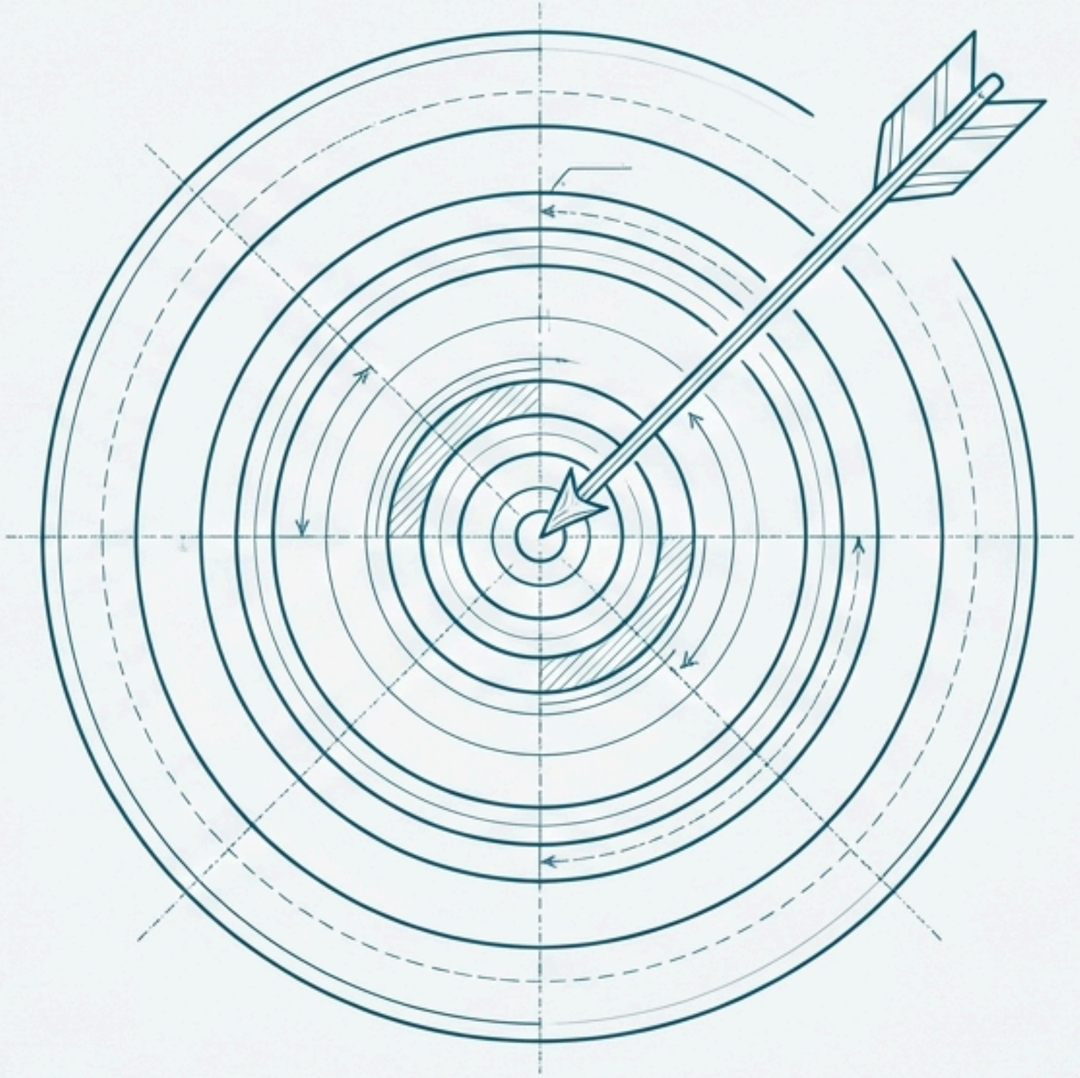
From Raw Data to Actionable Intelligence

A Rigorous Analysis of Customer Subscription Prediction Using a Decision Tree Model

The Objective: Build a Predictive Model to Identify Potential Subscribers

- The primary goal is to analyze customer demographic and behavioral data from a bank marketing campaign.
- We will build and validate a Decision Tree model capable of predicting whether a customer will subscribe to a new product or service.
- This model aims to provide interpretable insights to inform future marketing strategies.



The Foundation: An Overview of the Bank Marketing Dataset

41,188

Customer Records

20

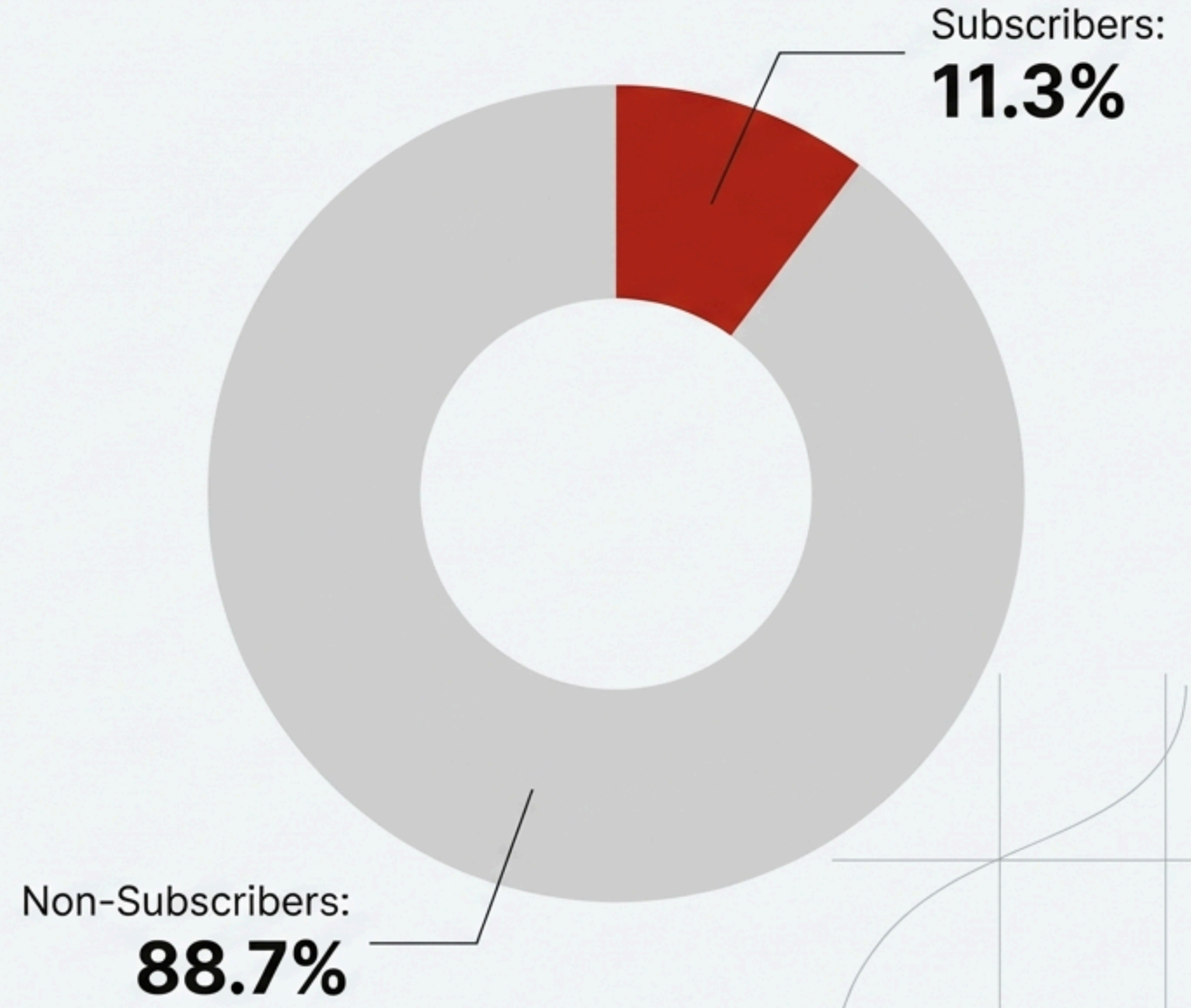
Input Features (demographic and behavioral attributes)

1

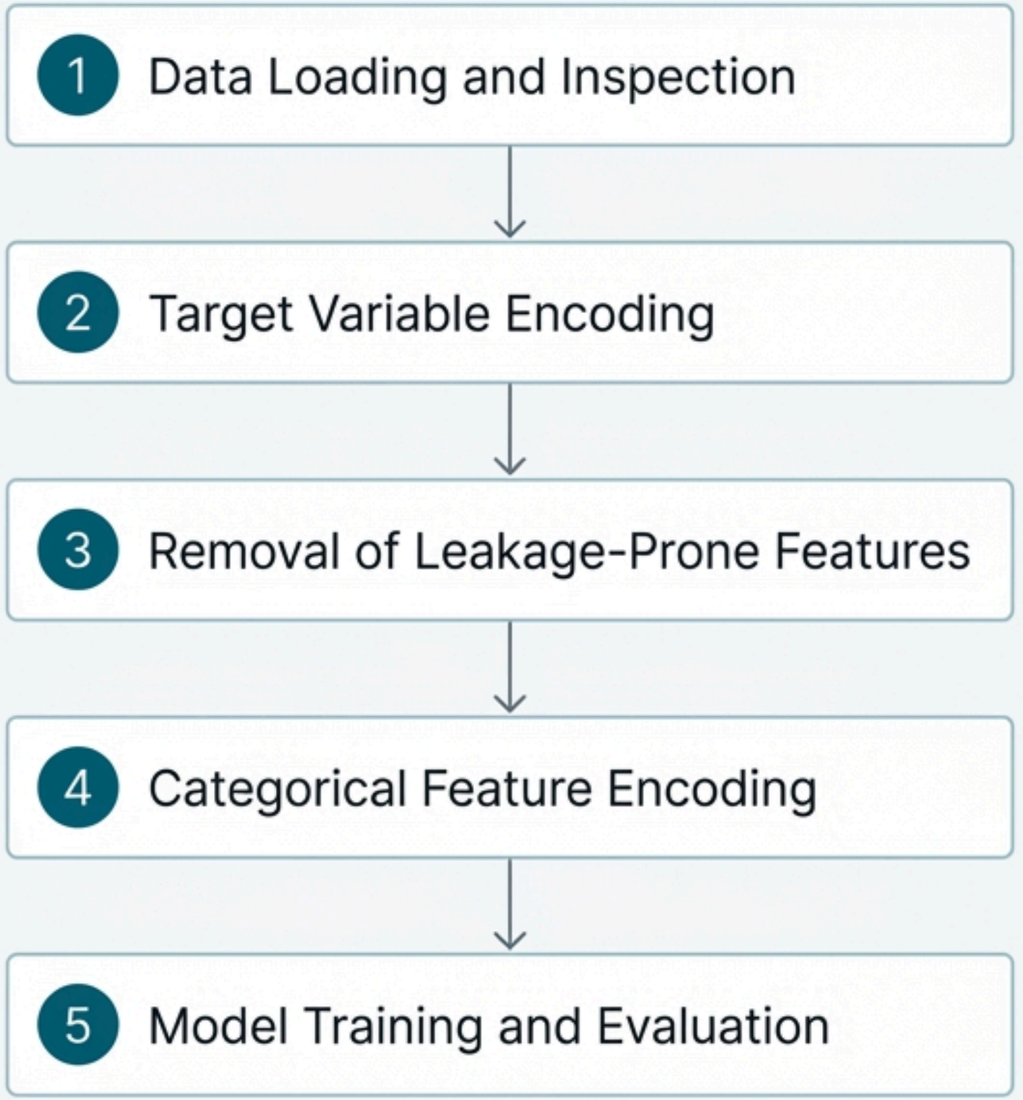
Target Variable (‘y’): Subscription (Yes/No)

The Core Challenge: A Highly Imbalanced Dataset

The most significant characteristic of the data is the severe imbalance between the two classes. Failing to address this would lead to a biased model that defaults to predicting ‘No Subscription.’



A Disciplined Workflow for Reliable and Repeatable Results



Data Preparation: Mitigating Bias and Ensuring Integrity

Leakage Prevention

Removed the `duration` column. This feature is unknown before a call is made and would introduce post-event bias, creating an unrealistically optimistic model.

Feature Engineering

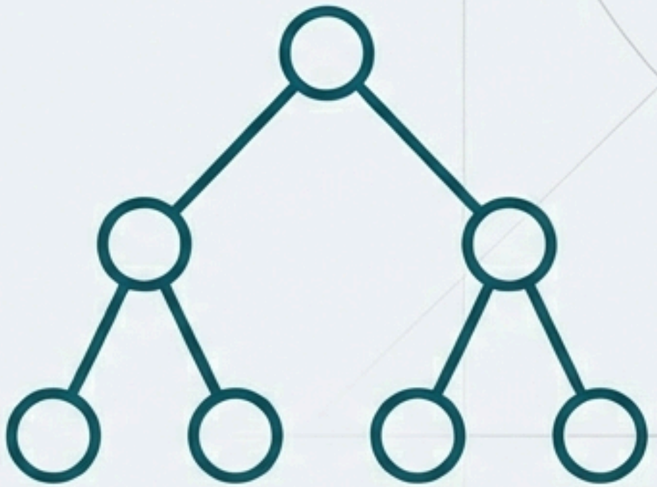
Encoded all categorical variables using One-Hot Encoding, expanding the feature space to 196 distinct features.

Preserving Balance

Used an 80/20 stratified train-test split to ensure the 11.3% subscriber rate was maintained in both the training and testing sets.

Model Selection: Balancing Predictive Power with Interpretability

Decision Tree Classifier



Justification

- **Interpretability:** Provides clear, human-readable decision rules.
- **Flexibility:** Natively handles non-linear relationships in the data.
- **Insight Generation:** Allows for direct extraction of feature importance.

Key Parameters

- **Max Depth:** Limited to prevent overfitting on the training data.
- **Class Weights:** Set to ‘balanced’ to force the model to pay more attention to the minority ‘Subscriber’ class.

The Verdict: Model Achieves 81% Overall Accuracy

The trained model demonstrates strong overall performance in classifying customers. However, accuracy alone doesn't tell the full story, especially with an imbalanced dataset.

81%

Accuracy

67%

Recall (Subscribers)

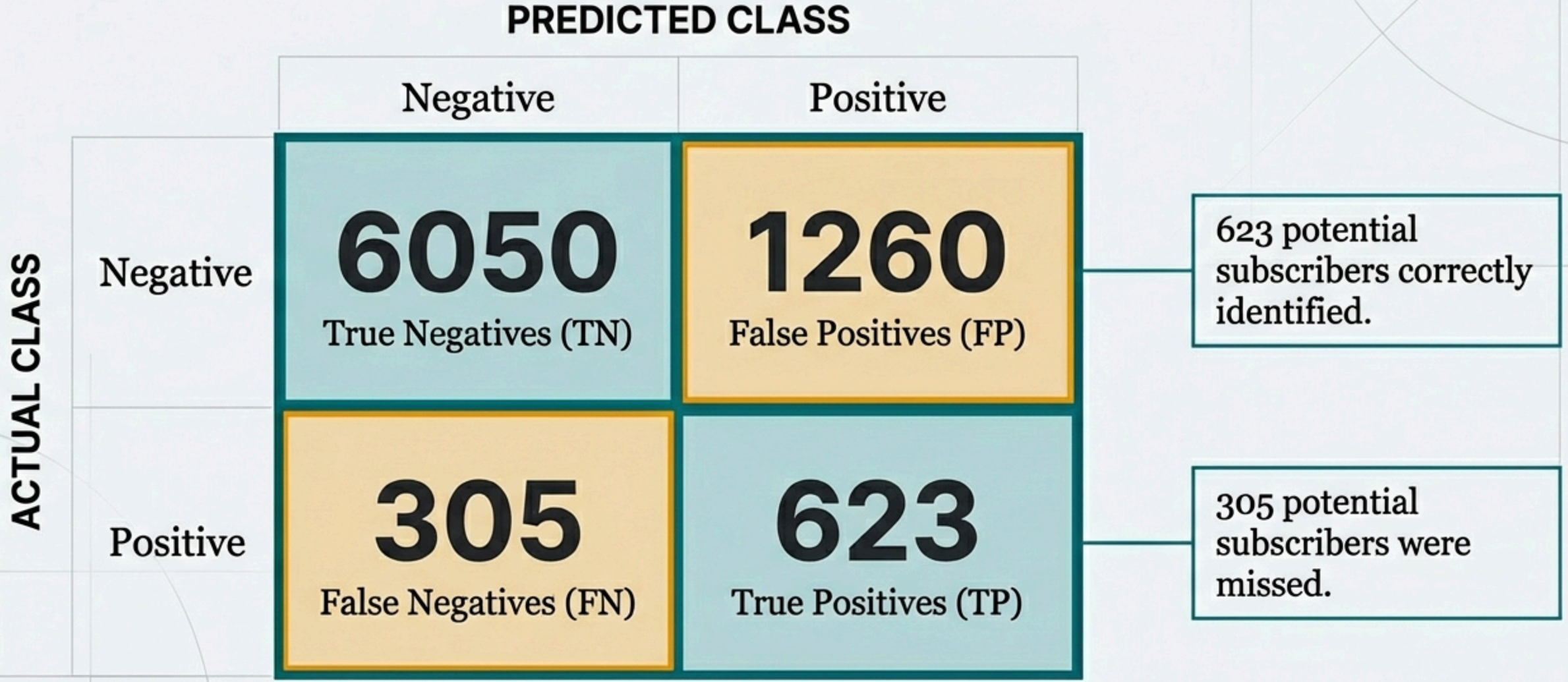
33%

Precision (Subscribers)

0.44

F1-Score (Subscribers)

Dissecting Predictions: A Look Inside the Confusion Matrix



The Strategic Choice: Prioritizing Recall to Maximize Lead Generation

For a marketing campaign, it is more valuable to identify a larger pool of potential subscribers (high Recall), even if it means contacting contacting some non-subscribers (lower Precision). Missing a potential customer (a **False Negative**) is more costly than reaching out to a non-prospect (a False Positive).

Our model is optimized for this business need, successfully identifying **67%** of all actual subscribers in the test set.

Key Drivers: Economic Conditions and Past Engagements Most Influence Subscription

The model's feature importance results reveal that macroeconomic indicators and a custom-npr's prior interaction history are the most critical factors in their decision to subscribe.

Putting the Model to Work: A Sample Prediction

To illustrate its practical application, the trained model was used to evaluate a single customer instance from the test data.

Input

Test Customer Instance

Decision Tree Model

Output

Predicted Outcome:

Customer will subscribe → **NO**

Conclusion: An Effective Model Delivering Actionable Intelligence

- **Accurate Prediction:** The model effectively predicts customer subscription behavior with 81% accuracy and high recall for the target class.
- **Clear Interpretability:** The Decision Tree provides transparent decision rules and highlights the most influential factors.
- **Actionable Insights:** Feature importance results offer clear guidance for marketing strategy, emphasizing economic factors and customer history.

This implementation successfully completes **Task 3** as specified.