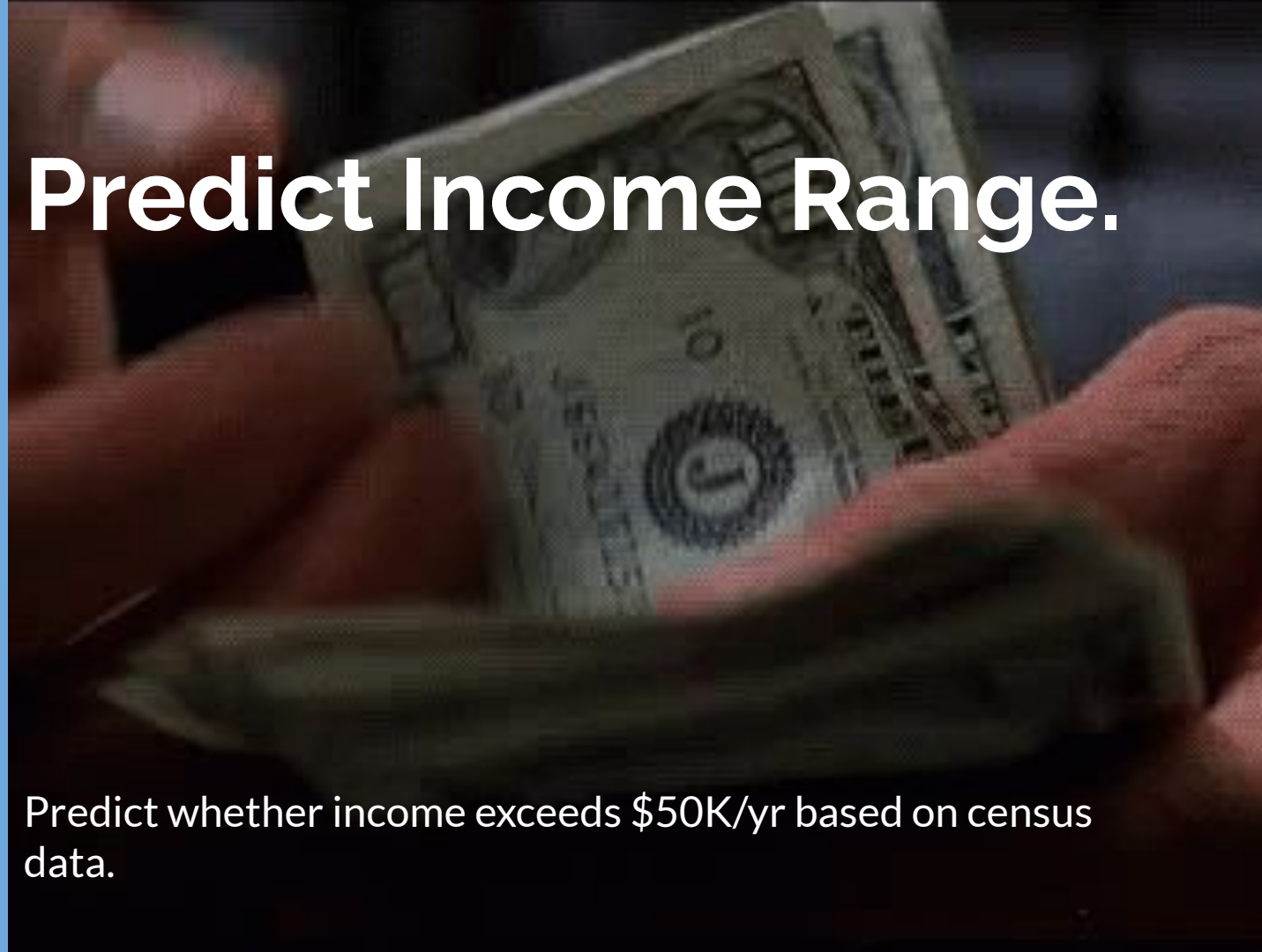


Predict Income Range.

Predict whether income exceeds $50K/yr based on census data.

I. Diaz, 02/2023

Project Description

The purpose of this project is to predict whether the income of a person exceeds 50K/yr based upon the census data. The data includes information related with age, higher level of education achieved, occupation, sex, hours-per-week, etc. The census was carried out in 1994.

The census data was extracted from the Census Bureau Database, and was uploaded to UCI Machine Learning repository with some modifications described into the project file.

Despite the reason for needing to estimate whether a subject earns more than 50K per year has not been disclosed, there may be many reasons for doing it, for example: Prevent tax evasion, planing better general public policies.

Data Description

Source:

UCI Machine Learning Repository

Number of Instances:

48.842

Number of Attributes::

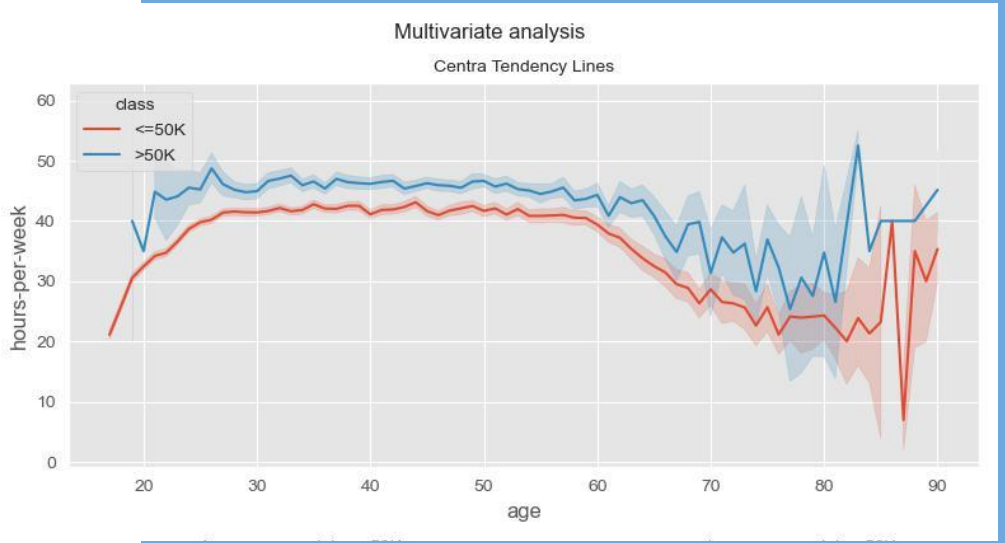
14

Feature Description

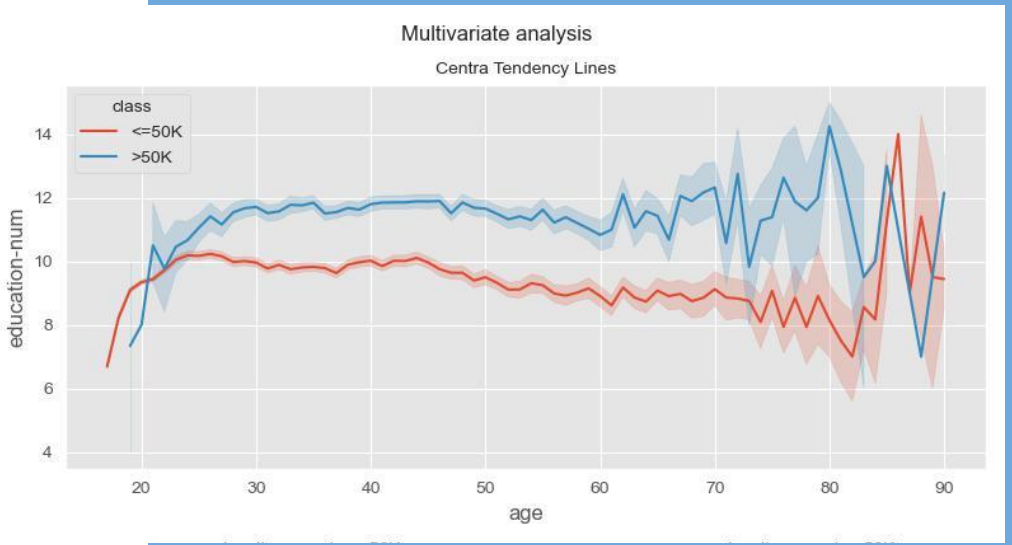
- **age**: continuous.
- **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: Final Weight, continuous variable calculated based upon certain criteria, described in the project file.
- **education**: The highest level of education achieved for that individual. This is nominal attribute. Preschool < 1st-4th < 5th-6th < 7th-8th < 9th < 10th < 11th < 12th < HS-grad < Prof-school < Assoc-acdm < Assoc-voc < Some-college < Bachelors < Masters < Doctorate.
- **education-num**: Highest level of education in numerical form.
- **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: Female, Male.
- **capital-gain**: Capital gains recorded. continuous.
- **capital-loss**: Capital Losses recorded. continuous.
- **hours-per-week**: continuous.
- **native-country**: country of origin of the subject.
- **class**: >50K, <=50K

Key findings

Age vs Hours per Week



Age vs Education



Model Evaluation

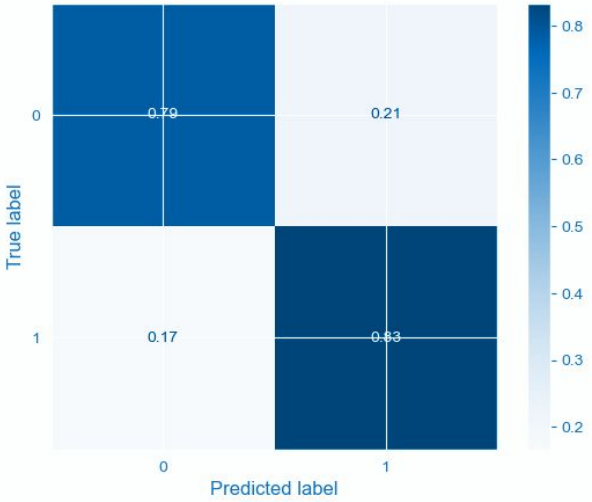
Results on Test Data:

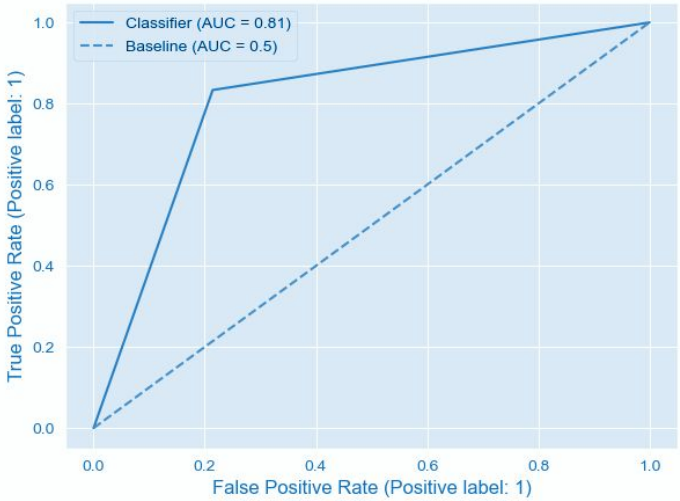
C: 1

Penalty: L2

Logistic Regression

Accuracy	Precision	Recall	F1
0.80	0.55	0.83	0.66



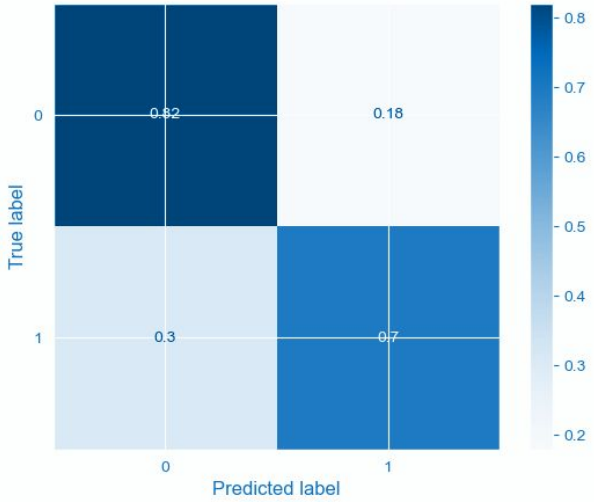


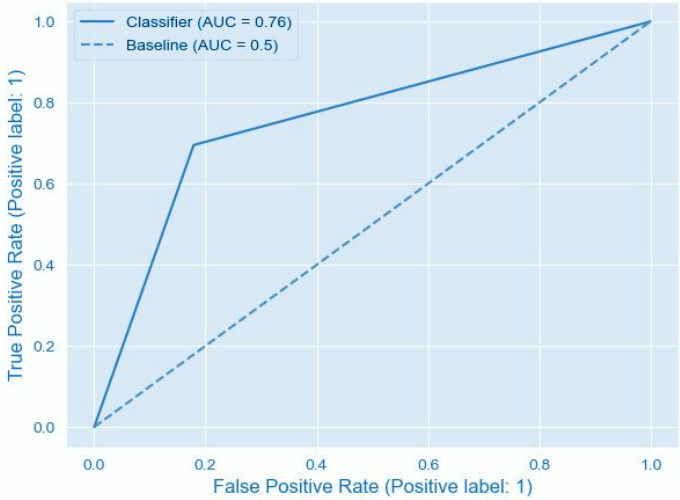
Results on Test Data

n_neighbors: 9

K Neighbors Classifier

Accuracy	Precision	Recall	F1
0.79	0.55	0.70	0.61





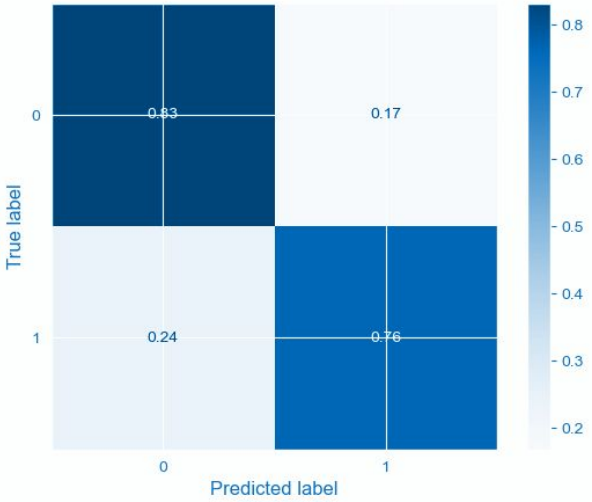
Results on Test Data

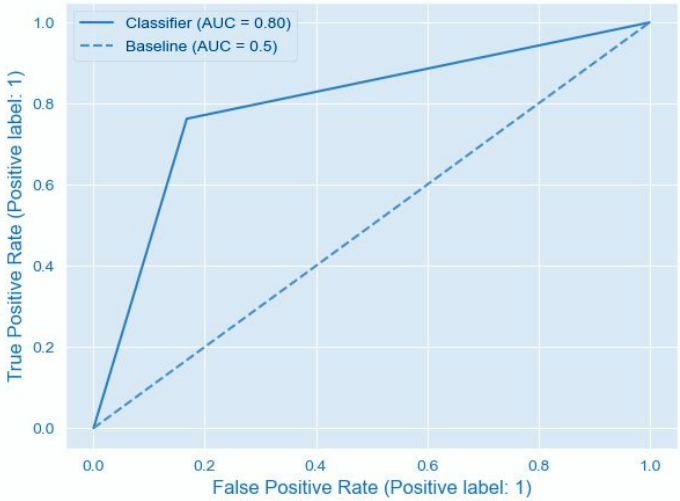
booster: gbtree

max_depth: 5

XGBoost

Accuracy	Precision	Recall	F1
0.82	0.59	0.76	0.66





The most important thing is hit the mark of the positive class or >50K, this will answer the question. The most important metric here over the Accuracy is Recall, or what is the same to say how many of the of the positive cases were predicted correctly over all the positives.

Having said this, the model that performed better is **Logistic Regression**

Recommendations

- For time constraints in training the models, wasn’t possible to tune more parameters, then it is highly recommended do a fine tune using more parameters.
- Train a Neural Network to compare with the already existing models.