Predicting Term Deposit Subscriptions with Machine Learning: Combined Supervised and Unsupervised Approach

Author: Kira Hou
Smith College SDS/CSC 293 Machine Learning

Introduction

How can a bank know which clients are most likely to say “yes” to a marketing call? In this project, I explored that question using **machine learning** on a **real-world dataset** from a Portuguese bank’s term deposit campaigns.

The goal was to predict whether a client would subscribe, using information about their demographics, finances, and past interactions. I combined **supervised models** to make predictions with **unsupervised methods** to explore hidden patterns in behavior. This two-part approach not only helped build stronger models, but also revealed deeper insights into what drives client decisions—insights that go beyond accuracy scores alone.

The Data

This dataset comes from a **Portuguese banking institution’s marketing campaigns**, which aimed to encourage clients to subscribe to a **term deposit**. The campaigns were conducted primarily through phone calls, often involving multiple contacts with the same client. The goal is to predict whether a client will say **“yes” or “no”** to subscribing.

Data Source

We used the bank-additional-full.csv file, which contains **41,188 records** and **20 input variables**, collected between **May 2008 and November 2010**. This version is the most complete and widely used in academic analysis (e.g., Moro et al., 2014).

Preprocessing

- No missing values were reported.
- Categorical variables were encoded using either label encoding or one-hot encoding, depending on the task.
- For clustering and PCA, all features were standardized to ensure fair comparison across different scales.
- A sample of 10,000 rows was taken for unsupervised learning tasks (e.g., K-means) to reduce computation time

Feature Categories

Client Attributes
- age: Age of client
- job, marital, education: Demographic info
- default, housing, loan: Financial standing and loan status

Campaign Contact Info
- contact, month, day, duration: Info on the last contact attempt
- Campaign History
- campaign: Number of contacts in this campaign
- pdays: Days since last contact (or -1 if never contacted)
- previous: Number of contacts in previous campaigns
- poutcome: Outcome of the last previous campaign

Target Variable
- y: Whether the client subscribed to a term deposit (yes or no)

The Model

To understand patterns in client behavior and predict subscription decisions, I used a combination of supervised and unsupervised learning methods.

Supervised Learning

I implemented and compared three classification models to predict whether a client would subscribe to a term deposit:
- **Logistic Regression:** A baseline linear model that estimates probabilities based on weighted input features.
- **Decision Tree:** A non-linear model that splits data into subgroups using feature-based rules.
- **Random Forest:** An ensemble method combining multiple decision trees to improve stability and predictive performance. Hyperparameter tuning for the Random Forest was performed using RandomizedSearchCV.

Based on overall predictive performance across evaluation metrics, I selected Random Forest as the final model for further analysis.

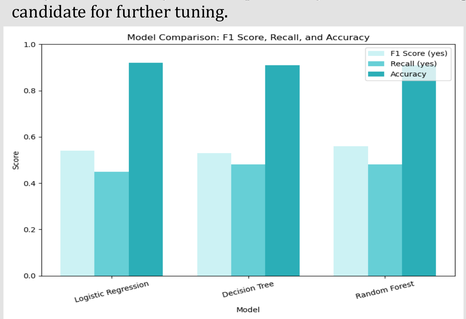
Unsupervised Learning

To explore the structure of the data beyond prediction, I applied:
- **K-means Clustering** to group similar clients without using the target variable.
- **Principal Component Analysis (PCA)** to reduce dimensionality and enable 2D visualization of client groupings. Clustering was performed on a 10,000-row sample for efficiency, and PCA was used to support visualization and reduce noise.

Results

Supervised Learning Performance

To predict whether a client would subscribe to a term deposit, I tested three classification models: logistic regression, decision tree, and random forest (RF1) using default settings.
- **Logistic regression** achieved high overall accuracy (91%) but struggled with recall (0.42) and F1 score (0.51) on the "yes" class.
- **Decision tree** had the highest F1 score for "yes" (0.57) but is prone to overfitting, making it less reliable despite matching 91% accuracy.
- **Random forest (RF1)** offered the best balance—maintaining 91% accuracy with more stable performance (F1 = 0.56, recall = 0.48). Its consistency and interpretability made it the strongest candidate for further tuning.

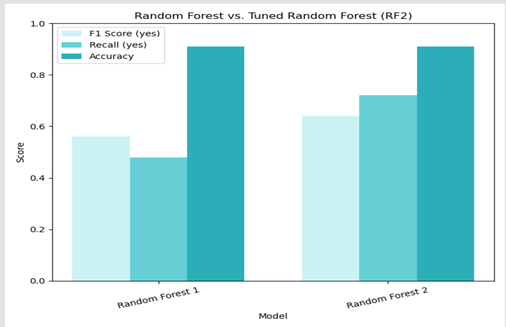


While not perfect, RF1 offered better generalization, less overfitting than the tree, and higher minority-class sensitivity than logistic regression. Based on this stronger performance, I selected Random Forest for further tuning.

Improvements with Tuned Random Forest (RF2)
Using RandomizedSearchCV, I tuned key hyperparameters to optimize performance:
- n_estimators = 157
- max_depth = 19
- min_samples_leaf = 1
- max_features = 'log2'

In addition, I lowered the classification threshold to 0.4, increasing the model’s sensitivity to positive cases. These changes led to a significant performance gain:
- Accuracy: 91.1%
- F1 Score (yes): 0.64
- Recall (yes): 0.73

Compared to RF1, RF2 showed a 17% improvement in recall and an 8-point gain in F1-score for the minority class, making it much more effective at identifying true positives while maintaining high overall accuracy.

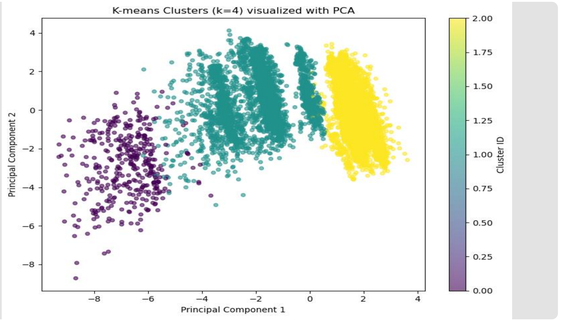


The feature importance plot confirms that call duration is the most predictive variable, followed by euribor3m, nr.employed, and age—highlighting how both individual behavior and macroeconomic context shape client decisions.

Unsupervised Learning Insights

To explore natural groupings within the client base, I applied K-means clustering (k = 3) on a 10,000-row sample of the dataset. The clusters were visualized using the first two principal components derived from Principal Component Analysis (PCA), which together capture a meaningful portion of the variance in the data.

The PCA plot below shows that the three clusters are fairly distinct in 2D space, indicating that the data contains underlying structure that K-means was able to capture. Each cluster represents a group of clients who share similar demographic, financial, and campaign-related characteristics.



Cluster Interpretation
K-means clustering (k=3), visualized through PCA, revealed three distinct client segments with meaningful differences in both campaign engagement and subscription behavior.
Cluster 0 represents the most responsive group, with a subscription rate of 60.7%. Clients in this cluster had the longest call durations (average ~305 seconds) and were contacted the fewest number of times, suggesting that high-quality, low-frequency contact is effective. These clients also showed the lowest macroeconomic confidence indicators, yet were most likely to subscribe.

- Cluster 1 had a moderate subscription rate of 16%. Clients were slightly younger, had slightly shorter calls than Cluster 0, and were contacted more frequently. Despite similar demographic and educational backgrounds, this group responded less strongly, suggesting that contact volume alone does not drive conversions.
- Cluster 2 was the least responsive group, with only 5% subscription rate. Clients in this cluster experienced higher employment and euribor3m rates—indicators of a stronger economy—but had the shortest call durations and the highest number of contacts. Their resistance may stem from campaign fatigue, higher skepticism, or less perceived need for the product.
- All three clusters shared similar characteristics in terms of education (university degree), marital status (mostly married), and housing loan status, yet differed sharply in behavior. This highlights that interaction quality and macroeconomic context may play a larger role than basic demographics alone.

Discussion

Looking ahead, several steps could enhance the current analysis. One promising direction is to incorporate PCA-transformed components into supervised models such as logistic regression or Random Forest, which may reduce redundancy in the feature space and improve generalization. Additionally, cluster membership derived from the unsupervised analysis could be used as an input feature or even as a basis for training separate models within each subgroup. This segmented modeling strategy could improve performance in regions where global models underperform. Future work could also involve experimenting with alternative clustering methods, such as hierarchical clustering or density-based approaches, which might uncover more complex subgroup structures.