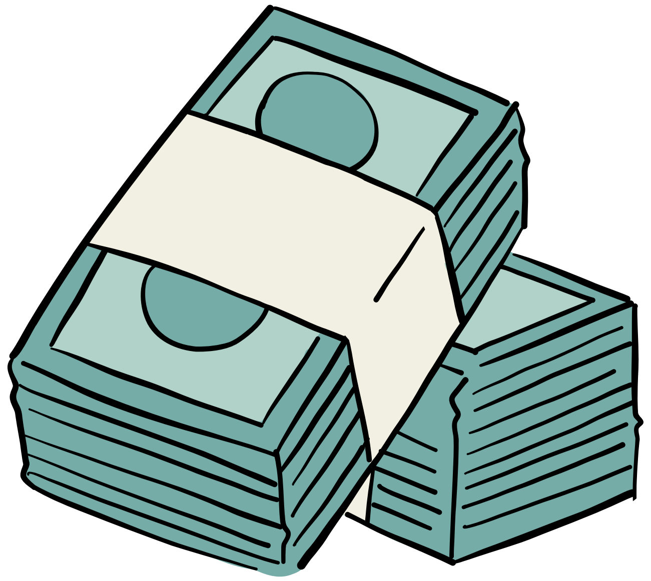
United States® Census Bureau

Predicting High Income Using US Census Data

Presented by:
Vadim Malz.

Date Presented:
November 19th, 2025

Problem Statement



01

Build a predictive model to classify individuals into >50k vs ≤50k income.

02

Understand socioeconomic drivers behind income differences.

03

Deliver an interpretable, business-ready solution.

DataSet Overview

- ~300k rows, mix of numerical and categorical features.
- Target is highly imbalanced (~7% earn >50k).
- Merged both Datasets into 3 separate block
    - Train Set : 70%
    - Val Set : 10%
    - Test Set : 20 %
- Target feature : “Income”

Feature Engineering (Numerical Features)

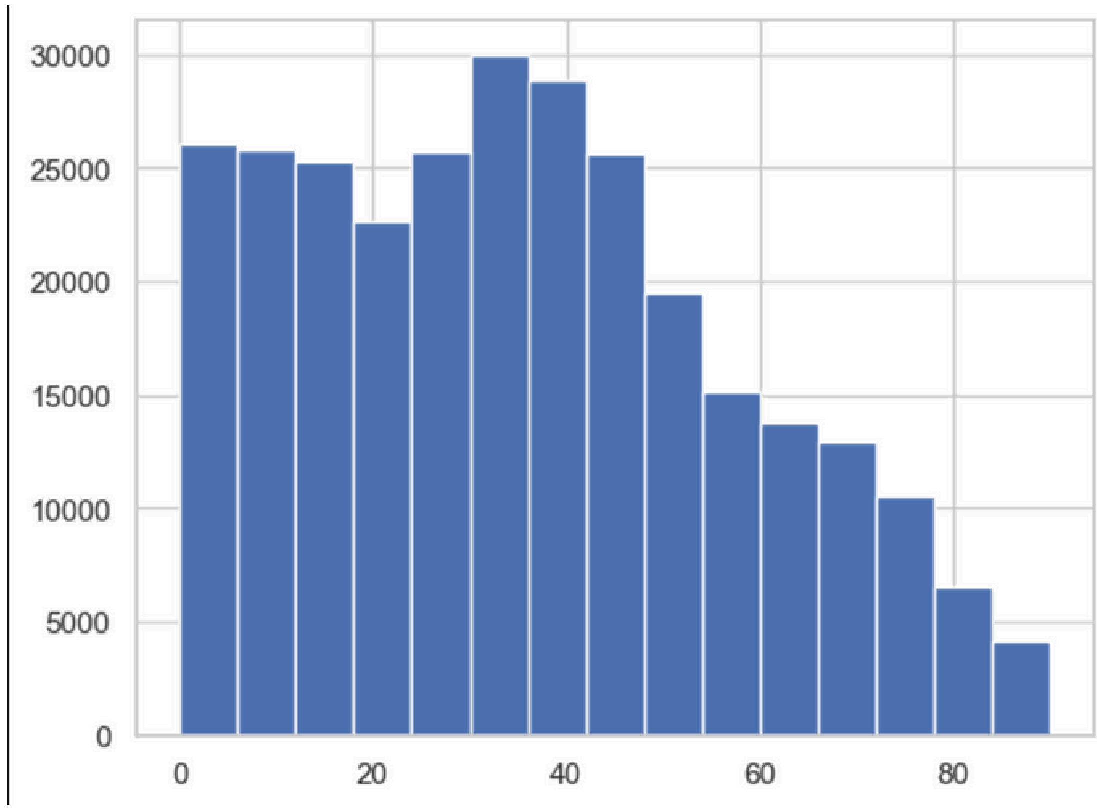
- 🔧 New Feature Created => captures overall capital financial activity in one variable.

$$TotalCapital = CapitalGains - CapitalLosses + DividendsFromStocks$$

Note on Tree-Based Models ⚠

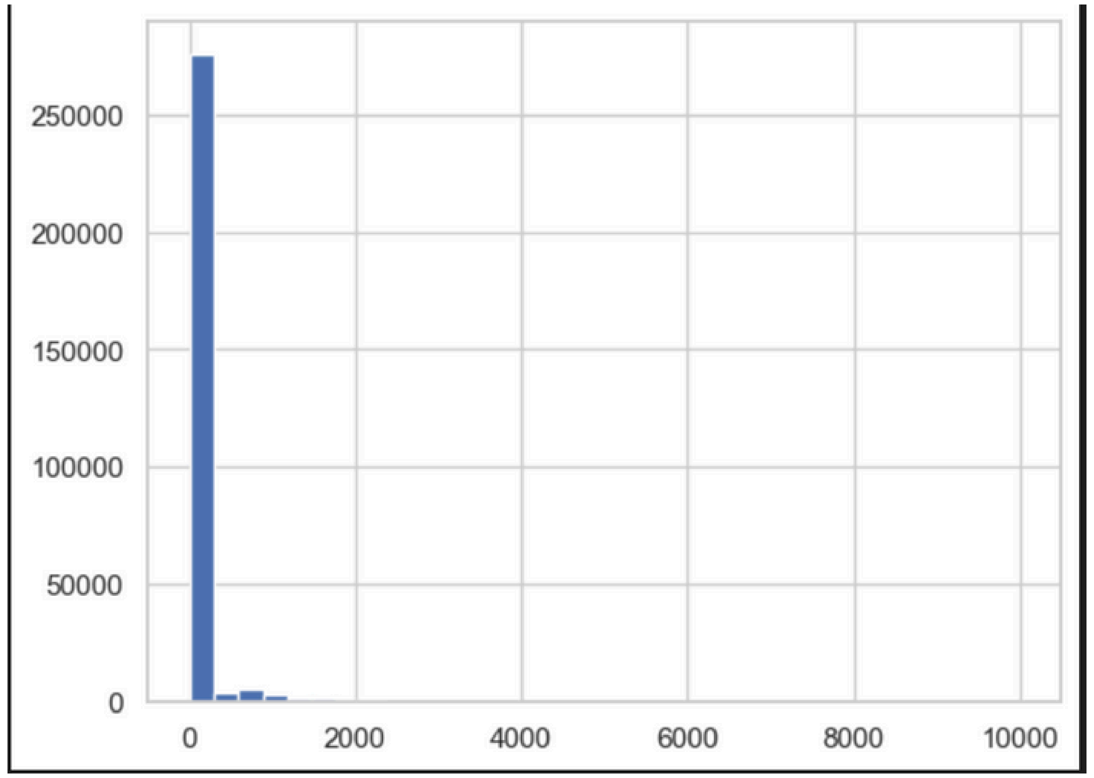
- For Random Forest and XGBoost, scaling (e.g., StandardScaler) does not improve and can reduce performance,
- Scaling was included only inside pipelines used for linear models, not for trees.

1. Approximately Normal Variables



- Applied Standard Scaler
  - weeks_worked_in_year
  - age
  - num_persons_worked_for_employer

2. Highly Right-Skewed Monetary Variables



- Applied Log1p
  - wage_per_hour
  - capital_gains
  - dividends_from_stocks

Feature Engineering (Categorical Features)

01.Dropping Irrelevant Features

a) Problem Understanding Logic

b) Redundancy of Information

c) Missing Rare/ Imbalanced

02. Binning / Grouping to Reduce Cardinality

- class_of_worker → grouped similar work classes into broader buckets
- marital_stat → simplified into fewer relationship categories
- (e.g., married / not married / separated)

03. Encoding Strategy

1. Low-cardinality features (≤10 unique values)

- One-Hot Encoding
  - Preserves interpretability
  - Safe for models with limited category counts

2. High-cardinality features (>10 unique values)

- Target Encoding
  - Avoids exploding feature dimensionality
  - Captures relationship between category and target income

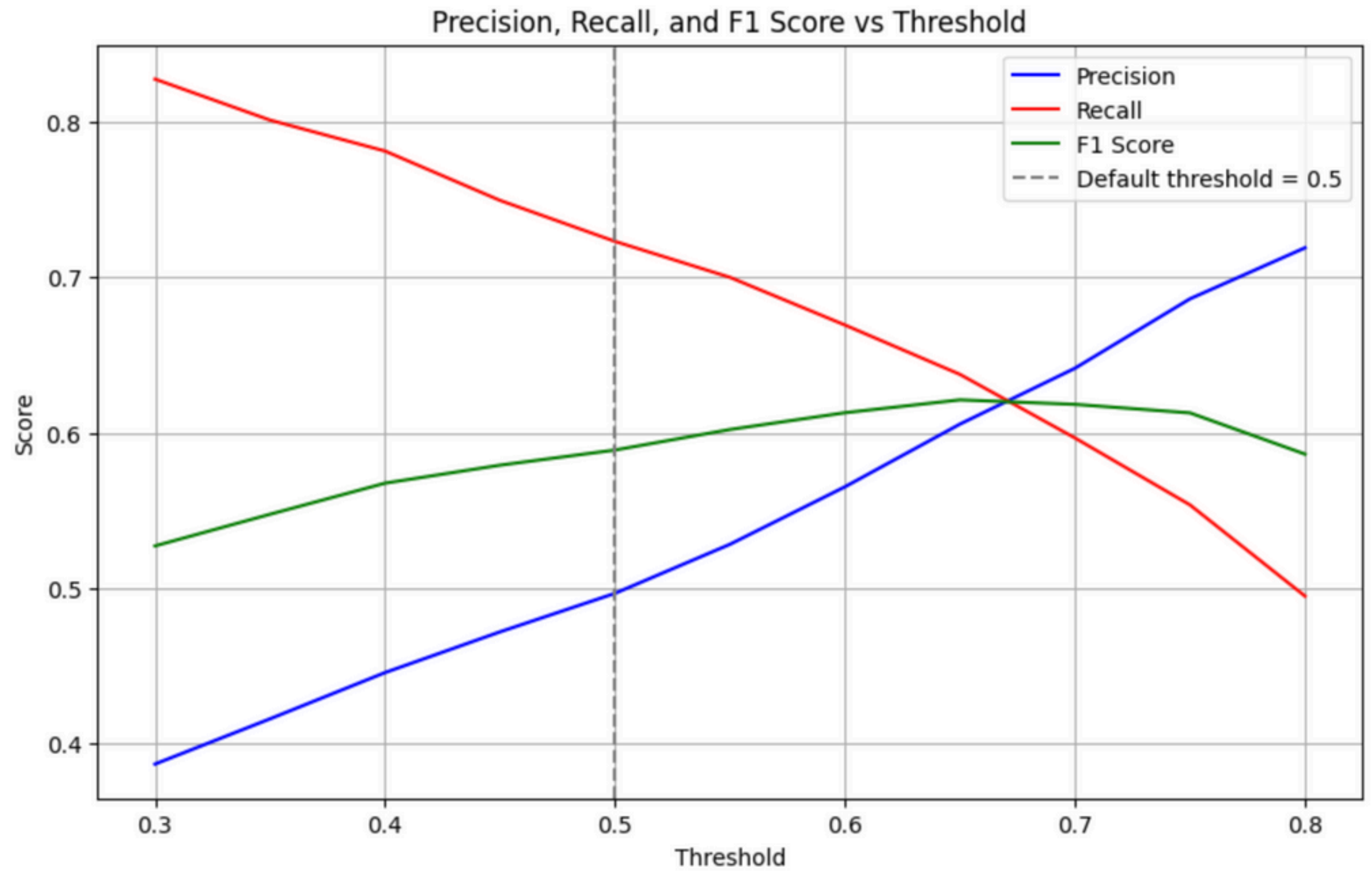
Model Evaluation (Val Set)

	Minority Class Precision	Minority Class Recall	Minority Class F-1 Score
LogisticRegression (default)	0.71	0.37	0.48
RandomForrest (Hyperparameter tuned)	0,76	0.44	0,56
XGBoost (Hyperparameter tuned)	0.50	0,72	0,59

Metrics & Threshold Optimization

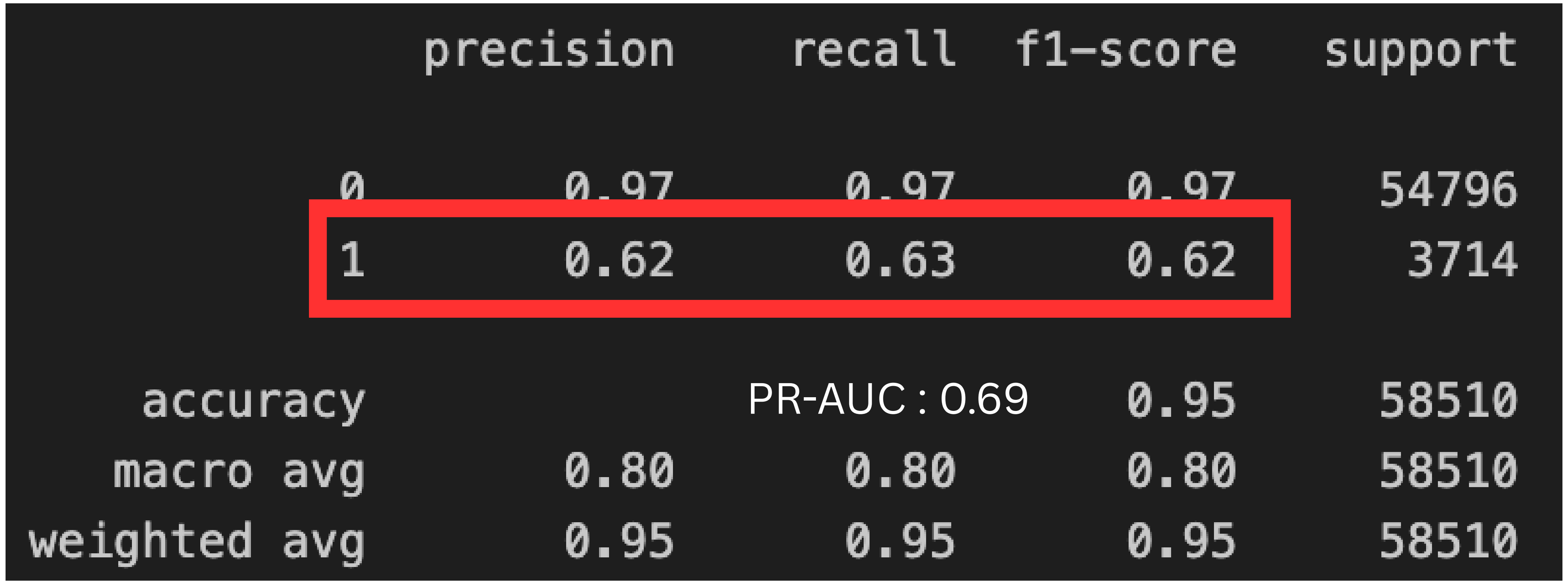
- Since the dataset is heavily imbalanced, need scoring to account for that
  - Accuracy is not enough → predicting False everytime gives ~93%
- Recall: how many high-income individuals the model correctly identifies.
  - Missing high-income individuals (low recall) is costly.
- F1-score: balances Precision & Recall → good for imbalanced data.
- PR-AUC: best summary metric when the positive class is rare.
- Of course, depending on the client's requirement the score during the training can alter

Threshold Optimization



- Default Threshold is not optimal for imbalanced set
- Increasing threshold increases the recall, but too low reduces the precision and therefore f-1 score
- Best F1 = 0.6224 at threshold = 0.6596

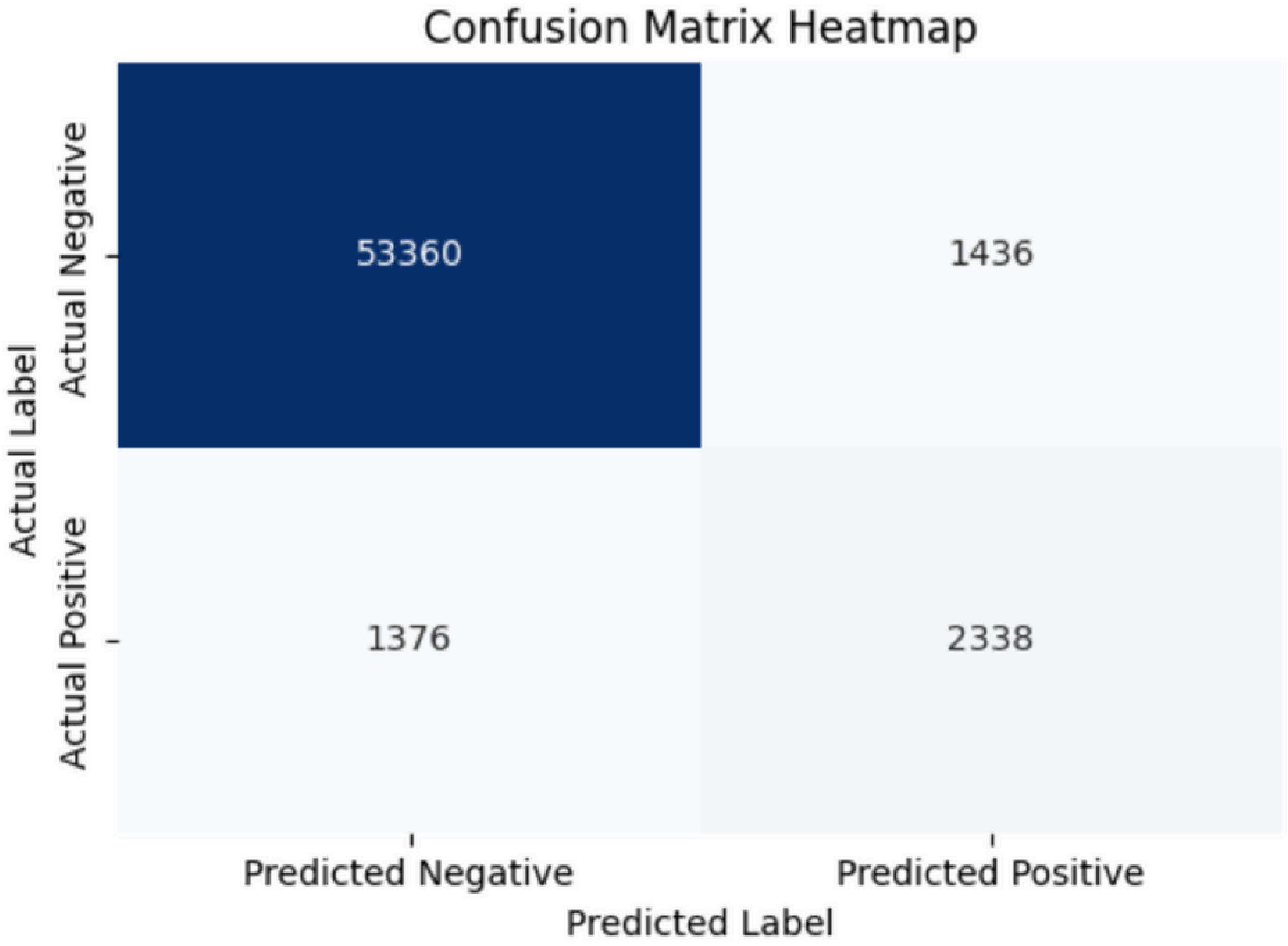
Final Model Evaluation (Test Set)



```
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     54796
           1       0.62      0.63      0.62      3714

    accuracy            PR-AUC : 0.69      0.95     58510
   macro avg       0.80      0.80      0.80     58510
weighted avg       0.95      0.95      0.95     58510
```



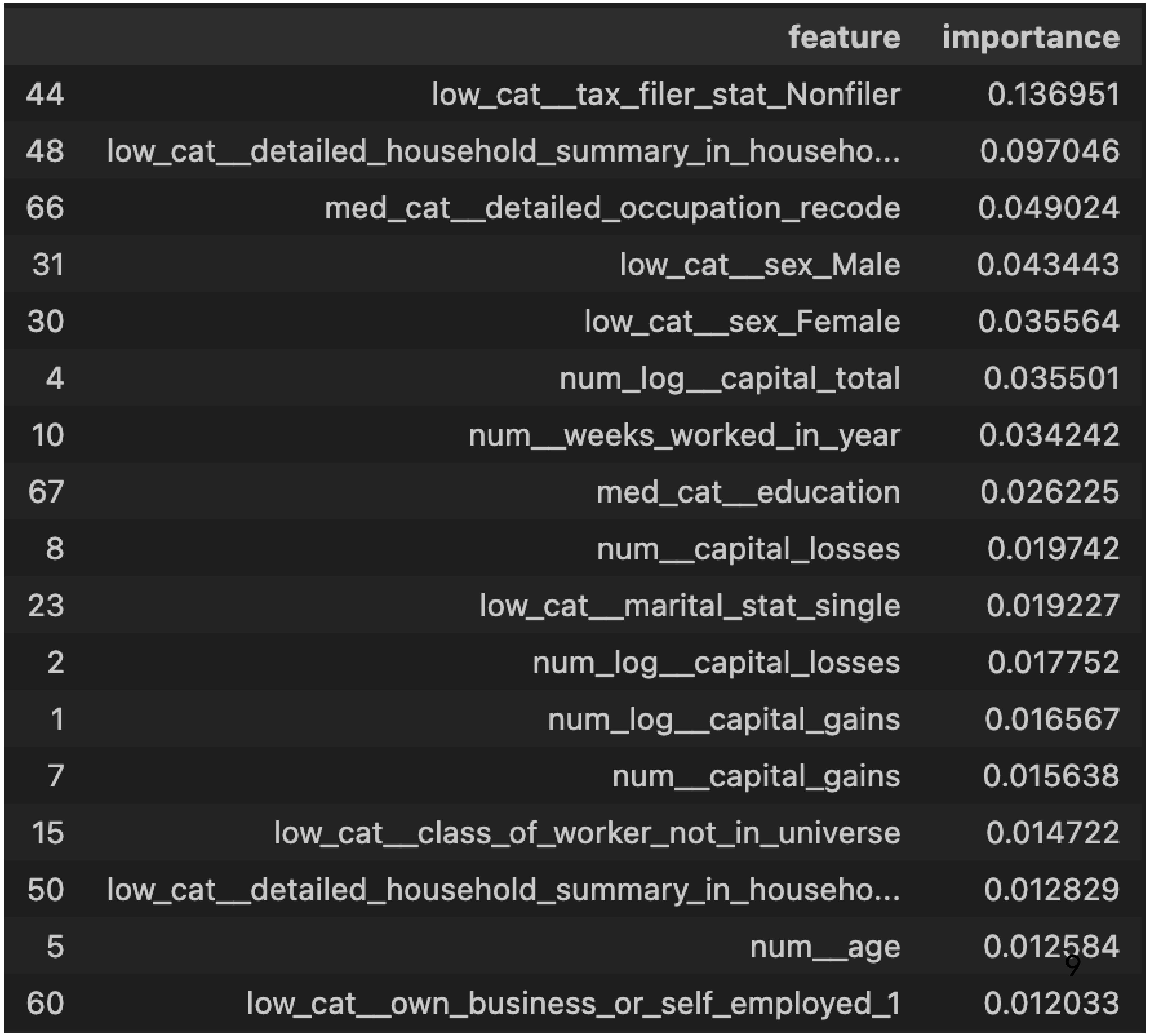
Confusion Matrix Heatmap

Actual Label	Predicted Negative	Predicted Positive
Actual Negative	53360	1436
Actual Positive	1376	2338

- **Strong Generalization**
  - Test-set performance is very close to train/validation results, no overfitting
- **Minority Metrics + PR-AUC**
  - The model is effective at detecting the rare high-income class without producing excessive false positives.
- **Confusion Matrix**
  - Most predictions fall into True Negatives and True Positives, showing the model correctly separates ≤50k and >50k incomes.
  - False Positives and False Negatives are relatively low, indicating balanced performance and good recall of the minority (>50k) class.

Feature Importance

- **Tax_Filer_Status**
  - Strongest predictor: individuals who do not file taxes overwhelmingly fall into the ≤50k group.
  - Provides a clear separation between stable vs unstable income patterns.
- **Detail_Summary_HouseHold**
  - Household role (e.g., child under 18) strongly indicates lower-income class membership.
- **Capital_Total**
  - Engineered feature combining gains, losses, and dividends.
  - Highly predictive of >50k income and more informative than capital_gains or losses alone.
- **Sex**
  - Meaningful contributor for both classes; reflects observed wage differences in the EDA and real-world income patterns.
- **Overall Insight**
  - Model relies on a combination of demographics, household structure, and economic indicators.
  - These features align logically with the socioeconomic factors that distinguish higher-income individuals.



	feature	importance
44	low_cat__tax_filer_stat_Nonfiler	0.136951
48	low_cat__detailed_household_summary_in_househo...	0.097046
66	med_cat__detailed_occupation_recode	0.049024
31	low_cat__sex_Male	0.043443
30	low_cat__sex_Female	0.035564
4	num_log__capital_total	0.035501
10	num__weeks_worked_in_year	0.034242
67	med_cat__education	0.026225
8	num__capital_losses	0.019742
23	low_cat__marital_stat_single	0.019227
2	num_log__capital_losses	0.017752
1	num_log__capital_gains	0.016567
7	num__capital_gains	0.015638
15	low_cat__class_of_worker_not_in_universe	0.014722
50	low_cat__detailed_household_summary_in_househo...	0.012829
5	num__age	0.012584
60	low_cat__own_business_or_self_employed_1	0.012033

Future Work and Limitations

Future Work

- Explore Deep Learning Models
- Expanded Feature Engineering
- Advanced Explainability (SHAP)

Limitation

- Class Imbalance
- No SMOTE / Resampling Techniques