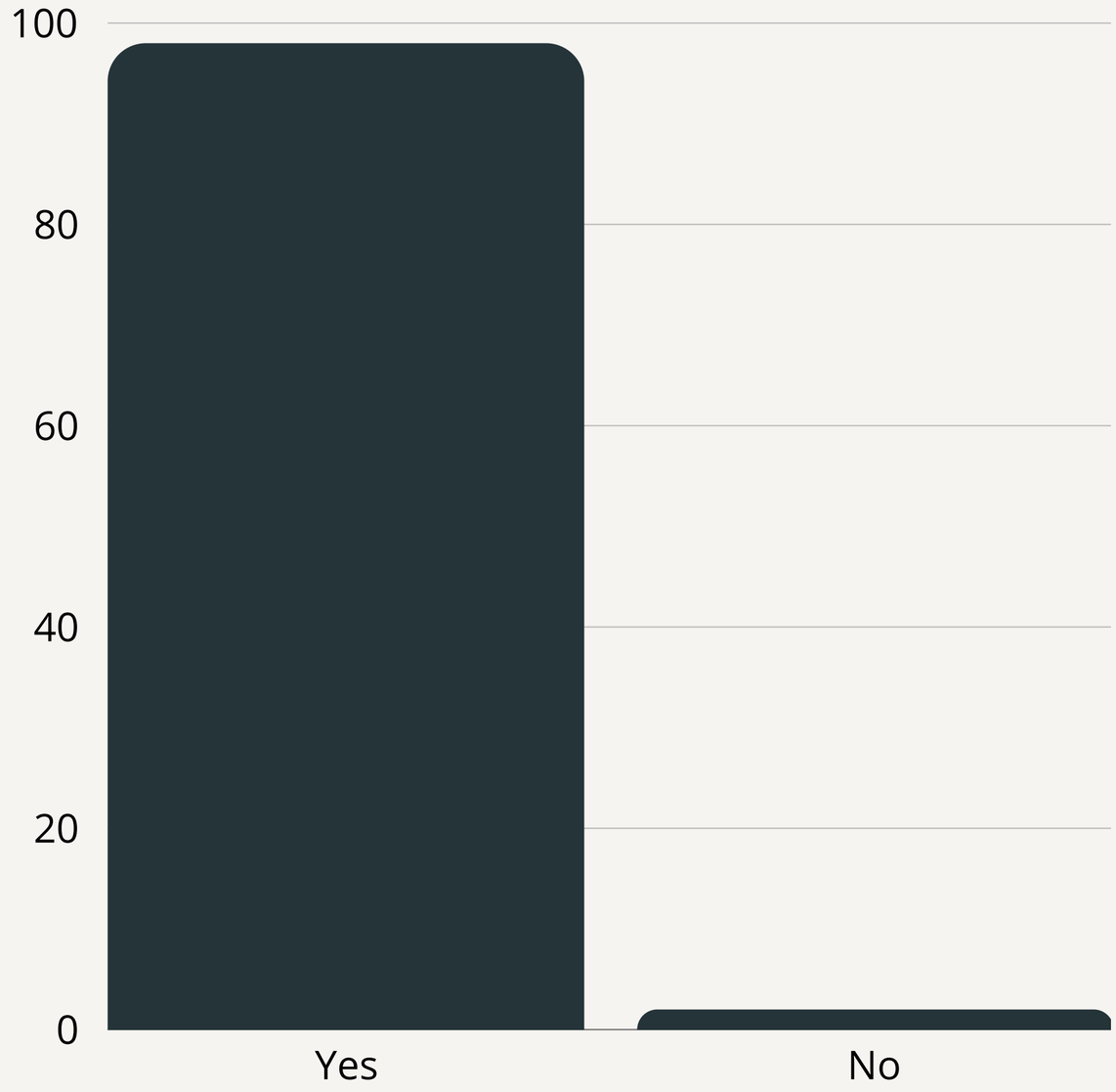
SMOTE lang tayo!

Handling Imbalance in Data

AIM MSDS 2024
ML1 Final Project Presentation
Jeremiah Dominic Soliman

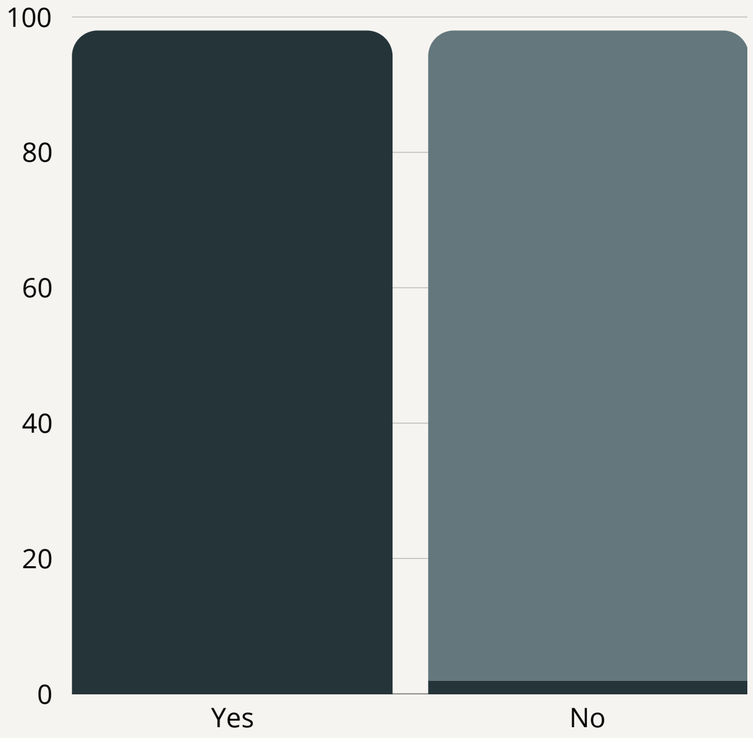


IMBALANCE

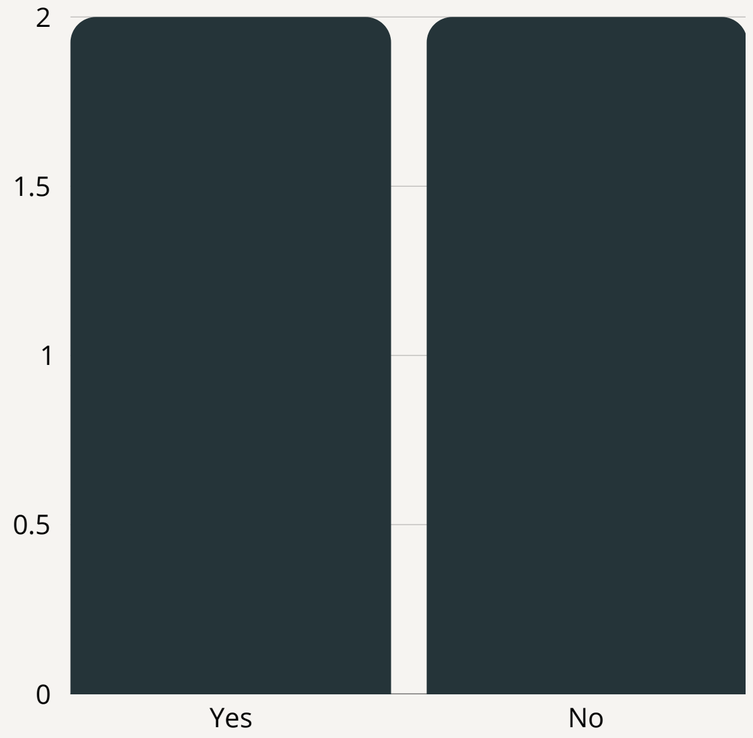
98% - 2%

PCC - 96.08%

RESAMPLE



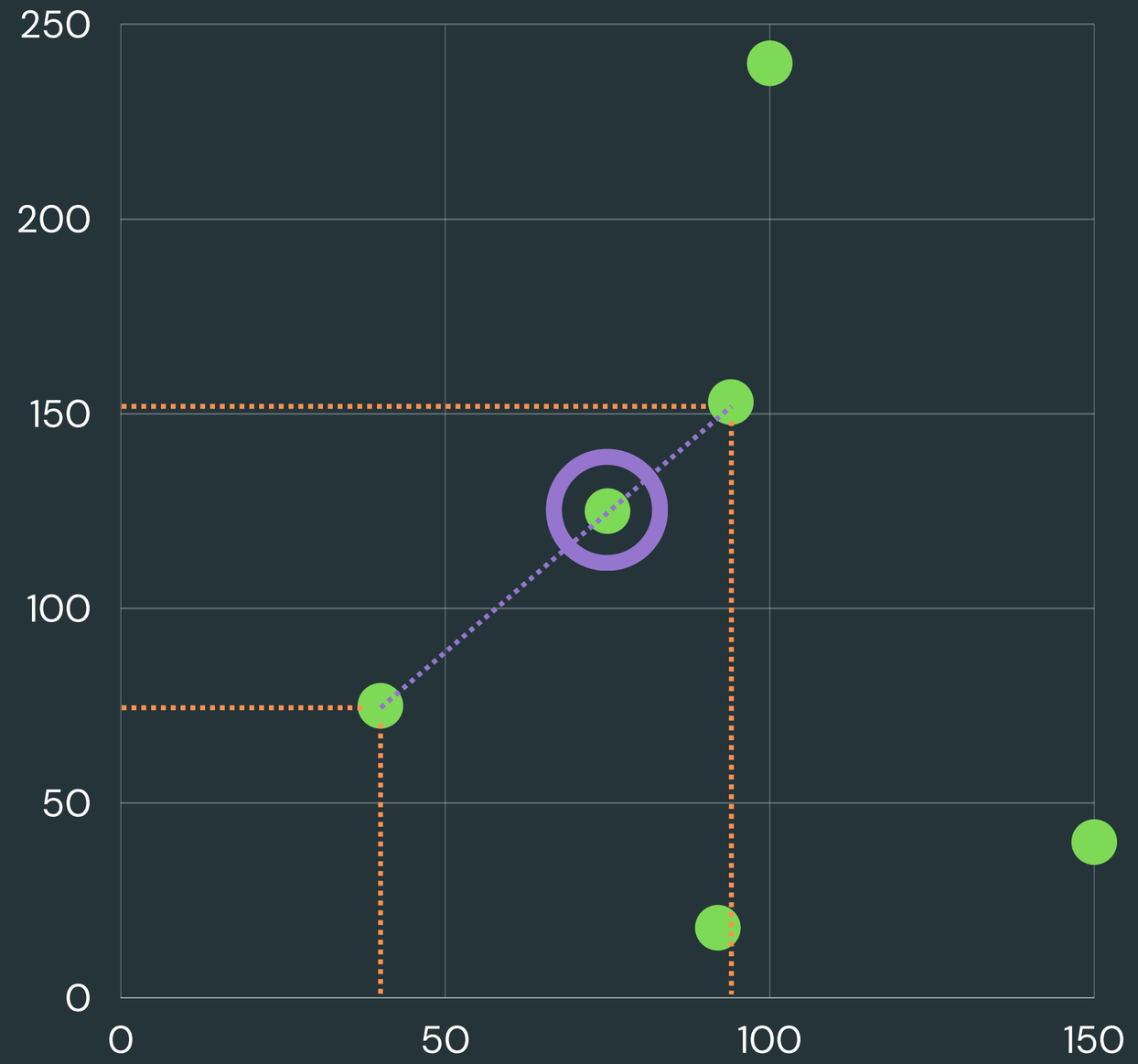
50 vs 50
Oversampling



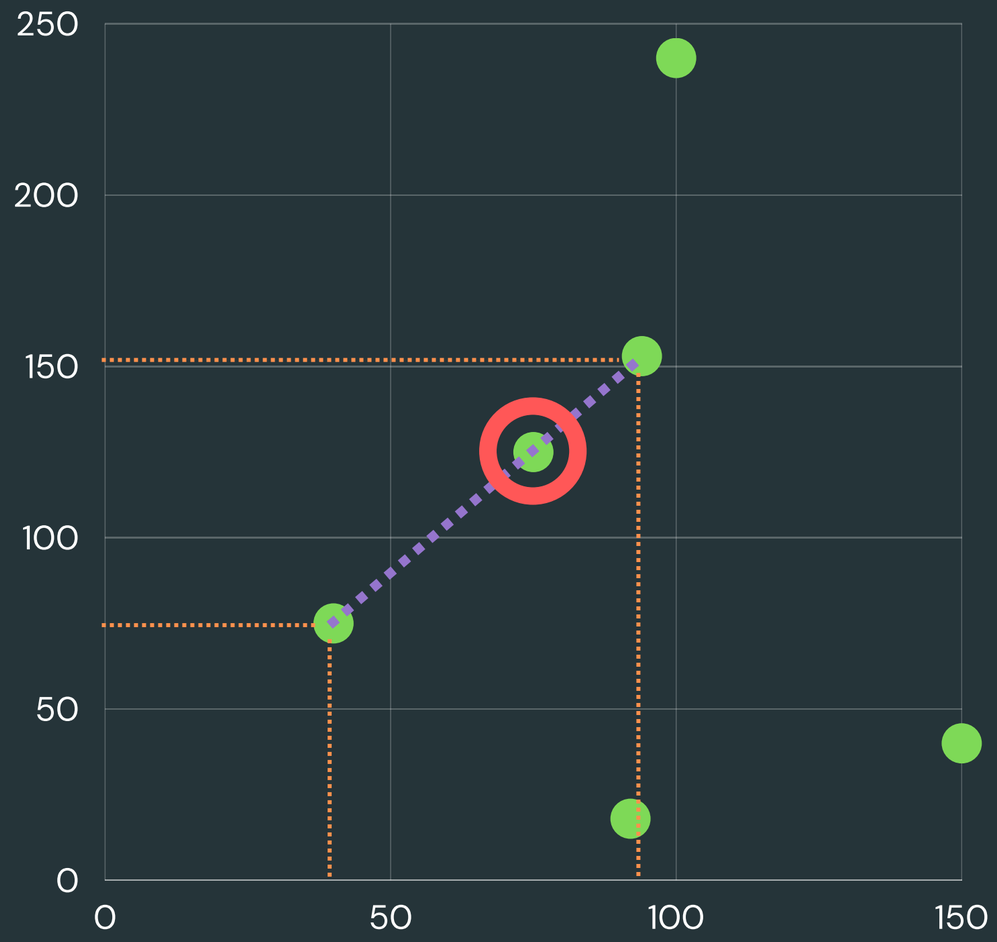
50 vs 50
Undersampling

SMOTE

Synthetic
Minority
Oversampling
TEchnique



SMOTE



1. **Select a minority class sample** from the original dataset.
2. Find its **k nearest minority class neighbors** in the feature space.
3. **Randomly select** one of the k nearest neighbors.
4. Generate a new synthetic sample by **interpolating** between the selected minority class sample and the randomly selected neighbor.
5. Repeat steps 1-4 until the desired number of synthetic samples is generated.

*New Sample = Original Sample - factor * (Original Sample - Neighbor)*

SMOTE

SMOTEN/C

Synthetic Minority Over-sampling Technique for Nominal and Continuous.

BORDERLINE

Detect Borderline Samples to use for generating new synthetic samples

KMEANS

Apply a KMeans clustering before to over-sample using SMOTE.

SVMSMOTE

Use an SVM algorithm to detect sample to use for generating new synthetic samples

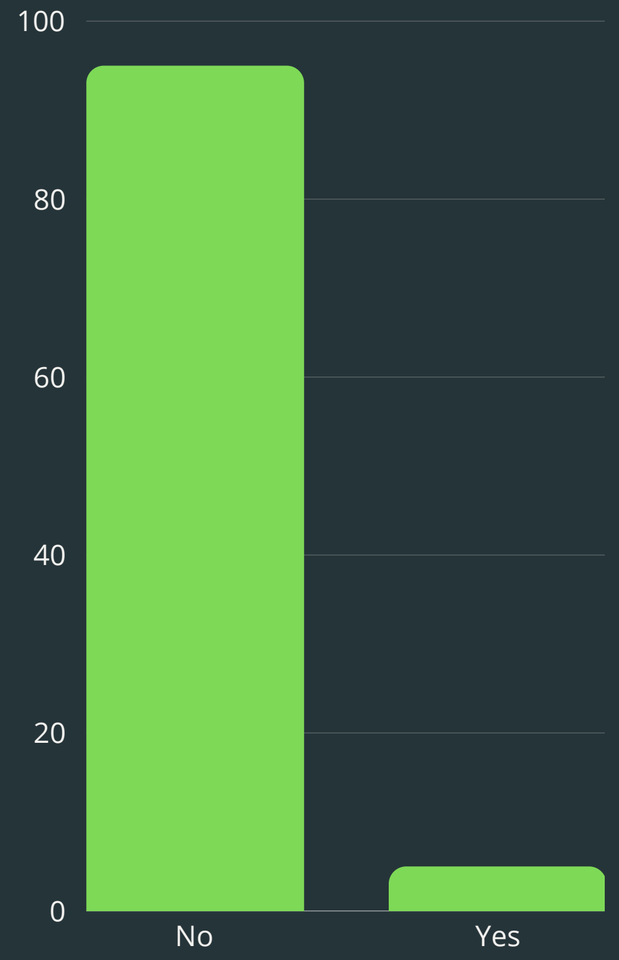
DATASET 1

Stroke Prediction Dataset
from ***Kaggle***

3,411 x 11

11 clinical features for predicting stroke events

'gender', 'age',
'hypertension',
'heart_disease',
'ever_married',
'work_type',
'Residence_type',
'avg_glucose_level', 'bmi',
'smoking_status', **'stroke'**



METHODOLOGY

DATA COLLECTION

- ACQUIRING DATASET FROM KAGGLE

DATA CLEANING

- REMOVING NULLS AND UNDER REPRESENTED UNIQUE VALUES

DATA QUESTIONS

EXPLORATORY DATA ANALYSIS

- DATA EXPLORATION
- DATA VISUALIZATION

MACHINE LEARNING

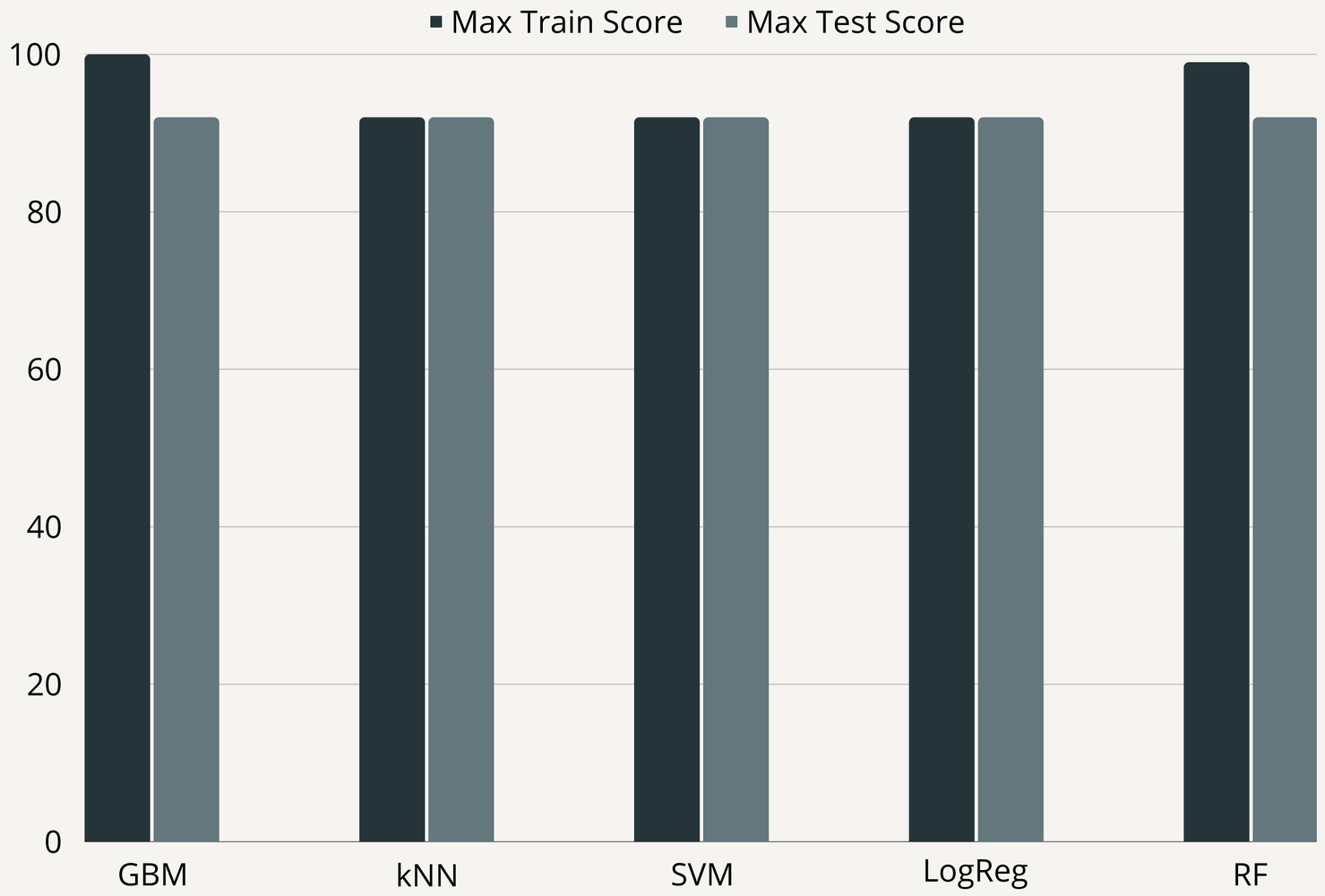
- GRIDSEARCHCV
- CLASSIFIER

RESAMPLE SMOTE

- UTILIZE SMOTE
- GRIDSEARCHCV

RESULTS AND INTERPRETATION

- EXPLAINING DATA INSIGHTS
- MAKING RECOMMENDATIONS



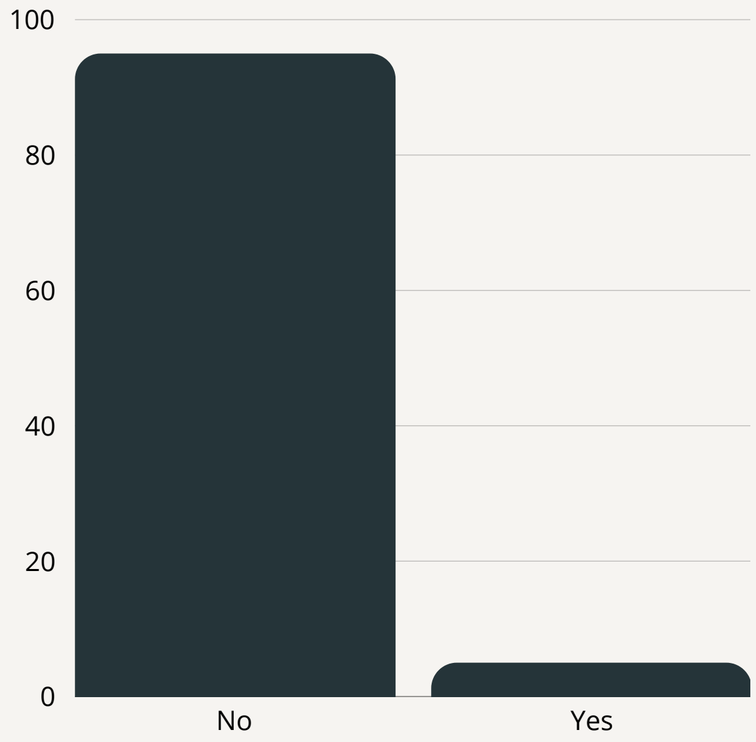
GridSearchCV without SMOTE

Confusion Matrix

Classification Report

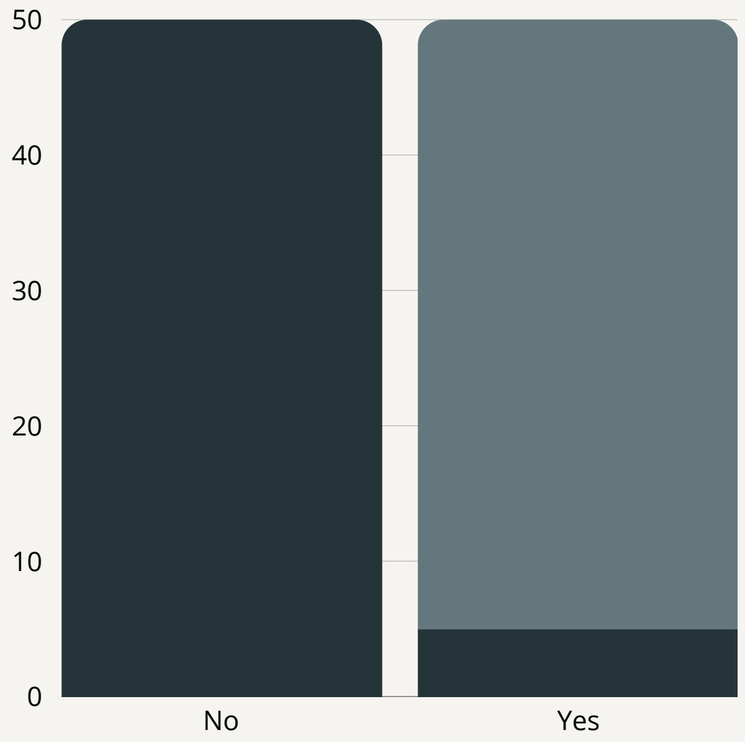
	Precision	Recall	F1 Score
0	**95%**	**100%**	**97%**
1	**0%**	**0%**	**0%**

OVERSAMPLING

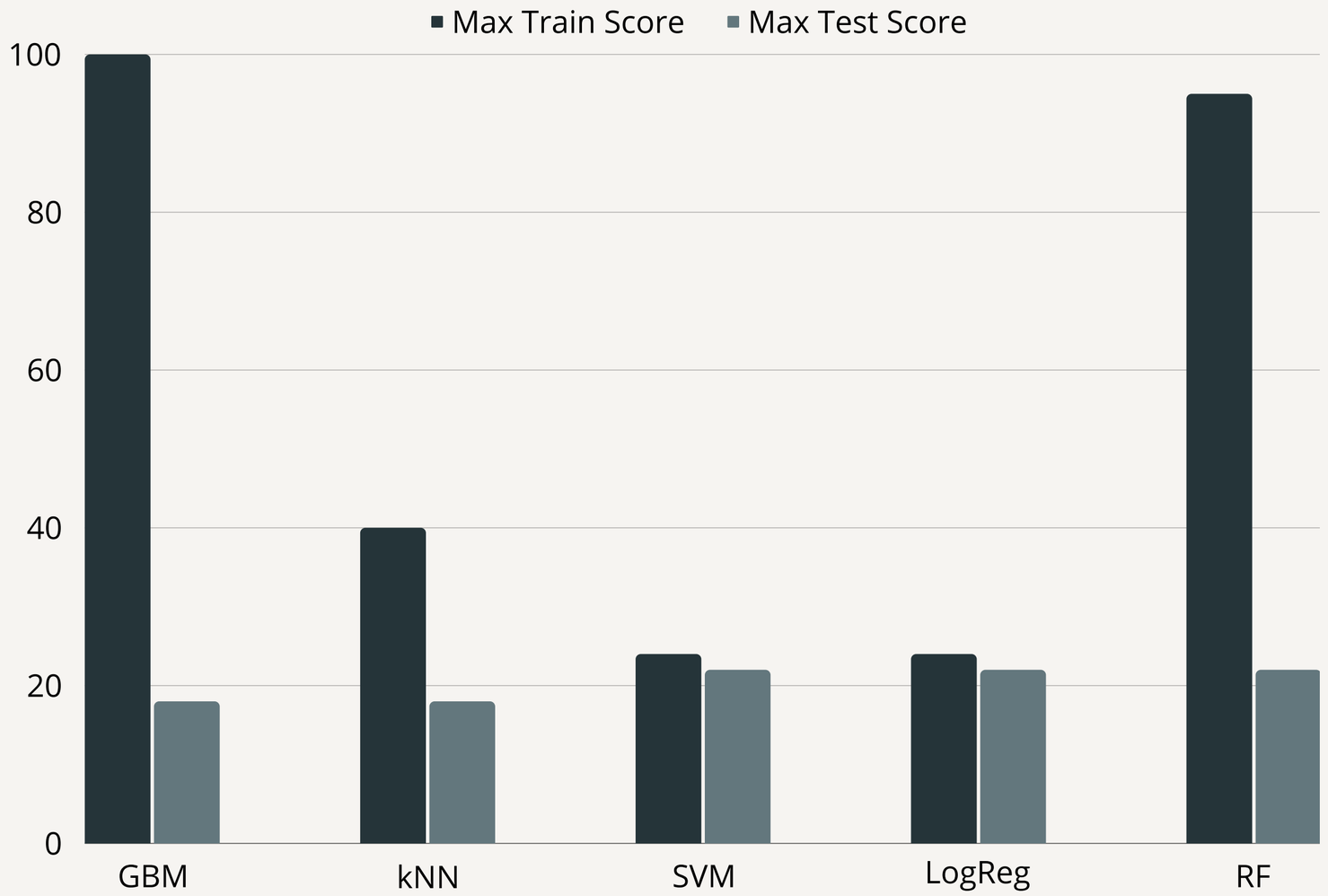


95% vs 5%

SMOTE

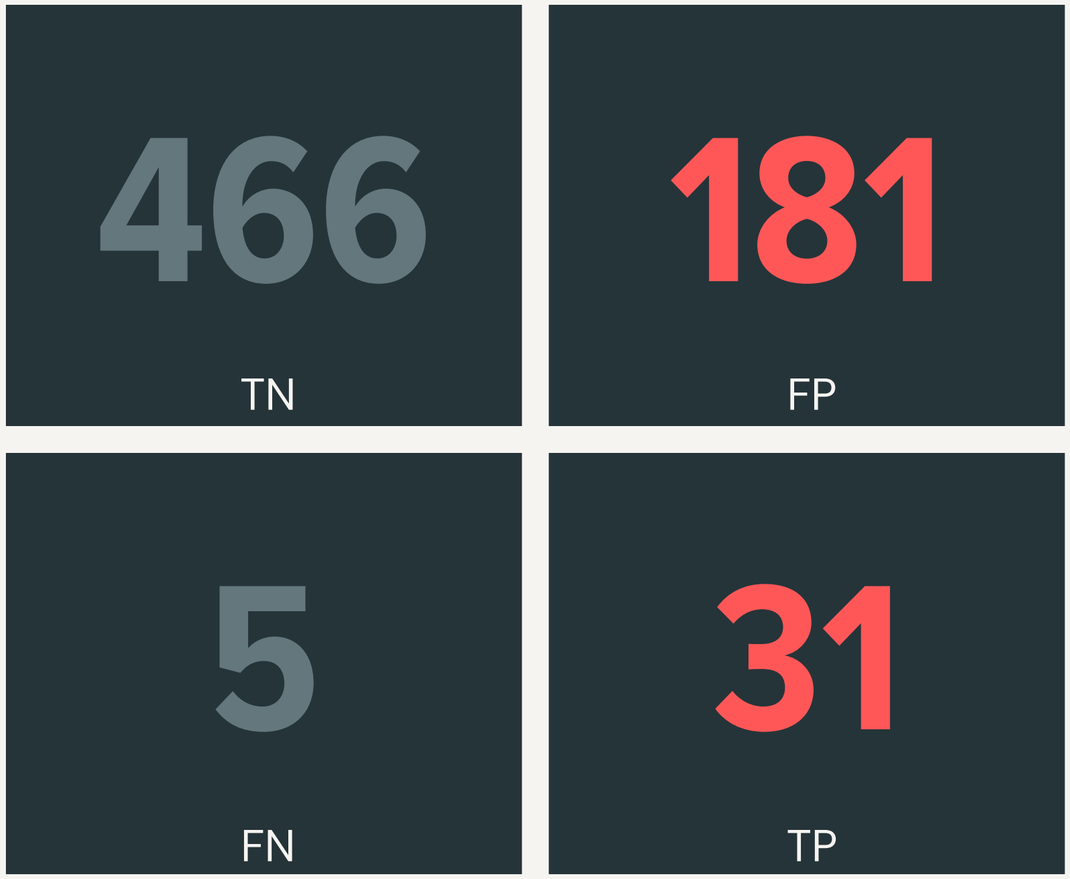


50% vs 50%



GridSearchCV with SMOTE

Confusion Matrix



Classification Report

	Precision	Recall	F1 Score
0	99%	72%	83%
1	15%	86%	25%

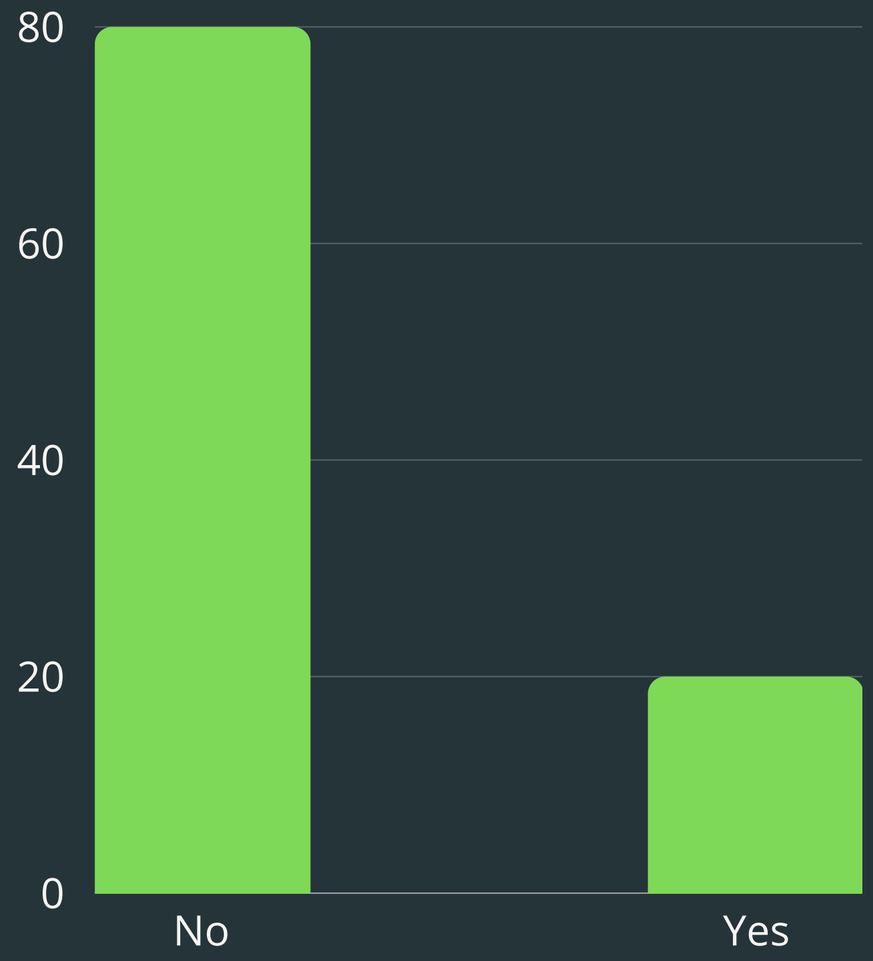
BUSINESS QUESTION

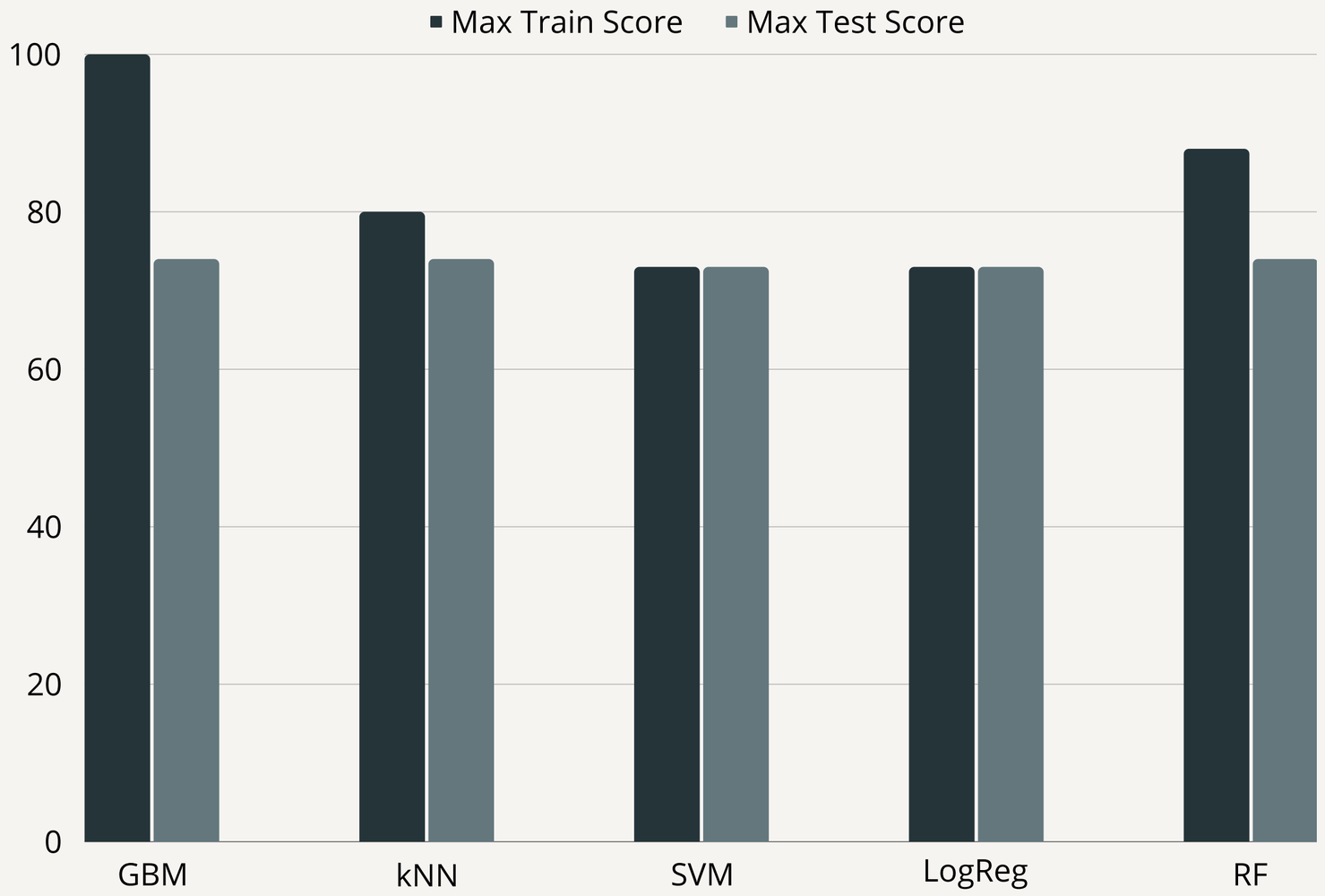
DATASET 2

Loan Default Dataset
from ***Kaggle***

11 features
- 'annual_inc', 'short_emp', 'emp_length_num', 'dti', 'last_delinq_none', 'last_major_derog_none', 'revol_util', 'total_rec_late_fee', 'od_ratio', **'bad_loan',** 'term_int'

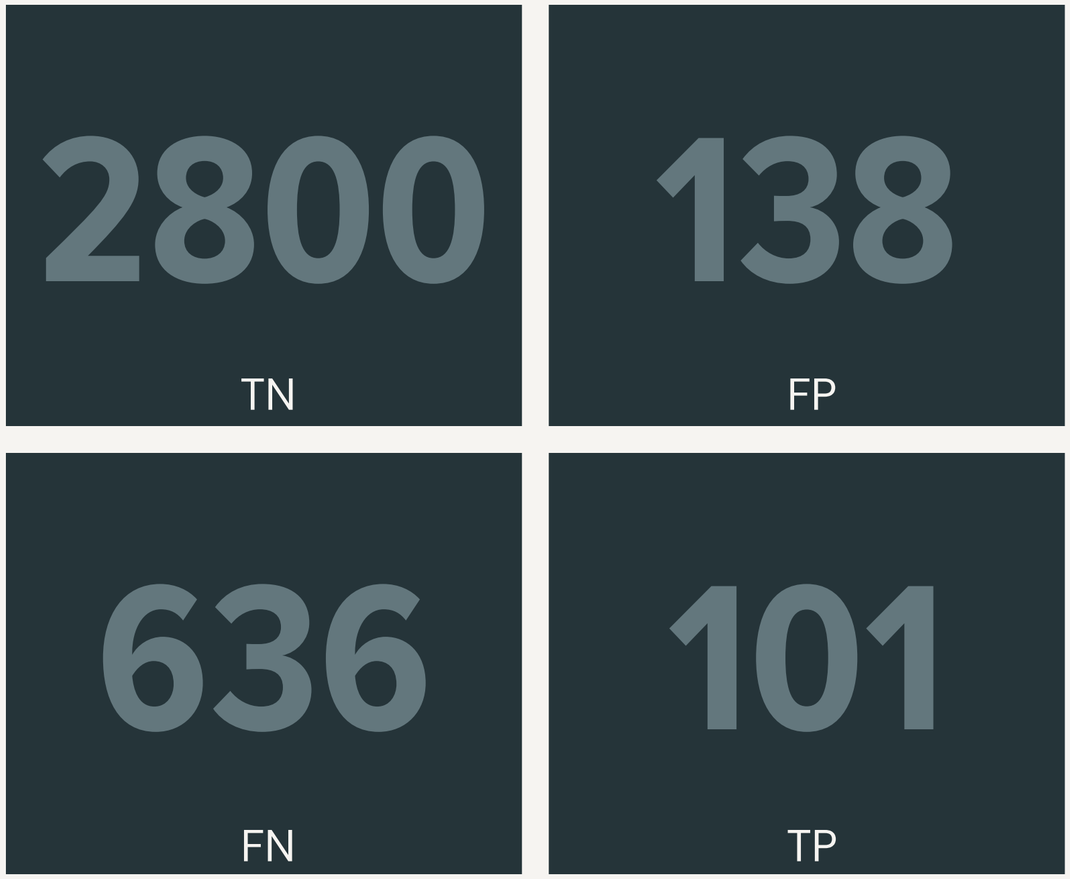
18, 371 x 11





GridSearchCV without SMOTE

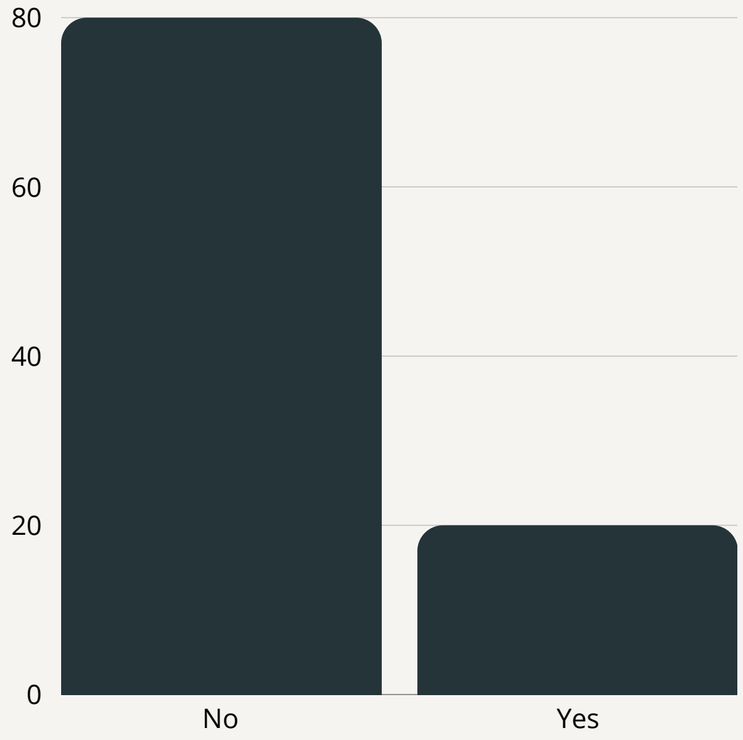
Confusion Matrix



Classification Report

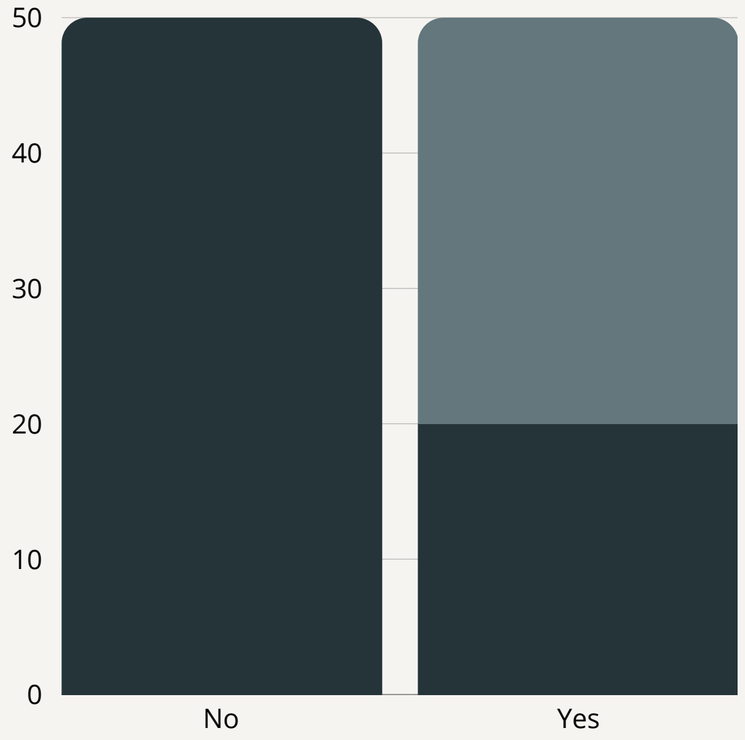
	Precision	Recall	F1 Score
0	**81%**	**95%**	**88%**
1	**42%**	**14%**	**21%**

OVERSAMPLING

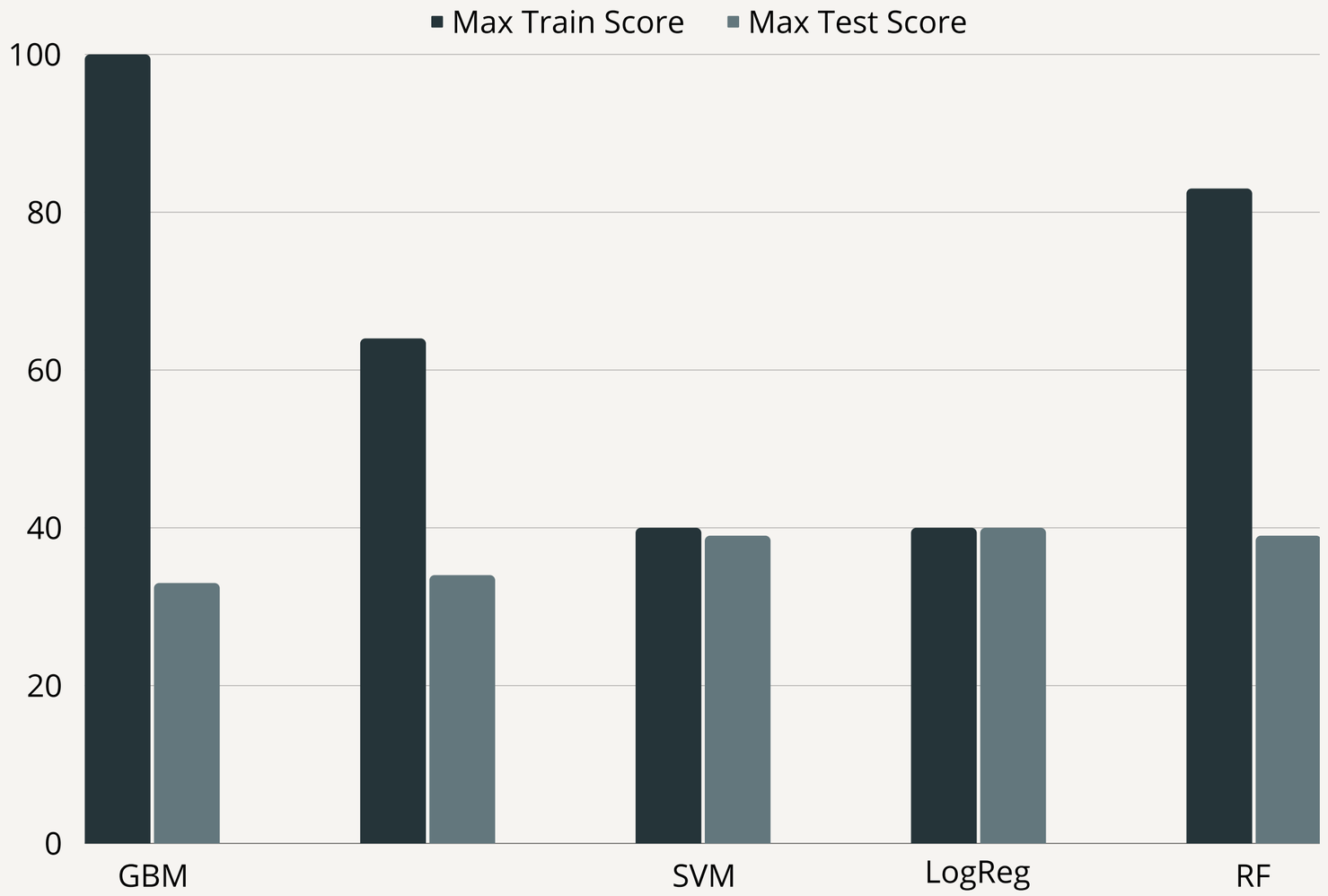


80% vs 20%

SMOTE

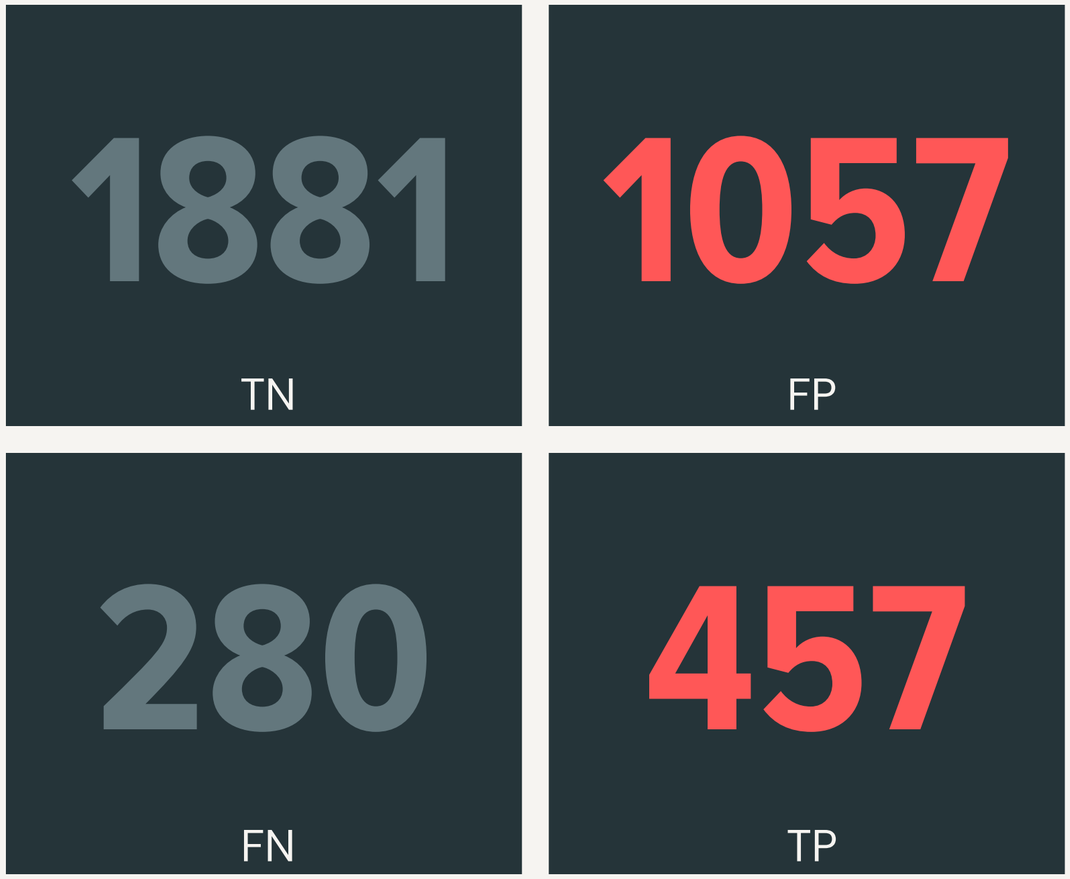


50% vs 50%



GridSearchCV with SMOTE

Confusion Matrix



Classification Report

	Precision	Recall	F1 Score
0	87%	64%	74%
1	30%	62%	41%

CONCLUSION

THANK YOU!