EASTERN INTERNATIONAL UNIVERSITY
BECAMEX BUSINESS SCHOOL

Course: MIS 451 - Machine Learning for Business

Bank Customer Segmentation via Classification Analysis

Lecturer: Mr. Dang Thai Doan & Ms. Huynh Gia Linh

Prepared by: Group 1

Quarter 1: 2025-2026

Team Member	IRN	Assigned Tasks
Le Nguyen Tam Nhu	**2132300065**	EDA (Visualizations)
Huynh Thuy Bao Tram	**2132300228**	Data Cleaning and Preprocessing
Huynh Trung Hau	**2132309003**	Model Development & Model Evaluation

1. Business Context and Objectives 3
1.1 Business Context 3
1.2 Project Objectives 3
2. Data Collection and Characteristics 3
2.1 Dataset Description 3
3. Exploratory Data Analysis (EDA) 4
3.1 Dataset Overview 4
3.2 Data Cleaning 4
3.3 Exploratory Data Analysis (After Cleaning) 5
3.3.1 Univariate Analysis 5
3.3.2 Outlier Exploration 5
3.3.3 Bivariate Analysis 6
4. Data Cleaning and Transformation 7
4.1. Separating Target and Features 7
4.2. Handling Missing Values (Imputation) 7
4.3. Feature Encoding and Scaling 8
4.4. Dimensionality Reduction (PCA) 8
4.5. Train-Test Split 8
5. Model Development (Classification) Hau 8
5.1 Evaluation Methodology 8
5.2 Model Selection Rationale 9
6. Model Evaluation and Comparison 9
6.1Model Comparison and Selection 9
6.2 Quantitative Results 9
7. Interpretation and Business Insights 10
7.1 Strength: High Cost-Efficiency 10
7.2 Limitation: Coverage Volume 10
7.3 Strategic Recommendation 10
8. GitHub Repository Links: 10

1. Business Context and Objectives

1.1 Business Context

This project focuses on improving the effectiveness of direct marketing campaigns conducted by a Portuguese retail bank. The bank regularly contacts customers by phone to promote a term deposit product. These efforts are costly, time consuming and often inefficient because many contacted clients are not likely to subscribe. Traditional marketing approaches treat all customers in a uniform manner, resulting in unnecessary outreach and low conversion rates.

To address this problem, machine learning classification techniques can be used to predict whether a client is likely to subscribe to a term deposit before they are contacted. By leveraging historical campaign data that includes demographic information, financial behavior and past interaction records, the bank can make more informed decisions about whom to target. This leads to better resource allocation, reduced contact costs, and higher subscription rates.

1.2 Project Objectives

The objective of this project is to build and compare multiple classification models to predict whether a customer will subscribe to a term deposit. The project aims to develop a robust analytical workflow that includes performing exploratory data analysis to understand customer characteristics, preparing the dataset through encoding and feature scaling, and training at least three traditional machine learning models along with one deep learning model as required by the MIS 451 guidelines. Model performance will be evaluated using metrics such as F1-score, recall and confusion matrices to address the strong class imbalance. The final goal is to interpret the most influential features and translate the results into practical business insights that can support more effective and targeted marketing strategies.

2. Data Collection and Characteristics

2.1 Dataset Description

This project uses the Bank Marketing dataset provided by the UCI Machine Learning Repository. The data was collected from telephone-based direct marketing campaigns conducted by a Portuguese bank between 2008 and 2010. Each record represents an interaction with a customer and contains demographic details, financial attributes and campaign-related information. The dataset used in this analysis includes >45,000 observations and 17 input variables.

The attributes are grouped and described as follows:

a. **Client Demographic Variables**
   - age - numerical value representing the customer's age
   - job - occupation type (e.g., admin, technician, blue-collar, retired, student).
   - marital - marital status (married, single or divorced)
   - education - education level (primary, secondary, tertiary or unknown).
   - default - indicates whether the customer has credit in default.

b. **Financial Attributes**
   - balance - average yearly account balance in euros.
   - housing - whether the customer has a housing loan.

- loan - whether the customer has a personal loan.

c. **Current Campaign Interaction**

- contact - communication channel used (telephone, cellular, unknown).
- day - day of the last contact.
- month - month of the last contact.
- duration - length of the last phone call in seconds

d. **Previous Campaign History**

- campaign - number of contacts performed in the current campaign.
- pdays - days since the last contact from a previous campaign (-1 indicates no previous contact).
- previous - number of contacts prior to the current campaign.
- poutcome - result of the previous campaign (success, failure or other).

e. **Outcome Variable (for evaluation only)**

- **y** - whether the client subscribed to a term deposit (yes or no).

3. Exploratory Data Analysis (EDA)

3.1 Dataset Overview

The dataset contains 45,211 marketing records with 17 attributes describing customer demographics, financial information, contact history and campaign outcomes. All variables are correctly loaded, and the structure shows no formatting or completeness issues.

The data includes a mix of numerical variables such as age, balance, day, duration, campaign, pdays and previous, and categorical variables including job, marital status, education, default, housing, loan, contact type, month, poutcome and the target variable y. Although several categorical fields contain the value unknown, these entries represent unavailable or uncollected information and are treated as valid categories rather than missing data.

No duplicate rows were identified. Summary statistics indicate notable right skewness in balance, duration, campaign, pdays and previous, where most customers fall within low ranges but a small group exhibits extremely high values. These distributions reflect genuine behavioral differences rather than data quality issues. Overall, the dataset is clean, consistent and ready for further exploratory analysis.

3.2 Data Cleaning

3.2.1 Handling the value "unknown"

Several categorical variables contain the label "unknown", including job, education, contact and poutcome. In this dataset, "unknown" does not represent missing data but reflects cases where information was not collected or customers did not provide it. These values capture meaningful behavioral groups and appear in large proportions, particularly in contact and poutcome. Removing or imputing these entries would distort the true distribution of categories and introduce bias into the model. Therefore, all "unknown" values were retained as valid categories rather than replaced.

3.2.2.Handling the special value pdays = -1

y	no	yes
pdays_flag		
contacted_before	76.928667	23.071333
never contacted	90.842669	9.157331

Figure 1: pdays_flag crosstab

The variable pdays uses the value −1 to represent customers who were never contacted in previous campaigns. Because this value reflects a distinct behavioral state rather than missing data, it was retained. A binary indicator pdays_flag was introduced to distinguish between never_contacted (pdays = -1) and contacted_before (pdays > -1). Comparing subscription behavior reveals a clear difference: previously contacted customers show more than double the subscription rate (~23%) compared to those never contacted (~9%). Additionally, the remaining pdays values display wide but valid variation, supporting that pdays reflects real campaign history rather than noise

3.3 Exploratory Data Analysis (After Cleaning)

3.3.1 Univariate Analysis

Univariate exploration was conducted to understand the distribution of each variable after cleaning. Categorical variables show clear dominance of specific groups, such as the high frequency of blue-collar, management, and technician jobs, as well as the large proportion of married clients. Education is mostly “secondary” or “tertiary”, while “default” almost always equals no. Contact types are primarily “cellular”, and most campaigns occurred in “May”, “July”, and “August”. The “poutcome” field is heavily skewed toward “unknown”, which matches the nature of the dataset.

For the numerical variables, age follows a moderate right-skewed distribution centered around middle-aged clients. Balance, duration, campaign, pdays, and previous show extreme right skewness with many low-value observations and a long tail of large values. These patterns reflect genuine customer and campaign behavior rather than data issues.

3.3.2 Outlier Exploration

Several numerical variables contain many extreme observations, especially balance, duration, campaign, pdays, and previous. Visual inspection using boxplots confirms that these are valid behavioral extremes rather than errors.

Examples include:

- Very high balances for wealthy clients
- Long call durations during successful marketing interactions
- High campaign counts for repeatedly contacted customers
- Large pdays gaps for past contacts in long intervals
- High previous counts for customers with extensive contact history

Removing or modifying these values would distort genuine patterns and may reduce model performance. For this reason, all outliers were retained.

3.3.3 Bivariate Analysis

a. **Categorical Variables vs Target (y)**

Chi-square tests for all categorical predictors return p-values < 0.001, indicating that each category is statistically associated with subscription outcome.

Key patterns observed:

- Job: Retired and student groups have noticeably higher subscription rates than blue-collar or services workers.
- Marital: Single clients respond better than married clients.
- Education: Higher education levels show slightly better conversion rates.
- Default / Housing / Loan: Clients with no default and no personal loan convert more frequently.
- Contact Method: Cellular contact performs significantly better than telephone or unknown.
- Month: May dominates contact volume but has lower success rates; certain months such as March and September show better conversion despite smaller samples.
- Poutcome: Previous “success” strongly predicts higher subscription rates, while “failure” predicts lower engagement.

These findings confirm that categorical attributes carry meaningful information for classification.

b. **Numerical Variables vs Target (y)**

Point-biserial correlation was used to evaluate the linear relationship between numerical predictors and the binary target variable. All correlations are statistically significant, although most are weak in magnitude.

Important observations:

- Duration shows the strongest positive correlation ($r \approx 0.39$). Longer calls tend to result in successful subscriptions.
- Pdays and previous have small positive correlations, suggesting that clients with prior interactions behave differently.
- Campaign is negatively correlated, indicating diminishing returns when a client is contacted too frequently.
- Balance shows a small but positive association with subscription.
- Age and day have minimal influence.

Boxplots and density plots further illustrate these patterns, showing that successful clients generally experience longer call duration and slightly different interaction histories.

3.3.4 Correlation Heatmap for Numerical Features

A correlation heatmap was used to identify multicollinearity among the numerical predictors. The correlations are generally weak, with the only notable relationship between pdays and previous (~ 0.45). This suggests that these two variables reflect related aspects of interaction history, but the correlation is not strong enough to require removal. Overall, there is no multicollinearity concern, and all numerical variables can be retained for modeling.

4. Data Cleaning and Transformation

To prepare the dataset for optimal performance in classification models, our group implemented a rigorous data preprocessing workflow. This process includes handling missing values, encoding categorical variables, scaling numerical features, and dimensionality reduction. The entire workflow was constructed using the Pipeline and ColumnTransformer classes from the Scikit-learn library to ensure consistency and prevent data leakage.

4.1. Separating Target and Features

Given that classification is a supervised learning task, the model inherently relies on labeled outcomes to identify underlying patterns. Consequently, the initial step involves partitioning the dataset into Input Features and the Target Variable. The input matrix comprises 15 independent variables - such as age, balance, and job type - which serve as the primary training data. The target variable (y), representing the client's subscription decision, was isolated and encoded into a binary format (1 for "yes", 0 for "no"). Crucially, we excluded the variable duration. Although it strongly correlates with the outcome, the call duration is only known after the interaction is completed. Therefore, including it would cause data leakage, making the model unrealistic for predicting outcomes before a call is made.

4.2. Handling Missing Values (Imputation)

Although our Exploratory Data Analysis (EDA) confirmed the absence of missing values (NaN) in the current dataset, we explicitly incorporated an imputation step within the pipeline to ensure robustness for production deployment. This mechanism acts as a critical safety net, allowing the system to handle potential incomplete entries in future data streams (e.g., from live API calls) without failure. Specifically, numerical features are imputed using the median, which offers a more stable estimate than the mean for skewed financial variables like balance. Meanwhile, categorical features are filled using the mode (most frequent value) to preserve structural integrity. It is important to note that the explicit string "unknown" is not treated as missing data; consistent with our EDA findings, it is retained as a meaningful category indicating unprovided information.

In the data preprocessing pipeline, we keep a SimpleImputer step for both numerical and categorical variables, even though the EDA results show that the current dataset does not contain any missing values (NaN). The main motivation is to ensure robustness and readiness for real-world deployment. If, in the future, new observations contain missing entries (e.g., due to data entry errors or incomplete data collection from an API), the pipeline will still run smoothly instead of crashing. In this sense, the imputer acts as a safety net for potential future issues, rather than a transformation that modifies the present training data.

Moreover, using the imputer does not introduce data leakage and does not contradict the EDA conclusions. The imputer only applies to actual NaN values, while EDA has confirmed that there are no NaNs in the current dataset. Therefore, in practice, this step does not change any existing observations. Importantly, values such as 'unknown' are not converted to NaN and are therefore preserved as a separate, meaningful category, which is then encoded by OneHotEncoder in line with the interpretation given in the EDA section.

Finally, the pipeline structure following the order *imputer → scaler/encoder* reflects a standard sklearn design pattern, which makes the preprocessing workflow clear, systematic, and easily reusable for future training runs or other datasets.

4.3. Feature Encoding and Scaling

Since machine learning algorithms - particularly distance-based classifiers such as SVM and KNN - are highly sensitive to data scale and incapable of processing raw text directly, we applied two critical transformations to prepare the dataset. First, numerical variables were standardized using StandardScaler to achieve a mean of 0 and a standard deviation of 1. This step prevents features with large magnitudes, such as balance, from disproportionately influencing the model's objective function, thereby ensuring faster convergence and superior performance. Simultaneously, categorical variables (e.g., job, marital) were converted into binary vectors via One-Hot Encoding. This technique effectively eliminates the risk of introducing artificial ordinal relationships (e.g., implying "married" is greater than "single") while correctly handling "unknown" entries as distinct, informative features.

Handling Special Values in 'pdays' variable. Before inputting data into the models, we conducted a specific check on the pdays feature (days passed since the client was last contacted). Our analysis revealed that a significant portion of the dataset contained the value -1. In this business context, -1 is not a numerical measurement of time but a symbolic placeholder indicating that the customer was not previously contacted. Treating -1 as a standard numerical value would be detrimental to distance-based algorithms (like KNN or SVM), as the model would interpret it as being numerically close to 0 (contacted today), whereas semantically, "never contacted" is a completely different state. To address this, we explicitly identified these cases and engineered a feature to distinguish between "newly contacted" customers and "previously contacted" ones, ensuring the model captures the correct behavioral patterns without numerical bias.

4.4. Dimensionality Reduction (PCA)

Following One-Hot Encoding, the feature space expanded substantially from 15 original features to 51 dimensions, resulting in a sparse matrix. To mitigate the "Curse of Dimensionality" and reduce the risk of overfitting, we applied Principal Component Analysis (PCA). The PCA was configured to retain 95% of the total variance, effectively compressing the data into 24 principal components. This step removes noise and multicollinearity while enhancing computational efficiency.

4.5. Train-Test Split

Since this is a supervised learning project focused on classification, we split the data into a training set (80%) and a test set (20%) with stratification (stratify=y). The stratification ensures that the class distribution (proportion of 'yes' vs. 'no') remains consistent across both sets, which is crucial for handling imbalanced data. The training set is used to train the classification models, while the test set is reserved to evaluate the models' generalization performance on unseen data using metrics like recall, precision, and F1-score.

5. Model Development (Classification) Hau

5.1 Evaluation Methodology

To ensure a robust and unbiased assessment of model performance, we implemented a rigorous comparative framework focusing on stability and business relevance. Given the significant class imbalance where "No" cases heavily outnumber "Yes" cases’ standard accuracy proved to be an insufficient metric. A trivial model predicting "No" for every client would achieve deceptively high accuracy while failing to identify potential subscribers. Consequently, we utilized the Macro

Average F1-Score as our primary selection metric. This approach ensures that the model's ability to correctly classify the minority target class is weighted equally with the majority class. Furthermore, we employed 10 fold cross-validation to validate model stability, mitigating the risk of overfitting to a specific data split and confirming that performance remains consistent across different subsets of the customer base.

5.2 Model Selection Rationale

To ensure a comprehensive evaluation, five algorithms were selected, ranging from simple linear models to advanced ensemble and deep learning approaches. Logistic Regression served as the baseline due to its interpretability and established effectiveness in binary classification.

K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) were included to capture non-linear relationships. KNN identifies local patterns without distributional assumptions, while SVM is effective in high-dimensional spaces and handles complex decision boundaries.

Finally, an Artificial Neural Network (ANN) was incorporated to model deeper non-linear interactions that traditional models may overlook. Its layered architecture adapts well to complex feature relationships and integrates seamlessly with the existing preprocessing pipeline, offering strong potential for enhanced classification performance.

6. Model Evaluation and Comparison

6.1Model Comparison and Selection

Four classification algorithms were evaluated: Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. Despite the updated evaluation, the Artificial Neural Network (ANN) remains the selected production model.

Although a configuration change caused Recall-yes to drop to 27.6% and F1-yes to 0.3678 (significantly lower than the original 54.9% and 0.5581), the ANN still records the highest F1-yes score (0.3678) among all tested models and maintains low overfitting risk. Given that the business objective is to maximize identification of term-deposit subscribers, the ANN continues to offer the strongest relative performance and is therefore retained as the recommended model for deployment, with the caveat that the earlier, higher-recall version should be recovered and used whenever possible.

6.2 Quantitative Results

The final Artificial Neural Network (ANN) was evaluated on the unseen stratified test set. The model achieved an overall accuracy of 89.8%. Detailed performance metrics are presented below:

Target Class	Precision	Recall	F1-Score	Support
0 (No – Non-Subscriber)	0.9100	0.9702	0.9391	7,985
1 (Yes – Subscriber)	0.5509	0.2760	0.3678	1,058
Accuracy			0.8890	9,043

Macro Average	0.7305	0.6231	0.6535	9,043
Weighted Average	0.8680	0.8890	0.8723	9,043

7. Interpretation and Business Insights

While the overall accuracy is high, the true operational character of the model is best understood through the Precision Recall trade-off for the target subscriber class.

7.1 Strength: High Cost-Efficiency

The model is now highly conservative. When it predicts “yes”, it is correct 55.1% of the time (Precision-yes), and it only ~5.9% of customers are flagged for contact. This delivers excellent cost efficiency and minimal marketing waste.

7.2 Limitation: Coverage Volume

Recall-yes has fallen to 27.6%, meaning the model now misses 73% (766 out of 1,058) of customers who would actually subscribe. Compared to the previous ANN version that captured 54.9% of subscribers, this configuration sacrifices substantial revenue potential for marginally higher precision.

7.3 Strategic Recommendation

- **Deploy the Artificial Neural Network (ANN) as the primary scoring model for term-deposit campaigns**, using the default classification threshold of 0.50. This configuration yields the highest F1-score (0.558) and Recall (54.9%) on the positive class among all tested algorithms, delivering an estimated 5–6× uplift in subscriptions compared to random or legacy targeting while contacting approximately 11% of the customer records.
- **Rank and prioritize the contact list according to the ANN’s predicted probability**. Execute outreach in descending order of score (≥ 0.70 → 0.50–0.69 → 0.35–0.49 if additional budget is available). This straightforward prioritization maximizes conversion rates and campaign ROI without requiring additional modelling effort.
- **Conduct a controlled A/B test in the next campaign cycle**. Allocate 20–30% of the campaign universe to the ANN-ranked list and the remainder to the current targeting approach. The four-week pilot will provide conclusive, real-world evidence of the uplift and facilitate full organisational adoption.

In summary, the selected ANN delivers the best combination of coverage and reliability among evaluated algorithms and is ready for immediate production deployment.

8. GitHub Repository Links:

Le Nguyen Tam Nhu: https://github.com/Le-Nguyen-Tam-Nhu/MIS-451

Huynh Thuy Bao Tram: https://github.com/Julie0203/MIS-451

Huynh Trung Hau: https://github.com/huynntrunghau2905-BA/Final_Project_MIS451