BANK MARKETING

ANALYSIS REPORT 0001

Hathim Azman

P153146

Farah Syahirah

P151357

Azrul Zulhilmi

P153478

Adam Suhail

P153109

EXPLORATORY DATA ANALYSIS

ABOUT DATASET

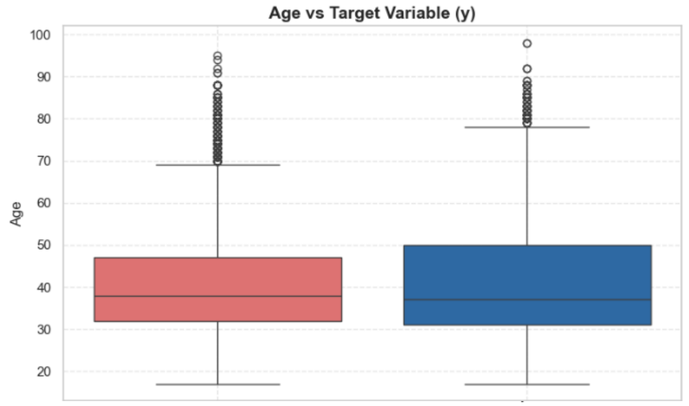
- Data was obtained from the UCI Machine Learning Repository.
- It reflects the results of direct marketing campaign conducted by banking institution.
- These campaigns were carried out through phone calls and include a wide range of customer attributes.
- The goal is to predict whether a customer will subscribe to a term deposit based on various attributes.

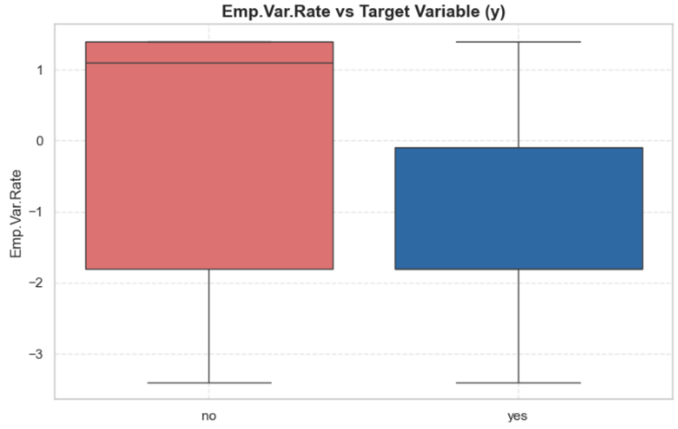
DATASET CHARACTERISTIC

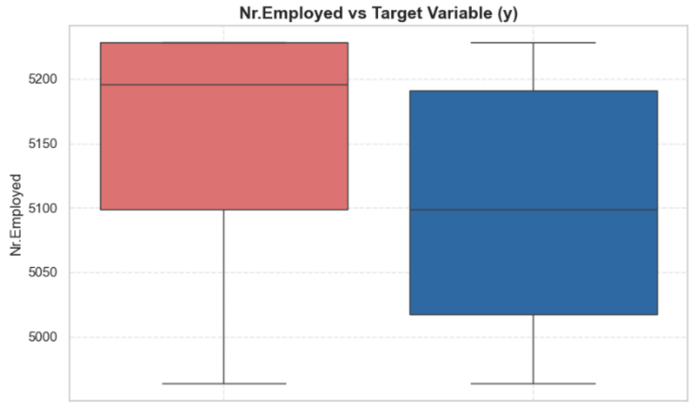
- Classification, multivariate
- 41,188 instances
- 20 features (19 predictors)
- 11 categorical variables including target variable
- 9 numerical variables
- y is our target variable, classification of both subscriber and non-subscriber to term deposit

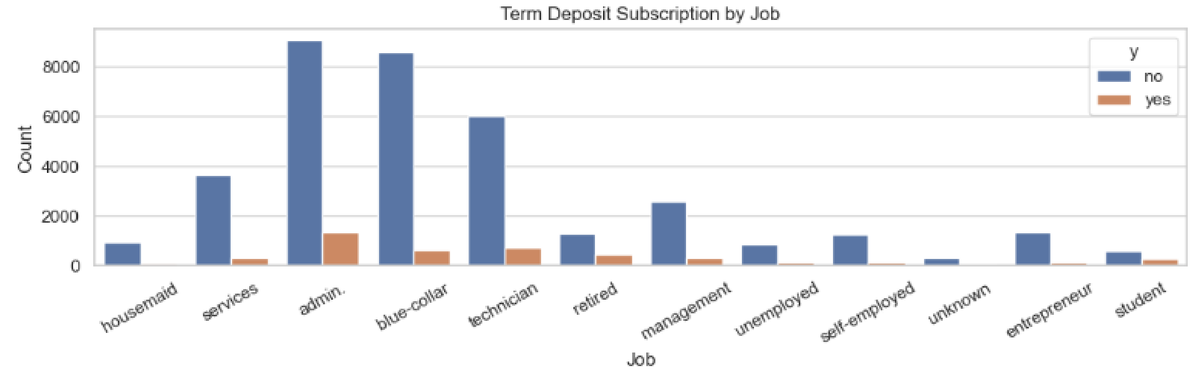
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

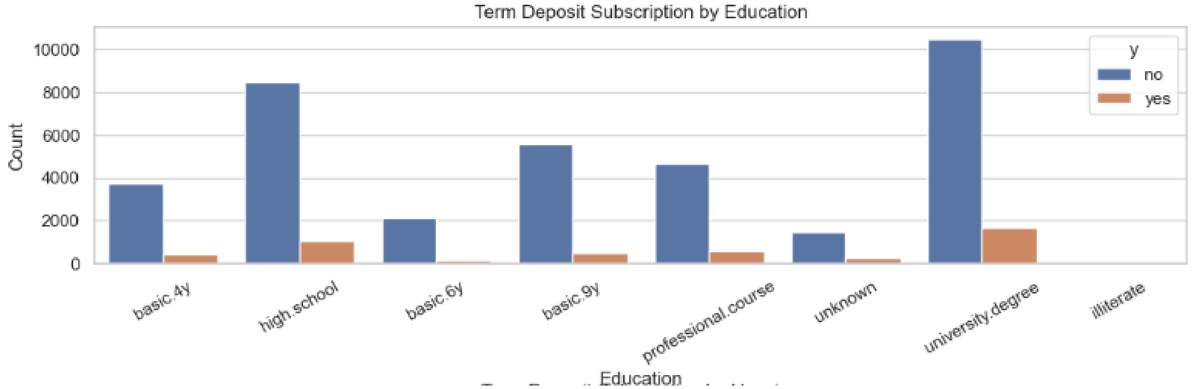
TARGET VS PREDICTOR

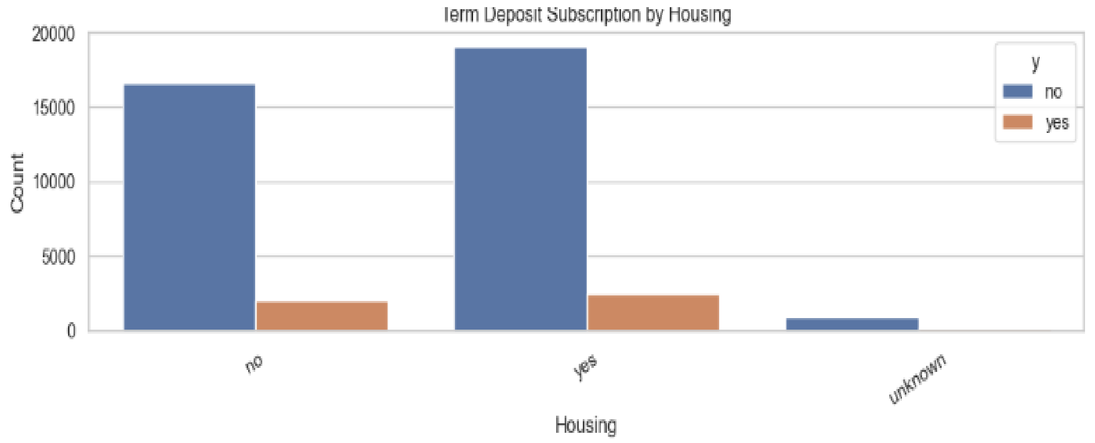


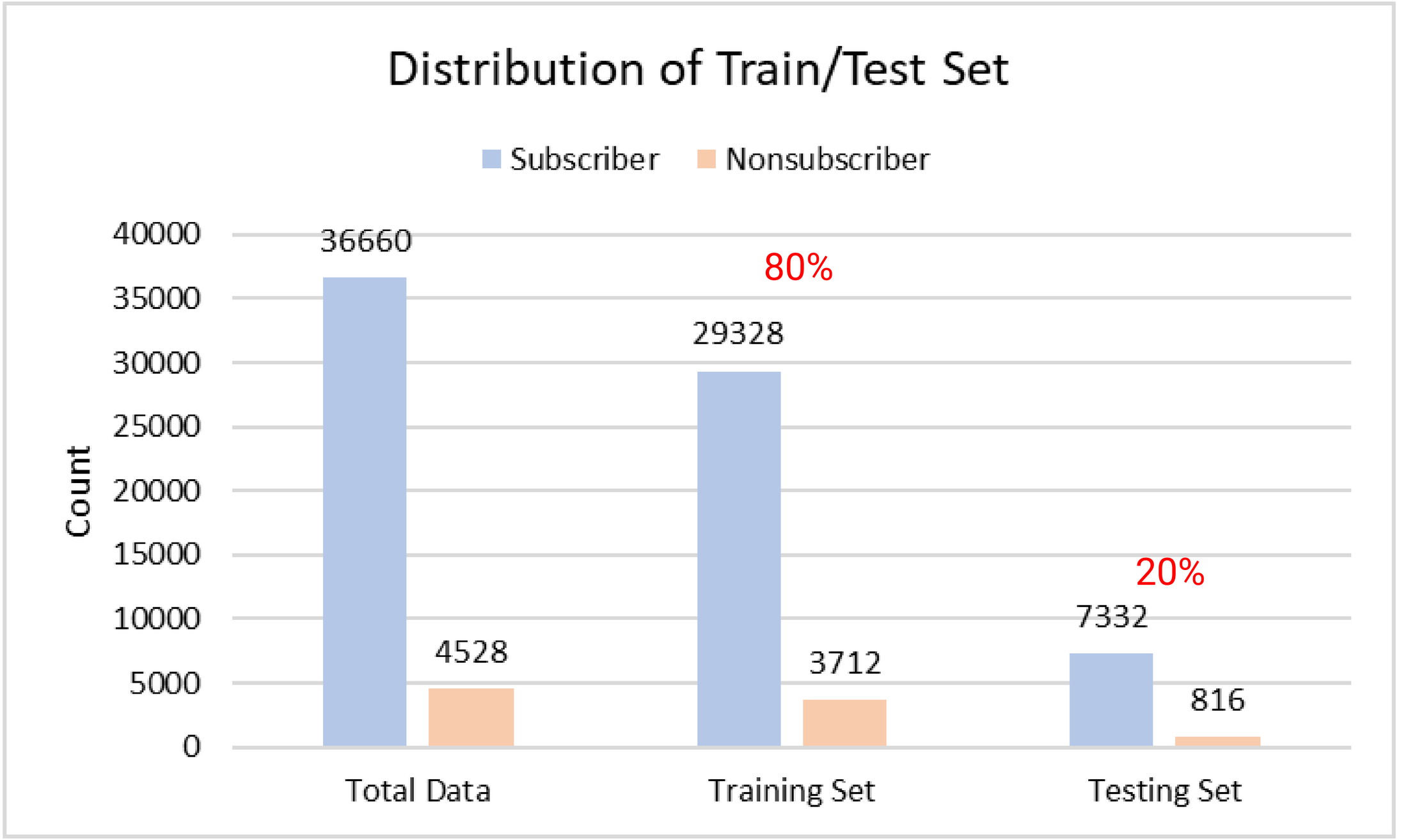










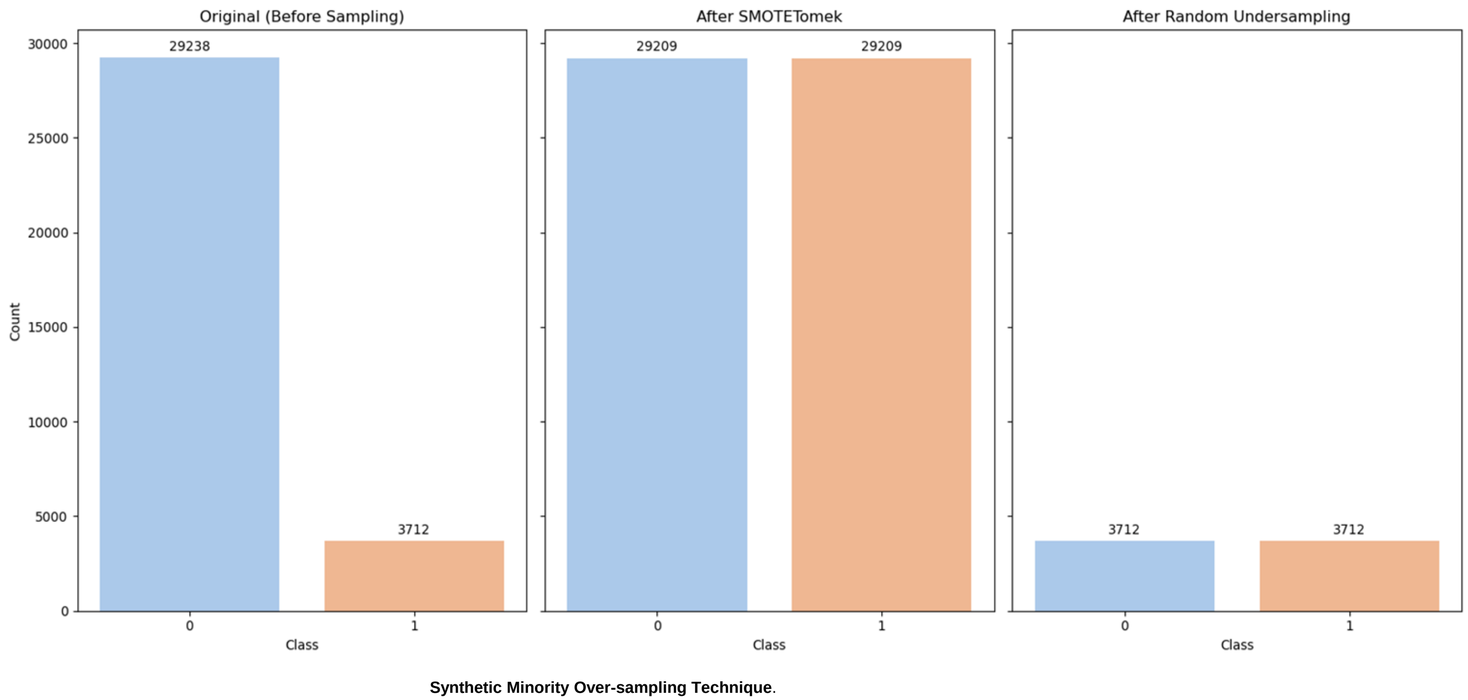

Distribution of Train/Test Set
Subscriber
Nonsubscriber
Count
40000
35000
30000
25000
20000
15000
10000
5000
0
36660
4528
80%
29328
3712
20%
7332
816
Total Data
Training Set
Testing Set

Addressing Imbalance

29,328 non-subscribers vs 3,712 subscribers
(approx 89% vs 11%)

Target Imbalance

The subscribed class (1) is rare compared to non-subscribers (0)



Random Undersampling (RUS)

Randomly removes examples from the majority class to equalize class sizes.

Synthetic Minority Over-sampling Technique.

SMOTE + Tomek Links

SMOTETomek combines SMOTE and Tomek Links to balance the classes and clean the data. It adds synthetic minority samples and removes borderline majority samples, reducing noise and improving model accuracy (Batista et al. 2004).

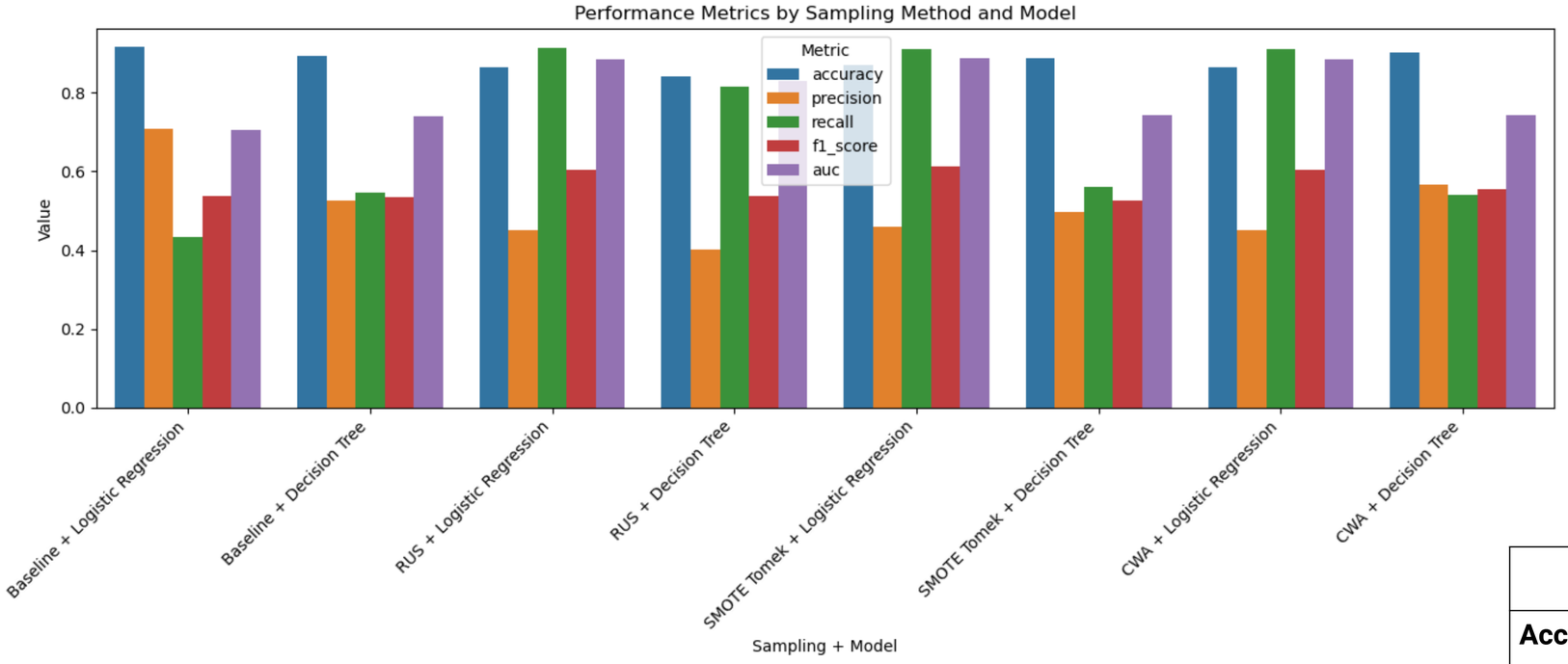
Class Weight Adjustment

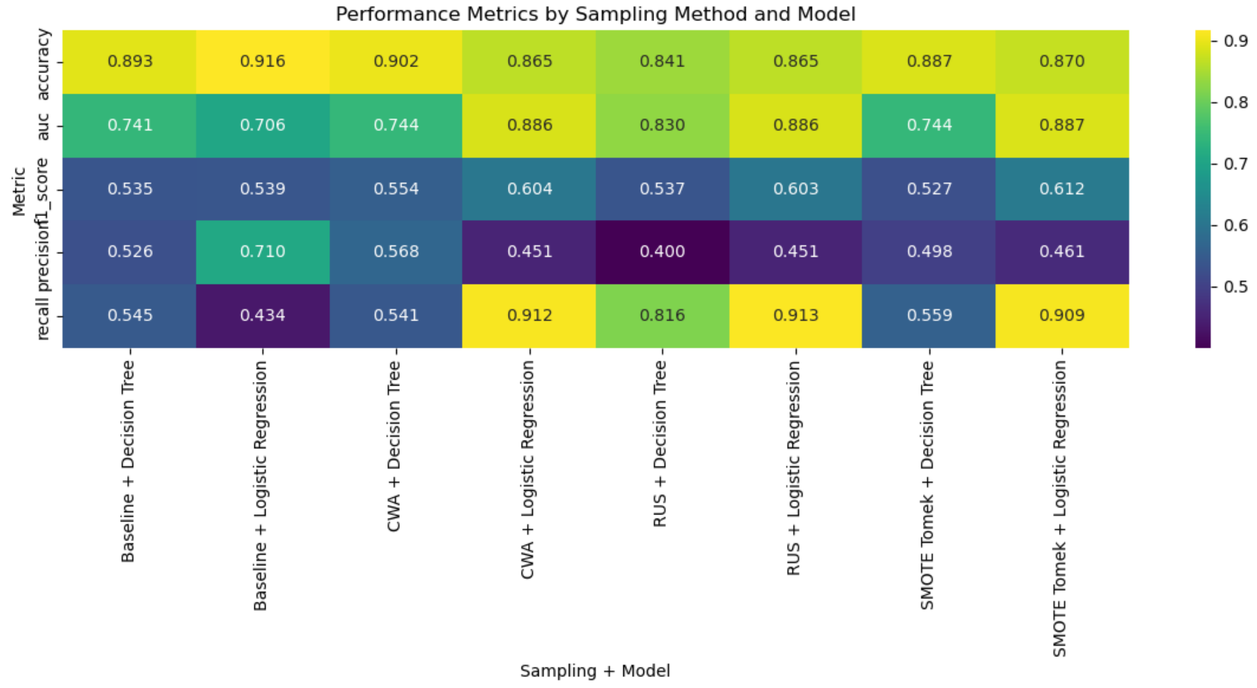
Reference: Batista, Gustavo & Prati, Ronaldo & Monard, Maria-Carolina. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations*. 6. 20-29.

PERFORMANCE METRICS OF MODELS + SAMPLING METHODS

#	Model	Sampling / Class Strategy	Classifier
1	Baseline + Logistic Regression	No Sampling	Logistic Regression
2	Baseline + Decision Tree	No Sampling	Decision Tree
3	RUS + Logistic Regression	Random Undersampling	Logistic Regression
4	RUS + Decision Tree	Random Undersampling	Decision Tree
5	SMOTE Tomek + Logistic Regression	SMOTETomek (Oversample + Clean)	Logistic Regression
6	SMOTE Tomek + Decision Tree	SMOTETomek (Oversample + Clean)	Decision Tree
7	CWA + Logistic Regression	Class Weight = 'balanced'	Logistic Regression
8	CWA + Decision Tree	Class Weight = 'balanced'	Decision Tree

Accuracy	Overall correctness of predictions (both 'yes' and 'no')
Precision	How many predicted 'yes' were actually correct — minimizes false positives
Recall	How many actual 'yes' were correctly predicted — minimizes false negatives
F1-Score	Harmonic mean of precision and recall — balances both concerns
AUC (ROC)	Area under the ROC curve — measures ability to rank positives above negative





Metric	Value
Accuracy	0.8699
Precision	0.4607
Recall	**0.9095** ✅
F1-Score	**0.6116** ✅
AUC	**0.8872** ✅

Best Model Based on Metric: SMOTE TOMEK + Logistic Regression

IMPLICATIONS THROUGH SIMULATION

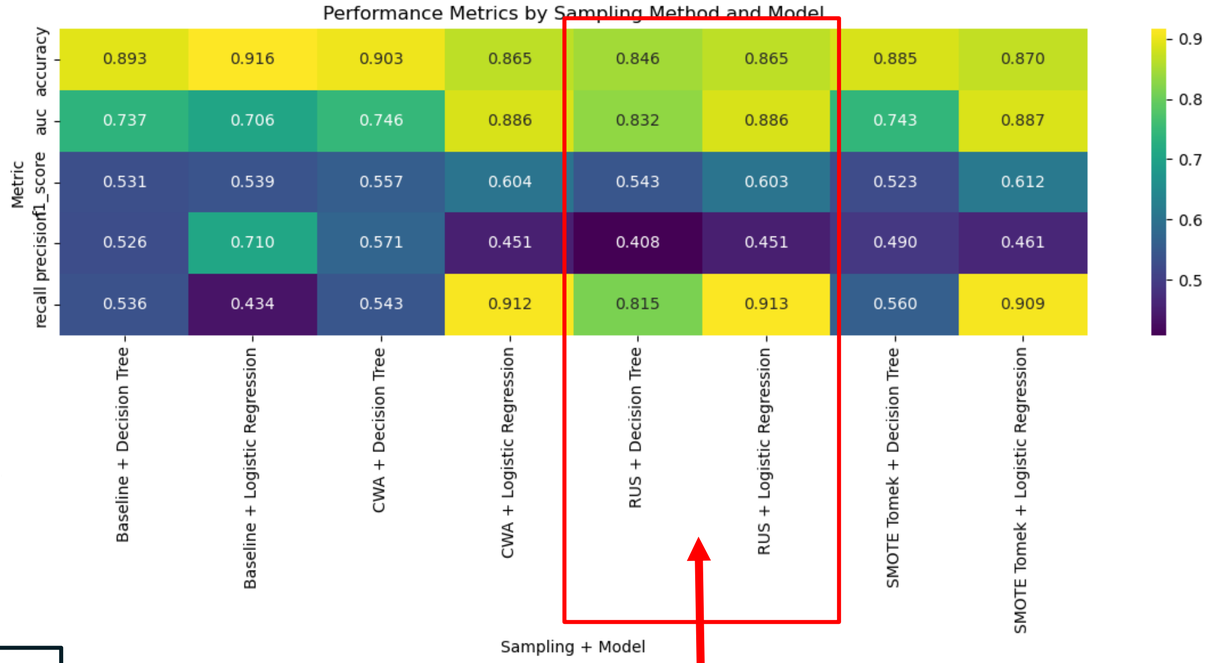
A **naive approach** is to contact all 100,000 customers **without adjusting** to possible True Positives or False Positives, and **compare them to different models.**

Simulation parameters:

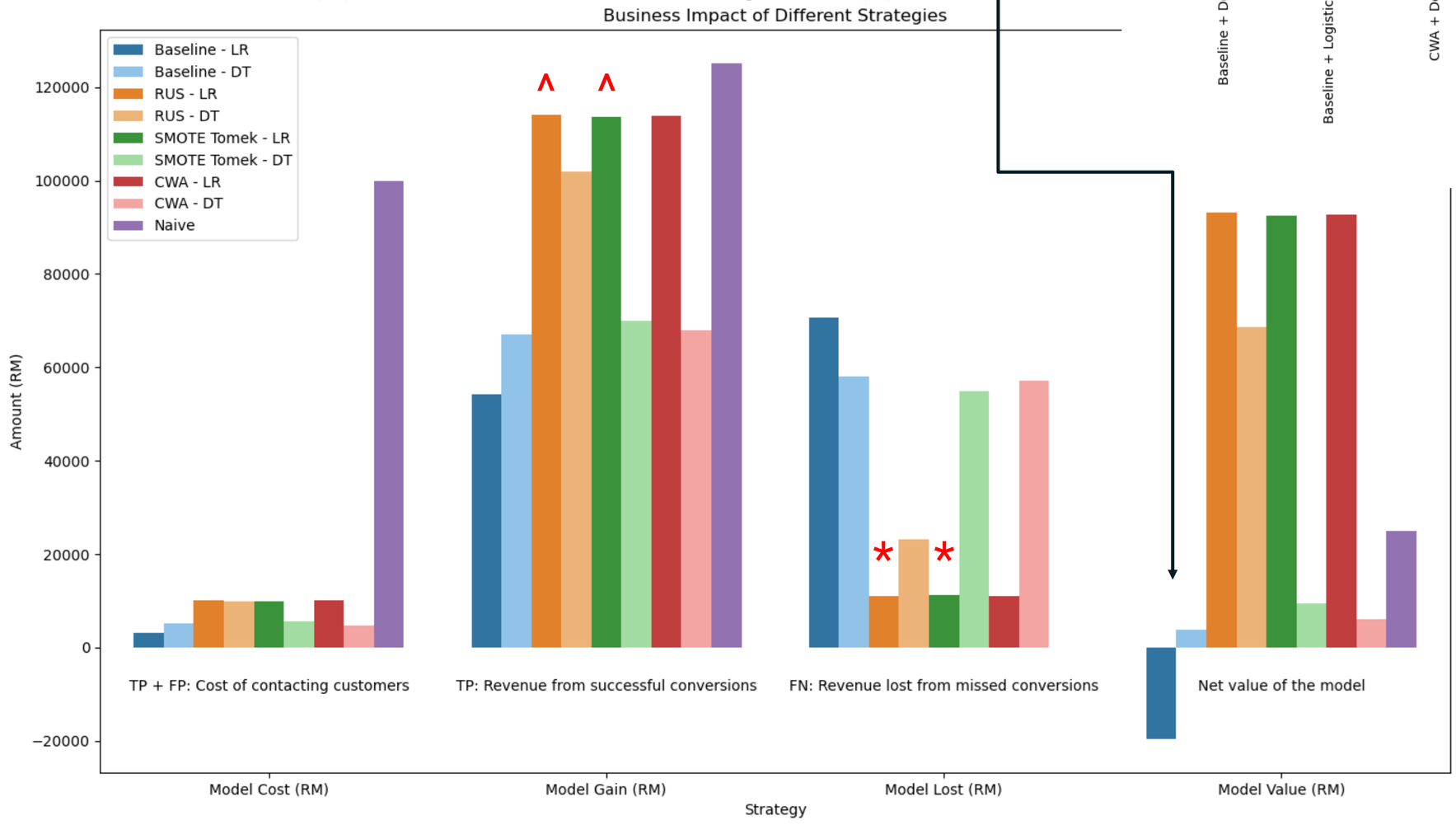
- 100,000 Simulated Customers
- Customer Conversion Rate : 5%
- Advertisement Cost for 1 Customer : RM 1
- Profit for Successful Conversion : RM 25

Metric	Formula	Description
Contacted	True Positive + False Positive	Total customers the model predicts as positive and are contacted
Model Gain (RM)**	True Positive * Profit for Successful Conversion	Revenue from correctly predicted conversions
Model Lost (RM)**	False Negative * Profit for Successful Conversion	Potential revenue lost from missed actual positives (not contacted)
Model Cost (RM)**	Contacted * Advertisement Cost	Total cost of contacting customers
Model Revenue (RM)	Model Gain - Model Cost	Revenue after deducting contact cost
Model Value (RM)**	Model Gain - Model Cost - Model Lost	Net business value considering both cost and missed opportunities

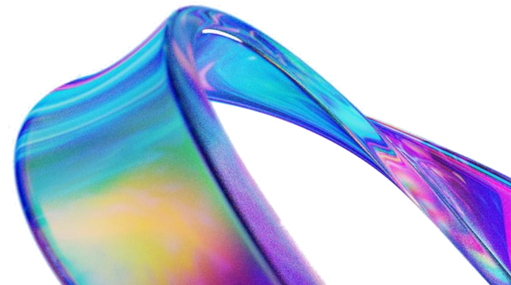
Baseline models may have high Accuracy and Precision but **perform worse than other models**!



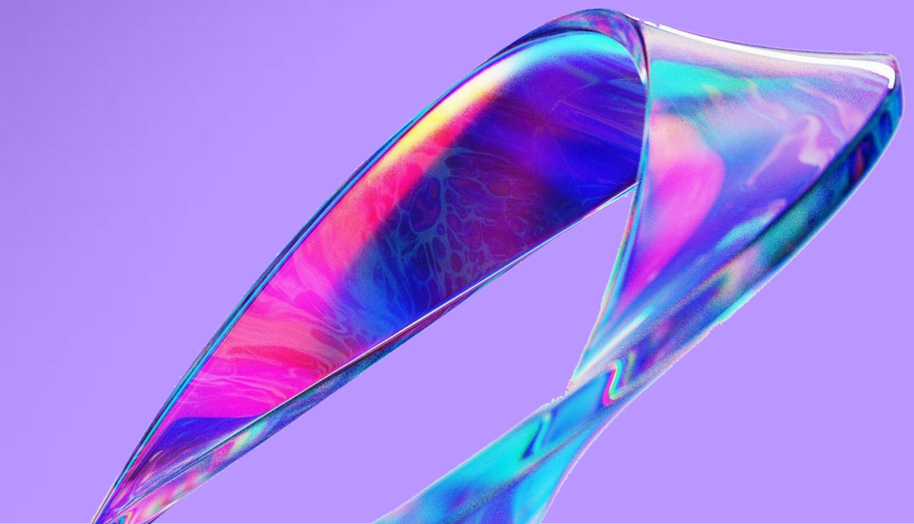
Adjusting for imbalance allows models to **recover minority signals**, leading to **higher true positives(^)** and much **lower false negatives(*)**.



Adjusting for class imbalance **improves precision and recall**. A balanced approach is F1-score / AUC. Improving Model Value.



WHAT DOES THIS MEAN?



Random Under Sampler - Logistic Regression shows best performance

1. Adjusting for class imbalance allows model to more accurately find true positives
2. No single metric is the best at identifying best model
3. Understand the objective of modelling, ie; lowest cost/highest profit/most accurate model.