Applying Supervised Machine Learning Models

By Natalia Quintero

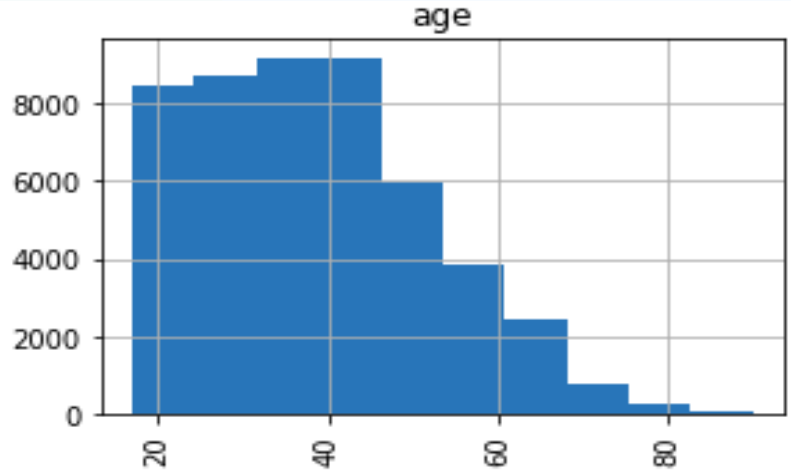
Data Set

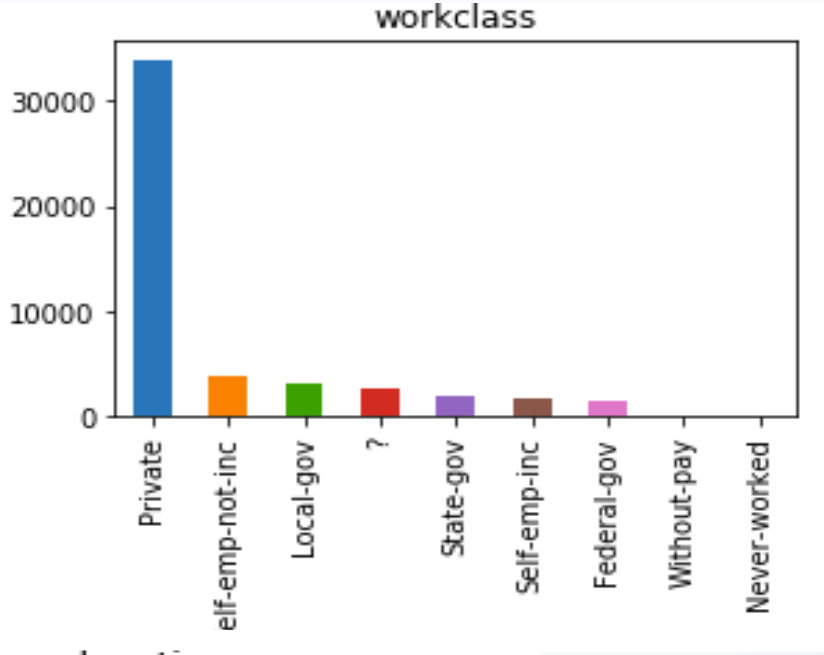
- Census Income data set from the 1994 census data
- Prediction task is to determine whether a person makes over 50K a year
- Machine Learning problem: classification
- The data set contains 14 attributes and 48,842 instances

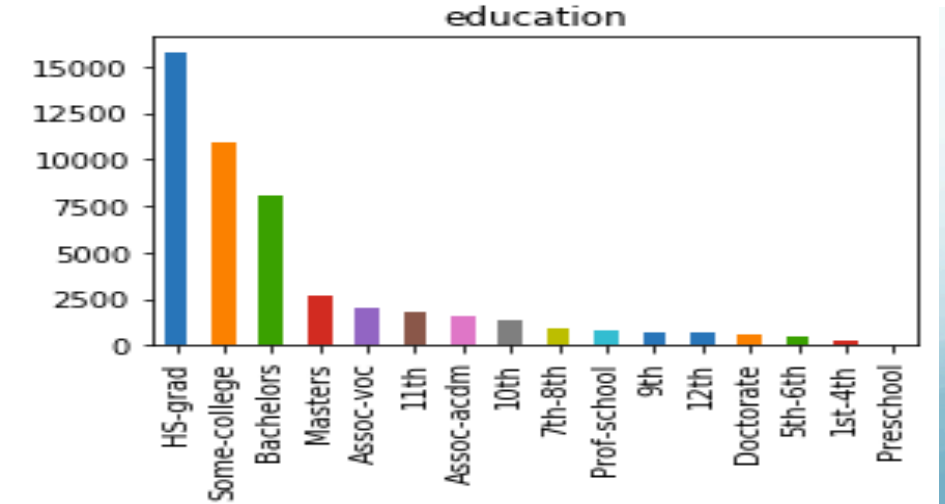
OSEM Methodology

- **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

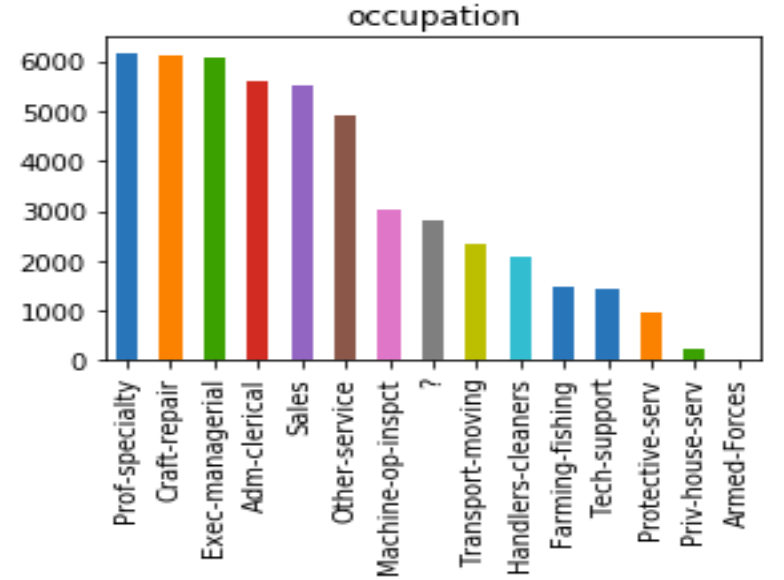
Some Attributes

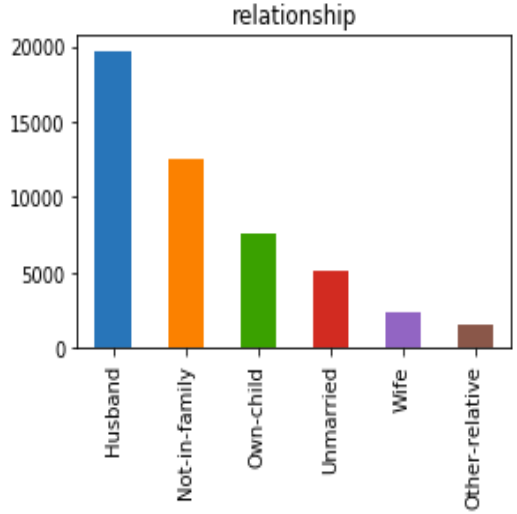


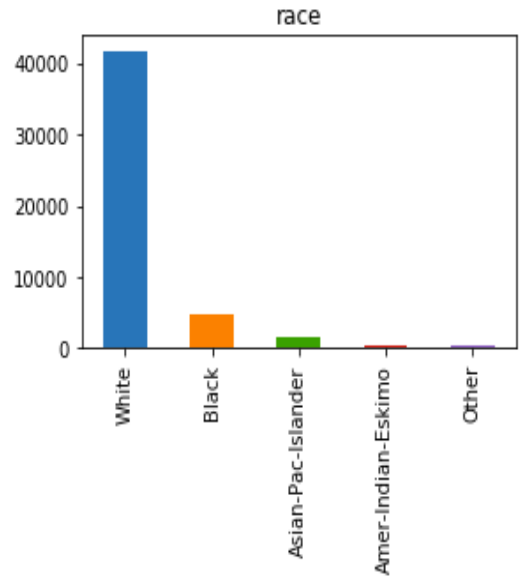


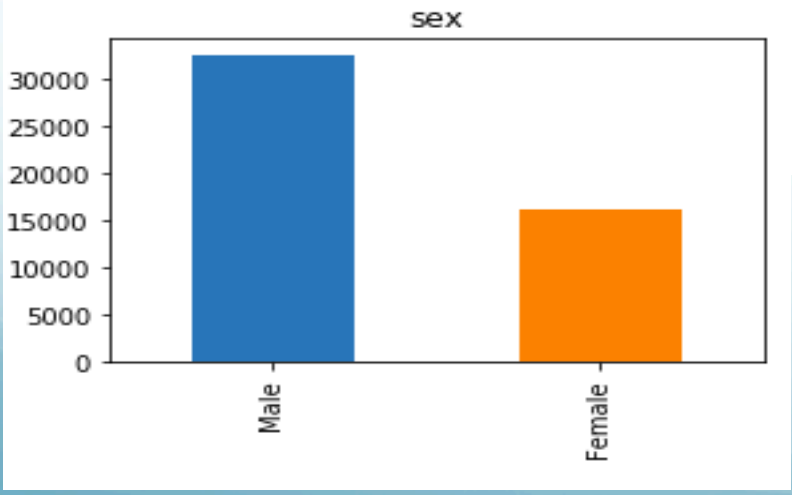


Some Attributes









Model Results

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

Results

Does a person makes more than 50K a year?

Yes, we can predict and classify income with the given data.

Thanks for your time