Optimizing Bank Marketing: A Predictive Model for Customer Subscription

Biplov Sharma

Data Science WiSe 2025

The Challenge: Moving from Mass Marketing to Intelligent Targeting

The Business Problem

Direct marketing campaigns are expensive and can fatigue customers if poorly targeted.

Goal: Identify clients most likely to subscribe to a term deposit to optimize marketing efforts.

Why This Matters

- **Reduce Marketing Costs:** Focus budget on high-potential leads.
- **Improve Campaign Effectiveness:** Increase the subscription rate per contact.

The Data Science Goal

- Build a robust predictive model that ranks clients by their likelihood of subscribing.
- The model should serve as a decision-support tool, not an automated system.

Our Structured Approach: From Raw Data to Actionable Insights

Data Understanding & EDA

Discovering the data's character and challenges.

Data Preparation

Forging clean, Forging clean, model-ready data without leakage.

Modeling

Training and tuning multiple algorithms to find the best fit.

Evaluation & Interpretation

Assessing performance through a business lens to drive value.

Understanding the Landscape: The Bank Marketing Dataset

Key Metrics

Source
UCI Machine Learning Repository

Observations
41,188 client contacts

Input Features
20

Time Period
2008-2010

A Critical Note on Data Leakage

The `duration` feature (call length) is a powerful predictor but is only known *after* the call.

Decision: It was excluded from all predictive models to ensure a realistic, deployable solution.

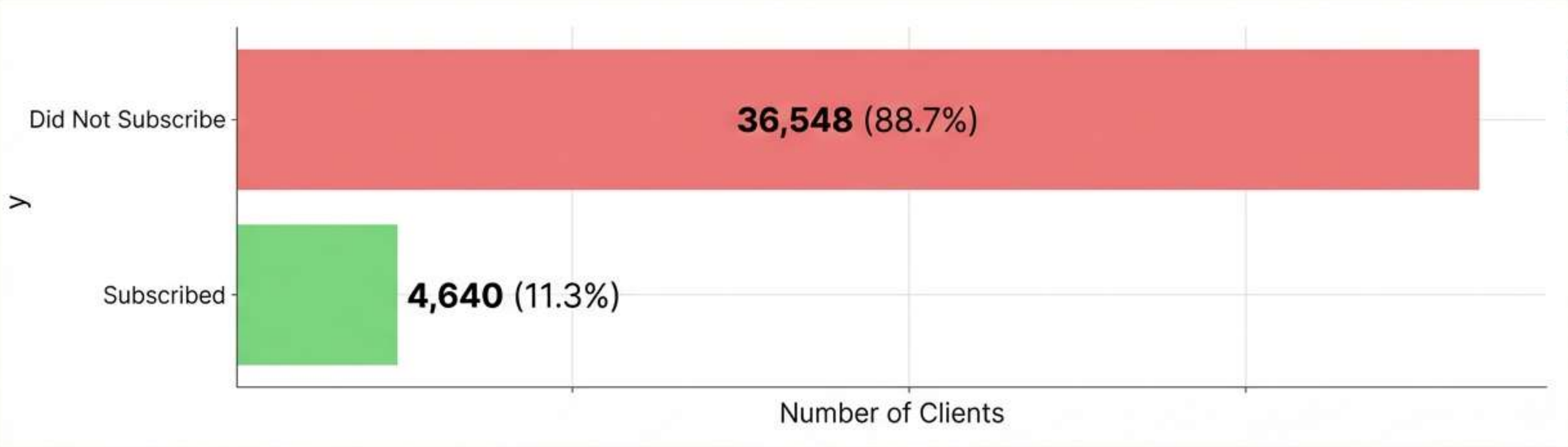
Feature Categories Breakdown

Client Attributes
`age`, `job`, `education`, `marital`, `default`, `housing`, `loan`

Campaign Contact Data
`contact`, `month`, `campaign`, `pdays`, `previous`, `poutcome`

Economic Context
`emp.var.rate`, `cons.price.idx`, `euribor3m`, `nr.employed`

The Core Challenge: Subscriptions are a Rare Event

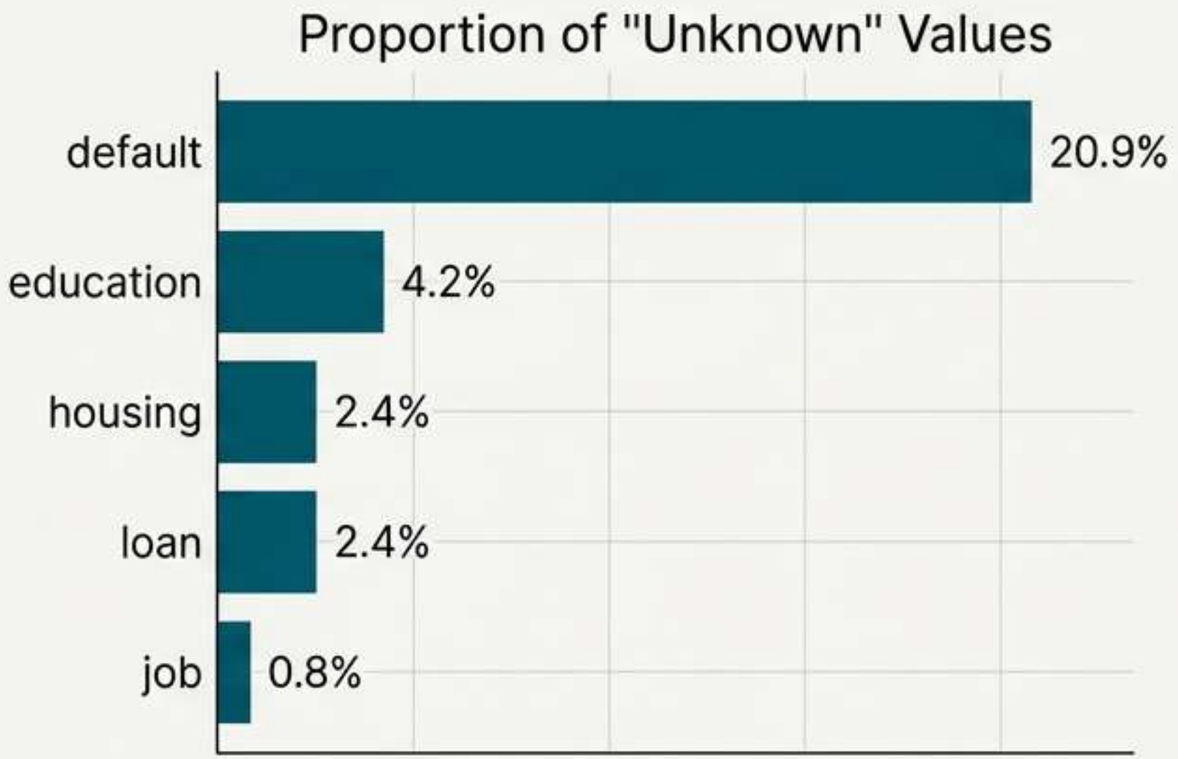


Implications for Modeling

- The dataset is highly imbalanced.
- Standard accuracy is a misleading metric. A model that always predicts 'No' would be ~89% accurate but commercially useless.
- **Our Focus:** Evaluation will prioritize metrics like ROC-AUC, Recall, and Precision that are robust to imbalance.

Uncovering Signal in Missing Information

The dataset contains no standard null values, but “unknown” appears in several categorical columns.



Why “Unknown” is a Feature, Not a Flaw

- **Hypothesis:** "Unknown" is not random missing data. It represents meaningful information (e.g., a client’s refusal to answer).
- **Decision:** Treat "unknown" as a distinct category during encoding.
- **Benefit:** This allows the model to learn if the absence of information is itself a predictive signal, preserving valuable data that would otherwise be dropped.

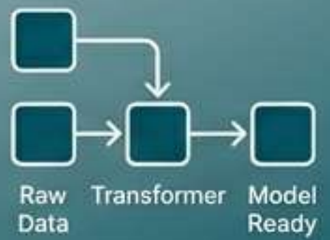
Engineering a Model-Ready and Leak-Proof Dataset

Domain-Informed Feature Engineering

- **pdays_transformed**: Converted 999 (never contacted) to -1 to treat it as a special case, not a large number.
- **contacted_before**: Created a binary flag from pdays to clearly separate new vs. existing contacts.
- **total_contacts**: Combined campaign and previous to capture a client's total exposure to marketing.

Systematic Preprocessing

- **Categorical Features**: One-Hot Encoded to convert text into numerical format.
- **Numerical Features**: Standard Scaled to normalize ranges and prevent feature dominance.



Ensuring Reproducibility and Preventing Leakage

All steps were encapsulated in a scikit-learn ColumnTransformer and Pipeline.

Benefit: This guarantees that the same transformations are applied consistently to training and testing data, preventing data leakage and making the entire workflow reproducible.

Selecting the Right Tool: A Multi-Model Approach

We trained and compared four distinct models to cover a range of behaviors and complexities.

1. **Logistic Regression**: A powerful and interpretable baseline.
2. **Decision Tree:** To capture non-linear rules and interactions.
3. **Random Forest**: An ensemble model for high performance and robustness against overfitting.
4. **Weighted Random Forest:** A specialized version of Random Forest with `class_weight='balanced'` to explicitly address the class imbalance identified in our EDA.

Training and Tuning Strategy

- **Data Split:** Used a stratified train-test split to maintain the target distribution in both sets.
- **Tuning:** Employed `GridSearchCCV` with 5-fold cross-validation to systematically find the best hyperparameters for each model.
- **Scoring Metric:** Optimized for **ROC-AUC**, as it effectively measures a model's ability to discriminate between classes, which is crucial for our ranking goal.

The Verdict: Performance Across Key Business Metrics

Model	ROC-AUC	Recall	Precision	F1-score	Accuracy
Random Forest Weighted	**0.813**	**0.638**	**0.414**	**0.502**	**0.857**
Decision Tree	0.810	0.273	0.622	0.379	0.899
Random Forest	0.810	0.206	0.721	0.320	0.902
Logistic Regression	0.801	0.224	0.696	0.339	0.902

Highest ROC-AUC: The Weighted Random Forest is best at ranking clients by subscription likelihood.

Misleading Accuracy: The non-weighted models show higher accuracy (~90%) by defaulting to the majority "No" class, but they fail to identify most actual subscribers (low Recall).

The Critical Trade-Off: Notice the inverse relationship between Recall and Precision. We will explore this next.


Model Performance Comparison Across Metrics
Metric Value
0
0.2
0.4
0.6
0.8
0.857
0.414
0.638
0.502
0.813
0.899
0.622
0.273
0.379
0.810
0.902
0.721
0.206
0.320
0.810
0.902
0.696
0.224
0.339
0.801
Random Forest Weighted
Decision Tree
Random Forest
Logistic Regression
Model
Metric
Accuracy
Precision
Recall
F1-score
ROC-AUC

Why the Weighted Random Forest is the Right Business Choice

High Precision

What it means: When the model predicts "Yes", it's very likely to be correct.

Business Cost: We miss a large number of potential subscribers (low Recall). **High cost of missed opportunity.**

High Recall

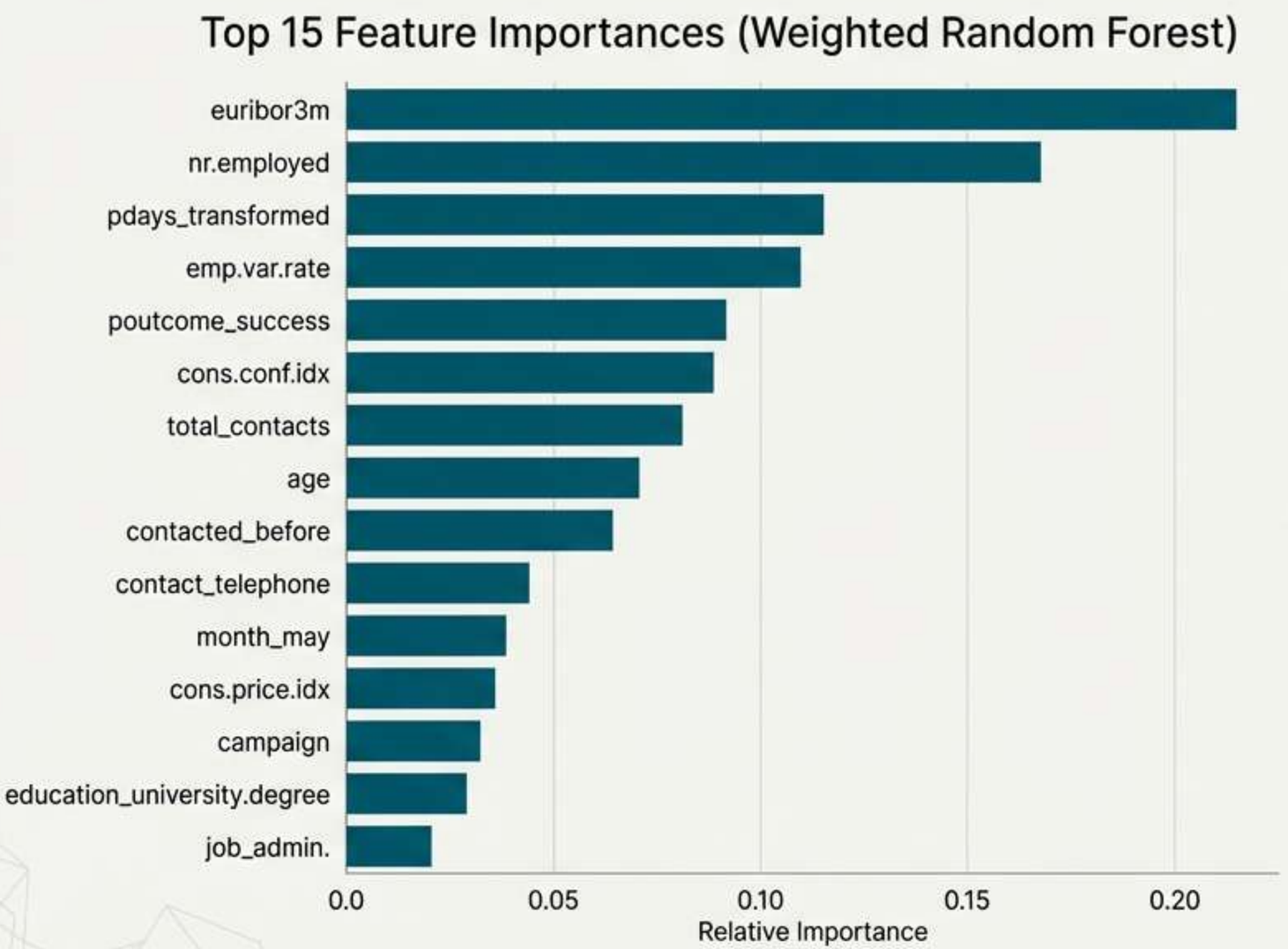
What it means: The model successfully identifies a large portion of all actual subscribers.

Business Cost: We contact some clients who won't subscribe (lower Precision). **Acceptable cost of marketing spend.**

For a growth-oriented marketing campaign, **missing a potential customer (False Negative) is more costly than contacting a non-subscribing one (False Positive).**

The Weighted Random Forest's high Recall (0.638) makes it the superior model for maximizing lead generation.

Decoding the Model: What Factors Drive Subscription?



Economic Context is King

Macroeconomic indicators are the most powerful predictors. Subscription is highly sensitive to the economic environment.

Client History Matters

How and when a client was contacted before are strong signals of future behavior.

Demographics Play a Supporting Role

'age' is moderately important, but less so than economic and behavioral factors.

Quantifying the Impact: A Look at Predicted vs. Actual Outcomes

The model successfully identifies **63.8%** of all potential subscribers in the test set.

For every 10 leads the model flags as positive, approximately **4** will convert. This trade-off was deemed acceptable to capture the majority of interested clients.

From Insights to Action: Strategic Recommendations

Target Based on Economic Climate

Insight:

Macroeconomic features are dominant predictors.

Action:

Prioritize and potentially increase marketing spend during periods with favorable indicators indicators (e.g., lower `euribor3m`, higher `nr.employed`).

Cultivate and Re-engage Past Contacts

Insight:

Previous campaign outcomes and contact history are highly predictive.

Action:

Develop targeted follow-up campaigns for clients with a `poutcome' of 'success' and re-engage previously contacted clients.

Implement a Ranked-Lead System

Insight:

The model excels at ranking clients by subscription probability.

Action:

Use the model's probability scores to create a tiered contact list. Focus top-tier marketing efforts (e.g., experienced agents) on the highest-probability leads to maximize ROI.

Thank You!

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences