Predictive Insights from Portuguese Bank Marketing Data

Gulnar Armour
Springboard
October 29, 2023

Introduction

- Importance of marketing in modern banking
- Study of Portuguese bank's direct marketing phone campaigns
- Goal: Determine likelihood of client subscribing to term deposit
- Use of data analysis and machine learning
- Examination of key factors impacting subscription likelihood

Dataset Overview

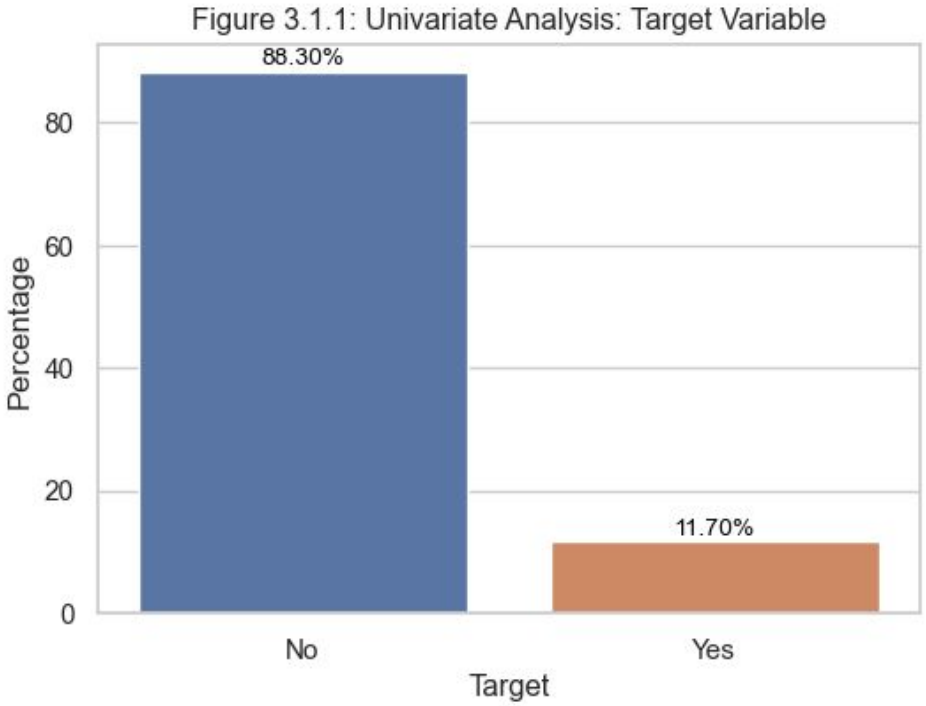
- Dataset from UC Irvine Machine Learning Repository
- Analysis of Portuguese bank's direct marketing campaigns
- Data wrangling process
  - Identification and labeling of variables
  - No missing values
  - 16 input variables

Dataset Overview

- The dataset consists of various input variables that can be categorized into three main sections:
    1. Bank client data (e.g., *age, job, marital status, education, default, balance, housing, loan)*
    2. Last contact of the current campaign (e.g., *contact, day, month, duration)*
    3. Other attributes (e.g., *campaign, pdays, previous, poutcome*)
- The output variable (desired target) is the client's subscription to a term deposit (binary: "yes", "no")

Univariate Analysis

- The target variable is significantly skewed towards non-subscribers, with 88.3% classified as "no", indicating positive subscriptions will be difficult to predict due to data bias
- The large class imbalance could negatively impact predictive model performance if not addressed, as models may become biased towards the majority non-subscriber class
- The univariate analysis establishes an initial baseline, highlighting the need for further exploration of other variables to better understand drivers of the minority subscriber class given target distribution challenges

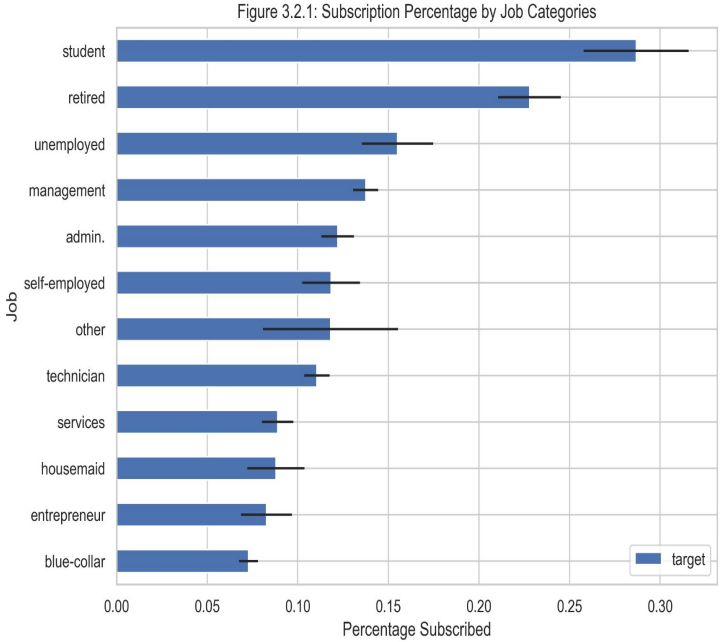


Bivariate Analysis - Job Categories

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

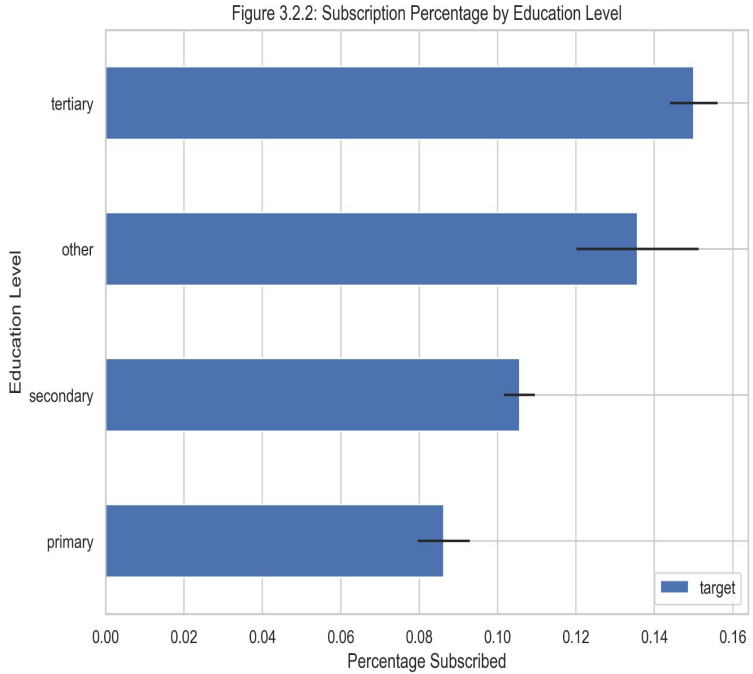
- Students/retirees may be more likely to subscribe due to their independence in finances

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



Bivariate Analysis - Education Level

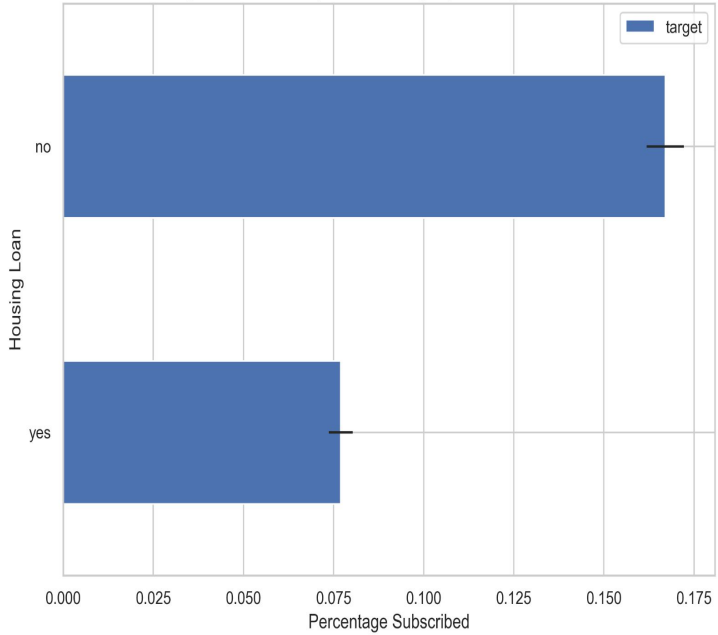
- Tertiary/advanced degree holders showed the highest subscription rate at 15.01%, followed by other education (13.57%) and secondary (10.56%)
- Higher education is associated with a greater likelihood to subscribe than other levels
- May be attributed to factors like increased financial literacy/knowledge gained from advanced studies



Bivariate Analysis - Housing Status

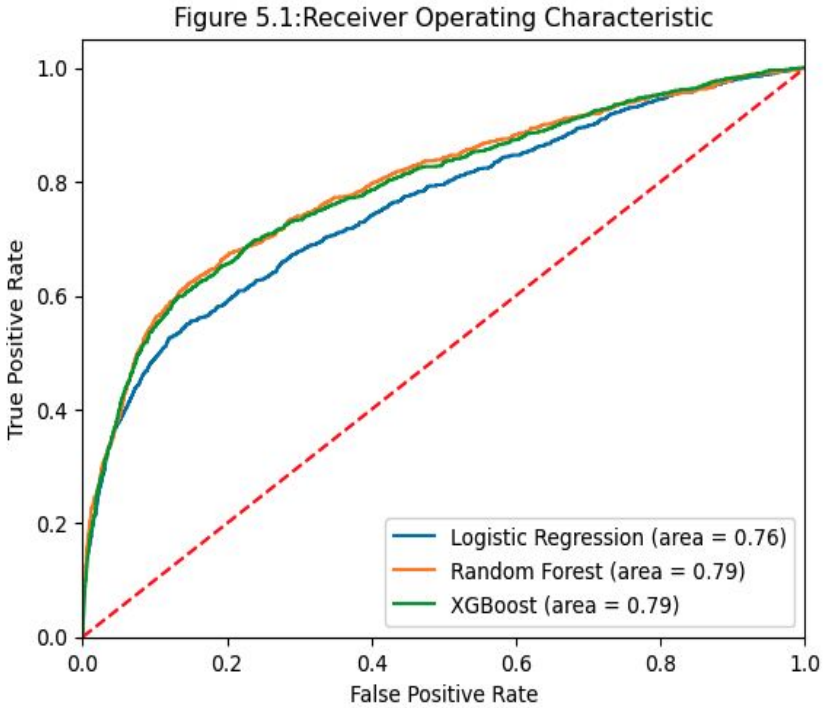
- A mortgage loan significantly lowered subscription rates from 16.70% to 7.70%
- Housing loan-free consumers subscribed more frequently than those with loans
- Considering the substantial disparity in rates, it is imperative to tailor marketing based on loan ownership status

Figure 3.2.4: Subscription Percentage by Housing Loan Status



Model Performance

- Based on predictive accuracy, Random Forest outperformed Logistic Regression (0.7575) and XGBoost (0.7899)
- According to their ROC AUC scores above 0.75, Random Forest performed slightly better than Logistic Regression and XGBoost
- The ROC AUC measures the ability of models to distinguish between positive and negative classes, but does not assess their calibration



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, suggesting that the model's performance is suboptimal.

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

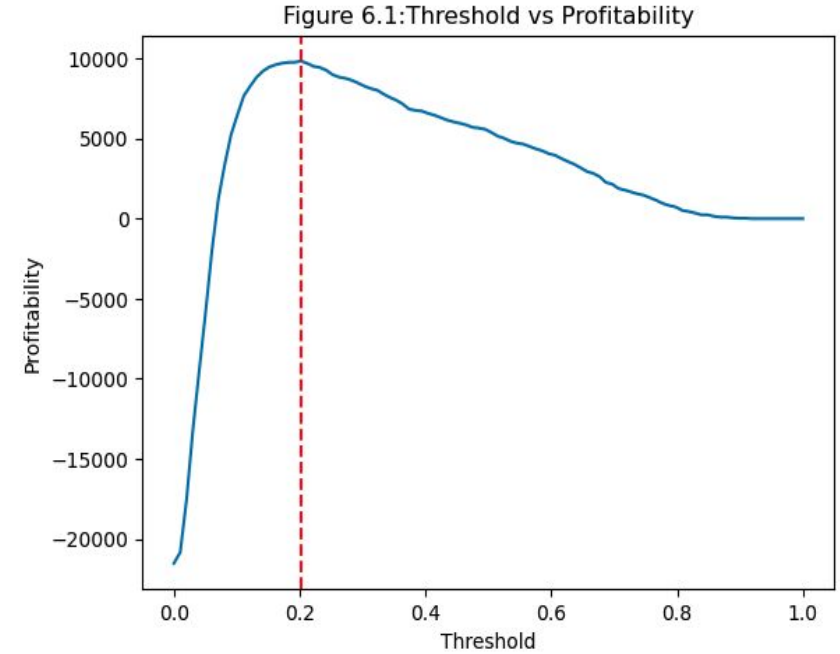
Feature Importance

- Random Forest identified *balance, age, day, campaign, and pdays* as the top 5 most important features for predicting subscriptions
- These align with expectations as 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

- Curve plots profit vs thresholds, finds optimal threshold for max returns
- Red line sets threshold at 0.2, balances outreach and success to widen reach
- Optimal threshold weighs predictions and profits, financial impact matters beyond accuracy



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 status, 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!