Applying Supervised Machine Learning Models

By Natalia Quintero

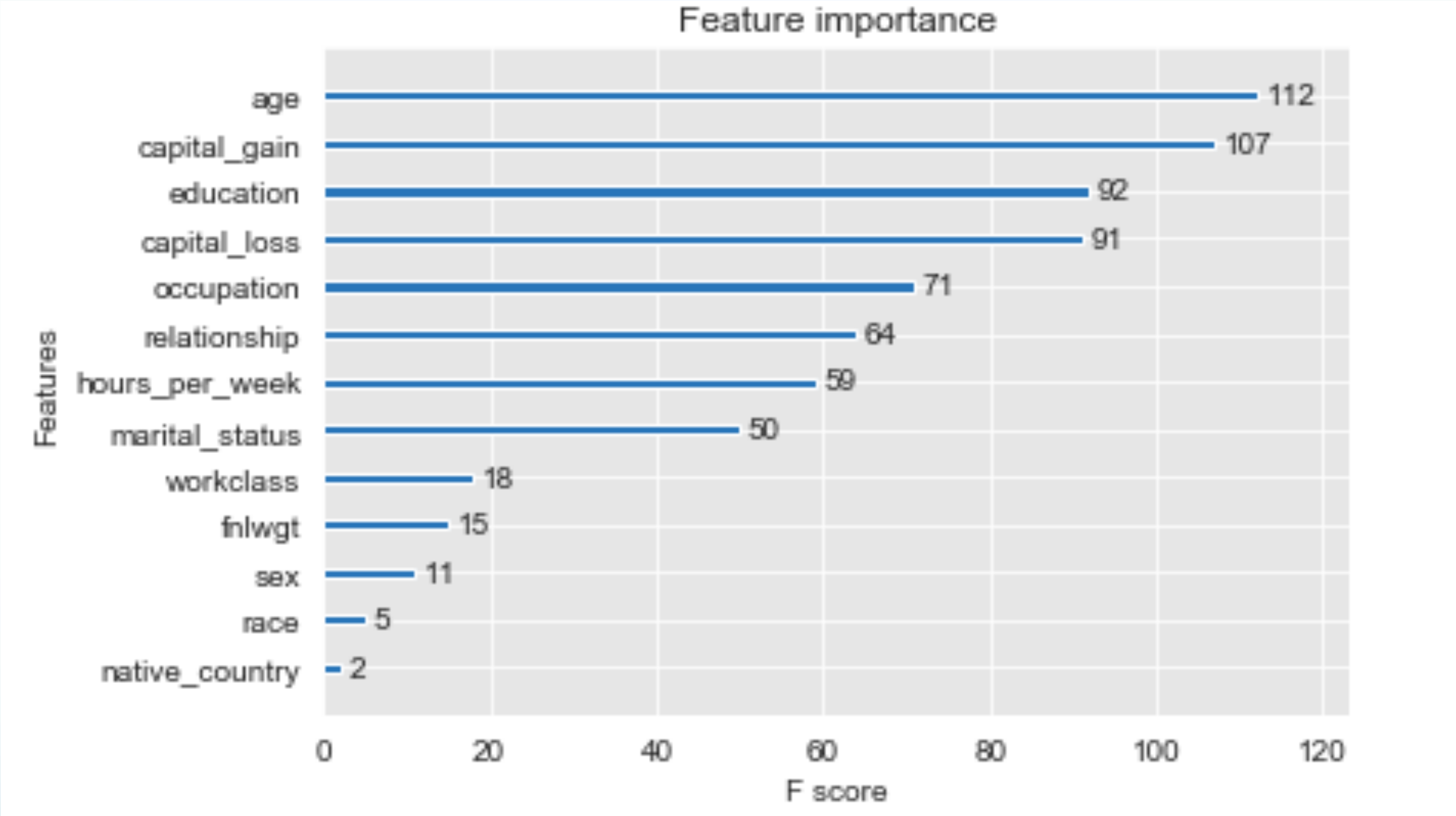
Data Set

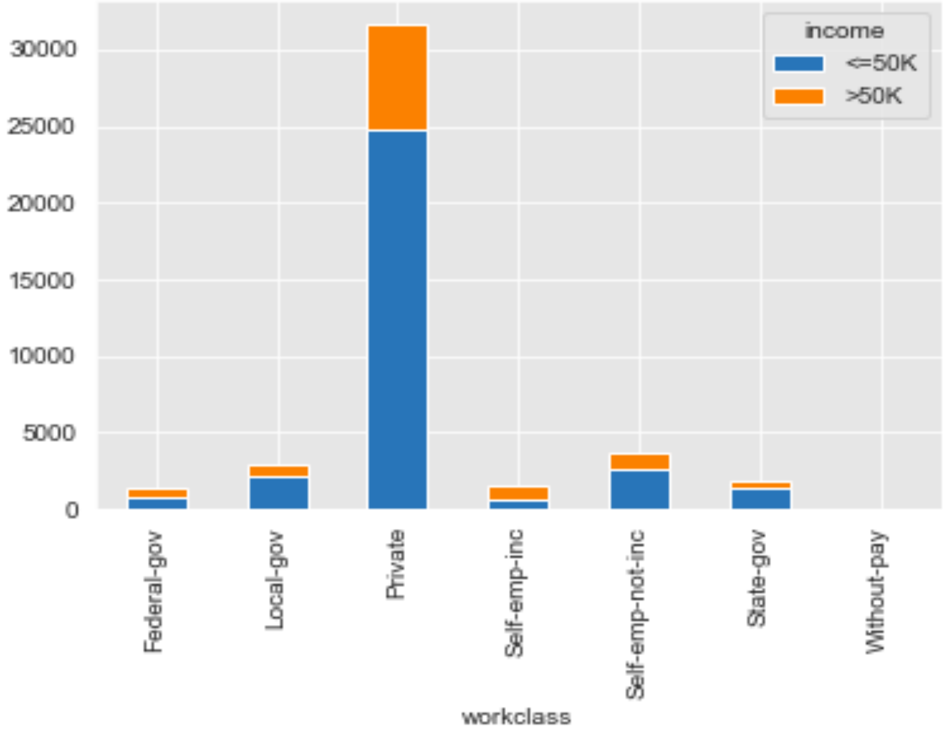
- Census Income data set from the 1994 census data
- Prediction task is to determine whether a person makes over 50K a year
- Machine Learning problem: classification
- The data set contains 14 attributes and 48,842 instances

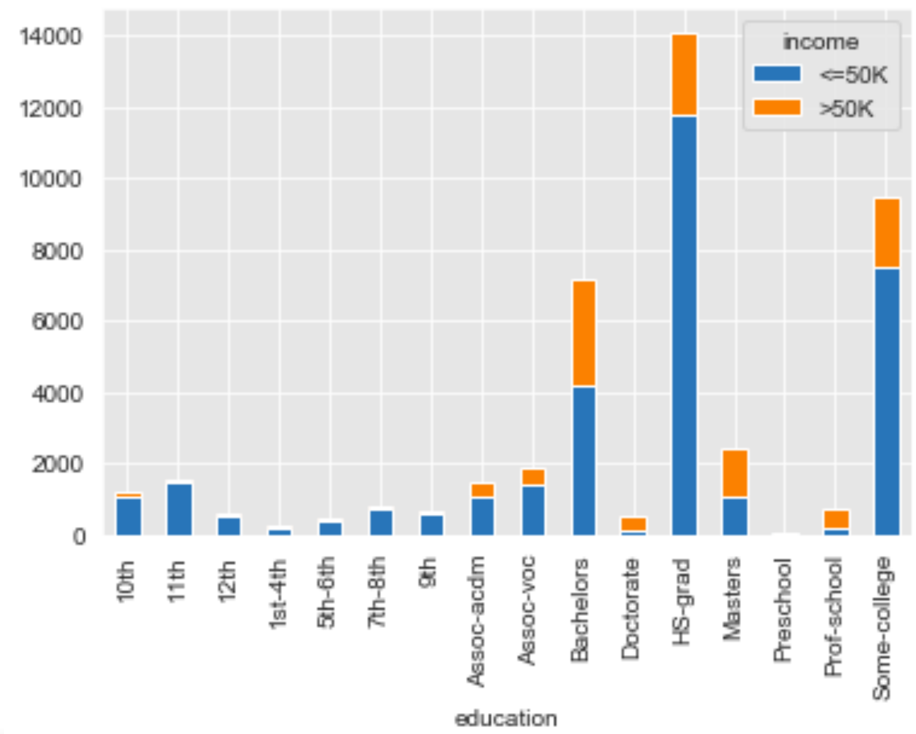
OSEMN Methodology

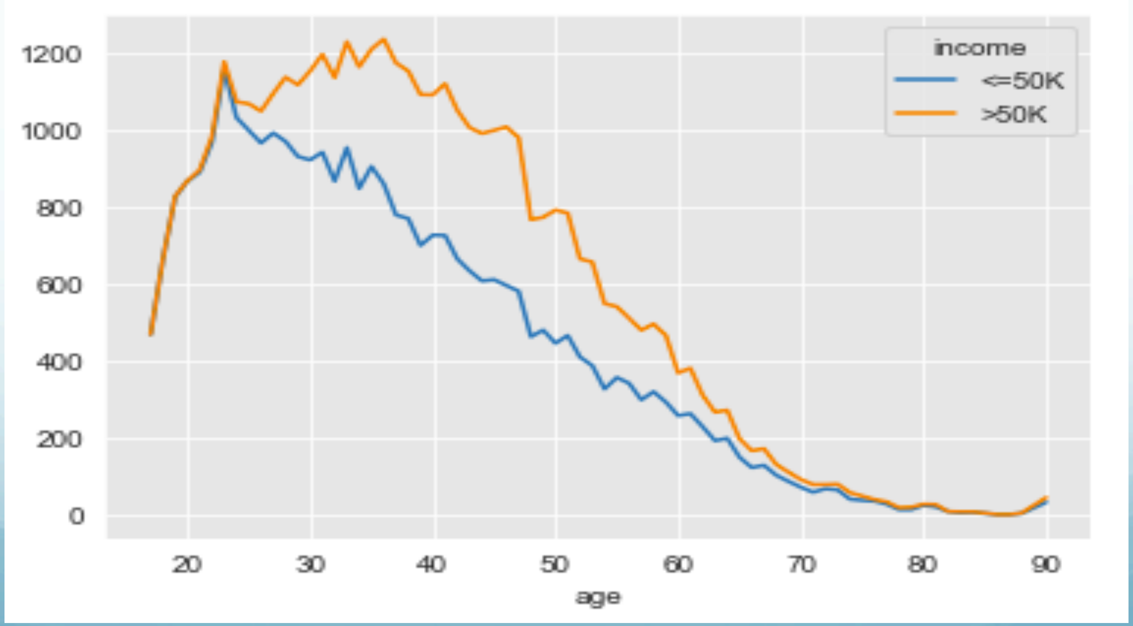
- **Obtain:** gather information, obtain the data http://archive.ics.uci.edu/ml/datasets/Census+Income)
- **Scrub:** remove data that is not needed, reduce noise
- **Explore:** set up the data, make sure the dataset meets what is necessary for the type of model to apply later on
- **Classification Models:** logistic regression and XGBoost
- **Interpret**: compare models and evaluate the results

Important Attributes




income
<=50K
>50K
30000
25000
20000
15000
10000
5000
0
Federal-gov
Local-gov
Private
Self-emp-inc
Self-emp-not-inc
State-gov
Without-pay
workclass


income
<=50K
>50K
14000
12000
10000
8000
6000
4000
2000
0
10th
11th
12th
1st-4th
5th-6th
7th-8th
9th
Assoc-acdm
Assoc-voc
Bachelors
Doctorate
HS-grad
Masters
Preschool
Prof-school
Some-college
education


income
<=50K
>50K
1200
1000
800
600
400
200
0
20
30
40
50
60
70
80
90
age

Model Results

- Logistic Regression Model: accuracy of 79%
- XGBoost Model: accuracy of 86%
- Better model for predictions: XGBoost

Interpretation of Results

- We can predict and classify income with the given data.
- The private sector is were most people work with higher salaries.
- Higher education means higher salaries.
- The peak income for those earning more than 50K is in between early 20’s and late 30’s.

Thanks for your time