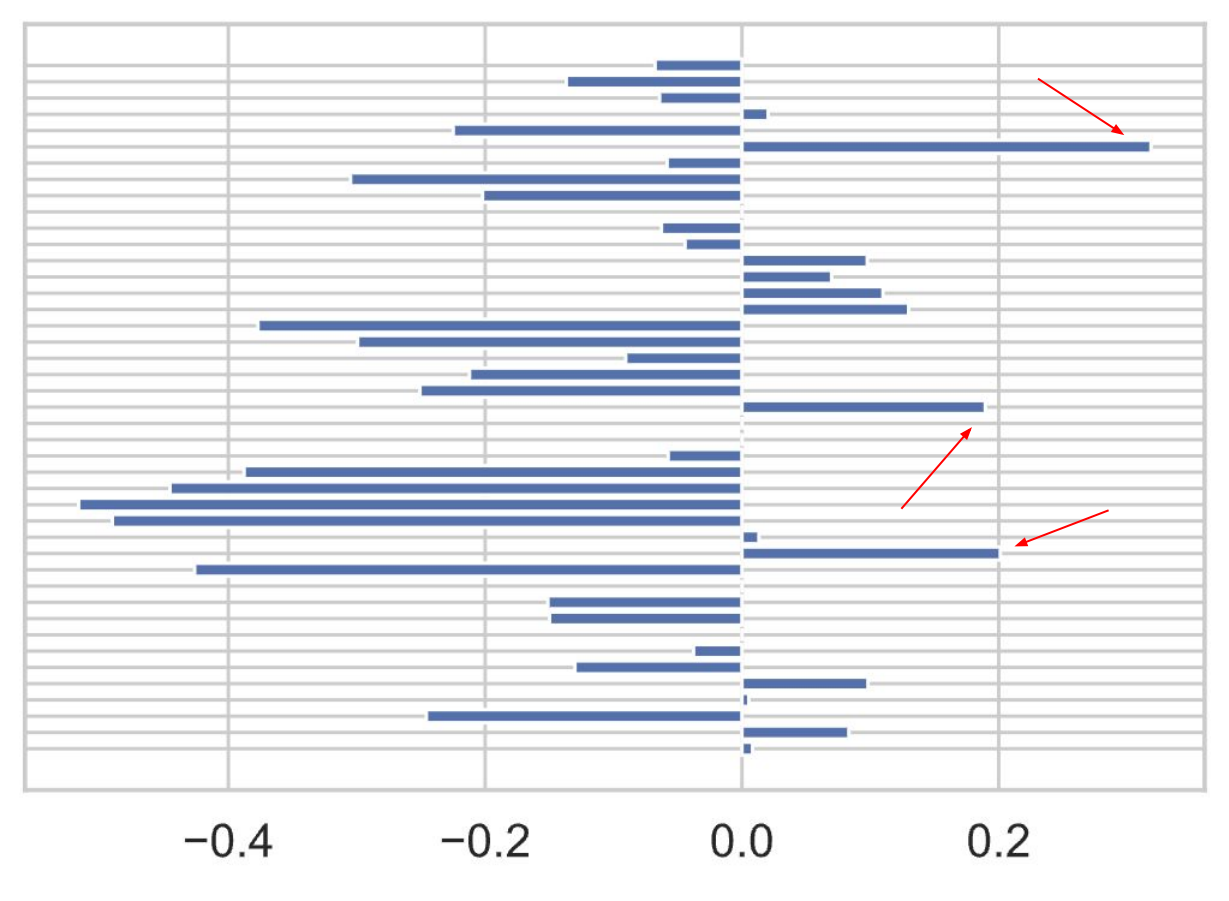

income
0
1
12000
10000
8000
6000
4000
2000
0
Count
40
50
60
70
hours per week


Family Relationship
Count
20000
15000
10000
5000
0
Husband
Not in family
Other relative
Child
Unmarried
Wife
0
1
Income

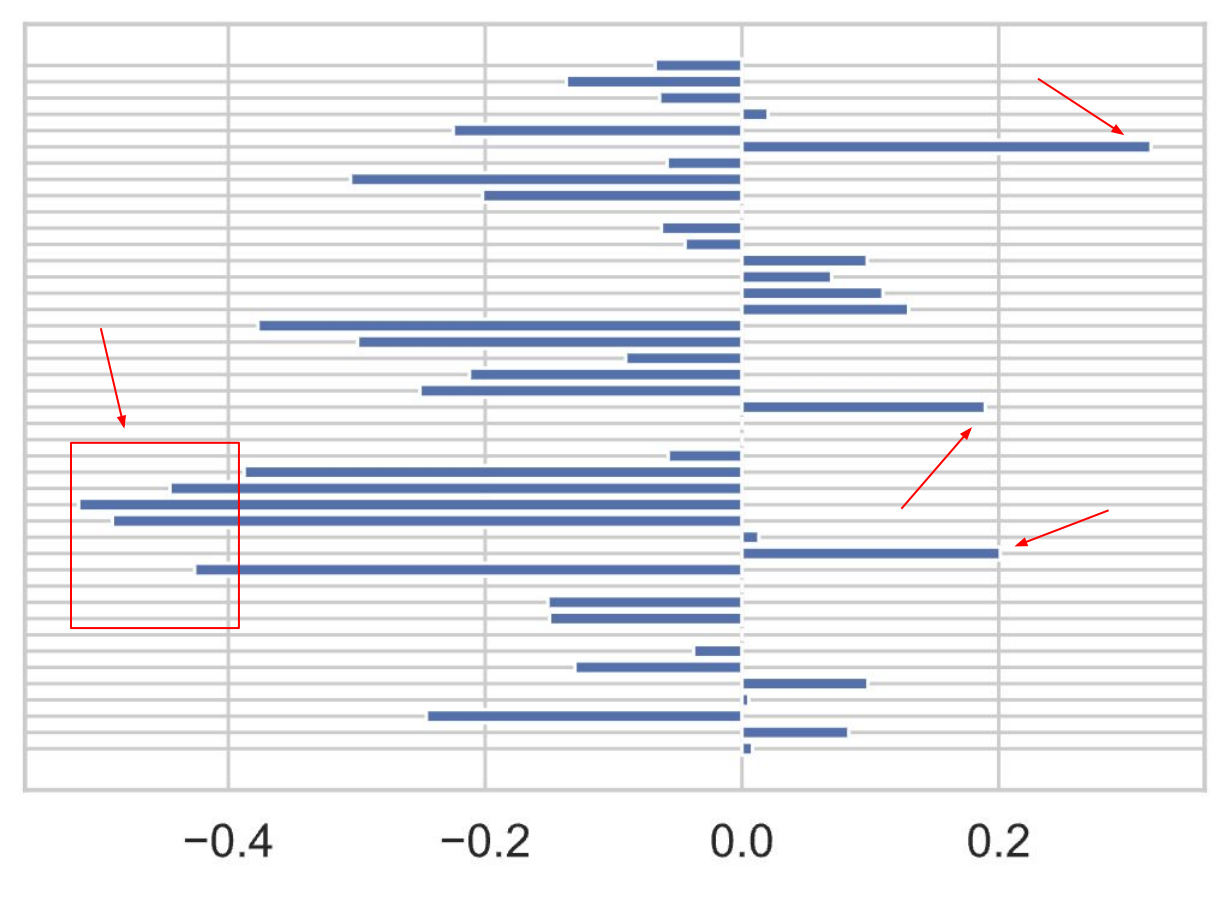
COEFFICIENTS- using divide by 4 rule



COEFFICIENTS- using divide by 4 rule



COEFFICIENTS- using divide by 4 rule



STANDOUT COEFFICIENTS

Relationship-Wife	0.322	Marital Status-Separated	-0.438
Occupation-Exec Manager	0.189	Marital Status-Never Married	-0.509
Workclass-Without Pay	0.206	Marital Status-Divorced	-0.419
		Marital Status-Married Spouse Absent	-0.483

MODELING

	LOGISTIC REGRESSION	LOGISTIC REGRESSION CV	RIDGE CLASSIFIER	LOGISTIC REGRESSION OVERSAMPLE	LOGISTIC REGRESSION CV OVERSAMPLE	RIDGE CLASSIFIER OVERSAMPLE
F1	0.641	0.638	0.632	0.671	0.638	0.632
ACCURACY	0.802	0.801	0.800	0.776	0.774	0.773

FINAL MODEL

LOGISTIC REGRESSION TEST SCORES:

Accuracy 0.796
F1 0.614

FUTURE WORK

- Add More Features
- Different Feature Engineering

THANK YOU