Classifying income.

* Dependent variable categorizes people if they make $50,000+/year(1) or not(0).
* Independent variables used:
  + Age – Age of the subject.
  + Workclass – What sector does the person work in?
    - 0 – private sector
    - 1 – Public/government.
    - 2 – Work without pay/volunteer or unemployed/never worked.
  + Sex – Person’s sex
    - 0 – Male
    - 1 – Female
  + Native country – Where the person is from.
    - 0 – USA
    - 1 – Canada
    - 2 – Central and South America
    - 3 – Europe
    - 4 - Asia
  + Education – The education level of a person.
    - 0 – HS graduate or GED.
    - 1 – Less than HS education.
    - 2 – Some College
    - 3 – Associates degree
    - 4 – Bachelors degree
    - 5 – Professional School
    - 6 – Masters degree
    - 7 – PhD.
  + Marital-Status – Marital status of the person.
    - 0 – Married
    - 1 – Never married
    - 2 – Divorced or separated
    - 3 – Widowed
  + Capital-gain – A continuous variable
  + Capital-loss – A Continuous variable
  + Hours-Per-Week – How many hours a week a person works.
* Removed education-num, occupation, relationship, race. All for being redundant to another variable. Race partially being redundant to “native-country” and for race having a VIF of 4.08. The reason race was removed is because it brought native-country’s VIF down from 5.418 to 2.260. In my opinion, in race’s case, on their own, neither case warrants removal, but with them together and trying to keep a smaller model, I feel race’s removal is justified in this case.
* Used XGBooster to analyze due to imbalanced dataset (0 (75%): 1 (25%)).
  + Altered cutoff probability to 0.69.
  + Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | **0 (< $50)** | **1 (> =$50)** |
| **0** | 11330 | 1105 |
| **1** | 1322 | 2524 |

* + 85% accuracy, 68% F1 score, 70% precision, 66% recall.
  + Variable importance plot (top 5, in order):
    - Age
    - Hours-per-week.
    - Capital Gain
    - Capital Loss
    - Sex (Male)
  + Cover plot’s importance (top 5, in order):
    - WorkClass (Unemployed)
    - Education (PhD)
    - Native-Country(Canada)
    - Education (Professional School)
    - Capital Gains
  + Gains plot’s importance (top 5, in order)
    - Marital\_Status (Married)
    - Education (< HS)
    - Capital Gains
    - Ed (HS-grad)
    - Ed(Some\_College)