A case study for US Census Bureau:

Predict person’s income is more or less than 50000 dollar

Xi Chen

Overview

- Background
- Usage of model considerations
- Data overview
- Exploratory data analysis (EDA) & Insights
- Modeling & Performance
- Conclusion
- Next step
- Backup slides

Background

What is US Census Bureau	• A government agency responsible for collecting and analyzing demographic and economic data • Conduct a nationwide census every 10 years to guide policy and funding decision
Why Census Data important	• Helps allocate billions of dollar for public services • Provide valuable demographic insights for economic and social planning
Objective	• Analyze US Census data to identify factors influencing income levels • Develop a data analysis and modeling pipeline to predict whether an individual earns more or less than $50000 per year

Usage of model considerations

It is important to understand the usage of the model, considering it has a big impact on the analysis:

- **Impact on predicting target**:
    - There might be better target to predict other than 50k depends on the usage. For example, to predict income bucket, or directly predict whether to give out benefit(Social benefit) or not instead of purly based on income is larger than 50k or not.

- **Impact on the evaluation approach**:
    - We might use different way to evaluate the model depends on the objective. i.e, for tax fraud might focusing on the >50k.
    - To help to understand how good the model's performance need to reach.

- **Impact on the Data:**
    - There might be more features can be gathered for specific target.
    - Data regulation.
    - Model bias based on race/gender.
    - The potential data sampling approach, data drifting issue

Data overview

- Data size: Total data size is 299285 records. 199534 (66.66%) for training and 99762 (33.34 %) for testing.
- Data collected year : 94 and 95
- Feature size: In metadata file, there are 45 different features in total, but in dataset only contains 42 features.
- Main feature groups:
    - Demographics: Age, Gender, Race, Citizenship
    - Employment: Occupation, Industry, Work Class, Hours Worked
    - Financial: Capital Gains/Losses, Dividends, Tax Status
    - Education & Family: Education Level, Marital Status, Household Composition
    - Geography & Migration: State, Region, Migration Patterns

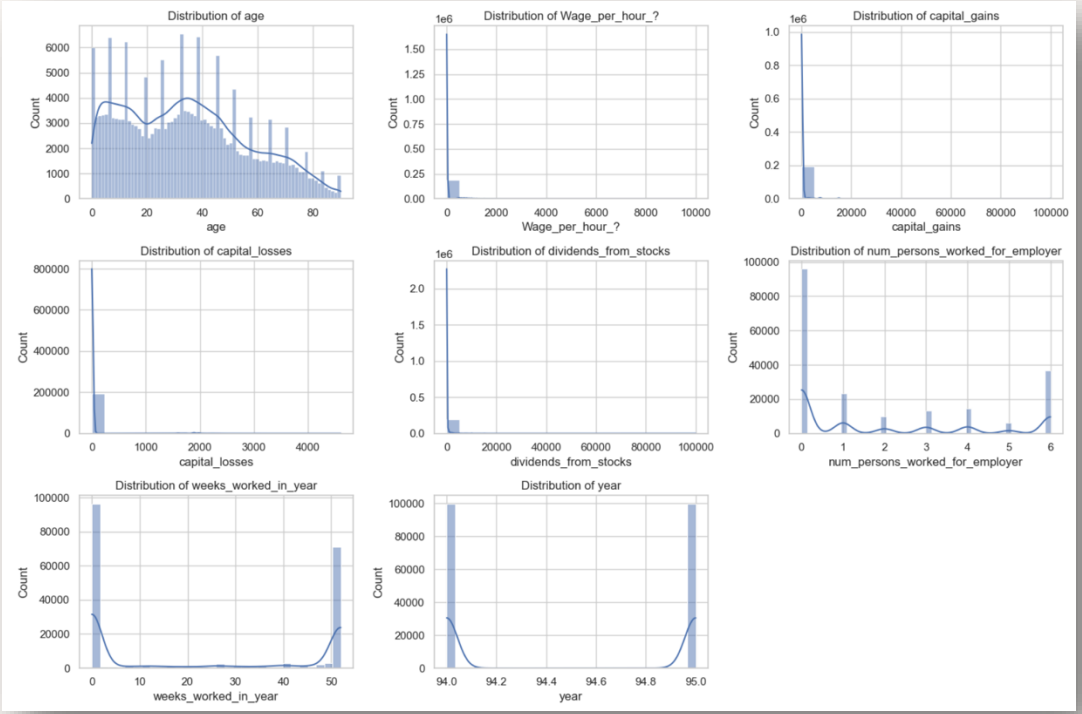
PS: There is no column name in the data, I matched it based on the feature order and the value report from metadata file.

Exploratory data analysis (EDA)

There are two groups of features: 8 Numeric features & 33 Categorical features.

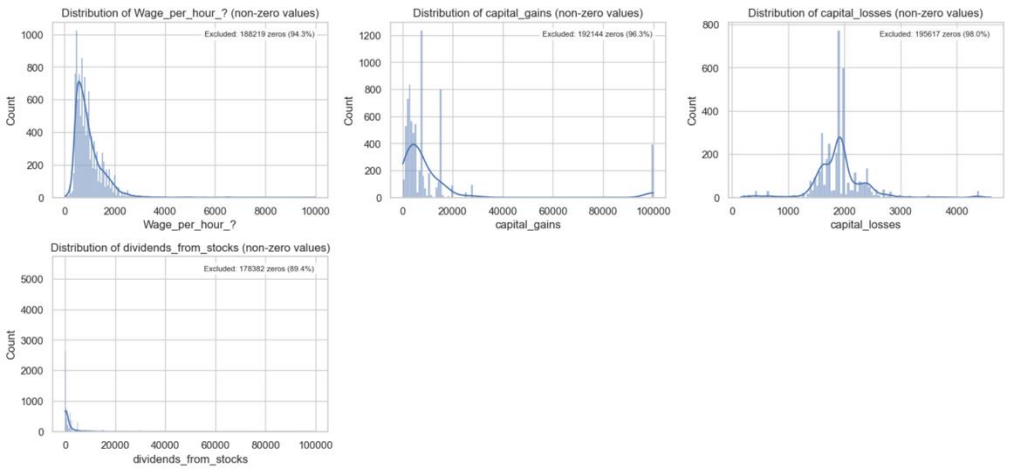
We’re mainly into the value distribution, frequency distribution to identify missing value, outliers, potential opportunity to extract more features for modeling purpose

EDA – Numeric feature



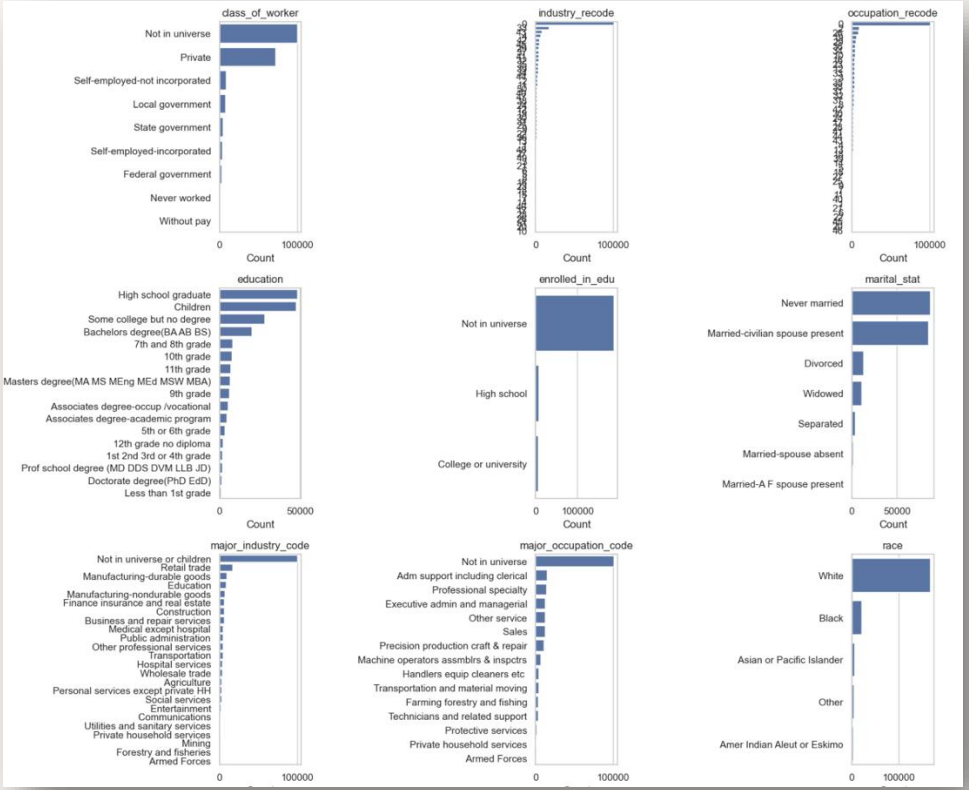
Feature	Findings
Age	There is a spike in each age group indictes there might have certain data sampling pattern.
Wage per hour Capital gains Capital losses Dividens	Majority are empty. Zoom in distribution(Remove 0) will show in next slide

EDA – Numeric feature



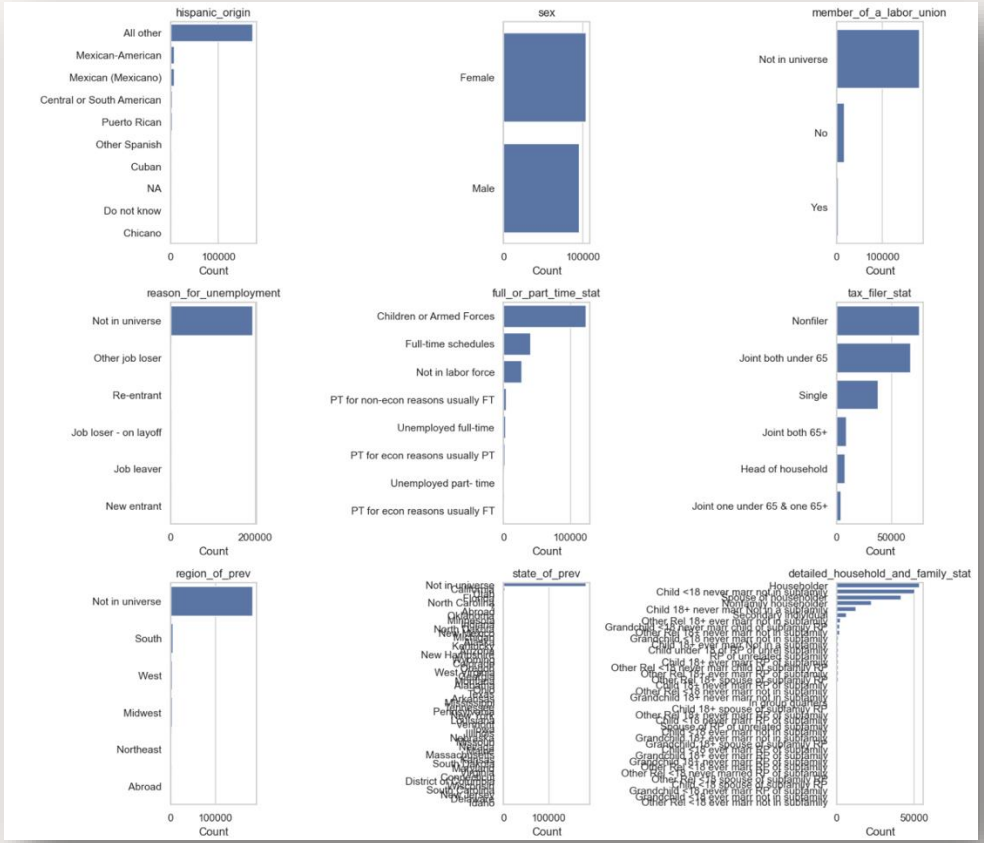
Feature	Findings
Wage per hour	This feature might be weeky income instead of hourly imcome. As the average is 800 dollar per hour, which is too high for hourly wage
Capital gains	There is a peak in 100000, there might be a cap on the declaration?
Capital losses	Compared with capital gain, the number of the capital losses is surprisingly small. I would expect the capital losses is also

EDA – Numeric feature



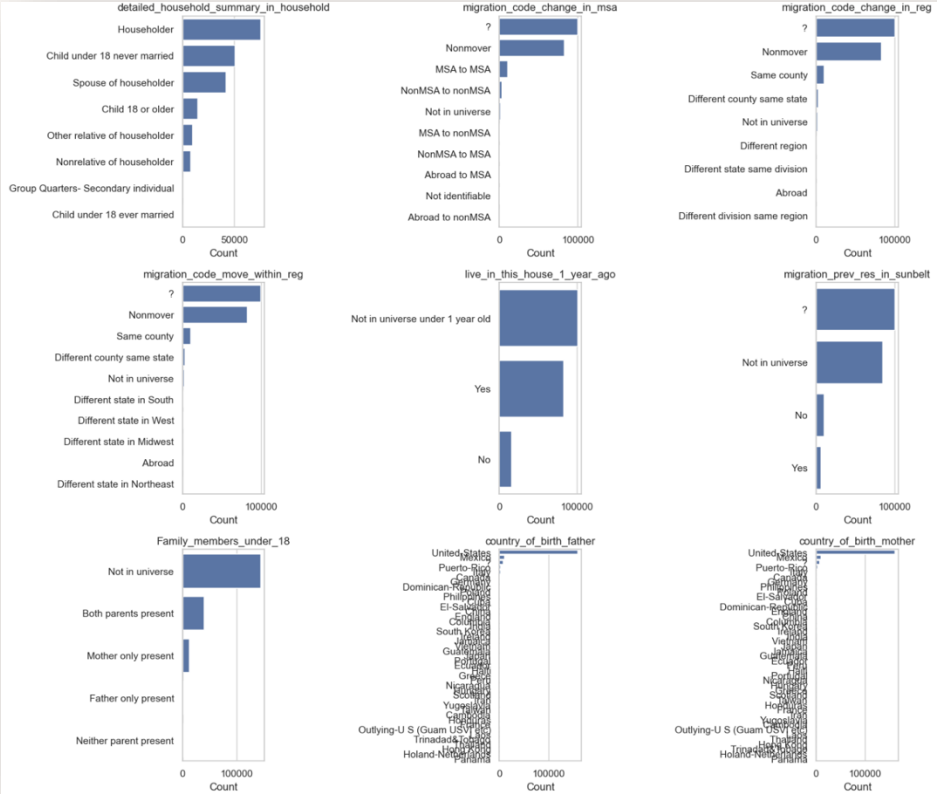
Feature	Findings
Class of worker Industry recode Occupation recode Major industry code	Class 0 might indicates students / children. The class of worker have a lot of private might need pay a bit attention. Might be caused by the sampling approach
Race	Majority are white, need to pay attention when it comes to race / gender bias if there is any ethical concerns might be raised by public

EDA – Categorical feature



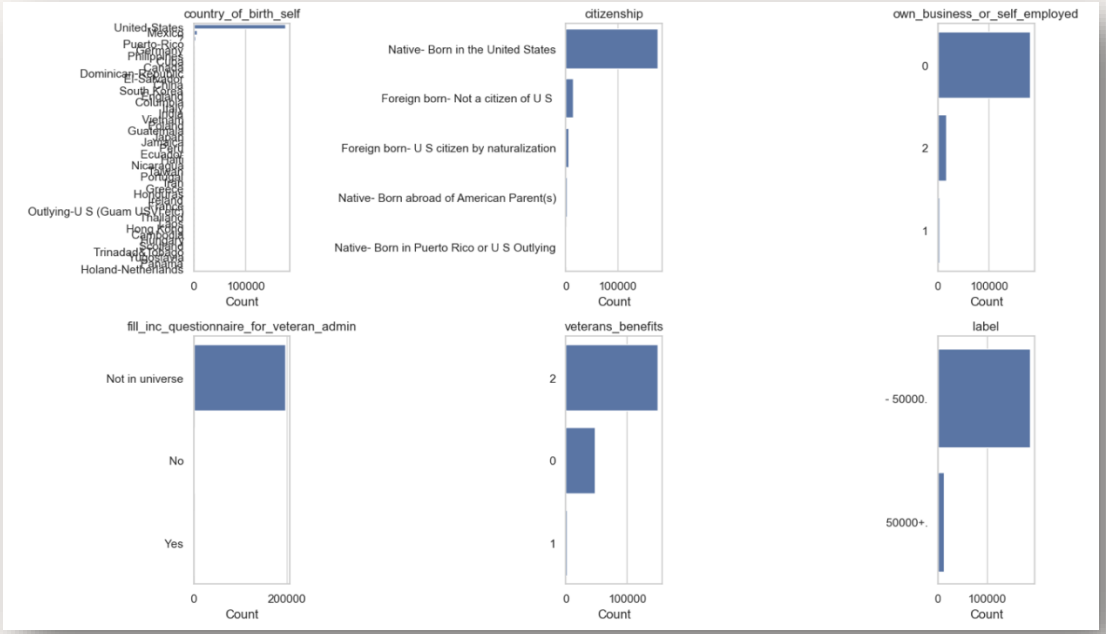
Feature	Findings
State_of_prev	Majority are the locals(who didn't move)
Household and family stat	This category might be able to break down to multiple categories

EDA – Categorical feature



Feature	Findings
Migration code change	What's the difference between "?" And Not in universe?
Country of birth father/mother/self	Majority are native Americans

EDA – Categorical feature



Feature	Findings
Label	The distribution of income is highly imbalanced

EDA – Insights

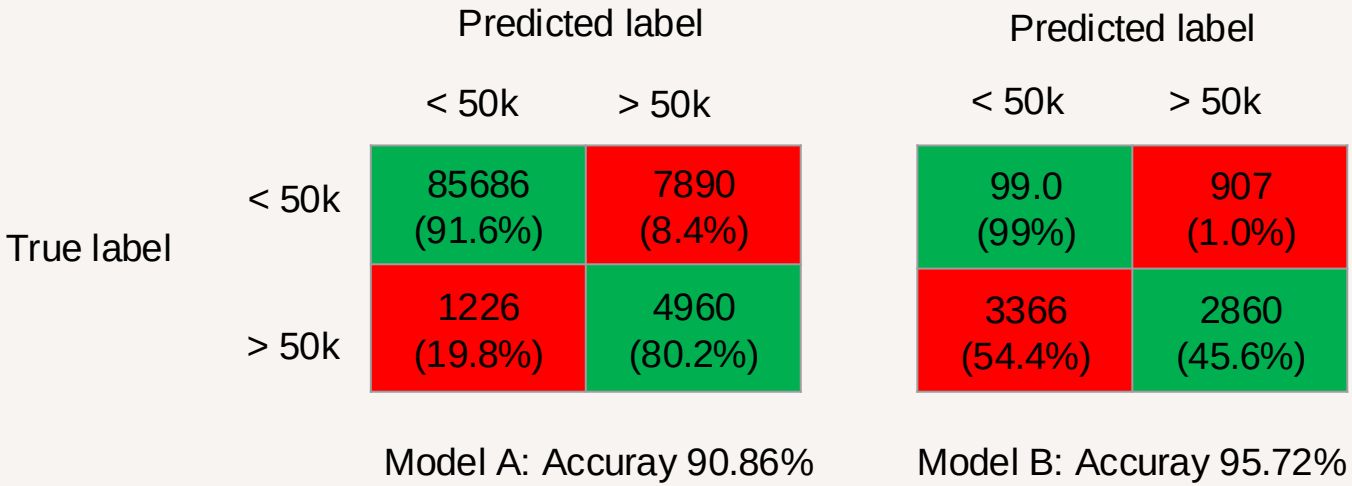
- **Data correctness**:
    1. No empty filed and obvious outlier
    2. Potential conflicted records: (I,e age vs education: below 18 with PhD, age vs marital status: below 18 and “Widowed”, Occupation vs. Industry: Teacher in construction industry. To be consolidated with business.)

- **Data Sampling**:
    1. There are spike in ages distribution. Need to find out why.
    2. There are lots of people under 18. For predicting purpose, would be better to have more people above 18 if the predicting target is income > 50k or not.

- **Data preparation for modeling**:
    1. Numeric feature: Some of the features distribution are skewed
    2. Geo related feature: We can extract geo / embedding for data modeling
    3. Feature engineering:
        - Expand the feature to multiple features such as household info
        - Statistics for the numeric value
    4. Tackle imbalanced dataset

Modeling – results

Two different models (XGBoost) with accuracy of: 90.86% and 95.72%



Model A: Accuray 90.86%

Model B: Accuray 95.72%

Even the model B has overall higher accuracy. But depends on the objective Model A might be a better choice. i.e Identify someone is under-declare income for tax benefit.

Conclusion

- The model can predict the income threshold of 50k with 95.72% accuracy
- Impact & benefit (Depends on the use case)
- Limitaions: The data sample is bit too old, and highly imbalanced data is challenging for modeling
- Domain knowledge: Some of the features, value and objective will need business to clarify

Next step

Business:

- Objective: Explore the other possible objectives for modeling target
- Data sampling: Evaluate if the sampling approach is representative or not
- Attributes: Identify if there is more attributes can be gathered
- Conflicting record’s value: Verify the rules and consolidate it for data cleansing

Technical:

- Feature engineering: Review it together with business
- Model enhancement: Tryout deep learning models (i.e deep learning model).
- Explore ensemble methods

Backup slides

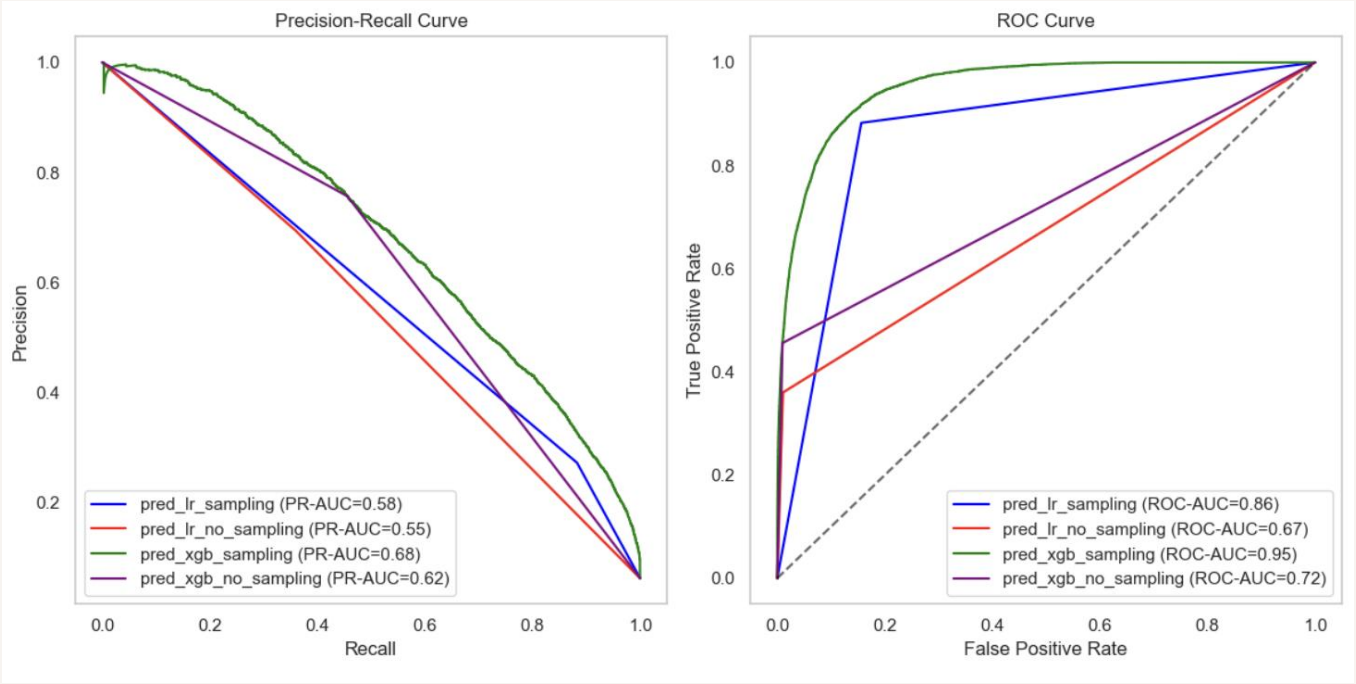
Modeling – setup (Backup)

Data prepartion:
- Encoding
- Standarlization
- Feature engineering

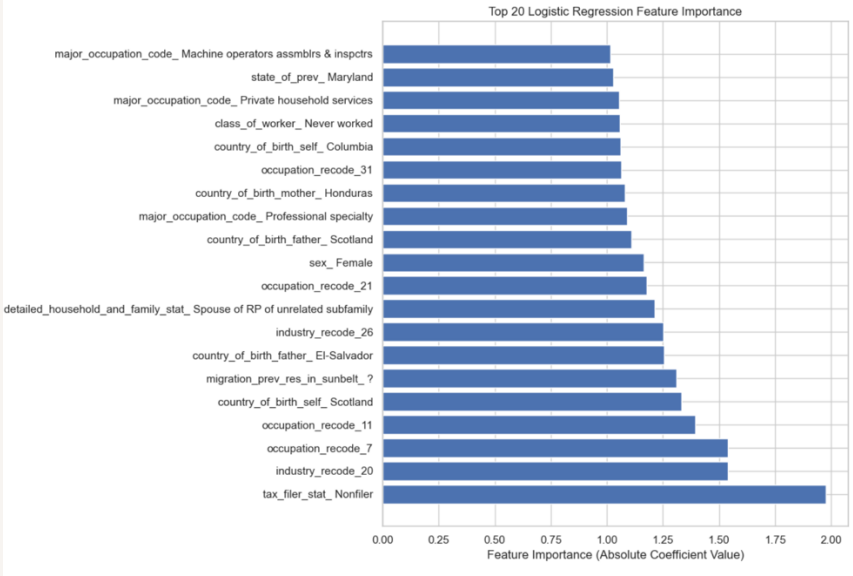
Hyperparameter tuning: Cross-validation

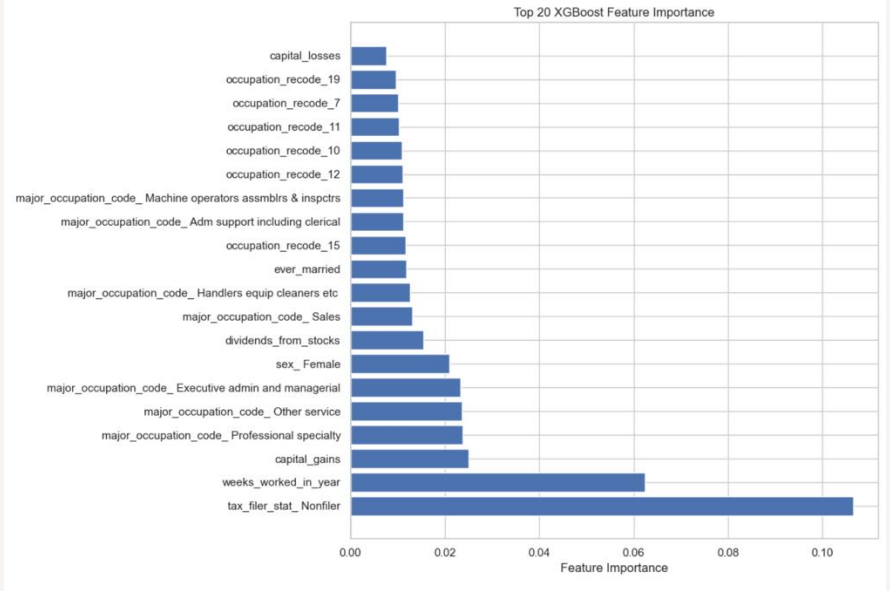
Model:
- LR
- XGBOOST
- Deep learning

Backup


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
pred_lr_sampling (PR-AUC=0.58)
pred_lr_no_sampling (PR-AUC=0.55)
pred_xgb_sampling (PR-AUC=0.68)
pred_xgb_no_sampling (PR-AUC=0.62)
ROC Curve
True Positive Rate
False Positive Rate
pred_lr_sampling (ROC-AUC=0.86)
pred_lr_no_sampling (ROC-AUC=0.67)
pred_xgb_sampling (ROC-AUC=0.95)
pred_xgb_no_sampling (ROC-AUC=0.72)

Backup


Top 20 Logistic Regression Feature Importance
major_occupation_code_ Machine operators assmblrs & inspctrs
state_of_prev_ Maryland
major_occupation_code_ Private household services
class_of_worker_ Never worked
country_of_birth_self_ Columbia
occupation_recode_31
country_of_birth_mother_ Honduras
major_occupation_code_ Professional specialty
country_of_birth_father_ Scotland
sex_ Female
occupation_recode_21
detailed_household_and_family_stat_ Spouse of RP of unrelated subfamily
industry_recode_26
country_of_birth_father_ El-Salvador
migration_prev_res_in_sunbelt_ ?
country_of_birth_self_ Scotland
occupation_recode_11
occupation_recode_7
industry_recode_20
tax_filer_stat_ Nonfiler
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Feature Importance (Absolute Coefficient Value)


Top 20 XGBoost Feature Importance
capital_losses
occupation_recode_19
occupation_recode_7
occupation_recode_11
occupation_recode_10
occupation_recode_12
major_occupation_code_ Machine operators assmblrs & inspctrs
major_occupation_code_ Adm support including clerical
occupation_recode_15
ever_married
major_occupation_code_ Handlers equip cleaners etc
major_occupation_code_ Sales
dividends_from_stocks
sex_ Female
major_occupation_code_ Executive admin and managerial
major_occupation_code_ Other service
major_occupation_code_ Professional specialty
capital_gains
weeks_worked_in_year
tax_filer_stat_ Nonfiler
0.00
0.02
0.04
0.06
0.08
0.10
Feature Importance