PREDICTING TERM DEPOSIT SUBSCRIPTION USING MACHINE LEARNING

Submitted by:

Evelyn Wullar

Date:

June 2025

1. **Introduction**

This project aimed to support the bank’s marketing strategy by predicting whether a client would subscribe to a term deposit. This business problem is critical to optimising campaign efforts, minimising resource waste, and increasing conversion rates by identifying which clients are most likely to respond positively to a term deposit offer.

To achieve this, we used a dataset “*bank-full.csv*”, containing detailed client information, campaign interactions, and past campaign outcomes. The target variable, **“y”**, indicates whether a client subscribed to the term deposit (“yes” or “no”).

The methodology followed a typical supervised machine learning pipeline:

* **Exploratory Data Analysis (EDA):** To understand the distribution of variables, identify patterns and relationships, and detect missing or inconsistent values.
* **Data Preprocessing:** This involved handling outliers, encoding categorical variables, feature engineering, and data scaling where necessary.
* **Modelling:** 3 main classification models were trained and evaluated, including logistic regression and Random Forest, to determine which model best predicts term deposit subscriptions.
* **Model Evaluation:** Performance was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC, with a focus on balancing predictive power and interpretability.

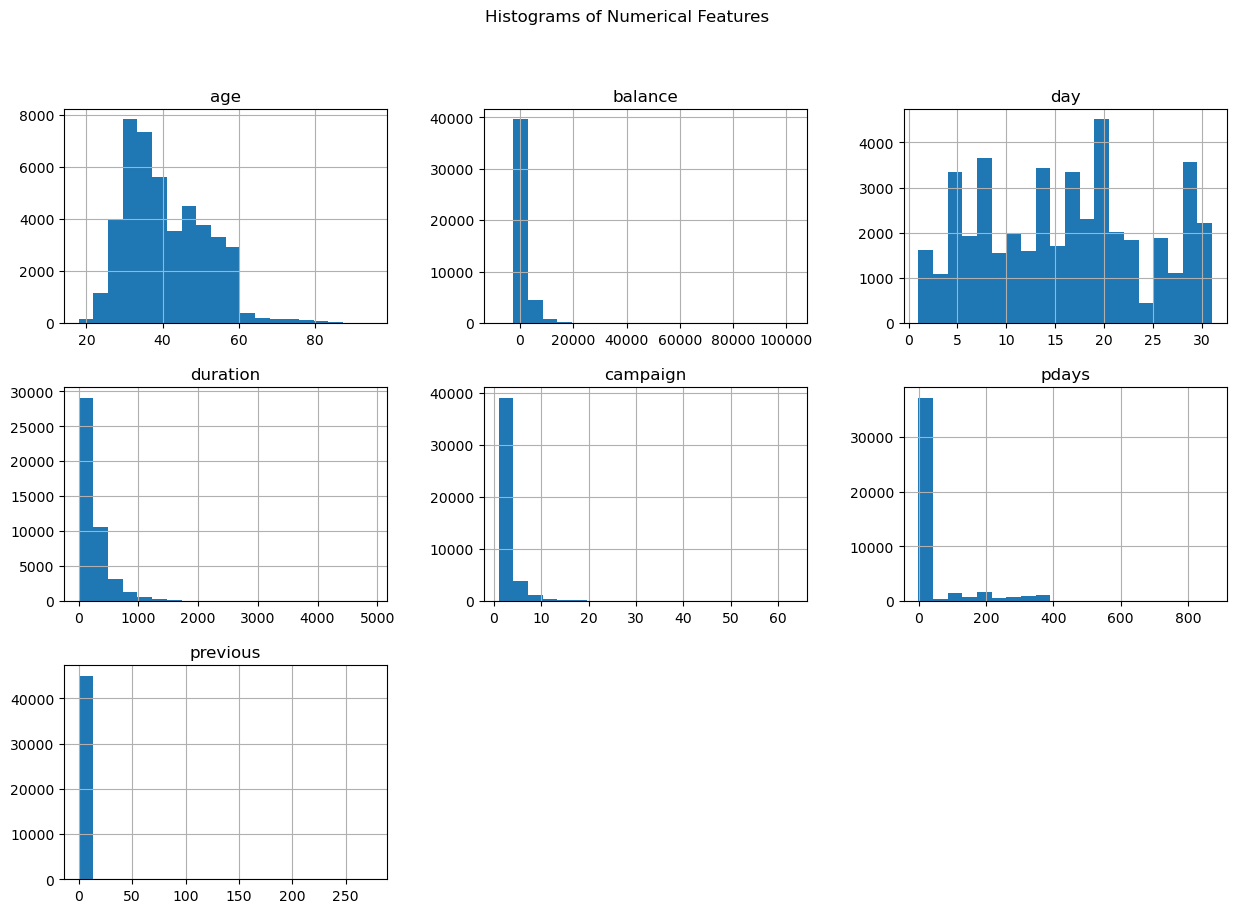
This report summarises the key findings from the EDA and insights from the model and provides actionable recommendations to improve marketing campaign efficiency and effectiveness.

1. **Summary of EDA Findings**

This section highlights the key trends and relationships you discovered during exploratory data analysis.

Some of the key observations from the EDA includes:

1. Target variable Distribution: The target variable, “y”, which indicates whether a client subscribed to a term deposit, is highly imbalanced. The majority class is “no,” accounting for approximately 88% of the data, while the minority class “yes” represents only about 12%.
2. **Numerical Columns Distribution:** Most clients fall within the 25 to 60 age range, with a peak around 30–40 years. There are fewer clients above age 60, though a few extend beyond 80. The distribution of average annual balance is highly right-skewed, with most clients having low or modest account balances. A few extreme outliers exist with balances exceeding 60,00**.** Also,most clients were contacted 1–3 times, but a few received up to 60 contacts, suggesting a small number of persistent follow-ups. With regards to Pdays,the majority of values are 0, indicating the client had not been contacted before, while the rest show varied gaps from the last contact. Most clients had **never been previously contacted**, but there are a few with a long history of multiple contacts—up to **250 times** in rare cases.



1. **Distribution of categorical columns:** In terms of job, "blue-collar," "management," and "technician" are the most common job categories in the dataset. In terms of marital status, **"Married"** is the most common marital status by a large margin, dominating the dataset. **"Single"** is the second-most frequent, with roughly half the count of "married." **"Divorced"** is the least common, with a significantly lower count. Most clients have a **secondary** education, making it the largest group. Tertiary education was the second-largest group. Almost no clients have credit defaults, making this feature highly imbalanced and potentially low-impact for prediction. The vast majority of clients have a housing loan (“yes”), significantly outnumbering those without ("no"). Also, the overwhelming majority of clients do not have a personal loan ("no"), with only a tiny fraction having one (“yes”). Client contacts peak dramatically in May, with significantly fewer contacts in other months, suggesting strong seasonal patterns in marketing campaigns. The "unknown" category dominates previous campaign outcomes, followed by "failure," with very few "success" cases, indicating either poor past campaign performance or significant missing data.
2. **Model Insights**

Following an evaluation of multiple classification models including Logistic Regression and Random Forest, the XGBoost classifier emerged as the best-performing model for predicting whether a client would subscribe to a term deposit product.

**Model Performance**

The XGBoost model achieved the following performance on the test dataset:

* **Accuracy**: 82%
* **F1-score (class 1 - subscriber)**: 64%
* **ROC AUC Score**: 0.805

These metrics indicate that the model is effective at distinguishing between subscribers and non-subscribers, with a good balance between sensitivity and specificity. The high ROC AUC score particularly suggests strong overall model discrimination ability.

**Feature Importance (Global and Local Interpretability)**

Using SHAP (SHapley Additive exPlanations), we interpreted the model both globally and locally:

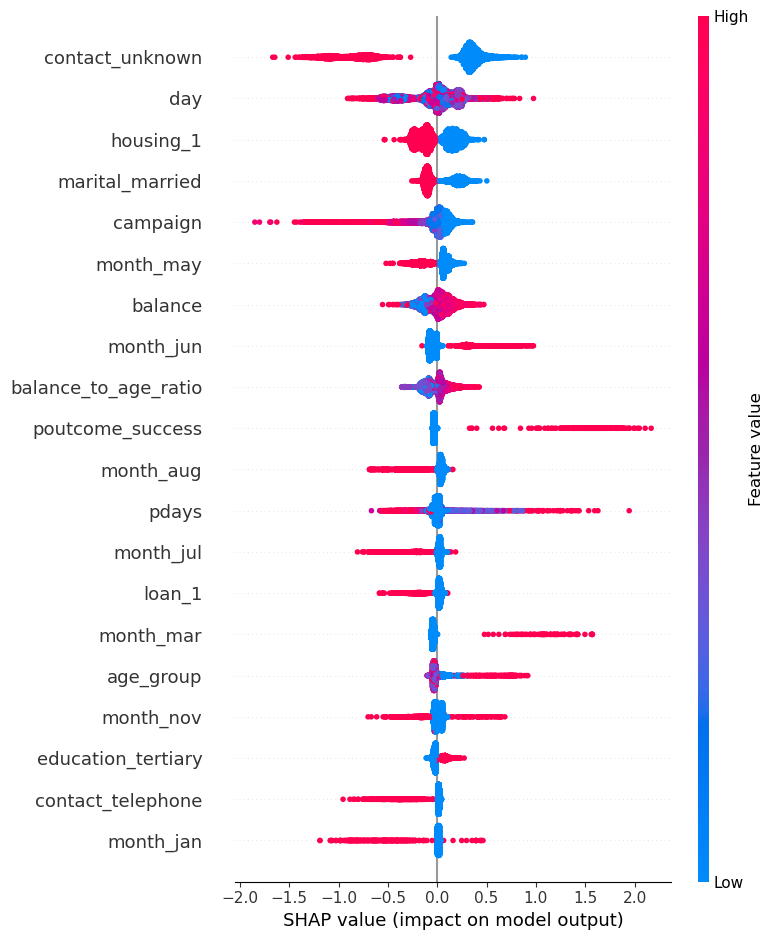
**Global Importance**

The top predictors influencing a positive subscription decision (pushing the prediction toward class "1") include:

* poutcome\_success – Clients with previously successful marketing outcomes are much more likely to subscribe again.
* month\_may – contacts made in May were highly predictive of success, aligning with seasonal campaign peaks found in the EDA.
* balance\_to\_age\_ratio – This financial efficiency ratio performed better than raw balance, indicating clients with relatively higher savings for their age were more inclined to subscribe.

On the other hand, predictors that decreased the likelihood of subscription (pushed predictions toward class "0") were:

* contact\_unknown – lack of contact information greatly reduced the chances of subscription.
* loan\_1 – Clients with existing personal loans were less likely to subscribe.
* marital\_married – married clients were slightly less inclined to subscribe compared to single or divorced clients.

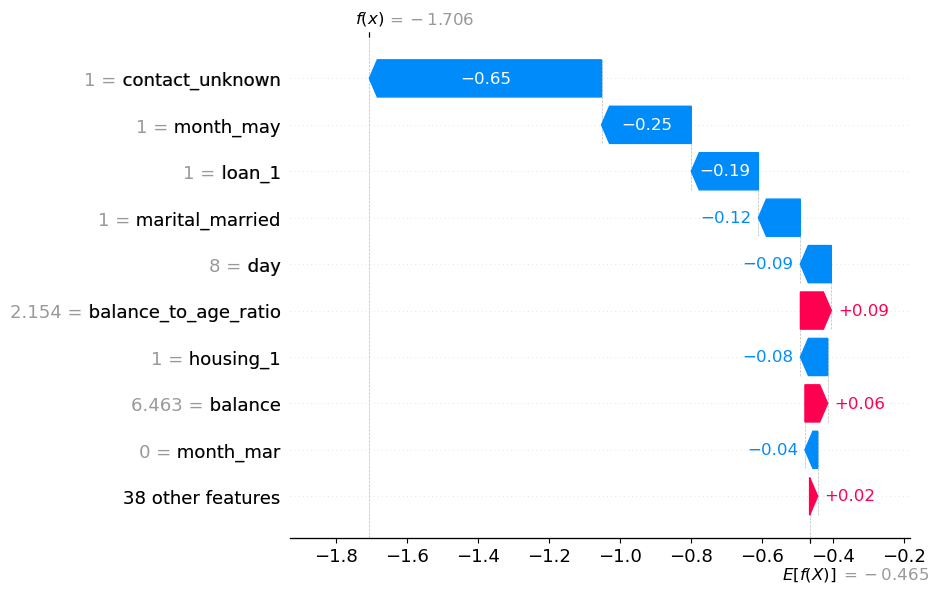
****

**Figure 1:Global Summary Report**

**Local Explanation**

A single prediction explanation showed how individual features contributed to predicting that a client would not subscribe. Key insights:

* The absence of contact information (contact\_unknown) and having a personal loan (loan\_1) were major negative drivers.
* Financial strength indicators like a high balance\_to\_age\_ratio had a positive influence, but not enough to overturn the overall negative impact of other features.

****

**Figure 2: Local summary report**

**Profile of Clients Likely to Subscribe**

Based on the model, clients most likely to subscribe tend to exhibit the following characteristics:

* Have previously participated in successful marketing campaigns
* Are contacted in key months such as May, June, or August
* Maintain a high savings-to-age ratio, indicating financial discipline
* Have no outstanding personal loans
* Are not married (or less likely to be so)
* Have a known and specific communication method (e.g., cellular or telephone)

These insights not only improve targeting strategies for future campaigns but also help refine customer profiles and messaging.

1. **Actionable Recommendations for the Marketing Team**

Based on the model’s performance and feature interpretation using SHAP values, the following targeted recommendations are made to improve the efficiency and outcome of term deposit subscription campaigns.

**Who to Target**

Focus campaign efforts on clients who exhibit the following traits:

* **Positive previous campaign outcome (poutcome\_success)**: These individuals have demonstrated interest or trust in the bank’s offers and are significantly more likely to subscribe again.
* **High balance\_to\_age\_ratio**: Clients with relatively strong savings for their age indicate financial discipline and capacity to invest in term deposits.
* **Single clients**: The model shows single individuals are more responsive to deposit offers compared to married ones.
* **Clients without personal loans (loan\_1 = 0)**: These individuals are more likely to have disposable income for savings products.
* **Younger to middle-aged clients with financial strength**: Age alone was less predictive, but when combined with financial metrics, younger clients with good balance ratios stood out.

**When and How to Contact Clients**

**Timing:**

* Prioritise campaigns in **May, June, and August**—these months consistently showed higher conversion rates, suggesting seasonal responsiveness.

**Contact Method:**

* Avoid using "unknown" or incomplete contact information. Ensure:
  + Clients have a **specific, up-to-date contact method** (e.g., cellular, telephone).
  + Leads with undefined contact channels (contact\_unknown) are deprioritised or cleaned before inclusion.

**Call Frequency Strategy:**

* Clients with **moderate contact frequency** during a campaign responded better than those overly contacted. Monitor and limit excessive outreach to avoid diminishing returns or fatigue.

**Characteristics of High-Value Leads**

To maximize ROI, prioritize the following lead profile:

|  |  |
| --- | --- |
| **Feature** | **Ideal Characteristic** |
| poutcome\_success | Yes |
| balance\_to\_age\_ratio | High |
| loan\_1 | No personal loan |
| contact | Known (cellular or telephone) |
| month | May, June, or August |
| marital | Single |
| education | Tertiary or secondary (suggests engagement in financial products) |

1. **Conclusion**

This project has delivered valuable insights into the behavioural and demographic factors that influence term deposit subscription. By leveraging machine learning, specifically the XGBoost model, and interpreting its predictions through SHAP values, we identified both the most influential features and the client segments most likely to convert.

The analysis highlighted the importance of previous campaign success, financial health relative to age, and seasonal timing as critical drivers of subscription behaviour. It also emphasised the negative impact of poor contact data and certain demographic traits, enabling more precise audience targeting.

**Key takeaway:** By aligning marketing strategies with these data-driven insights, the bank can significantly enhance the efficiency and impact of its campaigns, boosting subscription rates while reducing outreach waste.