Identify Key Socioeconomic & Demographic Factors to classify Medically Underserved Areas (MUAs) for Predictive Health Equity Modeling

Group 6

Ziyan Gao, Weiyu Qian, Saw Yee Tan,
Hannah Xu, Jiaying Zhong

Introduction & Datasets

01

02

Feature Selection

Prediction

03

04

Imbalance Data Detection

01

Introduction & Datasets

MUAs & Our Goals

What are MUAs? (Medically Underserved Areas/Populations)

- *Def*: regions where residents face a **shortage of primary healthcare services**.

Why It Matters:

- Scope: Over 82 million Americans live in MUAs.
- Impact: →worse health outcomes, e.g..
    - Higher rates of preventable diseases
    - Increased infant mortality, Shorter life expectancy

Our focus:

- **socioeconomic** and **demographic** factors for MUA designations.
- **Predictive** models to forecast underserved areas.
- Inform **equitable** healthcare interventions.

Datasets & Preparation

1 HRSA MUA/P

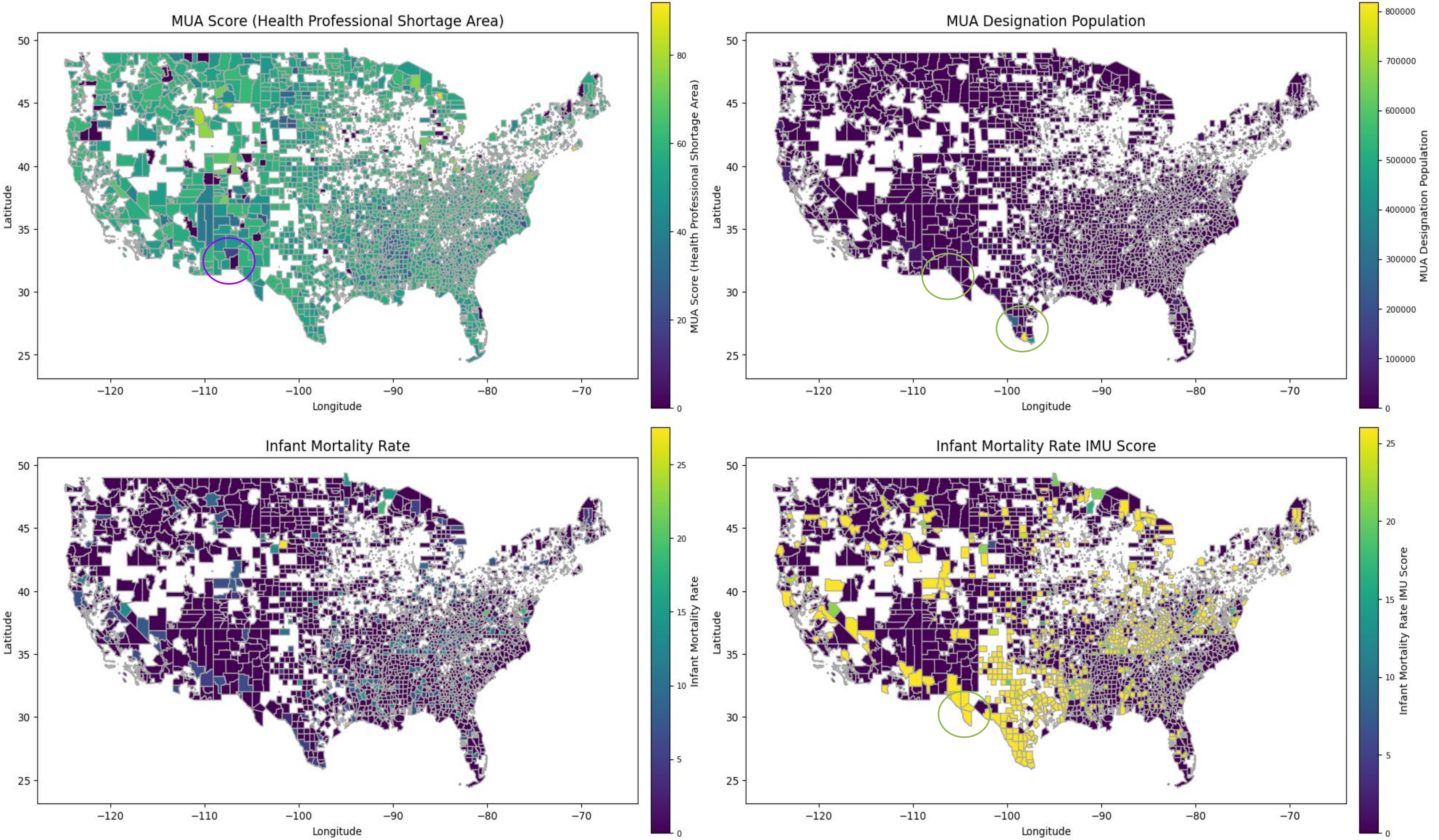
- Provider-to-population ratios, poverty levels, age distribution.
- Scope: Tract

2 ACS

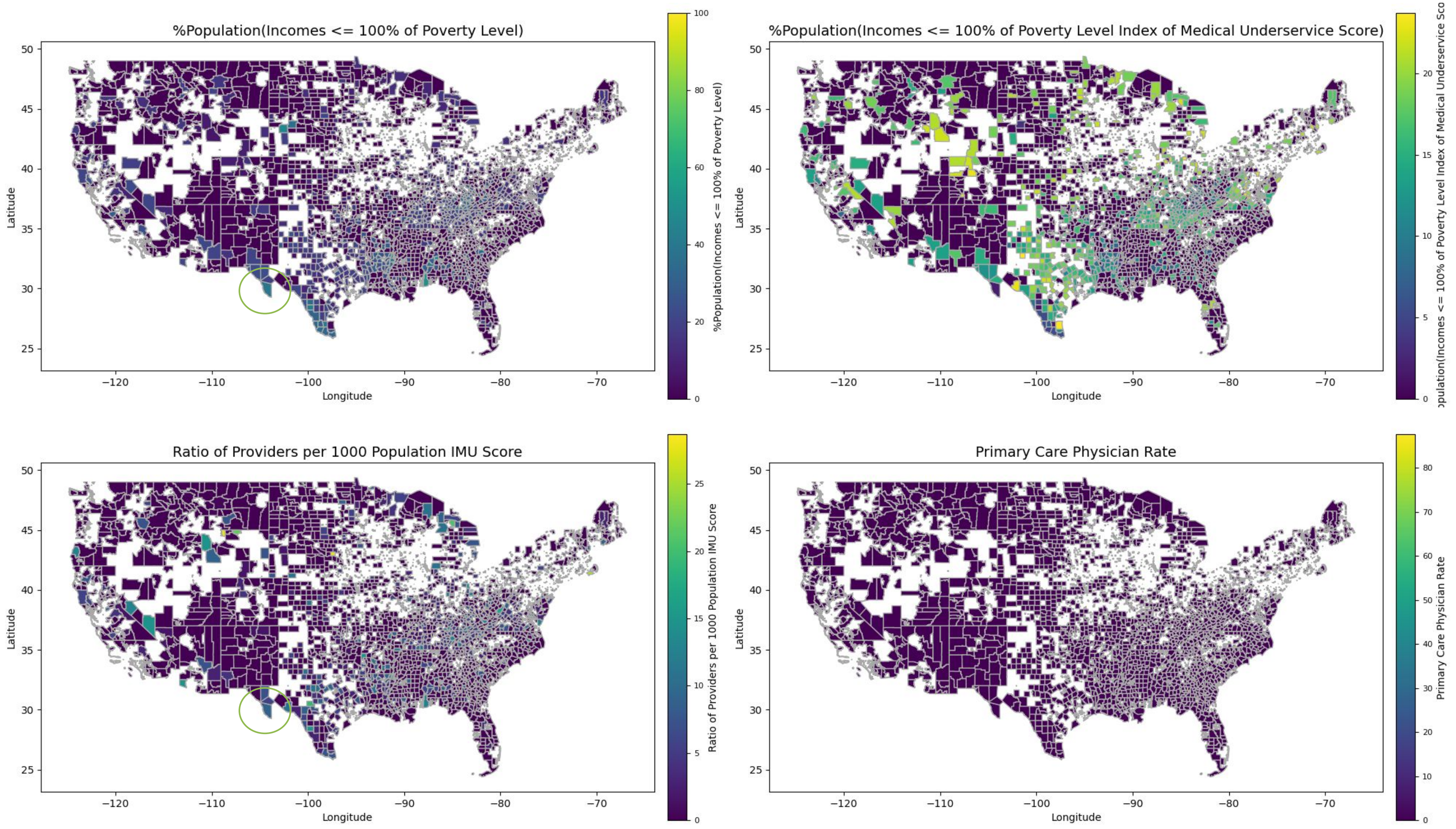
- Scope: ZIP Code
- Educational Attainment: (S1501).
- Health Insurance Coverage: Insurance status (S2701).
- Employment: Labor force participation and unemployment rates (S2301).

3 Integration

- Merge: Matched ZIP codes with census tracts
- Combined 2019–2022 data by tract to analyze MUA trends and predictors.



MUA Score (Health Professional Shortage Area)
MUA Score (Health Professional Shortage Area)
0
20
40
60
80
MUA Designation Population
MUA Designation Population
0
100000
200000
300000
400000
500000
600000
700000
800000
Infant Mortality Rate
Infant Mortality Rate
0
5
10
15
20
25
Infant Mortality Rate IMU Score
Infant Mortality Rate IMU Score
0
5
10
15
20
25
Latitude
25
30
35
40
45
50
Longitude
−120
−110
−100
−90
−80
−70



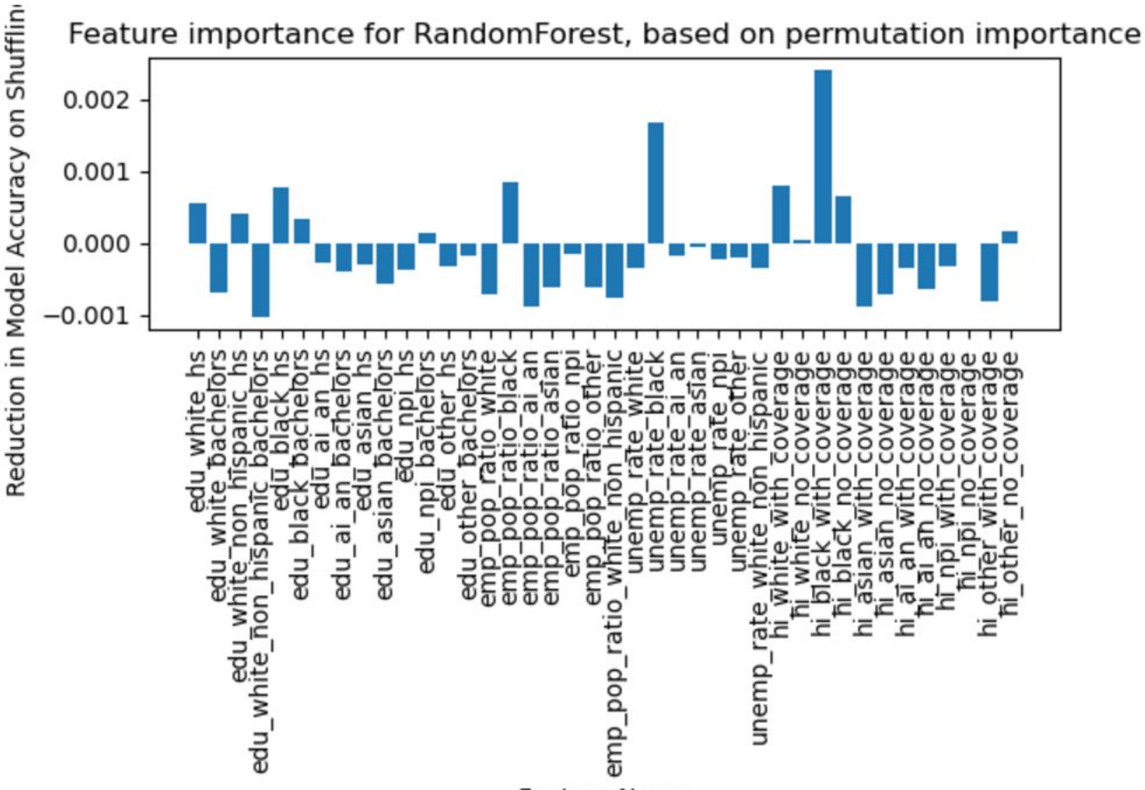
%Population(Incomes <= 100% of Poverty Level)
%Population(Incomes <= 100% of Poverty Level Index of Medical Underservice Score)
Ratio of Providers per 1000 Population IMU Score
Primary Care Physician Rate
Latitude
Longitude

02

Feature Selection

Method

1. Random Forest Classifier & Permutation feature Importance



2. Selected Features with Permutation Importance Greater than Zero

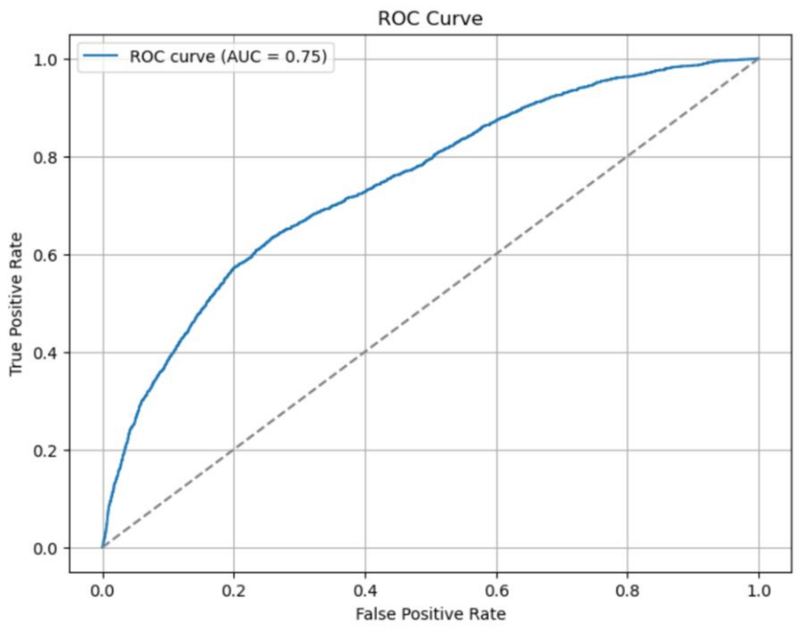
	feature	importance
30	hi_black_with_coverage	0.002399
22	unemp_rate_black	0.001671
15	emp_pop_ratio_black	0.000848
28	hi_white_with_coverage	0.000800
4	edu_black_hs	0.000776
31	hi_black_no_coverage	0.000657
0	edu_white_hs	0.000549
2	edu_white_non_hispanic_hs	0.000418
5	edu_black_bachelors	0.000346
39	hi_other_no_coverage	0.000167
11	edu_npi_bachelors	0.000131
29	hi_white_no_coverage	0.000036
37	hi_npi_no_coverage	-0.000012

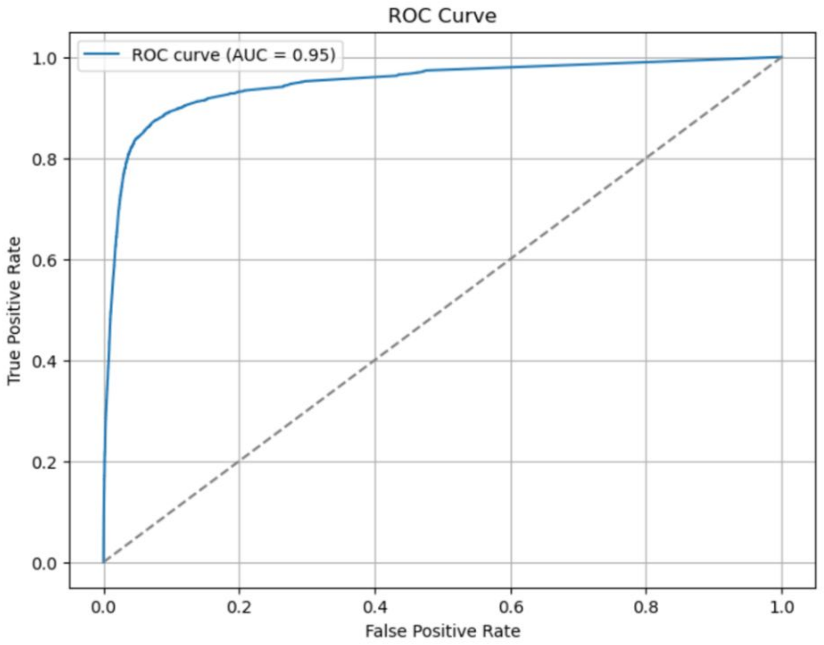
hi_black_with_coverage	The ratio of Black individuals with health insurance coverage.
unemp_rate_black	The unemployment rate among Black individuals.
emp_pop_ratio_black	The employment-to-population ratio for Black individuals.
hi_white_with_coverage	The ratio of White individuals with health insurance coverage.
edu_black_hs	The ratio of Black individuals with at least a high school education

Q: Why are there variables with permutation importance less than 0?

A: Some features are highly correlated with each other.

3. Retrain the Model with Selected Features

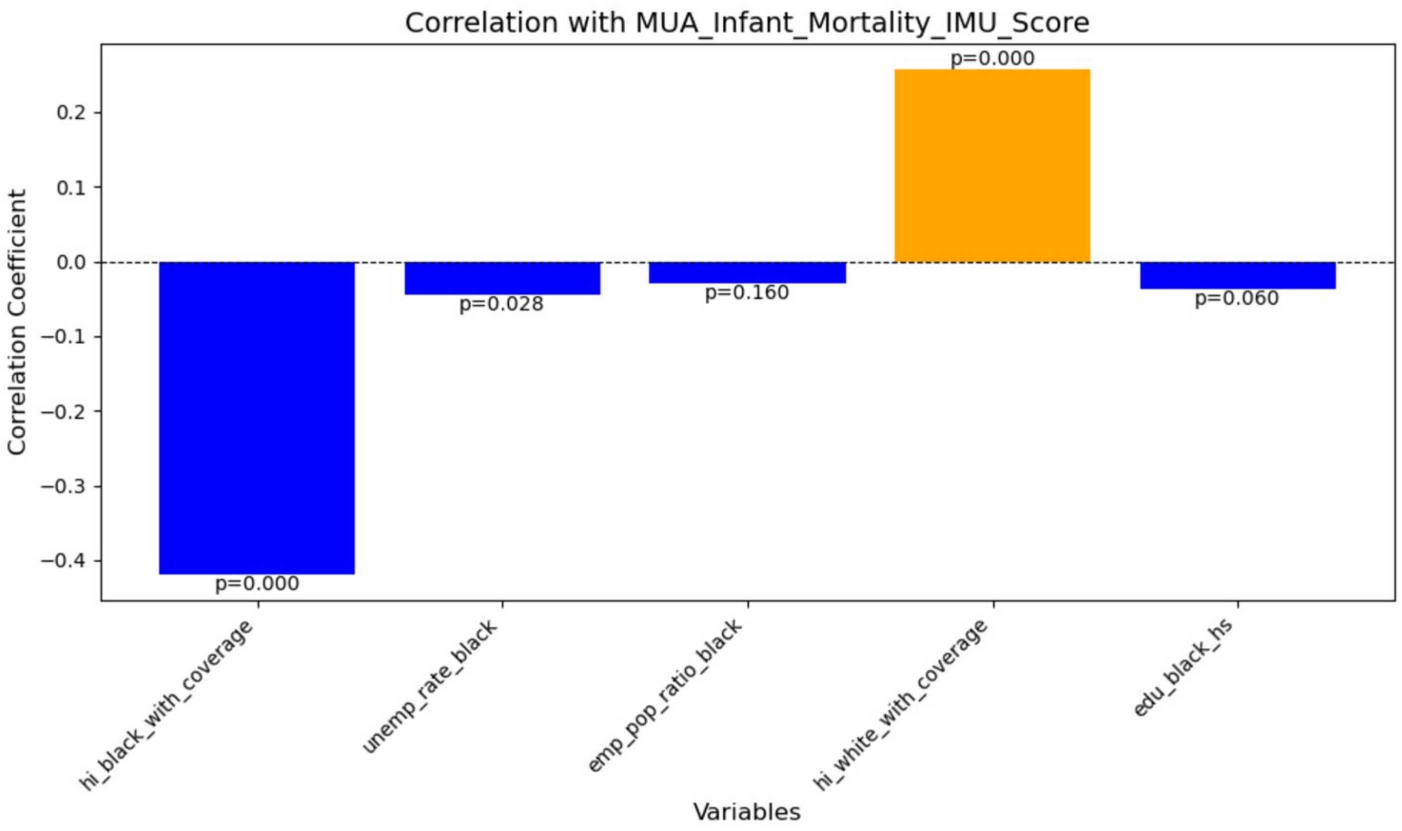




4. Interpretation

Conducted a correlation test with the top 5 variables and infant mortality rate.

hi_black_with_coverage	The ratio of Black individuals with health insurance coverage.
unemp_rate_black	The unemployment rate among Black individuals.
emp_pop_ratio_black	The employment-to-population ratio for Black individuals.
hi_white_with_coverage	The ratio of White individuals with health insurance coverage.
edu_black_hs	The ratio of Black individuals with at least a high school education



Correlation with MUA_Infant_Mortality_IMU_Score
Correlation Coefficient
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
p=0.000
p=0.028
p=0.160
p=0.000
p=0.060
hi_black_with_coverage
unemp_rate_black
emp_pop_ratio_black
hi_white_with_coverage
edu_black_hs
Variables

Correlations and P-values for Variables with MUA_Infant_Mortality_IMU_Score

Variable	Correlation (r)	P-Value	Interpretation
hi_black_with_coverage	-0.419	0.000000	Moderate negative correlation (significant)
unemp_rate_ black	-0.045	0.028474	Negligible negative correlation (significant)
emp_pop_ratio_black	-0.029	0.160080	Negligible negative correlation (not significant)
hi_white_with_coverage	0.257	0.000000	Weak positive correlation (significant)
edu_black_hs	-0.038	0.060319	Negligible negative correlation (not significant)

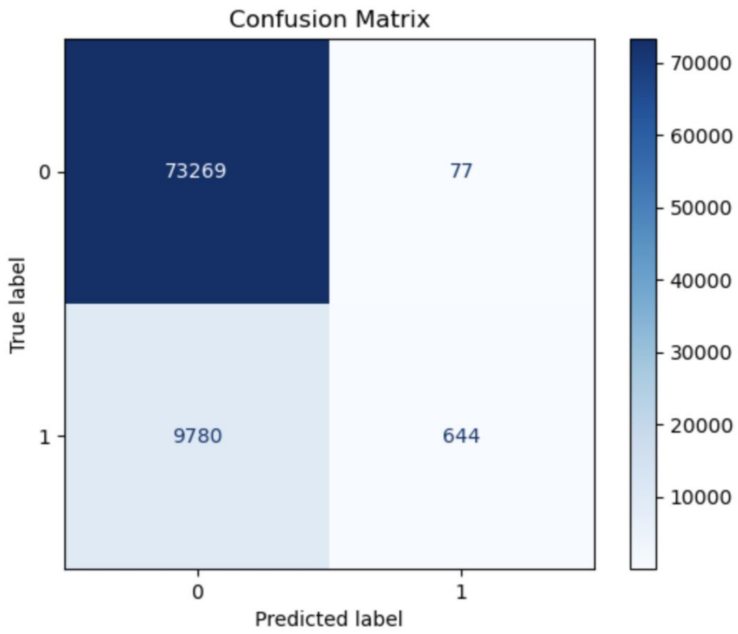
03

Prediction

Prediction

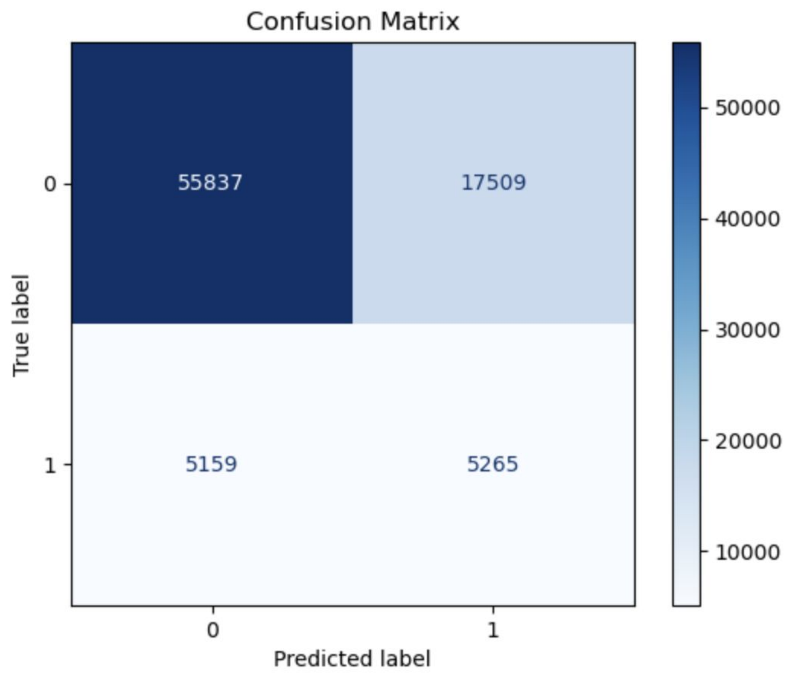
RandomForestClassifier

- Accuracy: 0.882 Precision: 0.893 Recall: 0.062
- Evaluation:
  The model performs exceptionally well in identifying negative cases but struggles significantly with detecting positive cases. Suitable for tasks where positive class detection is less critical, but not ideal for scenarios requiring high recall (e.g., critical anomaly detection).



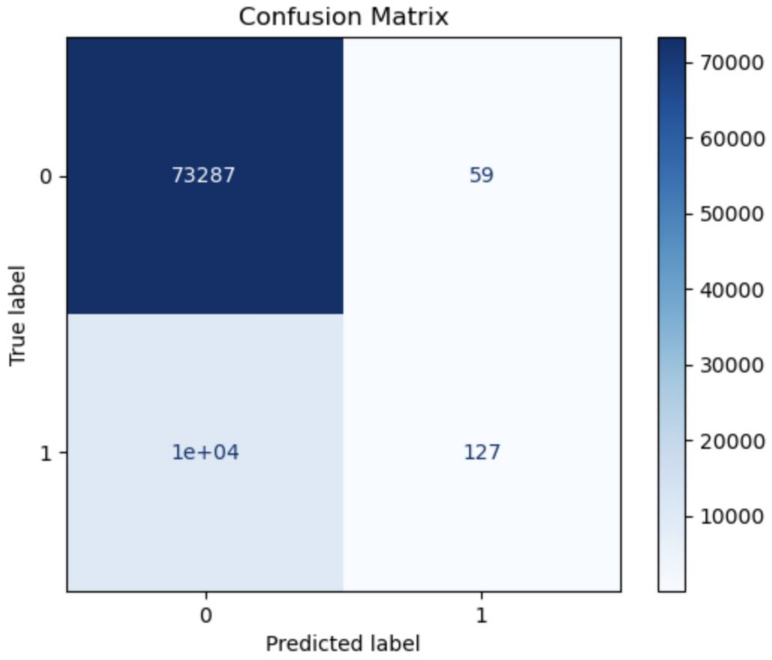
Logistic Regression

- Accuracy: 0.729, Precision: 0.231, Recall: 0.505
- Evaluation:
  Logistic Regression achieves a higher recall for the positive class, but its precision is low, indicating a tendency to overpredict the positive class.
  Suitable for tasks where higher recall is essential, but precision needs improvement for better reliability.



GradientBoostingClassifier

- Accuracy: 0.876, Precision: 0.683, Recall: 0.012
- Evaluation:
  The model performs well for negative cases but fails to recall positive cases, making it unsuitable for applications where detecting positive cases is critical.



04

Imbalance Data Detection

Check for imbalance

```
In train data, Class 0.0: 190832 samples (94.94%)
In train data, Class 1.0: 10168 samples (5.06%)
In test data, Class 0.0: 47615 samples (94.75%)
In test data, Class 1.0: 2636 samples (5.25%)
```

High percentage not MUA & Low percentage MUA

- MUA/P designation data is imbalanced.
- The accuracy value might be misleading.

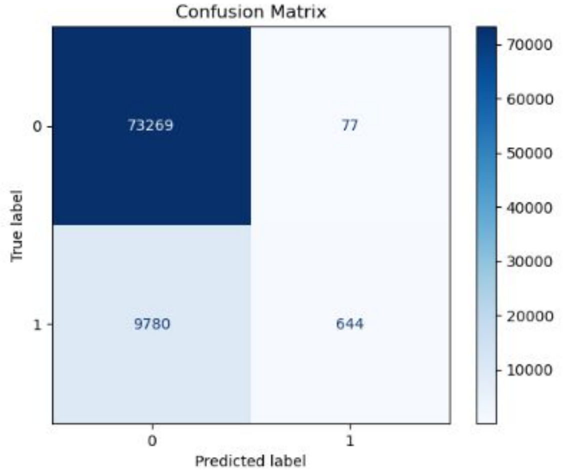
Resolve this by using resampling tech

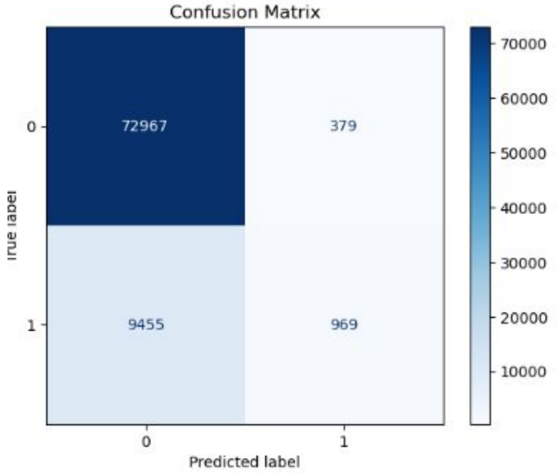
Accuracy: 0.882, Precision: 0.893, Recall: 0.062

After oversampling only the training data and get a balanced training data set.

We get improved precision, recall, and AUC for the 2023 prediction data.

Accuracy: 0.883, Precision: 0.719, Recall: 0.093





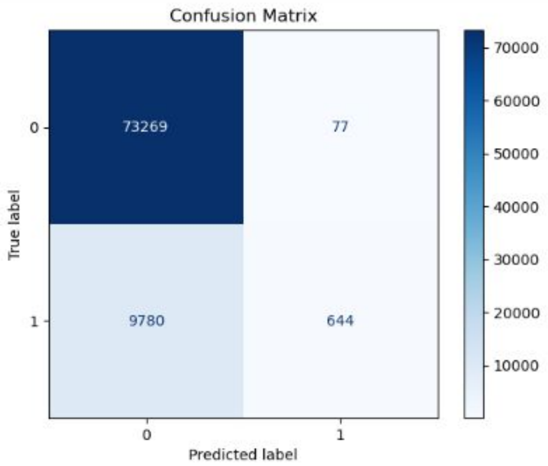
Resolve this by using resampling tech

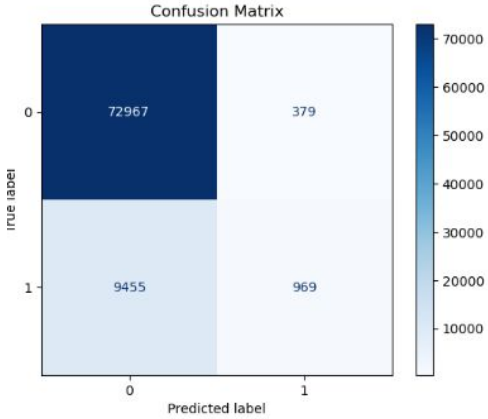
Accuracy: 0.882, Precision: 0.893, Recall: 0.062

After oversampling only the training data and get a balanced training data set.

We get improved precision, recall, and AUC for the 2023 prediction data.

Accuracy: 0.883, Precision: 0.719, Recall: 0.093





Thank You!

Next Group Please