EQUITY IN HEALTHCARE

Alain Siddiqui, Diego Maldonado, Faizah Khan,
Akari Kojima, Mariah N Cornelio

Metastatic Cancer

BACKGROUND AND OBJECTIVES

BACKGROUND

This project helps explore the challenges that healthcare inequity has posed to patients in the past. The dataset and project outcomes are useful to help understand if disparate treatments exist and the potential biases behind them that drive it.

OBJECTIVE

Utilizing a unique oncology dataset, developing a model to predict if patients received metastatic cancer diagnosis within 90 days of screening. Goal is to build models that detect relationships between patient demographics and likelihood of receiving timely treatment.

RESEARCH QUESTIONS

Does the demographic of a patient affect the likelihood of receiving a timely diagnosis?

METHODOLOGY

- Background/Objectives
- EDA - Exploratory Data Analysis
- Baselines
- Data Preprocessing
- Iterative Modeling
- Final Model
- Challenges
- Future Work/ Limitations

EDA: DATA INSPECTION

SHAPE:

- 12,906 rows
- 83 columns...

DATA TYPES:

- 11 object variables
- 72 integer/float variables

IMPORTANT FEATURES:

- DiagPeriodL90D: 1 = diagnosed in 90 days, 0 = not diagnosed in 90 days
- Population, density, age and income groups, race, insurance and payer type (Medicaid, commercial, etc.)

DESCRIBE STATISTICS:

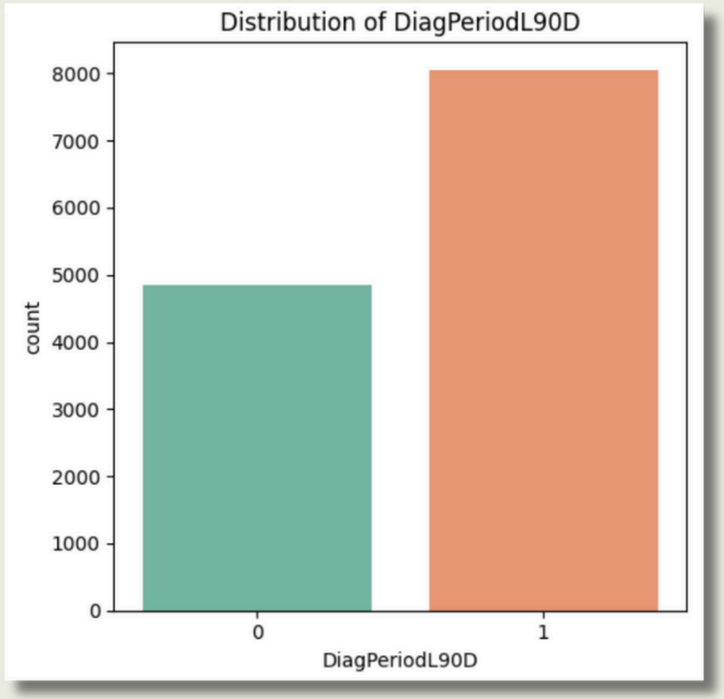
	patient_id	patient_zip3	patient_age	bmi	population	density
count	12906.000000	12906.000000	12906.000000	3941.000000	12905.000000	12905.000000
mean	547381.196033	573.754300	59.183326	28.984539	20744.441237	1581.950419
std	260404.959974	275.447534	13.335216	5.696906	13886.903756	2966.305306
min	100063.000000	101.000000	18.000000	14.000000	635.545455	0.916667
25%	321517.000000	331.000000	50.000000	24.660000	9463.896552	171.857143
50%	543522.000000	554.000000	59.000000	28.190000	19154.190480	700.337500
75%	772671.750000	846.000000	67.000000	32.920000	30021.278690	1666.515385
max	999896.000000	999.000000	91.000000	85.000000	71374.131580	21172.000000

MISSING VARIABLES:

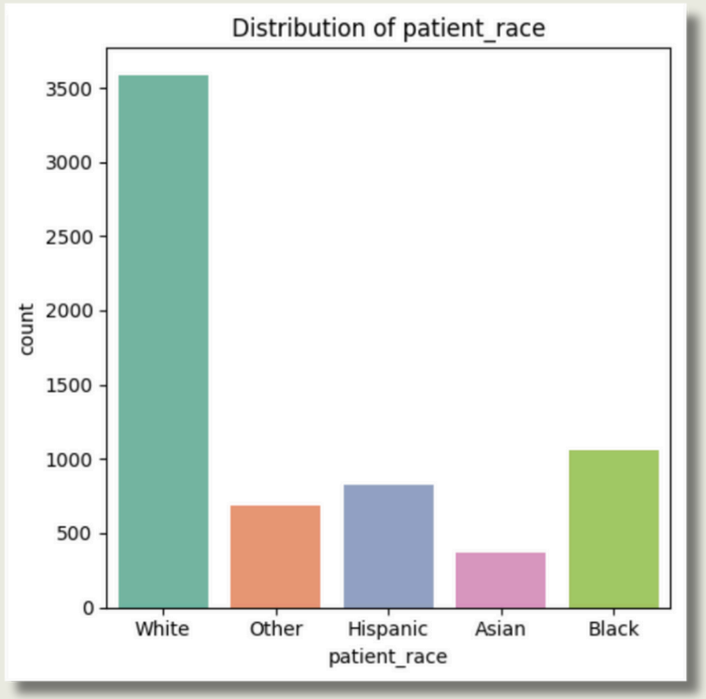
	Missing Values	Percent Missing
patient_id	0	0.00
patient_race	6385	49.47
payer_type	1803	13.97
patient_state	51	0.40
patient_zip3	0	0.00
patient_age	0	0.00
patient_gender	0	0.00
bmi	8965	69.46
breast_cancer_diagnosis_code	0	0.00
breast_cancer_diagnosis_desc	0	0.00
metastatic_cancer_diagnosis_code	0	0.00
metastatic_first_novel_treatment	12882	99.81
metastatic_first_novel_treatment_type	12887	99.81

EDA: DISTRIBUTIONS

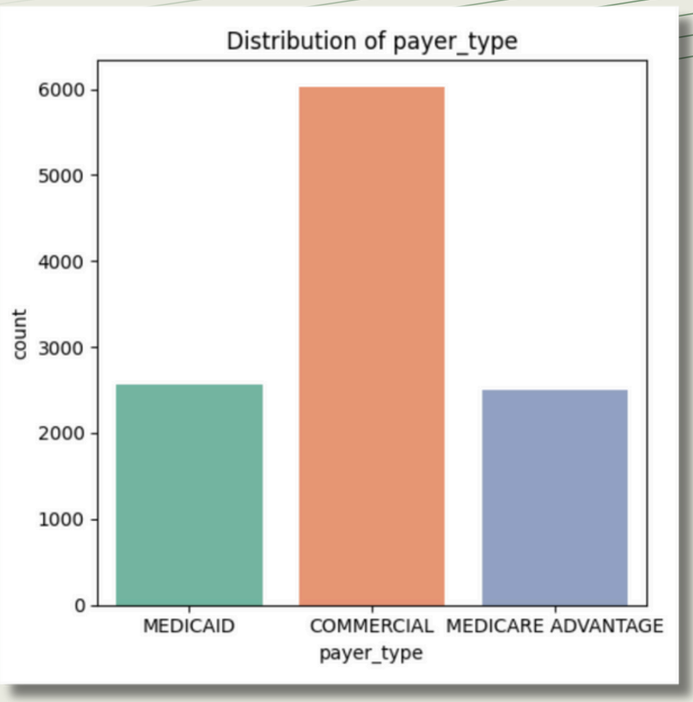
IMPORTANT CATEGORICAL VARIABLES



Our Target Variable



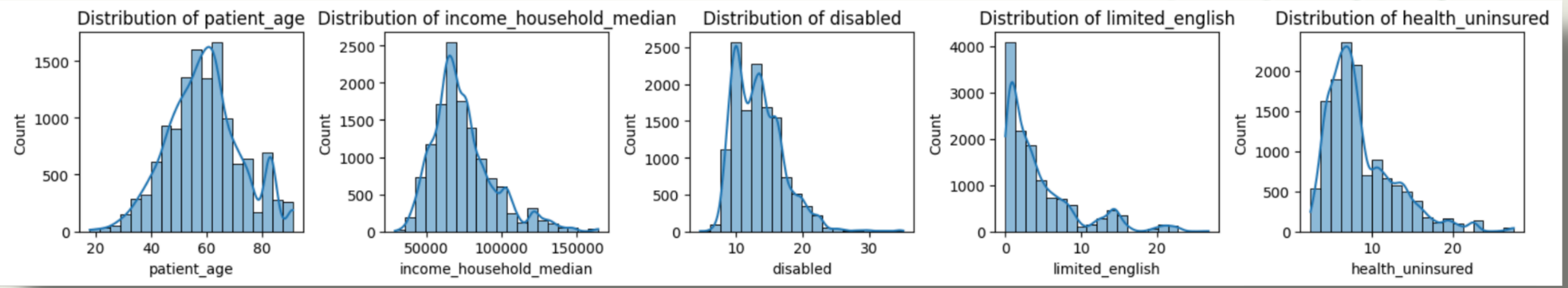
Race of Patient



Payer Type

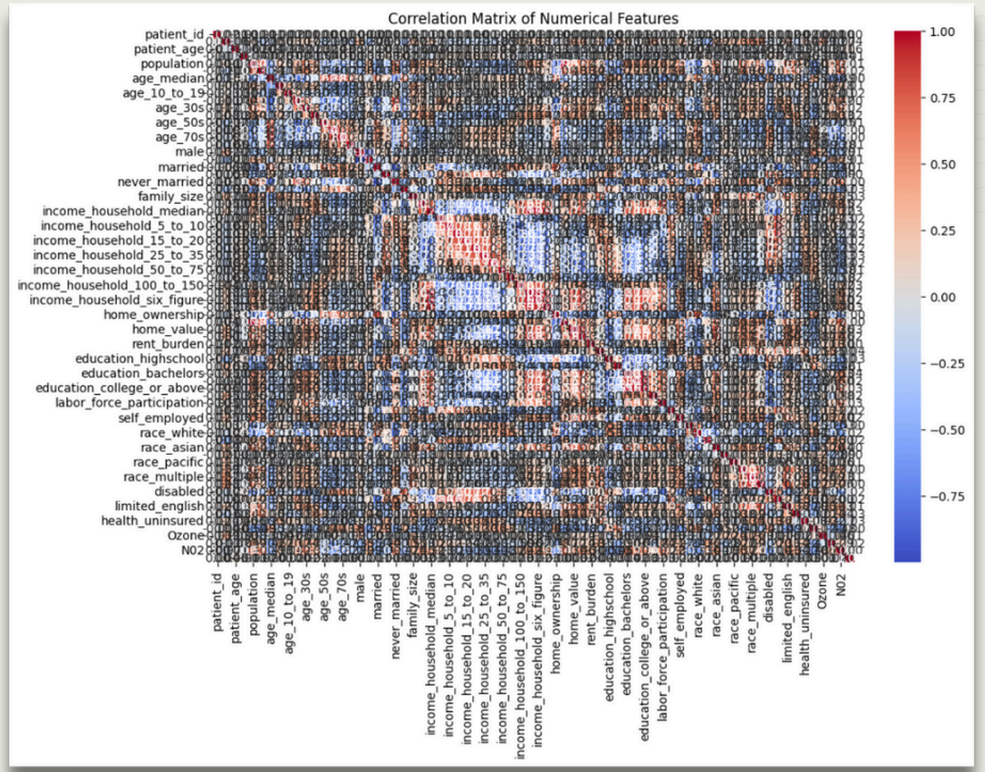
EDA: DISTRIBUTIONS

IMPORTANT NUMERICAL VARIABLES



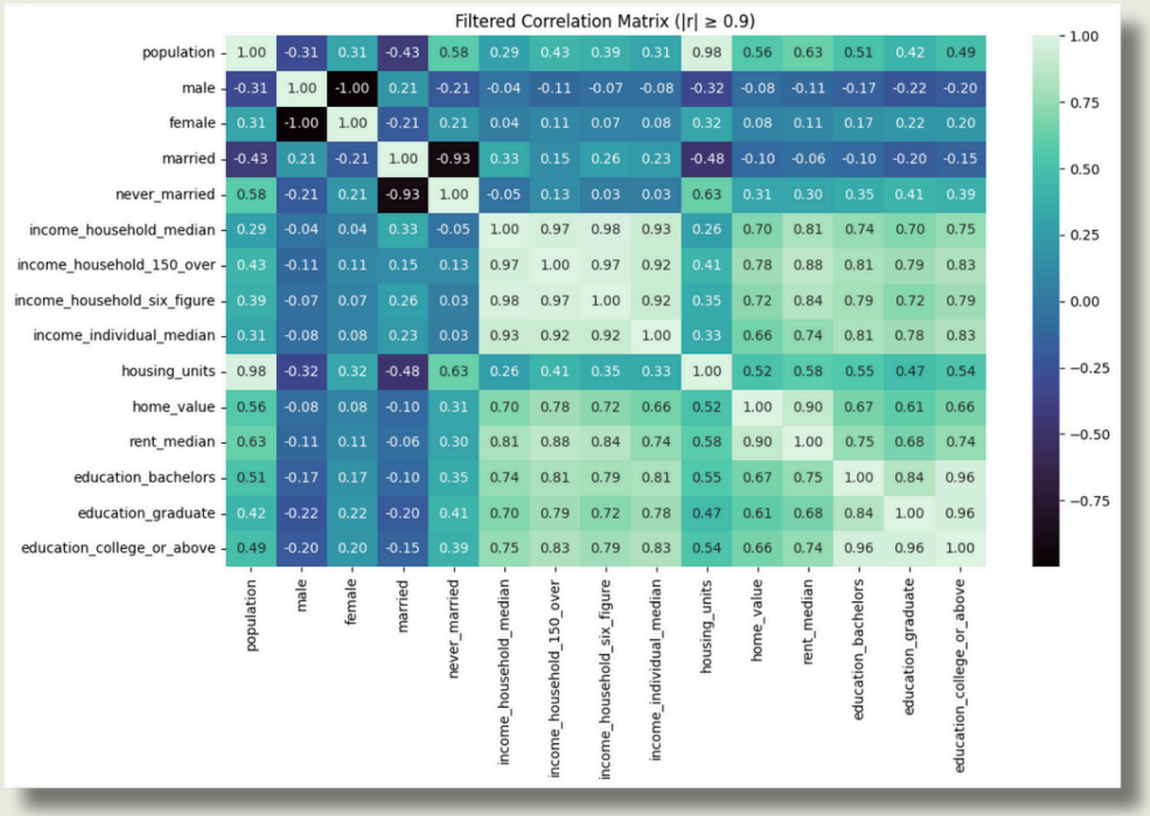
Some seem normally distributed, some are also skewed. This will help with data-preprocessing.

EDA: CORRELATION



EDA: CORRELATION

MULTICOLLINEARITY THRESHOLD: >= 0.8



Making it a little easier to read:

Feature 1	Feature 2	Correlation
female	male	−1.000000
housing_units	population	0.982847
income_household_six_figure	income_household_median	0.979407
income_household_six_figure	income_household_150_over	0.973406
income_household_150_over	income_household_median	0.966081
education_college_or_above	education_bachelors	0.961805
education_college_or_above	education_graduate	0.958673
income_individual_median	income_household_median	0.928126
never_married	married	−0.926330
income_individual_median	income_household_150_over	0.915547
income_individual_median	income_household_six_figure	0.915105
rent_median	home_value	0.903422
age_70s	age_median	0.887043

Note: For the sake of fitting the picture, the matrix on the left shows a threshold of >= 0.9

BASELINE MODEL

LOGISTIC REGRESSION

BEFORE TRAINING OUR BASELINE:

- Remove unnecessary columns like Patient_ID
- Get rid of any remaining null values
- Encode categorical data (label encoding vs frequency encoding)

METRIC HIERARCHY:

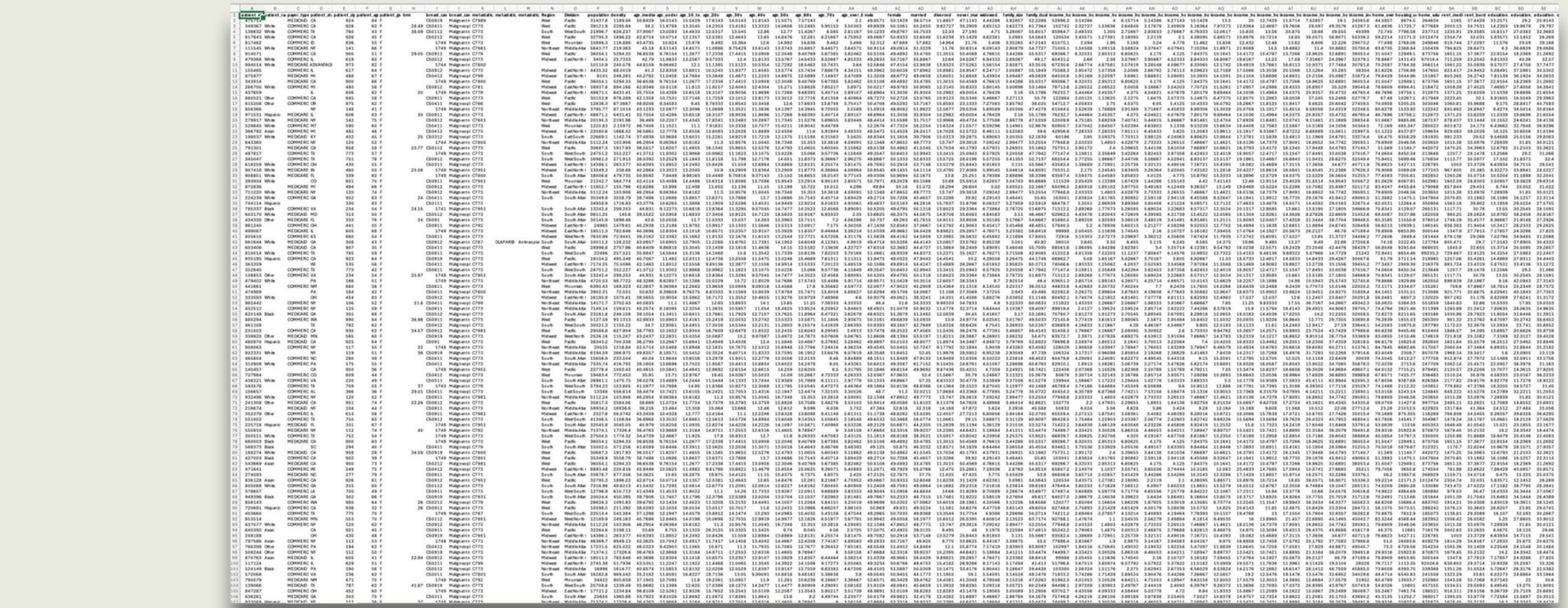
Recall > F1 > Accuracy > Precision

Note: Because our classes are unbalanced, we had to balance our class weight during each model training by stratifying by our target variable

Our baseline results show that...

- Overall model
  - Accuracy: 0.52
  - Precision: 0.64
  - Recall: 0.53
  - F1: 0.58
- Class 0
  - Precision:0.40
  - Recall: 0.51
  - F1: 0.45
- Class 1 (Performed better)
  - Precision: 0.65
  - Recall: 0.54
  - F1: 0.59

OUR DATA (ZOOMED OUT)...



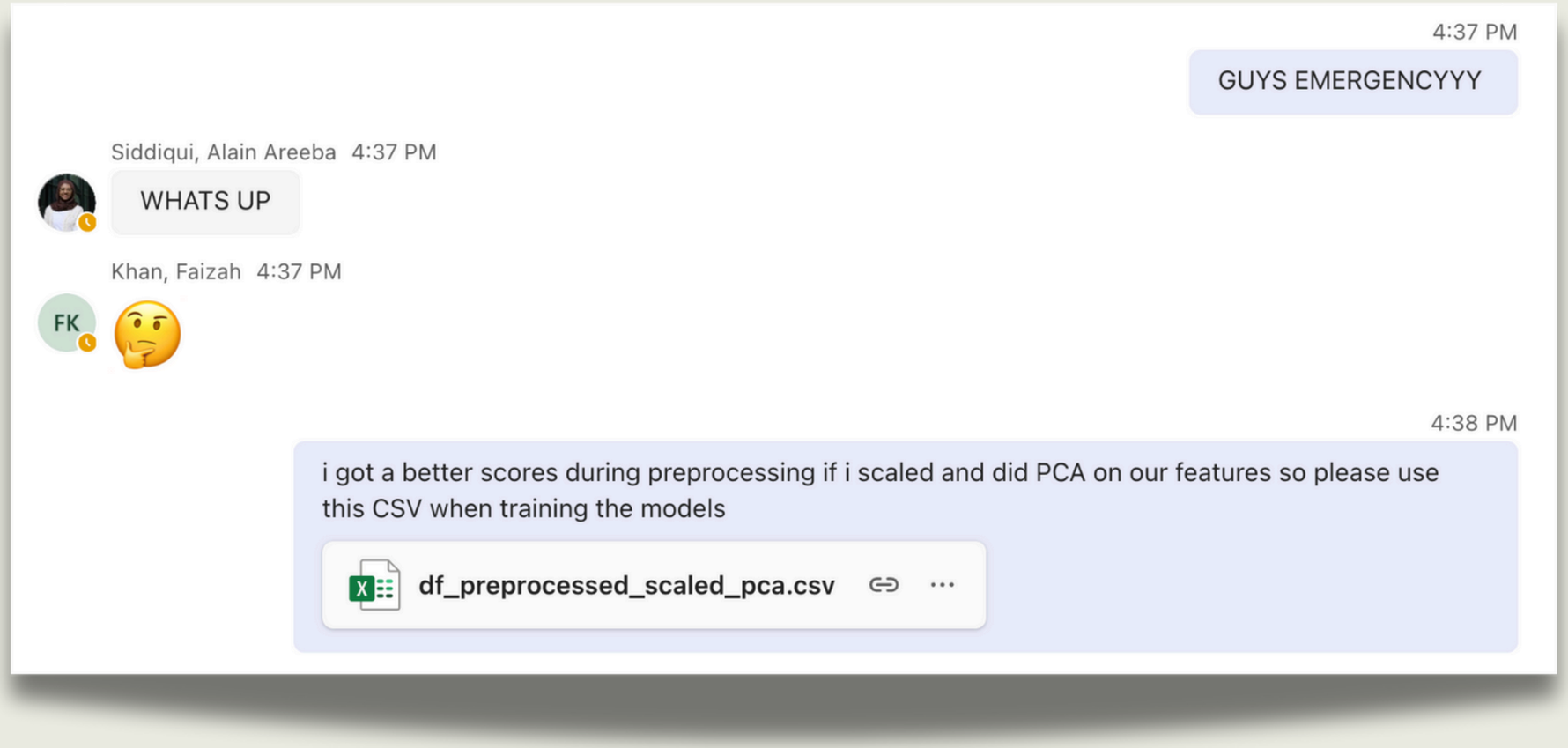
DATA PREPROCESSING

GOAL: USE DOMAIN KNOWLEDGE TO FIND RELEVANT AND NON-REDUNDANT FEATURES THAT RELATE TO OUR TARGET VARIABLE (DIAGNOSIS IN 90 DAYS)

TASK	***STRATEGY***
Feature Selection	Remove the columns with the most missing data and descriptive categorical values like Patient_ID and Breast_cancer_diagnosis_description.
Multicollinearity Handling	Remove redundant features that are multi-collinear with each other using our correllation matrix. Also remove features that have close to 0 correlation, but be careful because having a low correlation doesn't always mean that it is not important.
Outlier/Missing Value Handling	We had no outliers! Yay! But for the categorical data with missing value, we did not remove the rows. We just imputed with an "Unknown" category because these details are still important to our target variable.
Feature Engineering	Created a Poverty_Percentage, Older_Age_Percentage, and Low_Income_Percentage to show interactions between features.
Encoding	Ended up using label-encoding instead.
Scaling/Normalization	We did not use our Normalized dataset.

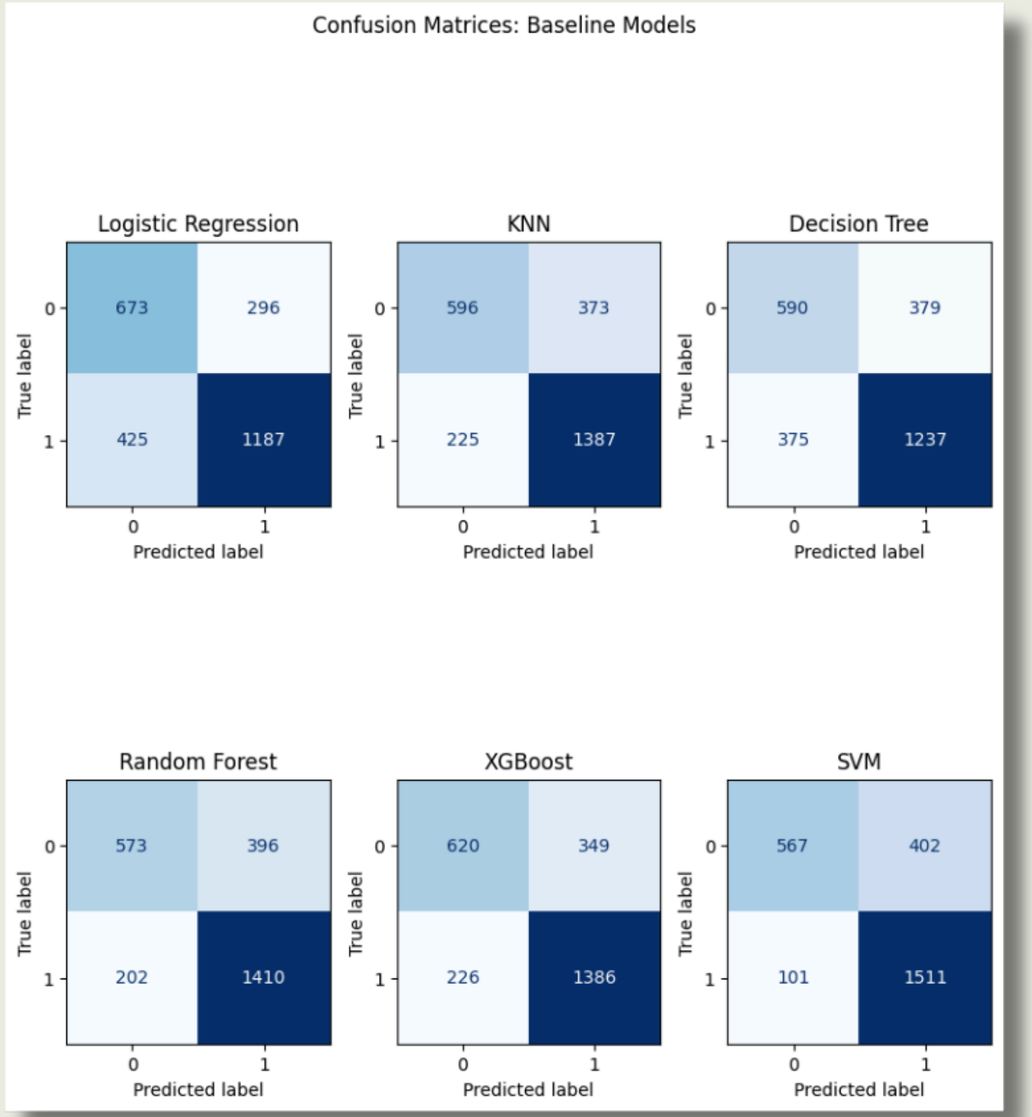
REDUCED FEATURES: 83 → 42

BUT WAIT!


4:37 PM
GUYS EMERGENCYYY
Siddiqui, Alain Areeba 4:37 PM
WHATS UP
Khan, Faizah 4:37 PM
FK
4:38 PM
i got a better scores during preprocessing if i scaled and did PCA on our features so please use this CSV when training the models
df_preprocessed_scaled_pca.csv

ITERATIVE MODELING

DiagPeriodL90D	proportion
1	0.624641
0	0.375359



Logistic Regression

	precision	recall	f1-score	support
0	0.61	0.69	0.65	969
1	0.80	0.74	0.77	1612

KNN

	precision	recall	f1-score	support
0	0.73	0.62	0.67	969
1	0.79	0.86	0.82	1612

Decision Tree

	precision	recall	f1-score	support
0	0.61	0.61	0.61	969
1	0.77	0.77	0.77	1612

Random Forest

	precision	recall	f1-score	support
0	0.74	0.59	0.66	969
1	0.78	0.87	0.83	1612

XGBoost

	precision	recall	f1-score	support
0	0.73	0.64	0.68	969
1	0.80	0.86	0.83	1612

SVM

	precision	recall	f1-score	support
0	0.85	0.59	0.69	969
1	0.79	0.94	0.86	1612

FIXING THE IMBALANCE

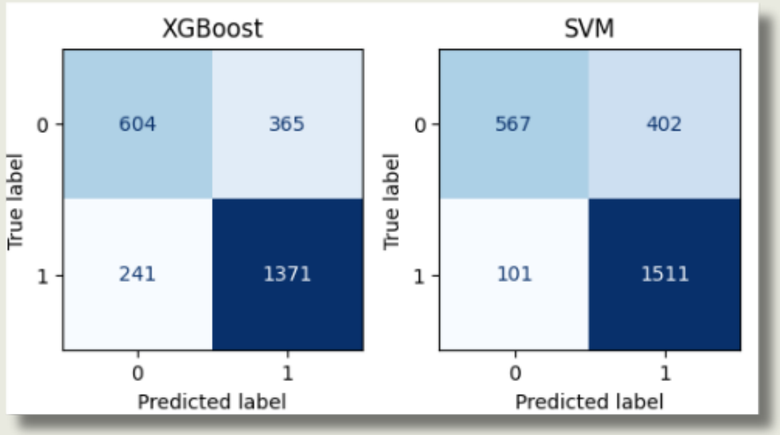
SMOTE

XGBoost

	precision	recall	f1-score	support
0	0.71	0.62	0.67	969
1	0.79	0.85	0.82	1612

SVM

	precision	recall	f1-score	support
0	0.85	0.59	0.69	969
1	0.79	0.94	0.86	1612



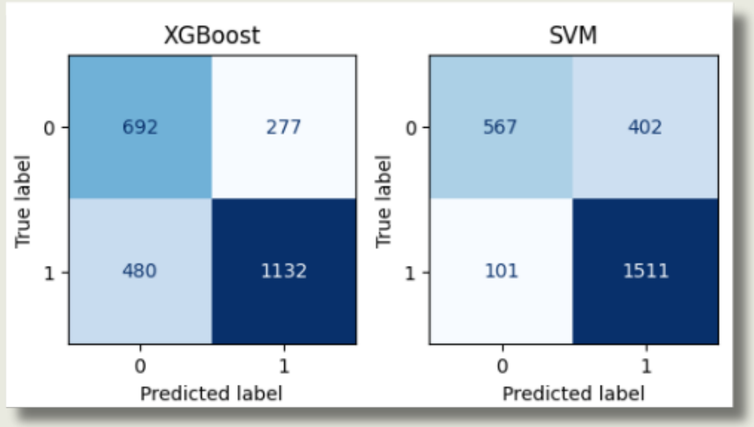
RUS

XGBoost

	precision	recall	f1-score	support
0	0.59	0.71	0.65	969
1	0.80	0.70	0.75	1612

SVM

	precision	recall	f1-score	support
0	0.85	0.59	0.69	969
1	0.79	0.94	0.86	1612



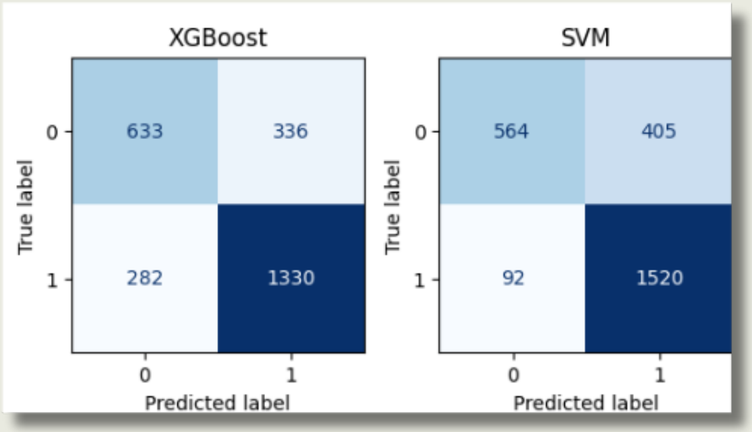
ADASYN

XGBoost

	precision	recall	f1-score	support
0	0.69	0.65	0.67	969
1	0.80	0.83	0.81	1612

SVM

	precision	recall	f1-score	support
0	0.86	0.58	0.69	969
1	0.79	0.94	0.86	1612



HYPERPARAMETER TUNING

XGBoost

n_estimators: Number of trees
max_depth: Depth of the trees
learning_rate: How much each tree contributes to the final model
scale_pos_weight: How much importance given to minority class

SVM

C: Trade-off between margin size and classification error
kernel: Type of function used to transform data
gamma: Influence of training samples
class_weight: Handles class imbalance

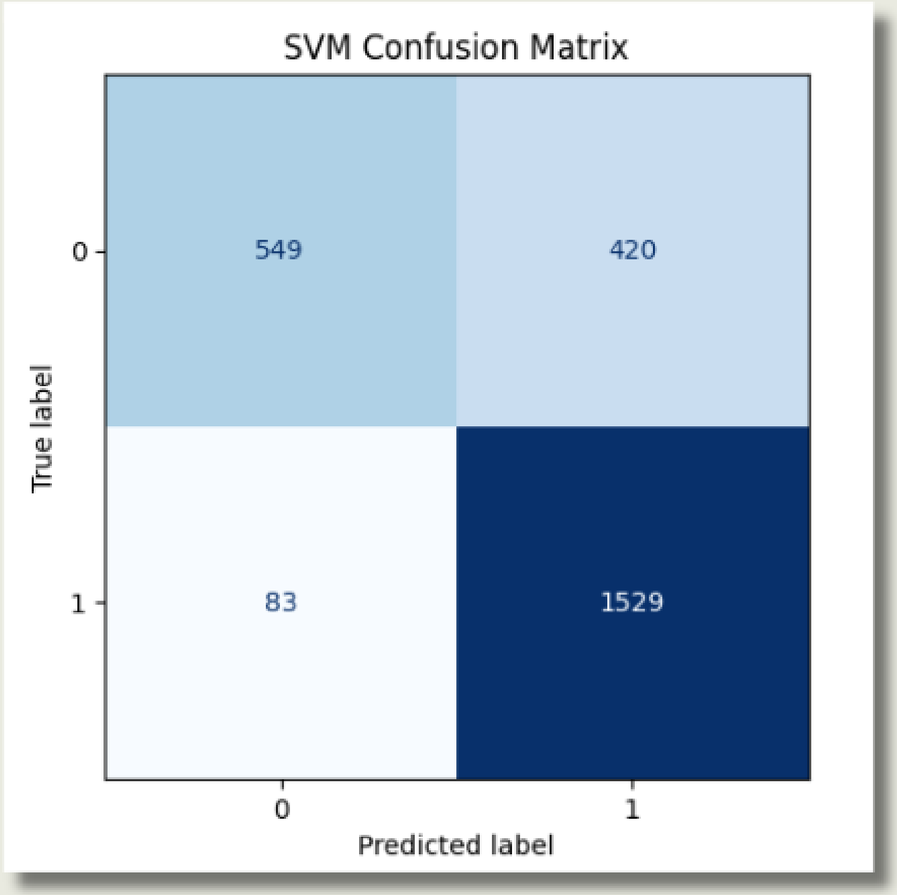
SVM (RUC + GRID SEARCH)

SVM

	precision	recall	f1-score	support
0	0.87	0.57	0.69	969
1	0.78	0.95	0.86	1612
accuracy			0.81	2581
macro avg	0.83	0.76	0.77	2581
weighted avg	0.82	0.81	0.79	2581

SVM Accuracy: 0.8051
SVM Precision: 0.7845
SVM Recall: 0.9485
SVM F1 Score: 0.8587

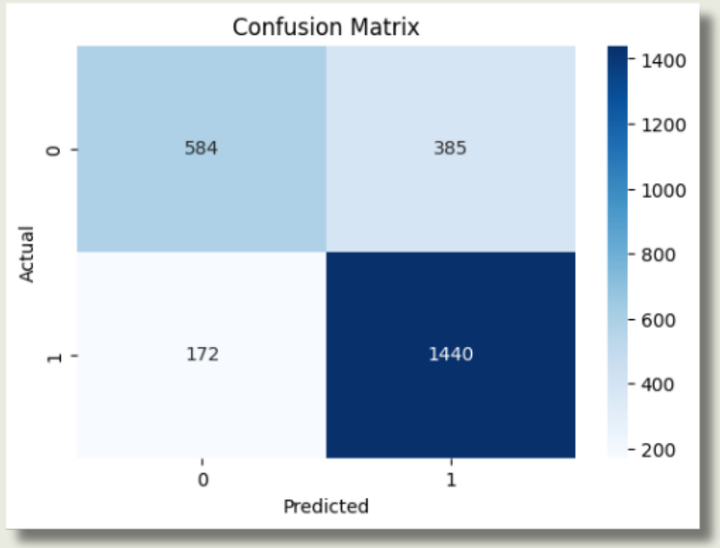
Not the best class 0 recall... but excellent recall overall



XGBOOST

- Oversampling
- New features
- Grid Search

	precision	recall	f1-score	support
0	0.73	0.64	0.68	969
1	0.80	0.86	0.83	1612

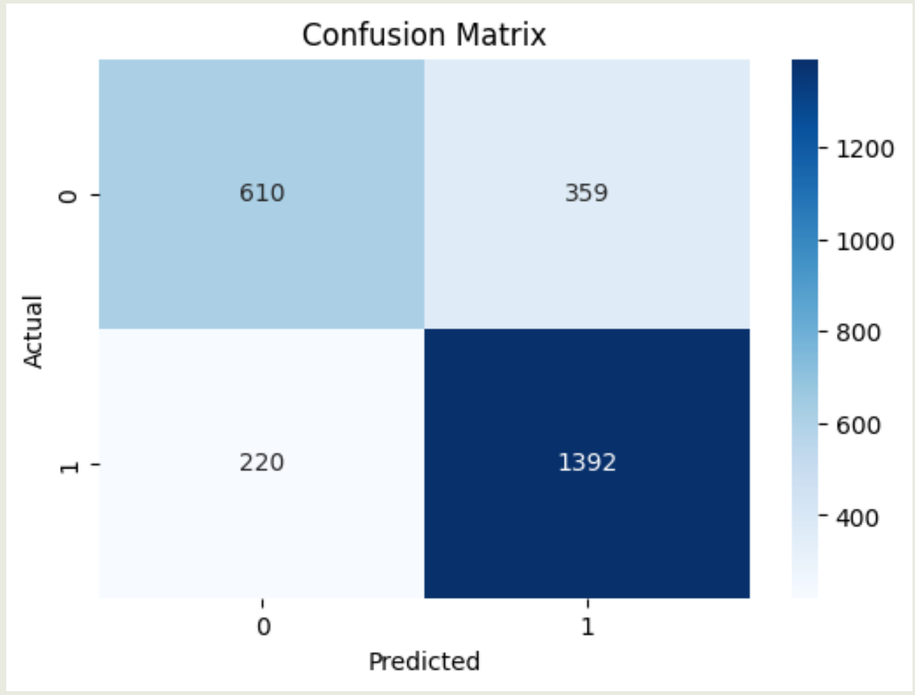



```
Best Parameters: {'colsample_bytree': 1, 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200, 'subsample': 0.8}
```

XGBOOST + GRIDSEARCH

	precision	recall	f1-score	support
0	0.62	0.71	0.66	969
1	0.81	0.74	0.77	1612

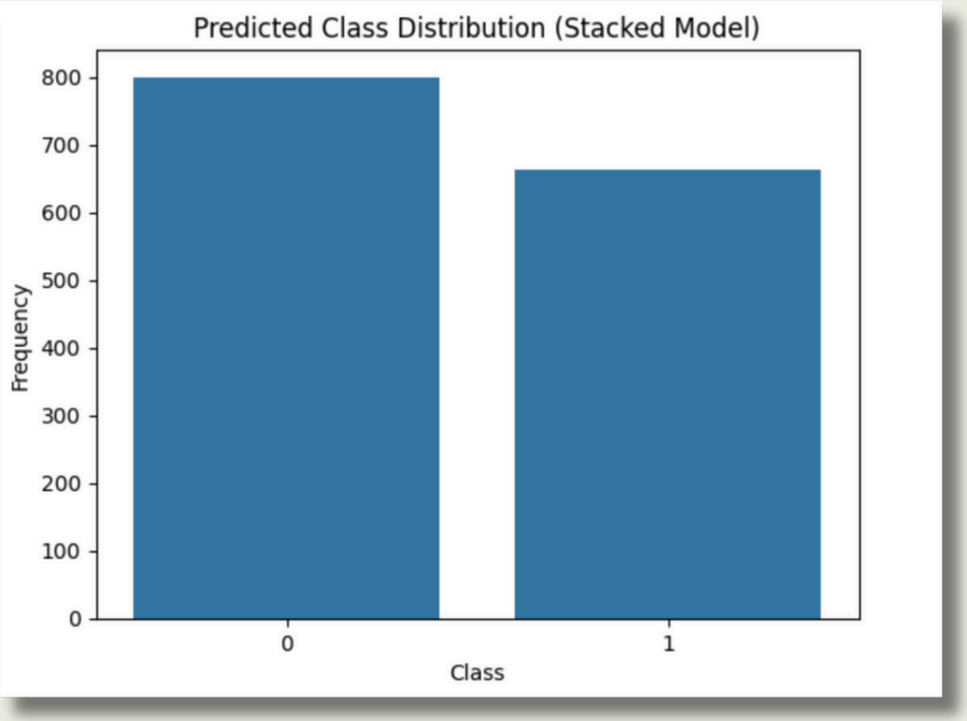
- Lowered Recall for class 1
- Raised class 0

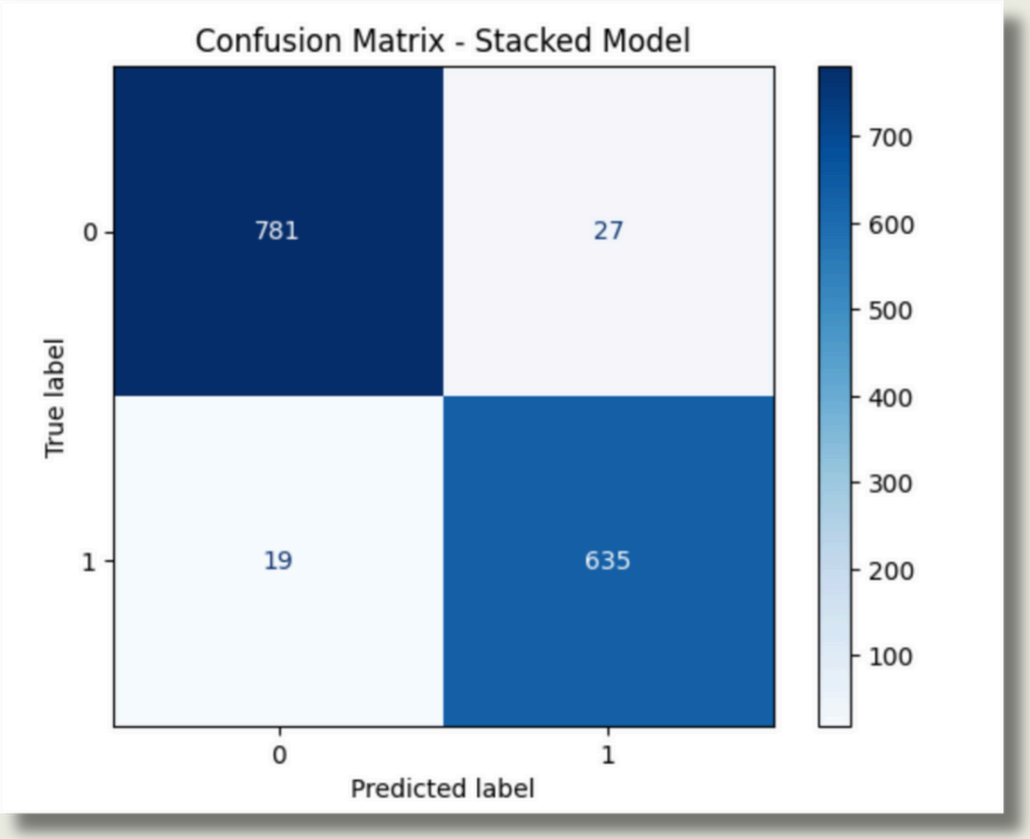


FINAL MODEL

Evaluating Stacked Model with SMOTEENN:

	precision	recall	f1-score	support
0	0.98	0.97	0.97	808
1	0.96	0.97	0.97	654
accuracy			0.97	1462
macro avg	0.97	0.97	0.97	1462
weighted avg	0.97	0.97	0.97	1462





CHALLENGES

CLASS IMBALANCE

The biggest challenge for this dataset was the severe class imbalance (2:1 ratio of class 1 to class 0 samples). This forced us to change our approach when creating the models, using techniques such as artificial sampling and stacking models to balance the dataset and increase the recall score of class 0.

AMOUNT OF FEATURES

The biggest problem in the preprocessing stage was feature selection. Significant time had to be devoted to reducing the number of features to avoid overfitting and multicollinearity, and even after cutting the number of features in half, we were still left with 42 columns to use for machine learning.

LIMITATIONS AND FUTURE WORK

Feature Selection/Engineering

Choosing a different combination of features could have given us different results. Changing the threshold for multicollinearity or keeping some of the low-correlated features could be done in the future to see how it affects results.

Synthetic Sampling Techniques

Testing different versions of synthetic sampling methods could have improved our models. They behaved unexpectedly at times, and it would’ve been interesting to see if multiple methods used in unison would give a better result

Encoding Categorical Variables

We chose label encoding for the categorical variables so we wouldn’t have to deal with even more features. One-hot encoding would’ve created significantly more features, but maybe it could’ve contributed to the model accuracy more than label encoding

Thank you.

Data Used: https://www.kaggle.com/competitions/widsdatathon2024-challenge1/data