Predictive Insights from Portuguese Bank Marketing Data

•••

Gulnar Armour

Springboard

October 29, 2023

Introduction

- Marketing in modern banking
- Direct marketing phone campaigns for a Portuguese bank
- Determine likelihood of clients subscribing to term deposits
- Data analysis and machine learning
- Factors influencing subscription likelihood

Dataset Overview

- UC Irvine Machine Learning Repository dataset
- Data wrangling process
  - Variable identification and labeling
  - No missing values
  - 16 input variables

Dataset Overview

- The dataset consists of various input variables that can be categorized into three main sections:

  1. Bank client data (*age, job, marital status, education, default, balance, housing loan, personal loan)*

  2. Last contact of the current campaign (*contact, day, month, duration)*

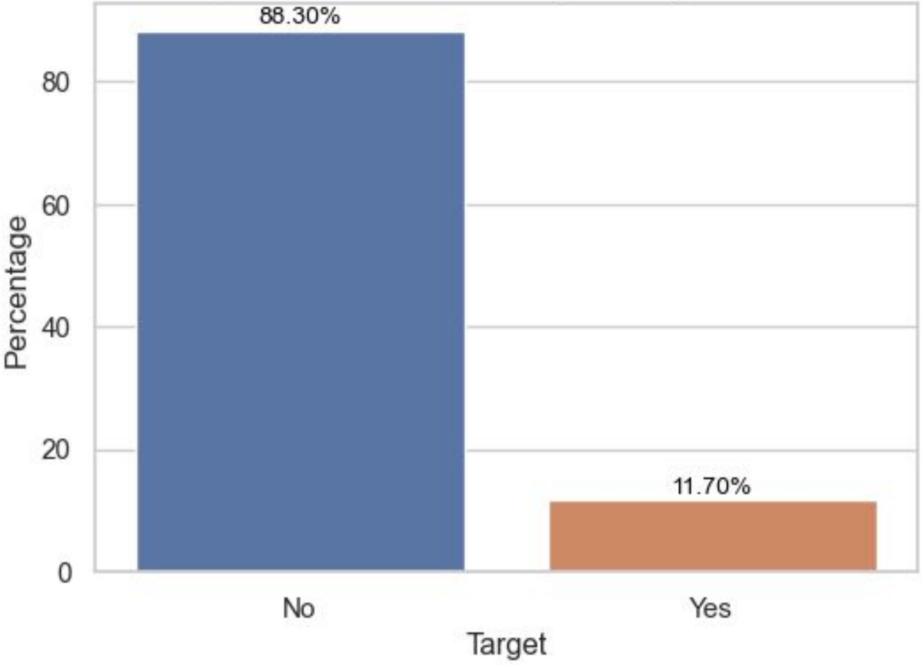
  3. Other attributes (*campaign, pdays, previous, poutcome*)

- The output variable (target) is the client's subscription to a term deposit (binary: "yes", "no")

Univariate Analysis

- The majority class is "no" at 88.3%
- The minority "yes" class only accounts for 11.7%
- This class imbalance can impact predictive model performance
- Models may be biased towards the majority class
- Oversampling techniques address this issue

Figure 3.1.1: Univariate Analysis: Target Variable



Bivariate Analysis - Job Categories

- Subscription rates varied significantly across job categories:
  - students (28.68%)
  - retirees (22.79%)
  - management (13.76%)
  - entrepreneurs (8.27%)

- Students and retirees are more inclined to subscribe due to their financial independence

- Subscription rates vary by job, showing the need for tailored marketing approaches

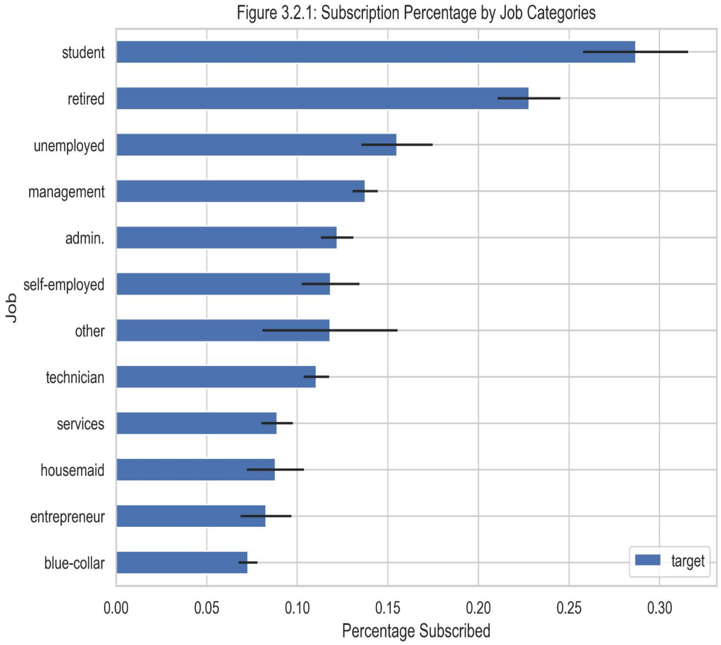


Figure 3.2.1: Subscription Percentage by Job Categories

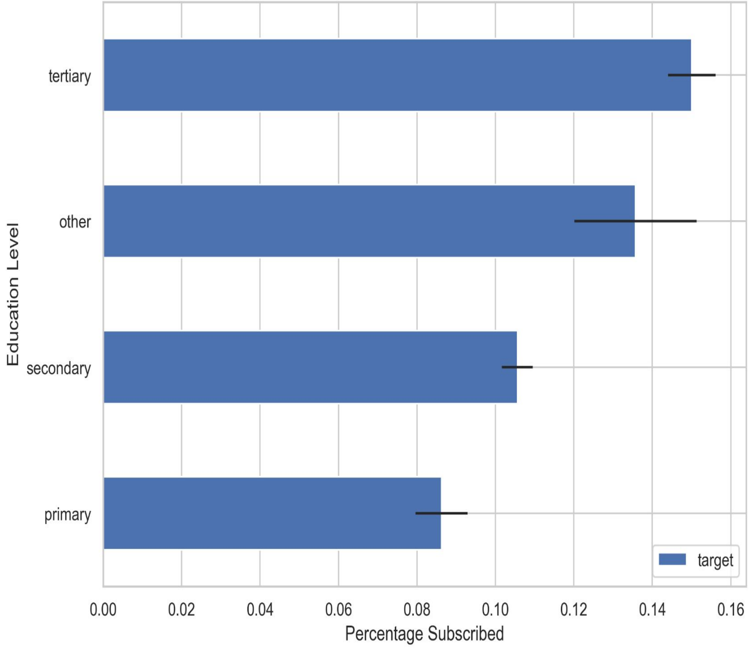
Bivariate Analysis - Education Level

- Tertiary/advanced degree holders showed the highest subscription rate at 15.01%, followed by other education (13.57%) and secondary (10.56%)

- Higher education is associated with a greater likelihood to subscribe than other levels

- May be attributed to factors like increased financial literacy/knowledge gained from advanced studies

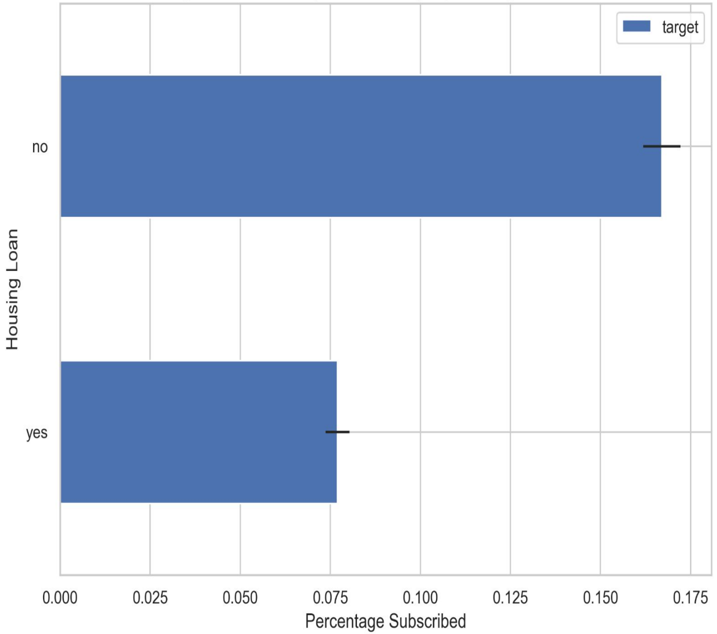
Figure 3.2.2: Subscription Percentage by Education Level



Bivariate Analysis - Housing Status

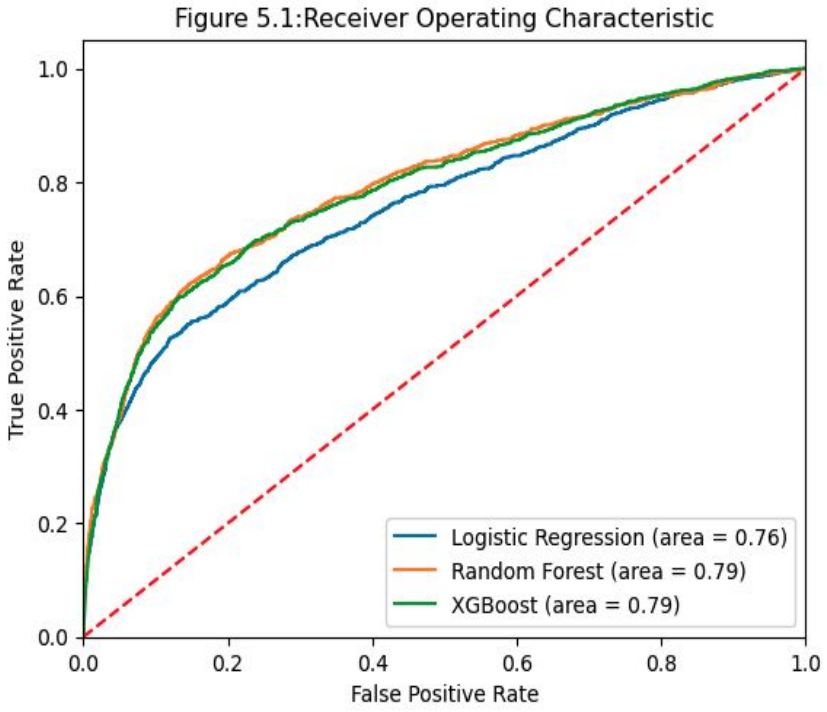
- A mortgage loan significantly lowered subscription rates from 16.70% to 7.70%
- Housing loan-free consumers subscribed more frequently than those with loans
- Considering the substantial disparity in rates, it is imperative to tailor marketing based on loan ownership status

Figure 3.2.4: Subscription Percentage by Housing Loan Status



Model Performance

- Based on predictive accuracy, Random Forest (0.7941) outperformed Logistic Regression (0.7575) and XGBoost (0.7899)
- The ROC AUC measures the ability of models to distinguish between positive and negative classes
- Selected model for the analysis: Random Forest



Classification Report: Random Forest

- The Random Forest model demonstrates high precision, recall, and F1-score for class "No", indicating strong performance in predicting this class

- Precision, recall, and F1-score for class "Yes" are noticeably lower

	Precision:	Recall:	F1-Score:	Support:
“No”	0.90	0.99	0.94	7952
“Yes”	0.70	0.23	0.35	1091

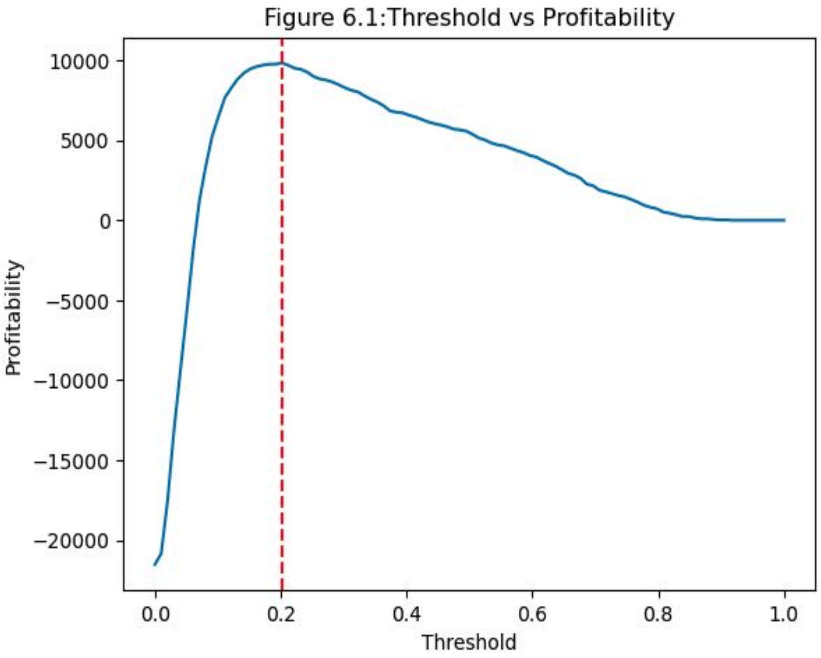
Feature Importance

- Model identified top 5 most important features for predicting subscriptions
- Factors like balance, age, recency/frequency of contact would reasonably impact customer responses
- However, other uncaptured variables may also influence outcomes, highlighting potential for additional feature engineering to further improve model performance

Top 5 Features				
balance	*age*	*day*	*campaign*	*pdays*

Threshold vs Profitability

- Threshold vs Profit curve: graphical representation of the relationship between decision threshold and resulting profit
- This was based on assumed values:
    - A deposit amount of €1000
    - A net investment margin of 3%
    - Revenue per subscription: €30
    - Cost per call: €6
- Red line shows the optimal threshold for maximum returns at 0.2
- Optimal threshold of 0.2 expands the reach of the marketing campaign



Conclusion

- A random forest model achieved an ROC AUC of 0.7941, demonstrating the ability to accurately predict client subscription to term deposits
- Bivariate analysis revealed patterns in subscription rates across job categories, education levels, housing loan, personal loan, and age groups
- Setting an optimal threshold of 0.2 on predictions maximized profitability by balancing outreach scope and successful subscriptions

Conclusion

- Additional factors not captured could further improve the model, and costs/revenues may vary in reality
- Analysis demonstrates how banks can leverage data analytics for more informed marketing decisions

Q&A

Thank you!