Bank‑Marketing Classifier

A CRISP‑DM driven machine learning pipeline that predicts whether a Portuguese

bank client will subscribe to a term deposit (y = yes | no).

We benchmark four tuned algorithms k‑NN, Logistic Regression, Decision Tree, and

SVM on the full Bank Marketing dataset.

Quick Result Snapshot

Model | Best Params | Test Accuracy | Test F1 (yes) | Train Time (s)

k‑NN | n\_neighbors 7, weights distance | 0.904 | 0.494 | 5.60

Logistic Regression | C 10, penalty l1 | 0.912 | 0.523 | 29.54

Decision Tree | criterion gini, max\_depth 5 | 0.915 | 0.571 | 8.12

SVM | kernel rbf, C 10, gamma scale | 0.910 | 0.539 | 2029

Insight  – The tuned Decision Tree delivers the best F1 for the minority class

(yes) while keeping training fast. Logistic Regression and SVM tie on overall

accuracy, but SVM is roughly seventy times slower to fit.

1 Business Understanding

Tele‑marketing calls are expensive. By predicting the likelihood that a client

will subscribe, the bank can focus on high‑probability leads and cut campaign

spend.

2 Data Understanding

Source bank-additional-full.csv (UCI ML Repo)

Size 41 188 rows, 20 input attributes plus y target

Imbalance 88 percent no, 12 percent yes

Leakage note Feature duration (call length) is highly predictive but unknown

until a call ends; kept here for benchmark but can be dropped for real‑time

models.

3 Data Preparation

Handle unknown

Columns with less than five percent unknown values are imputed with mode

Columns with five percent or more keep unknown as a category

Numeric features → StandardScaler

Categorical features → OneHotEncoder(drop='first')

Split seventy percent train thirty percent test with stratify=y

4 Modeling and Hyper‑parameter Tuning

k‑NN neighbors 3 5 7 11 weights ∈ uniform distance

Logistic Reg C range 0.01 … 100 penalty ∈ l1 l2

Decision Tree criterion, max\_depth, min\_samples\_split, min\_samples\_leaf

SVM kernel  linear rbf C  0.1 … 100 gamma ∈ scale auto

GridSearchCV (five‑fold) optimizes F1 with yes as the positive label.

5 Evaluation Details

KNN

Accuracy 0.90 F1 0.49

Confusion [[10600 365] [816 576]]

Logistic Regression

Accuracy 0.91 F1 0.52

Confusion [[10678 287] [797 595]]

Decision Tree

Accuracy 0.92 F1 0.57

Confusion [[10609 356] [694 698]]

SVM

Accuracy 0.91 F1 0.54

Confusion [[10593 372] [741 651]]

6 Environment Setup

python -m venv .venv

source .venv/bin/activate # Windows .venv\Scripts\activate

pip install -r requirements.txt # pandas scikit-learn>=1.3 numpy joblib

7 Run the Pipeline

python src/train\_models.py \

--data data/bank-additional-full.csv \

--drop-duration False \

--n-jobs 4

Outputs

Hyper‑parameter search logs for each model

results/comparison\_table.csv containing train‑time and accuracy summary

Pickled best estimators in models/

8 Next Steps

Train a production model without duration

Address class imbalance with class\_weight balanced or SMOTE

Experiment with ensemble learners RandomForest XGBoost

Tune probability threshold for higher recall on yes cases

Deploy via FastAPI and CI for scheduled retraining

9 Citation

Moro S. Cortez P. Rita P. 2014

A Data‑Driven Approach to Predict the Success of Bank Telemarketing

Decision Support Systems. DOI 10.1016/j.dss.2014.03.001