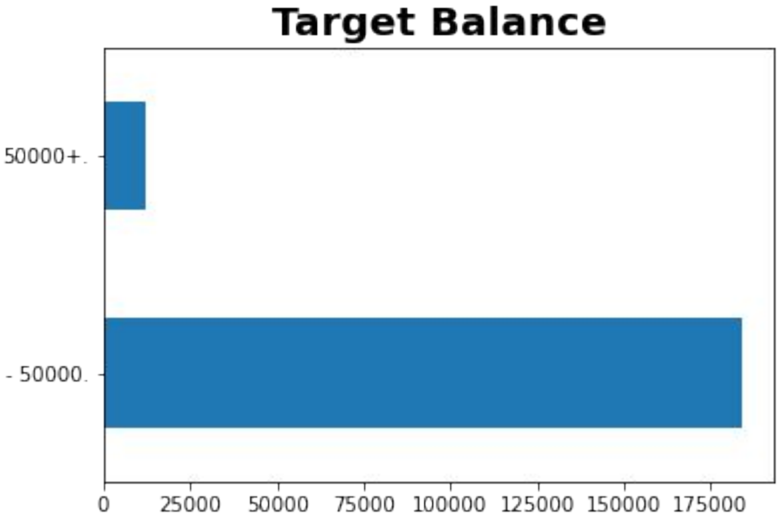
Census Income

Sibongile Toure

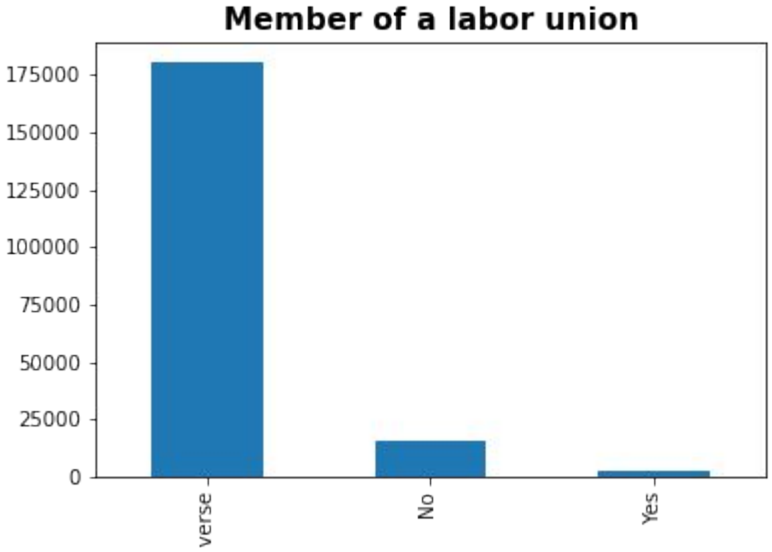
EDA

- Unbalanced Dataset
- Features
  - 33 Nominal
    - Marital Status
    - Class of Worker
  - 7 Continuous
    - Age
    - Number of Hours Worked

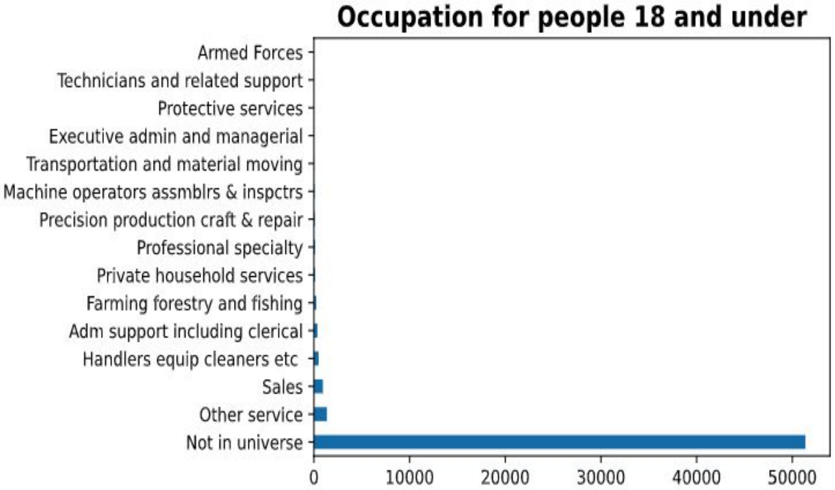


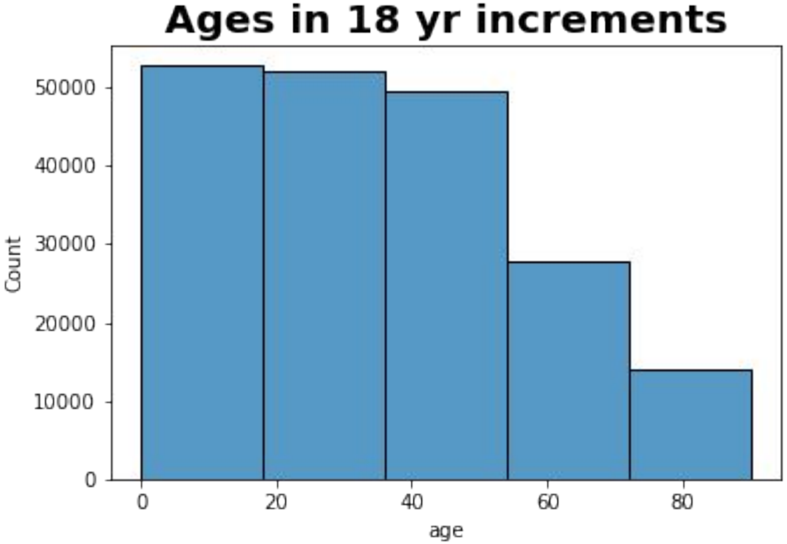
Feature Engineering

- Dropped features with ~50% NaN values
- Excluded data for people 18 and Under

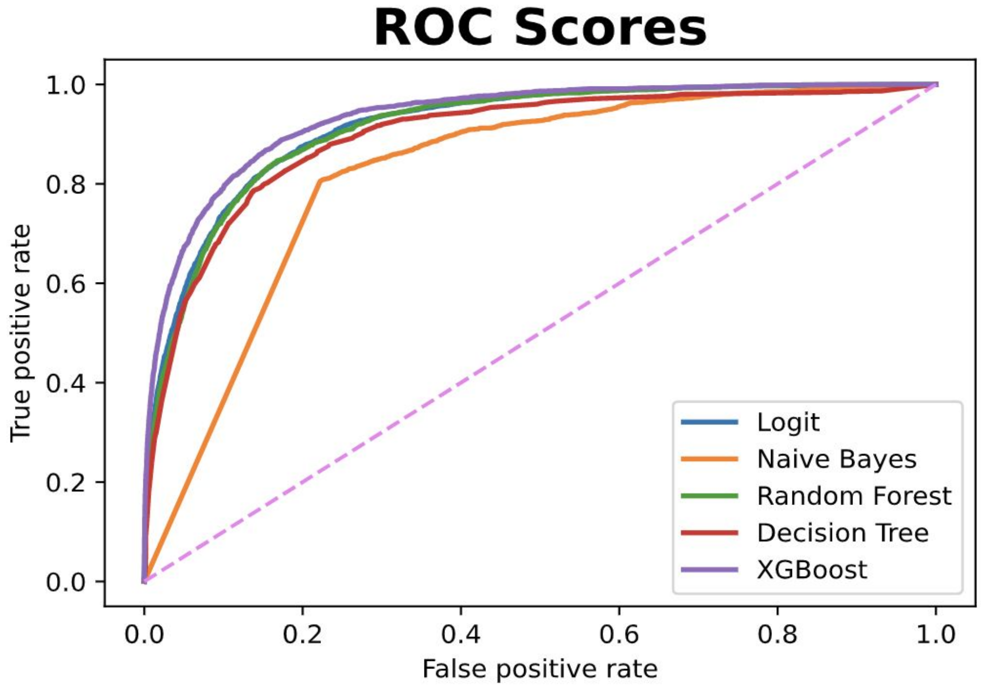


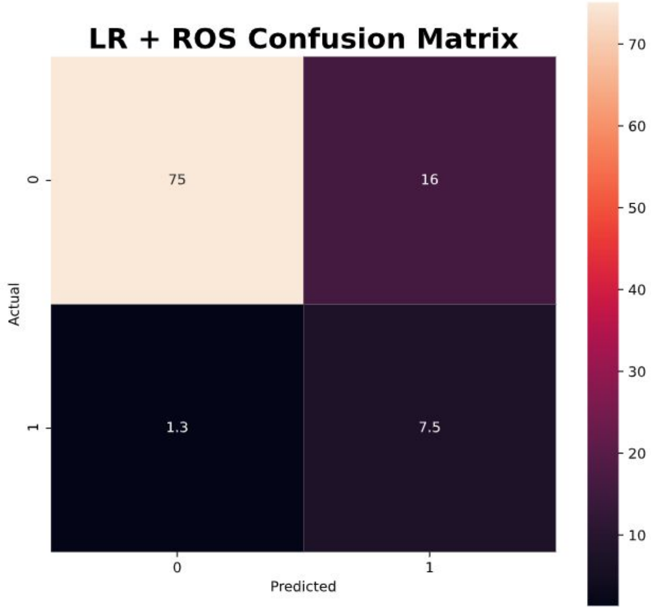
Feature Engineering



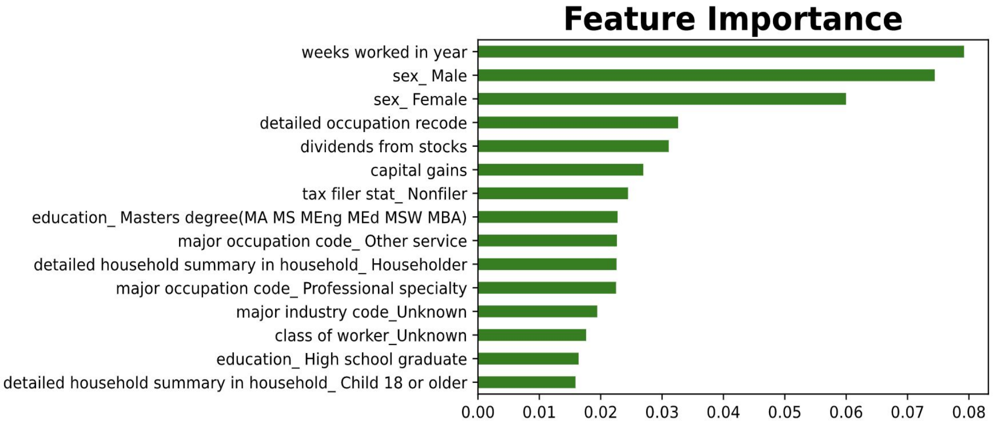


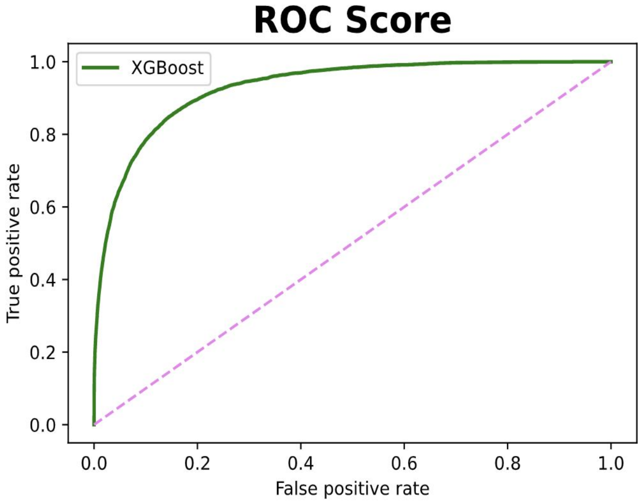
Modeling


ROC Scores
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Logit
Naive Bayes
Random Forest
Decision Tree
XGBoost


LR + ROS Confusion Matrix
Actual
0
1
75
16
1.3
7.5
Predicted
0
1
70
60
50
40
30
20
10

XGBoost


Feature Importance
weeks worked in year
sex_ Male
sex_ Female
detailed occupation recode
dividends from stocks
capital gains
tax filer stat_ Nonfiler
education_ Masters degree(MA MS MEng MEd MSW MBA)
major occupation code_ Other service
detailed household summary in household_ Householder
major occupation code_ Professional specialty
major industry code_Unknown
class of worker_Unknown
education_ High school graduate
detailed household summary in household_ Child 18 or older
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08


ROC Score
XGBoost
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Takeaways

- By segmenting out people under 18 we are able to get a clearer picture of the workforce
- Certain features such as how many weeks a person worked in a year and their sex has an impact on whether they make above or below $50000

Future Work

- Include more data from other sources
  - E.g. Neighborhood data can provide insights
- Try additional models