UNCOVERING THE SILENT EPIDEMIC: INSIGHTS INTO STROKE RISK FACTORS

Rain Jd Kim Francia Pao

GROUP 1 - TWICE

WHAT IS STROKE?

Stroke occurs when **blood flow and oxygen supply to the brain is suddenly interrupted** due to constricted blocked vessels or when they burst.[3]

Symptoms:

1. Paralysis / Numbness of arms, legs or face
2. Visual, verbal and mobility impairment
3. Headache and vomiting
4. Loss of consciousness

3 https://www.ninds.nih.gov/health-information/disorders/stroke

STATE OF THE NATION

TOP 3 CAUSES OF DEATH
from JAN-MAY 2022[1]

1	Heart Disease	18.6%
2	**Stroke**	**10.4%**
3	Cancer	9.8%
⋮		
7	COVID 19	5.3%

NEWS

Average Stroke Mortality: 63,800 / Year

"Out of 109 million Filipinos, only 5.7% are above age 65 years old. Despite the young population, stroke remains a fatal disease and the second cause of death."[2]

1 PSA: https://psa.gov.ph/content/2022-causes-deaths-philippines-preliminary-30-june-2022

2 Collantes ME, Navarro J, Belen A, Gan R. Stroke systems of care in the Philippines: Addressing gaps and developing strategies. Front Neurol. 2022 Nov 24

GAPS IN HEALTHCARE

Only ONE neurologist
for every 218,000 Filipinos

2/3 of neurologists
concentrated in urban centers

Overcrowding of public hospitals &
Inadequate CT scan machines

=

Poor stroke awareness

Uncontrolled risk factors

High medical expenses

Source: Collantes ME, Navarro J, Belen A, Gan R. Stroke systems of care in the Philippines: Addressing gaps and developing strategies. Front Neurol. 2022 Nov 24

MAIN OBJECTIVE

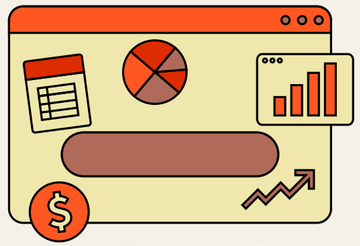
How can we apply machine learning to raise awareness on the major contributors to stroke?

SDG Goal 3 Target:

Strengthen developing countries' capacity for early warning, risk reduction and management of national and global health risks.

DATA SCIENCE PIPELINE

Data Collection



Data Cleaning & Preprocessing

Model Training

Evaluation & Interpretation

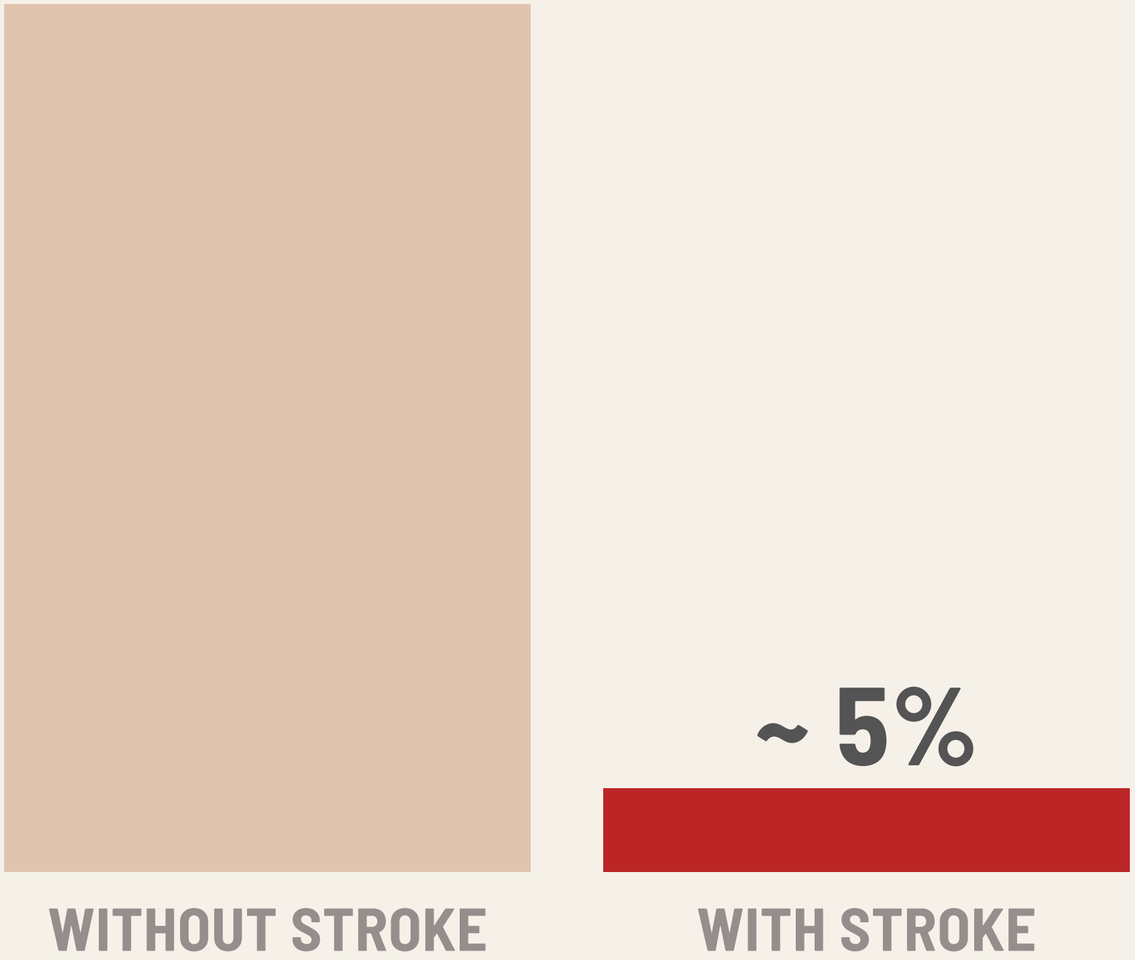
Deployment

DATA UNDERSTANDING

Dataset introduction and initial exploratory data analysis

Sprint 2 | Twice

WORKING WITH IMBALANCED DATA

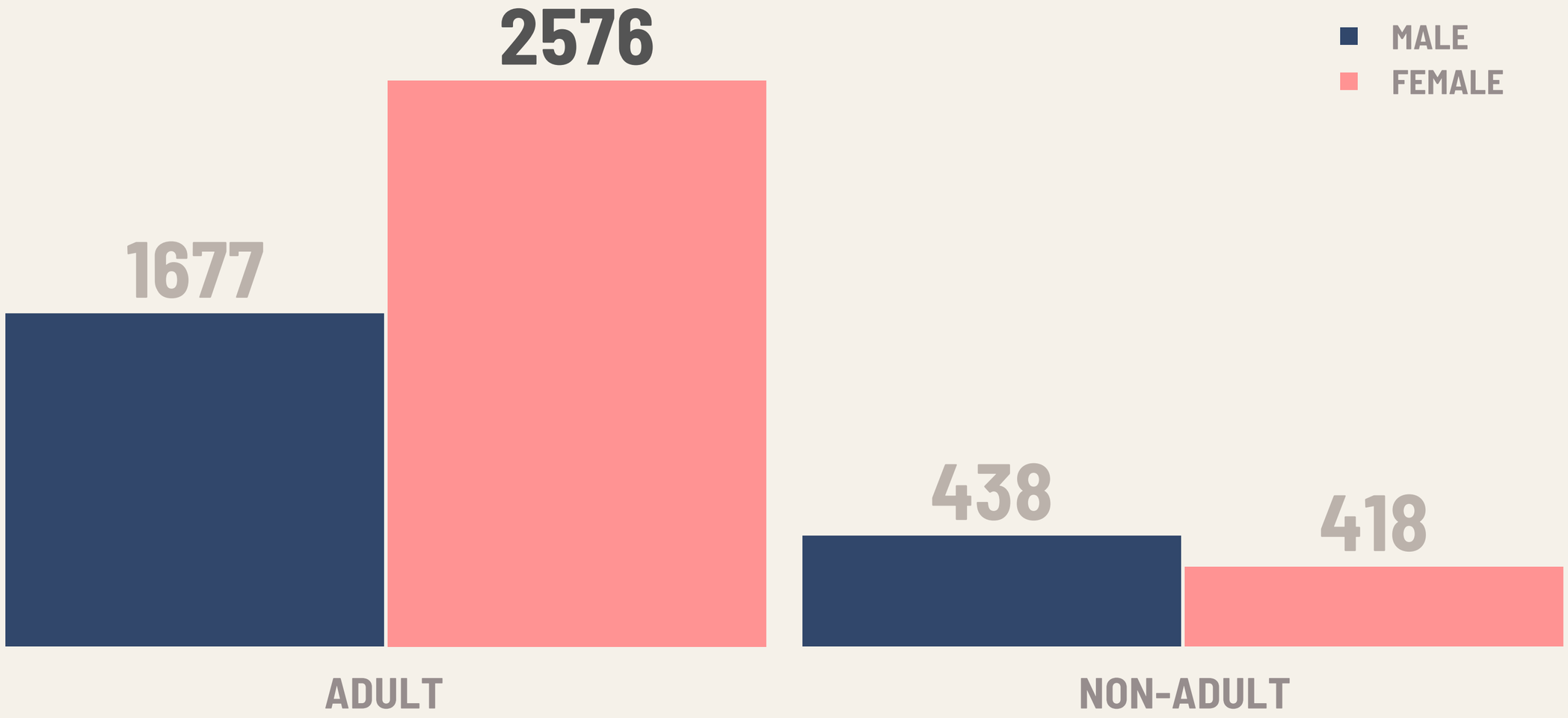


SOURCE: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

RAW DATA

FEATURES											TARGET
id	gender	age	hypertension	heart disease	marital status	work type	residence type	average glucose level	bmi	smoking status	stroke

OVERVIEW



CORRELATION MATRIX

1

UNSPECIFIED GENDER

201

MISSING BMI VALUES

DATA PREPARATION

A summary of processing steps

Sprint 2 | Twice

FEATURE PREPARATION STEPS

Kept as is

- hypertension
- heart_disease
- avg_glucose_level
- gender
- stroke

Engineered then dropped

- age
- bmi
- age_group
- weight_group

Subcategories renamed

- smoking_status

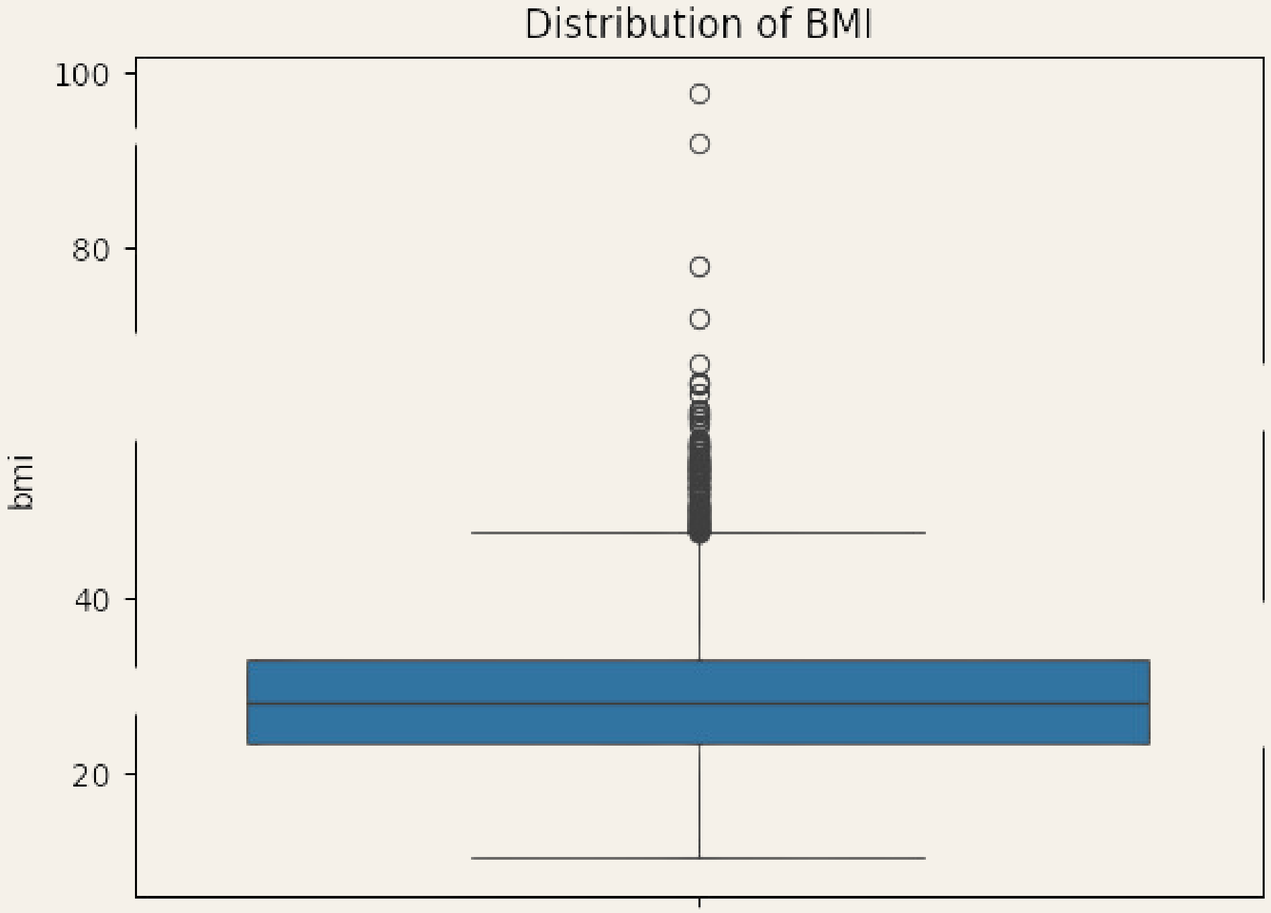
Dropped

- id
- work_type
- residence_type
- marital_status

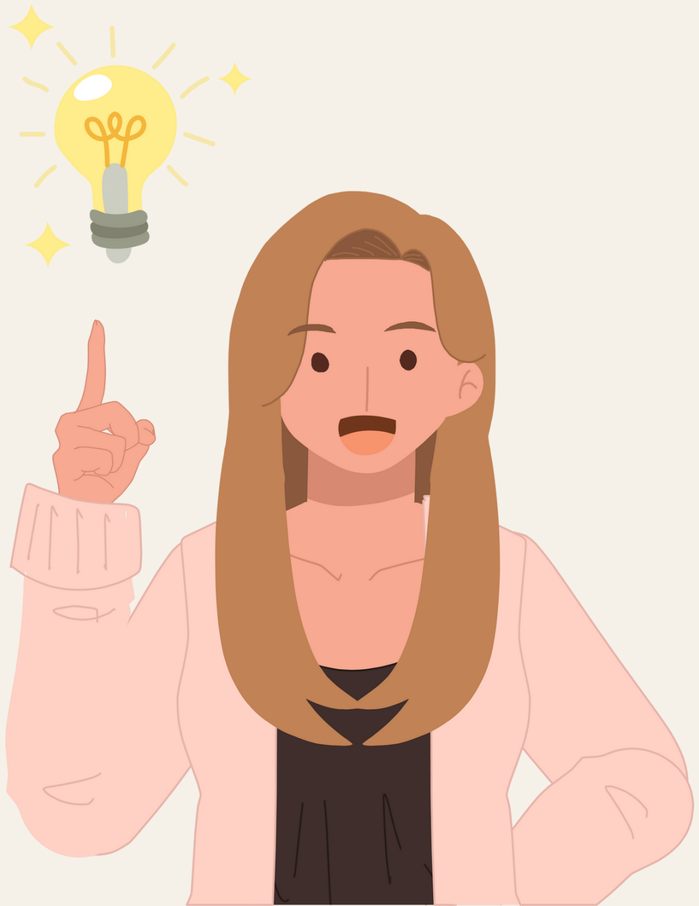
BMI TREATMENT

Outliers were **NOT** removed

NaN data points imputed with **median** BMI of entry's age group and gender



FEATURE ENGINEERING



Age

Age range	Age group
0	Infant
1 - 12	Child
13 - 17	Adolescent
18 - 65	Adult
65 above	Older adult

BMI

BMI range	Weight group
< 16.5	Severely underweight
16.5 - 18.4	Underweight
18.5 - 24.9	Normal weight
25 - 29.9	Overweight
30 - 34.9	Obese I
35 - 39.9	Obese II
40 <	Obese III

ONE-HOT ENCODING

> **Purpose.** To derive a numerical representation of categorical variables as input for ML models

gender	age_group	weight_group
Female	older adult	normal weight
Male	adult	obese i

gender_ Female	gender_ Male	age_group_ infant	age_group_ child	age_group_ adolescent	age_group_ adult	age_group_ older adult	weight_group_ severely underweight	weight_group _underweight	weight_group_ normal weight	weight_group _overweight	weight_group _obese i	weight_group_o bese ii	weight_group_ obese iii
1	0	0	0	0	0	1	0	0	1	0	0	0	0
0	1	0	0	0	1	0	0	0	0	0	1	0	0

*subset of dataset only
*not all features present on illustration

OVERVIEW OF PREPROCESSED DATASET

One-hot encoded

gender

- female
- male

age_group

- infant
- child
- adolescent
- adult
- older adult

weight_group

- severely underweight
- underweight
- normal weight
- overweight
- obese i
- obese ii
- obese iii

engineered features

medical

- hypertension
- heart_disease

non-one-hot encoded

numerical

- avg_glucose_level

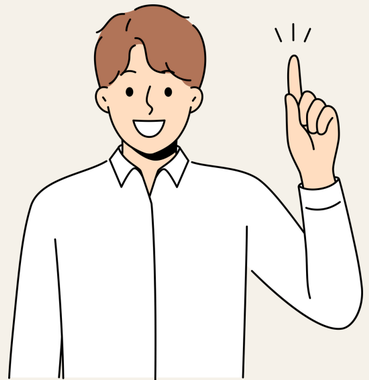
target

stroke

Shape

5109 rows

22 columns

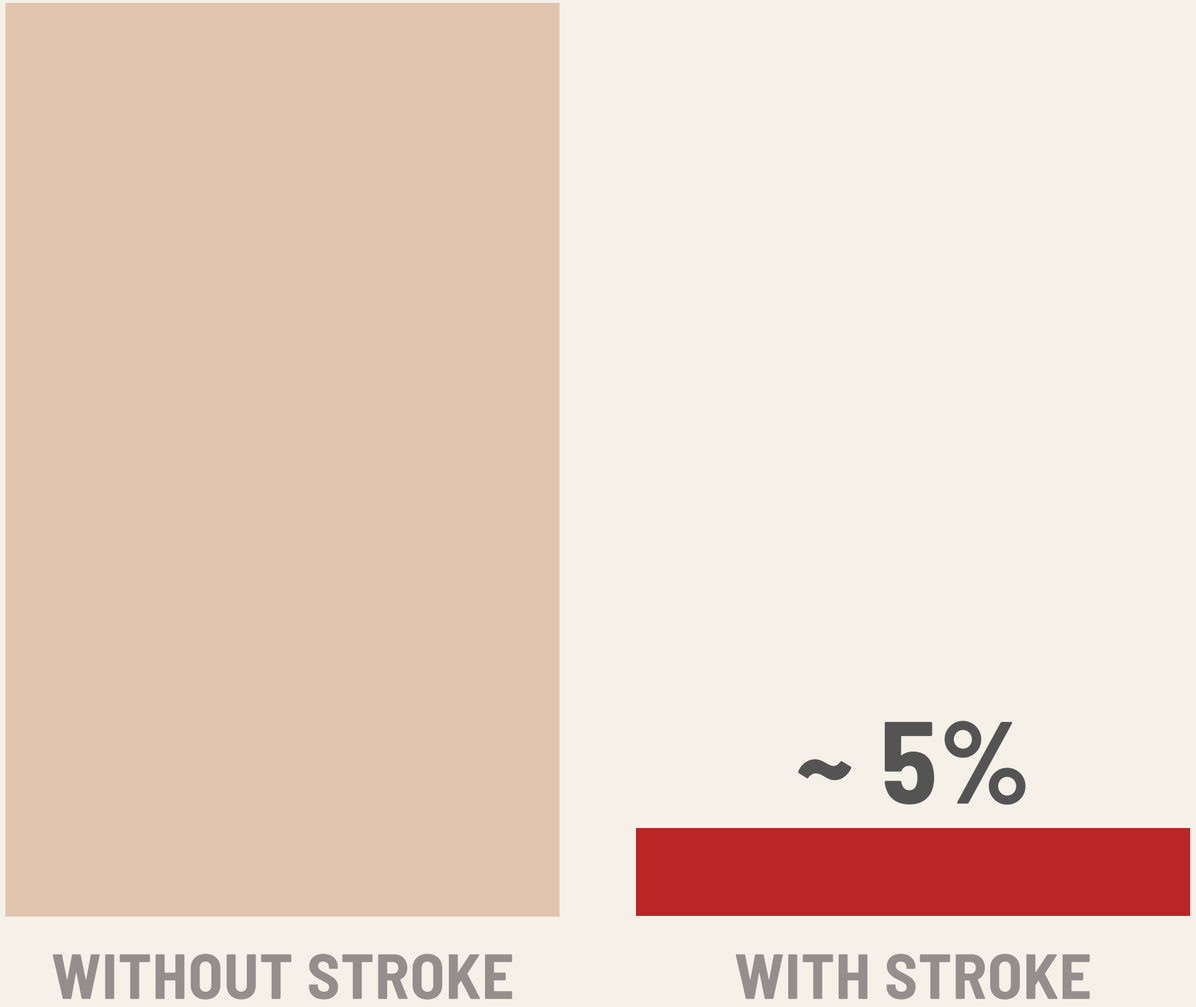


MODELING

Baselining, base models, resampling methods, and hyperparameter tuning

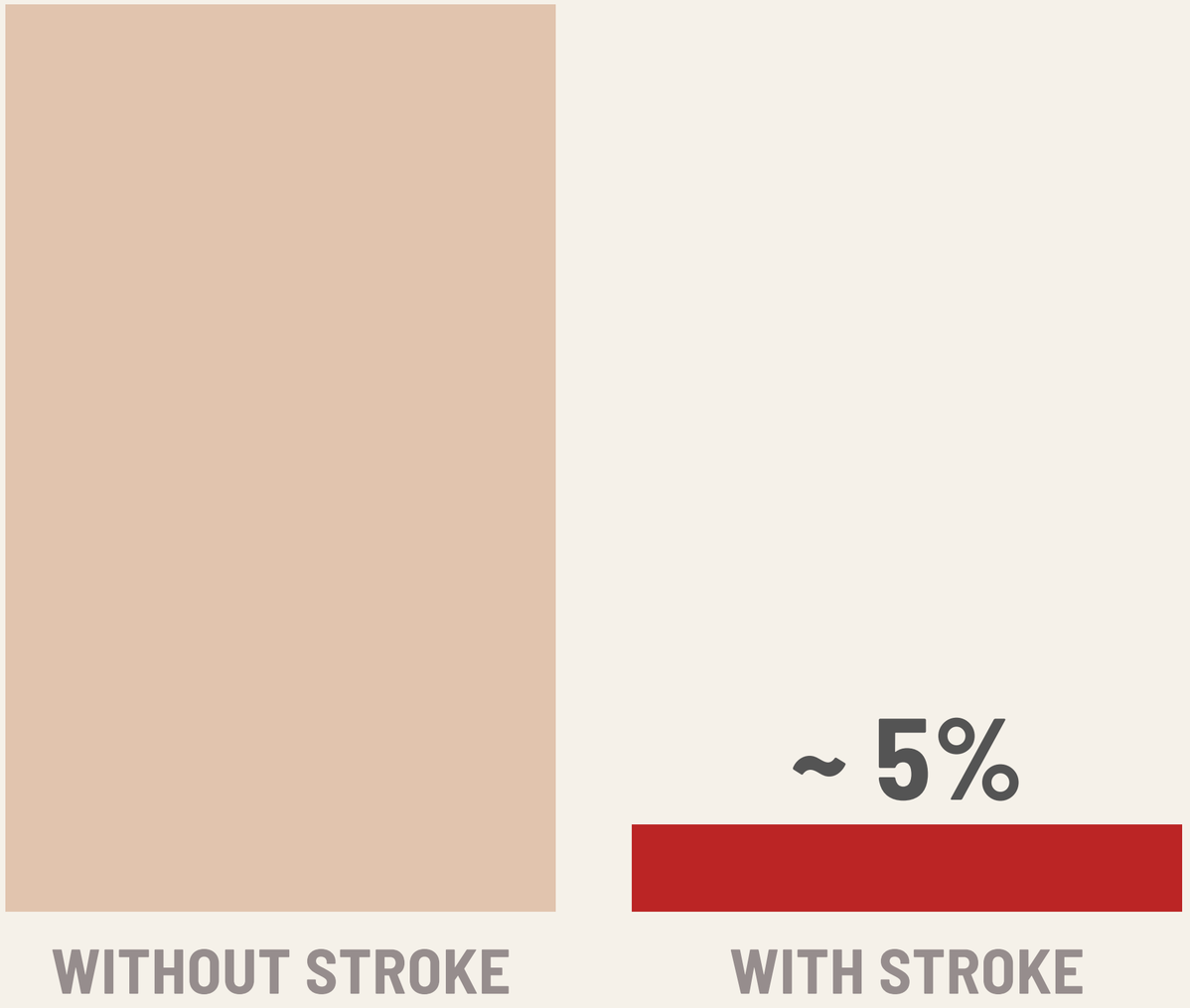
Sprint 2 | Twice

Can the model predict stroke better than random chance?



PROPORTION CHANCE CRITERION FOR PRECISION, RECALL, F1

1.25*PCC = 6.09%



TRAINVAL HOLDOUT

MACHINE LEARNING MODELS

K-Nearest Neighbor (KNN)

Logistic Regression

Decision Tree

Random Forest

Gradient Boosting

Adaptive Boosting

Extra Tree

Extra Gradient Boosting

RESAMPLING TECHNIQUES

OVERSAMPLING

SMOTE

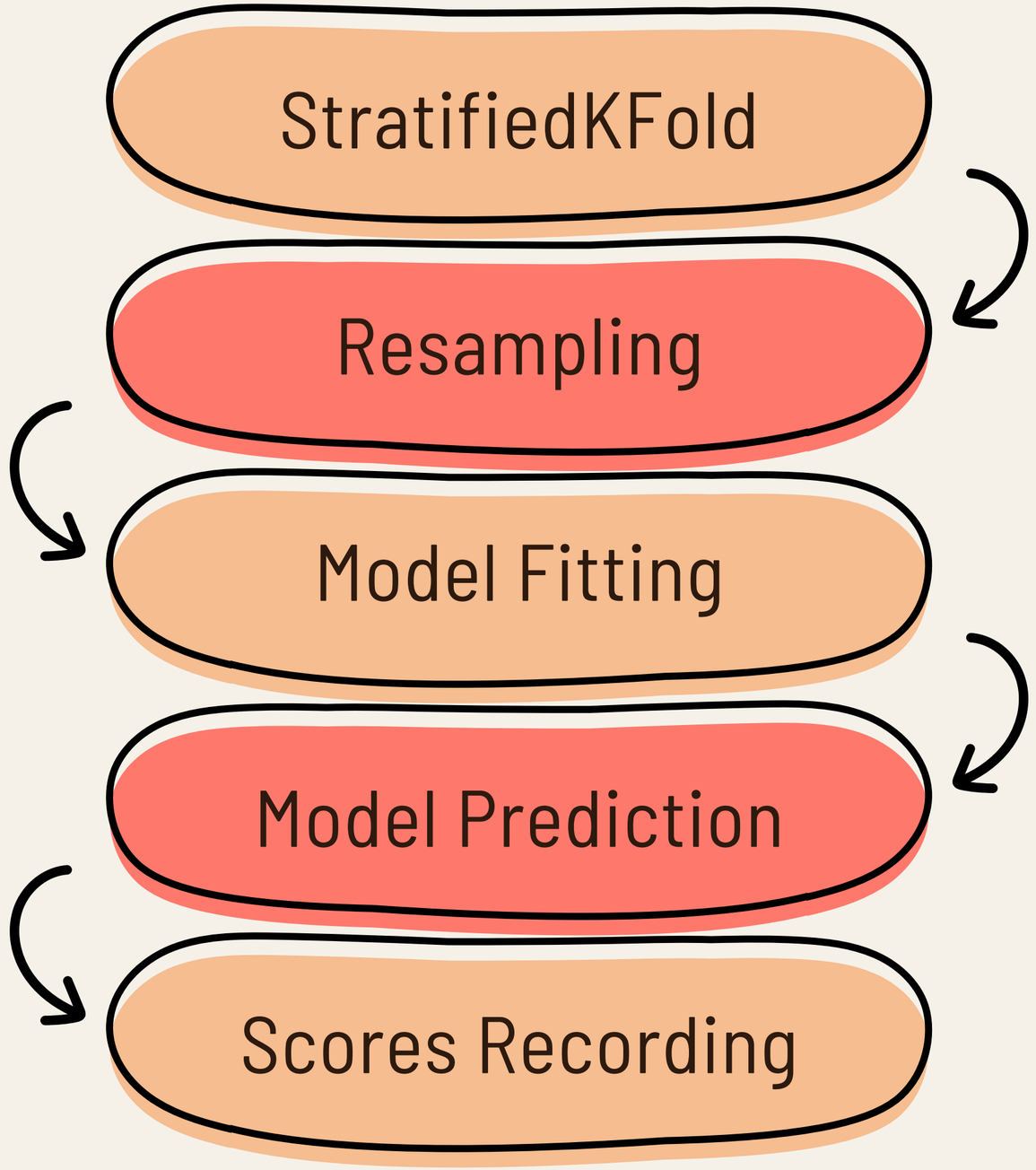
SMOTE-N

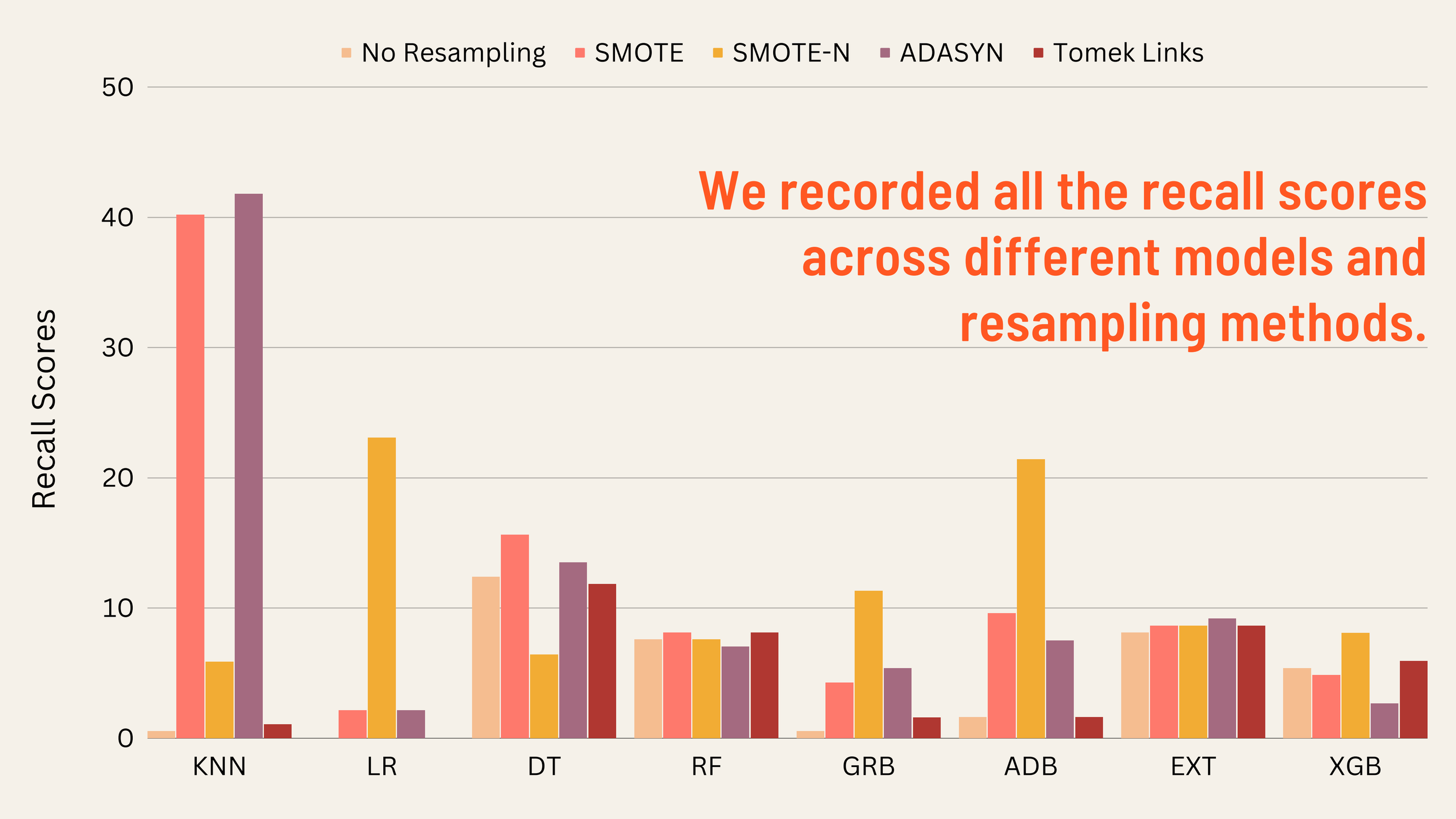
ADASYN

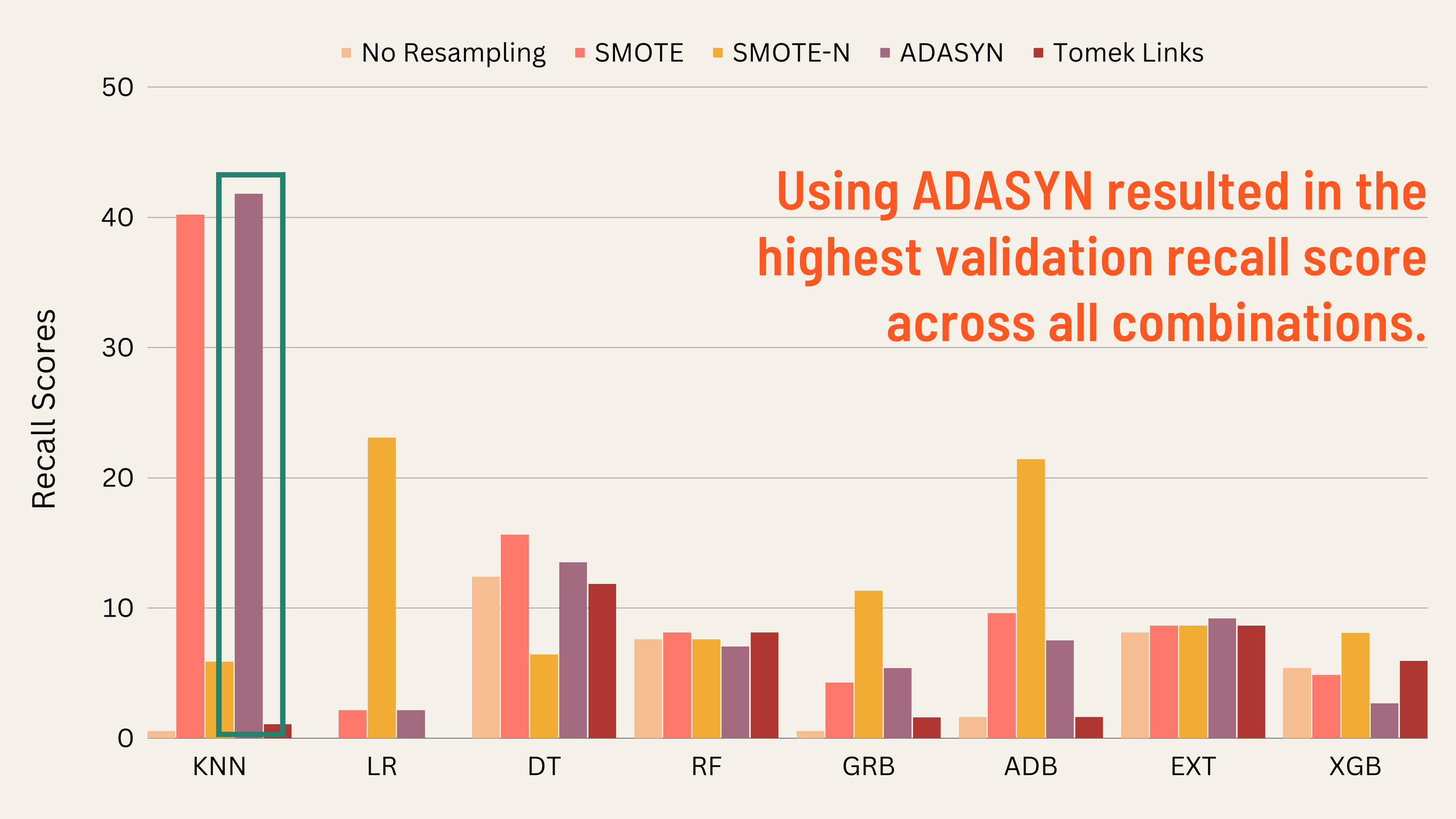
UNDERSAMPLING

Tomek Links

BASE MODEL CREATION

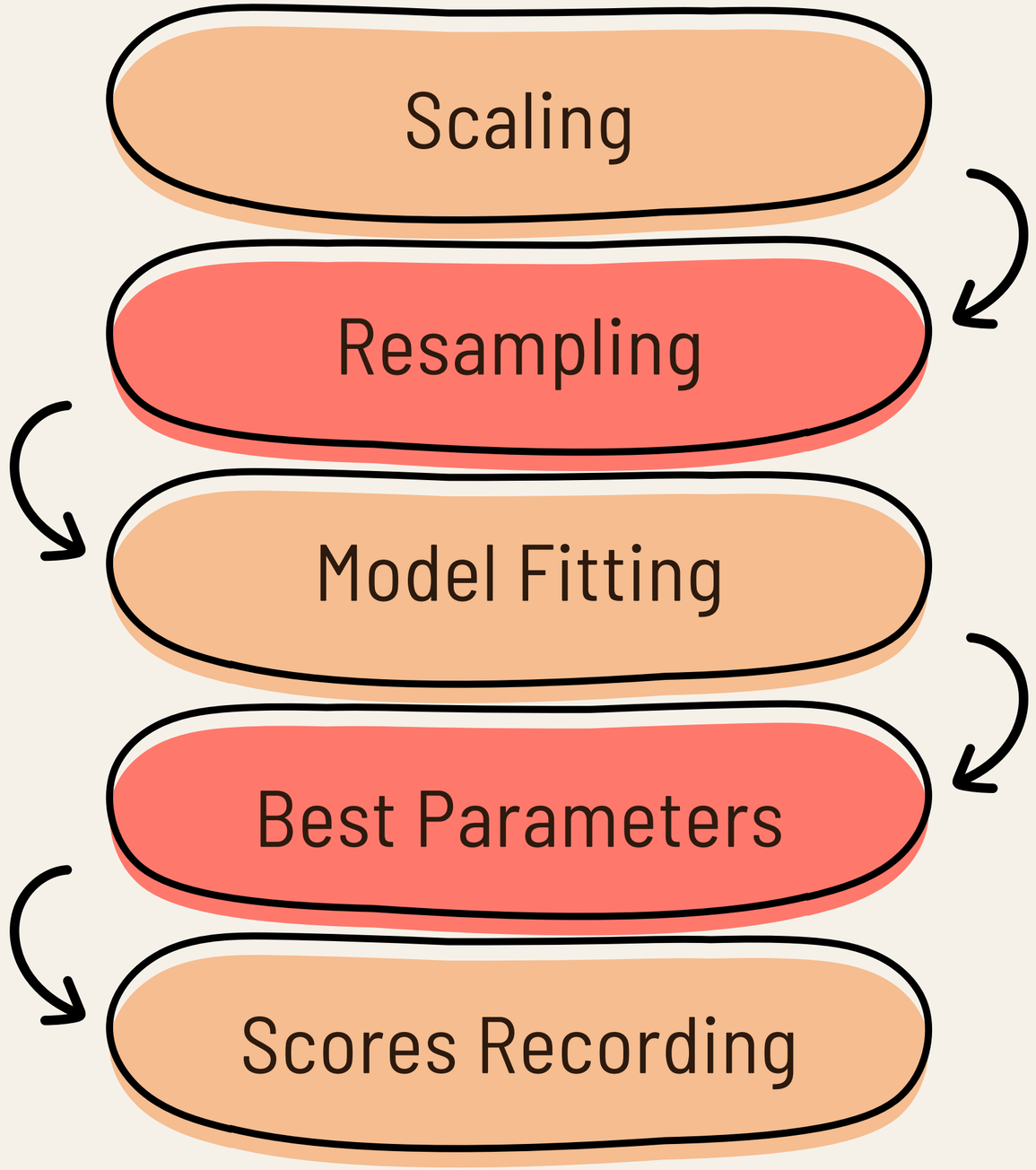


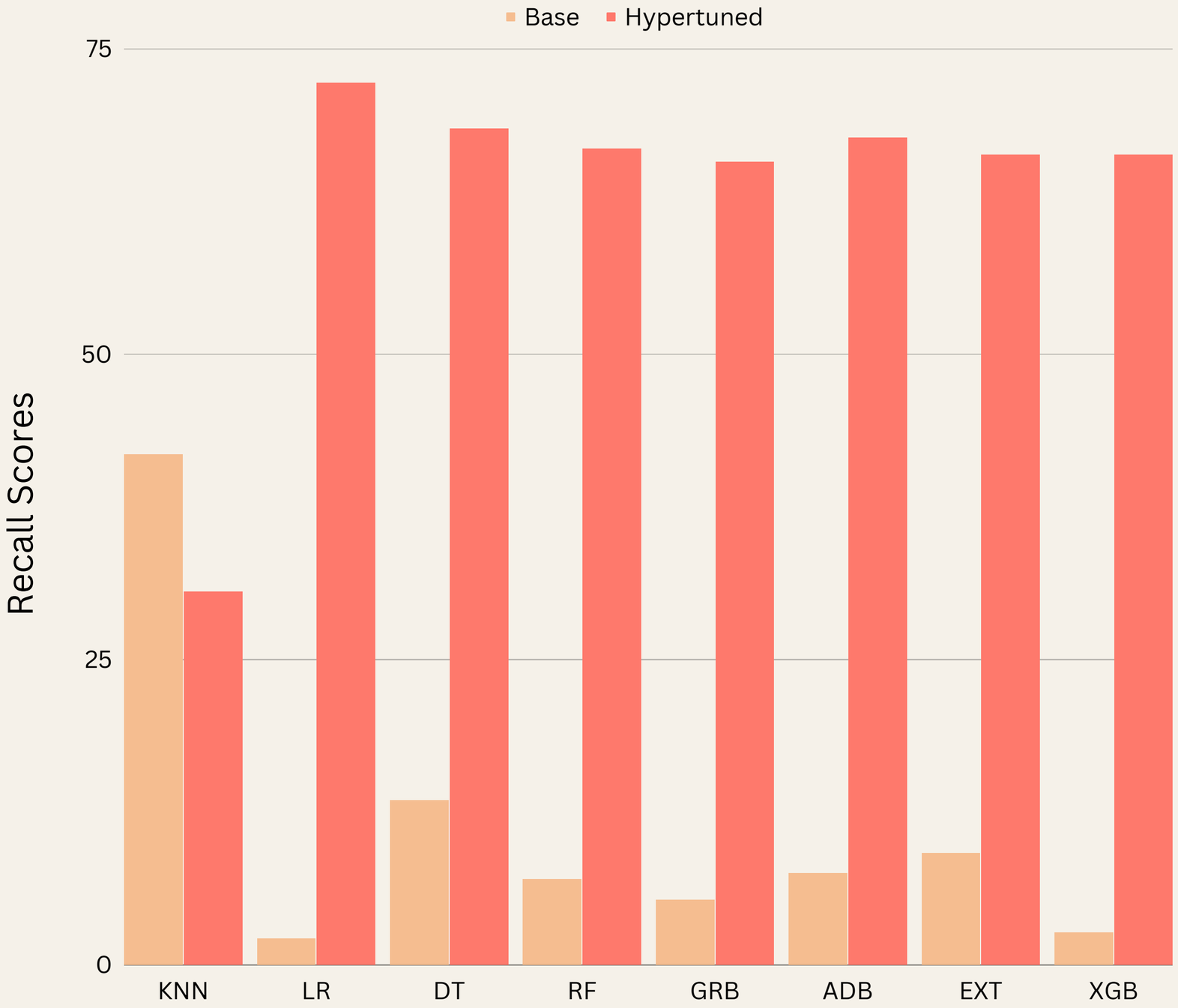

No Resampling
SMOTE
SMOTE-N
ADASYN
Tomek Links
50
40
30
20
10
0
Recall Scores
KNN
LR
DT
RF
GRB
ADB
EXT
XGB
We recorded all the recall scores across different models and resampling methods.


No Resampling
SMOTE
SMOTE-N
ADASYN
Tomek Links
50
40
30
20
10
0
Recall Scores
KNN
LR
DT
RF
GRB
ADB
EXT
XGB
Using ADASYN resulted in the highest validation recall score across all combinations.

HYPER PARAMETER TUNING

Using Grid Search

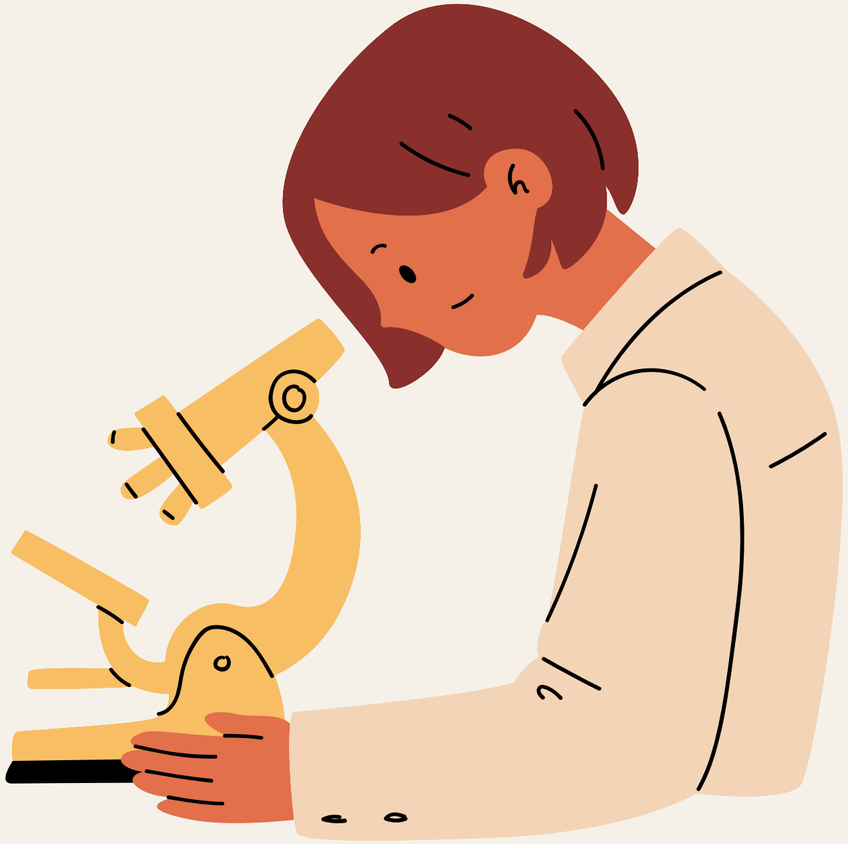




After the hypertuning process, model performance increased by up to ~33%

MODEL EVALUATION

Rank Based Scoring via relevant metrics



Sprint 2 | Twice

EVALUATION METRICS

Accuracy

number of correctly predicted instances

Precision

accuracy of positive predictions

Recall

minimizes false negative predictions

F1-Score

harmonic mean of precision and recall

CHOOSING THE BEST MODEL

BASE MODEL

POST-SAMPLING

HYPERPARAMETER TUNING W/ SAMPLING

EXTRATREESCLASSIFIER

8.08%

BASE VALIDATION RECALL

9.15%

POST-SAMPLING VALIDATION RECALL

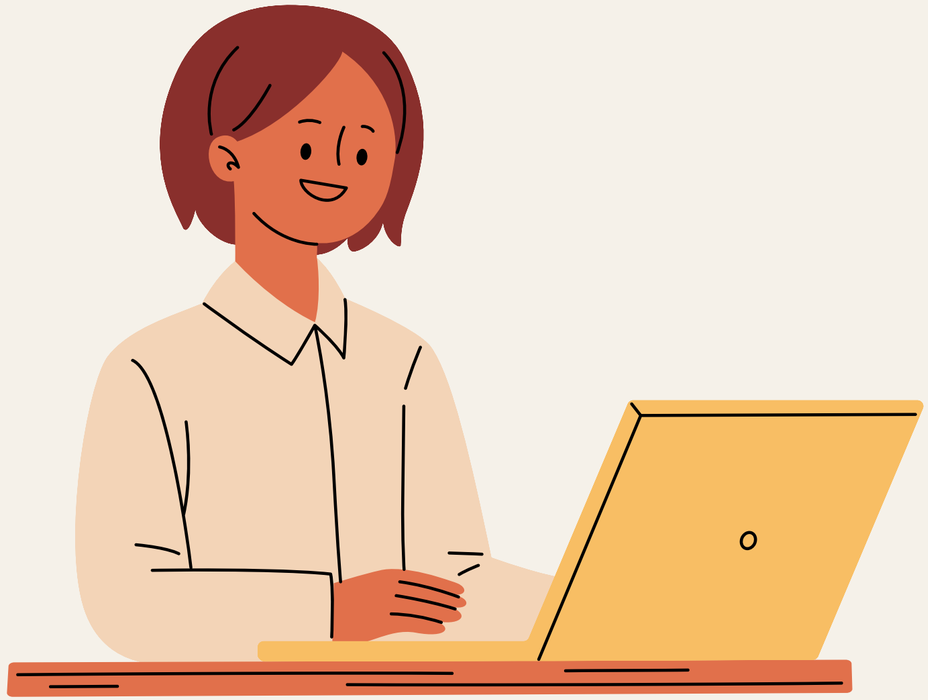
66.30%

HYPERPARAMETER TUNING W/ SAMPLING VALIDATION RECALL

In comparison to P (6.09%), the ExtraTreesClassifier model performed significantly better.

BEST MODEL & INTERPRETATION

Model Performance and Feature Importance



Sprint 2 | Twice

EXTRATREESCLASSIFIER RECALL SCORES

The model addresses the problem of overfit with the base model.

Training Recall	**66.88%**
Validation Recall	**66.30%**

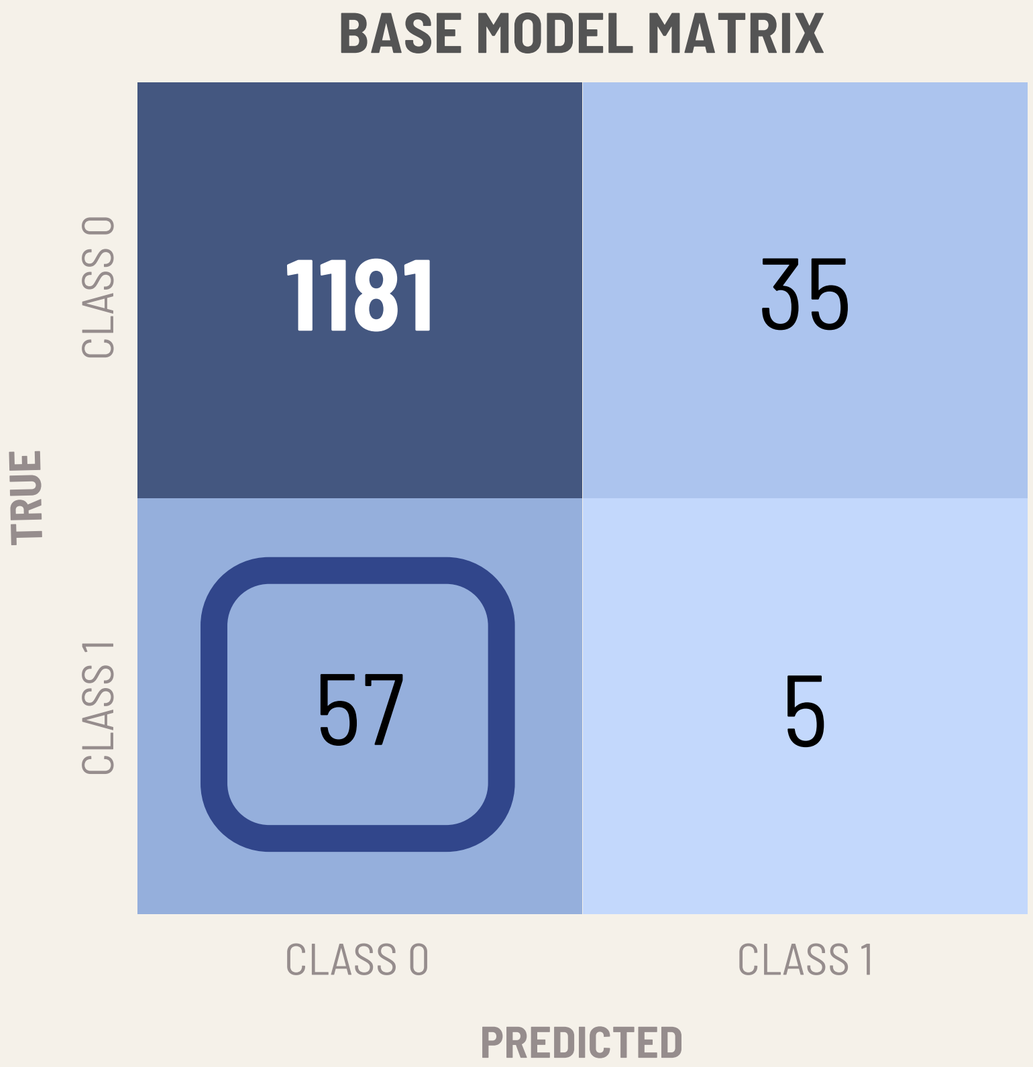
Base Model Train Recall	**99.60%**
Base Model Validation Recall	**8.08%**

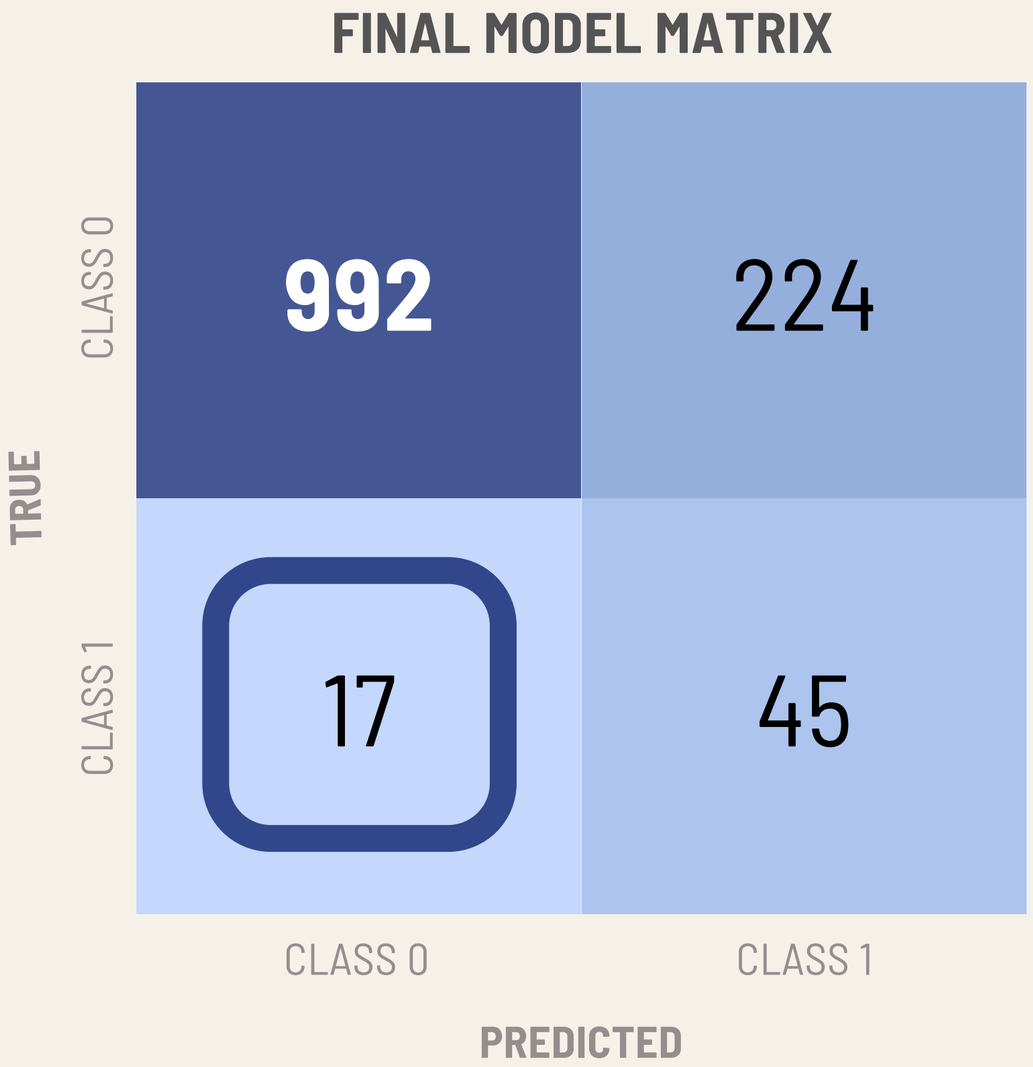
EXTRATREESCLASSIFIER RECALL SCORES

The model performed well in predicting the target for the holdout data.

Holdout Recall	**72.58%**
Training Recall	**66.88%**
Validation Recall	**66.30%**
Base Model Training Recall	**99.60%**
Base Model Validation Recall	**8.08%**

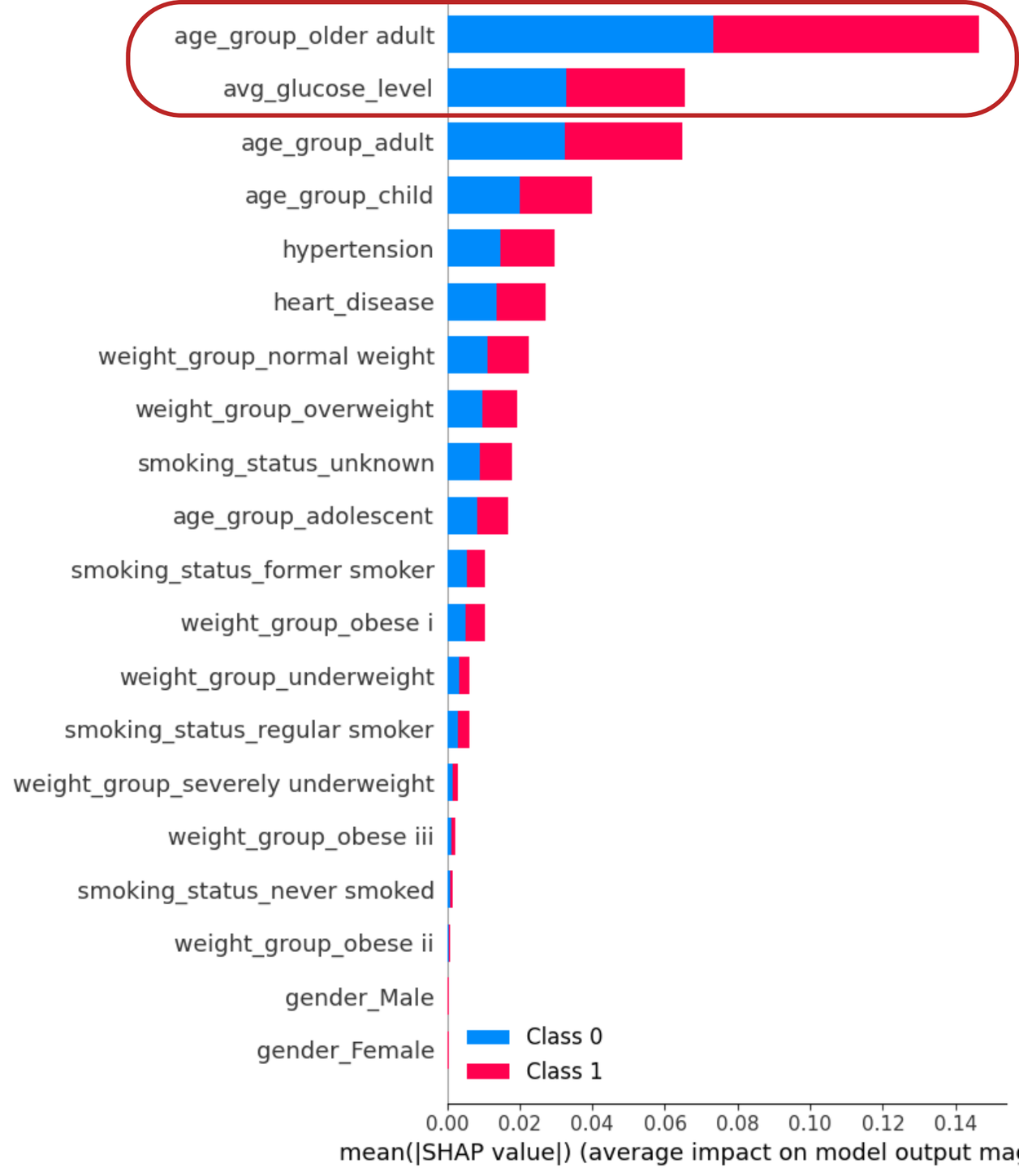
CONFUSION MATRIX (HOLDOUT DATA)



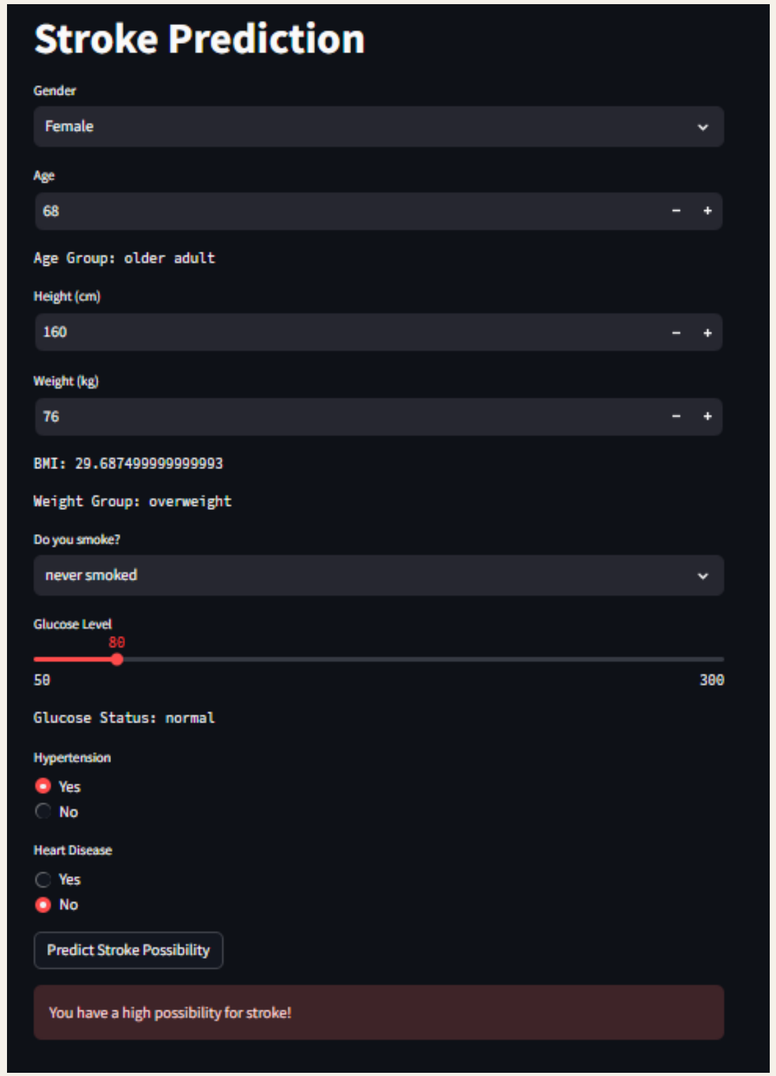


SHAP SUMMARY

‘Age’ and ‘Glucose Level’ are the main factors driving the decision of the model in predicting stroke



STREAMLIT DEMO



https://eskwelabs-sprintii.streamlit.app/

CONCLUSION

KEY RISK FACTORS TO STROKE

- ✓ Old Age
- ✓ High Glucose Levels

RECOMMENDATIONS

Increase Datasize & Diversity

- Model trained on limited dataset (5000 data points), with mostly adults
- Find Philippine Dataset

Add Other Features

- Explore other health indicators (blood pressure, physical activity, family stroke history)

THANK YOU!

Rain

Jd

Kim

Francia

Pao

GROUP 1 - TWICE