Exploratory Data Analysis of Portuguese Bank Data

Naomi Buell[1] and Richie Rivera[1]

[1]CUNY School of Professional Studies

September 28, 2025

1 Introduction

This analysis focuses on a bank marketing dataset that records the outcomes of a Portuguese bank's telemarketing campaign for term deposits. Our goal is to apply machine learning techniques to this dataset and figure out most effective tactics that will help the bank in next campaign to persuade more customers to subscribe to the bank's term deposit. We use `bank-additional-full.csv` with all examples, ordered by date (from May 2008 to November 2010) (Moro, Rita, and Cortez 2014). The dataset in enriched with the addition of five new social and economic features/attributes. We picked this version over the original because it was found that the addition of the five new social and economic attributes lead to substantial improvement in the prediction of a success.

2 Exploratory Data Analysis

The dataset contains 41,188 records and 21 attributes, including both client-level variables (e.g., age, job, marital status) and economic indicators (e.g., employment variation rate, consumer confidence). Initial correlation analysis shows that certain macroeconomic features are highly collinear. For instance, `emp.var.rate` is strongly correlated with both `euribor3m` (r = 0.97) and `nr.employed` (r = 0.95), suggesting redundancy and the need for dimensionality reduction (Figure 1).

Numeric variable distributions reveal skewness and outliers. For example, `campaign` and `duration` are heavily right-skewed, with extreme cases of over 50 calls to the same client. The `pdays` variable has a peculiar distribution with a spike at 999, indicating no prior contact—a special value requiring transformation.

Categorical variable analysis was initially done by plotting the proportional rates for each class. This analysis showed that there was significant variation in outcomes based on a few key factors. Specifically, `defaults` had the strongest impact, leading to a 100% rejection rate, while `contact` and the applicant's job and `poutcome` also showed clear predictive potential. In contrast, other variables like `marital`, `housing`, and `loan` had very little variation and were not strong indicators of the outcome. In order to double check our visual analysis, we used Weight of Evidence (WoE) and Information Value (IV) which are methods outlined in (Bhalla 2015). WoE calculates the predictive ability of a variable in relation to a dependent binary variable. Since our dependent outcome variable is a "yes" or "no," the method is applicable. The use of IV summarizes the WoE by using a weighed sum of the WoE values for each bin in the independent variable.

An advantage of using WoE and IV is that we can also find the predictive ability of numerical values by creating bins of each numerical value thus allowing us to also have a comparable outcome for numeric and categorical variables. By employing this method, we were able to verify that `poutcome` has a strong predictive ability and that `previous`, `cons.price.idx`, `month` were other strong predictors. Table 1 bins IV and summarizes how strong of a predictor each variable is. Specifically, looking at Table 2, it has a moderate predictive ability with a higher likelihood of subscribing when the client younger than 31 or older than 55. Now looking at Table 3 for `euribor3m`, we can see that the likelihood of an converting is highest when we are in the lowest bin

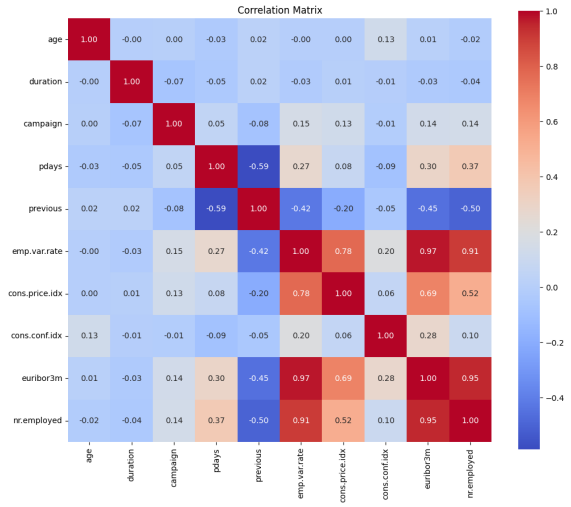

Figure 1: Figure 1. Correlation matrix

Information Value	**Variable Predictiveness**	**Variables in category**
Less than 0.02	Not useful for prediction	`pdays`, `loan`, `housing`, $\text{day}_o f_w eek$
& 0.02 to 0.1	Weak predictive Power	`marital`, `education`, `campaign`
0.1 to 0.3	Medium predictive Power	`default`, `age`, `job`, `contact`
0.3 to 0.5	Strong predictive Power	`previous`, `cons.price.idx`, `month`
>0.5	Suspicious Predictive Power	`poutcome`, `cons.conf.idx`, `euribor3m`, `emp.var.rate`, `nr.employed`, `duration`

Table 1: Table of Information Value and Variable Predictiveness with Variables in Category

Table 2: WOE and IV Analysis of Age Bins

Cutoff	**N**	**Events**	**% of Events**
(16.999, 28.0]	4216	736	0.158621
(28.0, 31.0]	5114	608	0.131034
(31.0, 33.0]	3679	394	0.084914
(33.0, 35.0]	3504	351	0.075647
(35.0, 38.0]	4662	434	0.093534
(38.0, 41.0]	3871	311	0.067026
(41.0, 45.0]	4311	348	0.075000
(45.0, 49.0]	3776	289	0.062284
(49.0, 55.0]	4473	428	0.092241
(55.0, 98.0]	3582	741	0.159698

where the `euribor3` month rate is less than 1.046 and typically decreases as the rate increases. We also found that some categorical variables had significant missingness. `pdays` (the number of days since last contact) had a `default` (whether or not the client has credit in default) has significant missingness. `education` (education level), `housing` (whether the client has a housing loan), `loan` (whether the client has a personal loan), `job` (type of job), and `marital` (marital status) have some minor missingness. We will address this missingness in the **Pre-processing section**.

3 Algorithm selection

Given the binary classification target (yes or no to subscription to the bank's term deposit), we decided to recommend a logistic regression model to predict outcomes. This model will be easy to interpret, computationally efficient, and provides probabilistic predictions for our binary outcome.

Table 3: WOE and IV Analysis of Euribor 3-Month Rate Bins

Cutoff	N	Events	% of Events
(0.633, 1.046]	4128	1895	0.408405
(1.046, 1.299]	4508	666	0.143534
(1.299, 1.41]	3926	494	0.106466
(1.41, 4.191]	4504	415	0.089440
(4.191, 4.857]	4989	156	0.033621
(4.857, 4.864]	3457	150	0.032328
(4.864, 4.96]	3983	235	0.050647
(4.96, 4.962]	4515	234	0.050431
(4.962, 4.964]	3662	181	0.039009
(4.964, 5.045]	3516	214	0.046121

To mitigate the colinearity present between predictors, we will perform dimensionality reduction in the **Pre-processing section**. We also considered Naive Bayes and KNN. We have enough observations that we do not need to rely on these models that tend to do well even with smaller sample sizes. Naive Bayes doesn't work well for datasets with a large number of continuous features, which we have 10 of here. KNN is more computationally intensive than a logistic regression.

4 Pre-processing

We addressed missingness by removing columns with high levels of missingness and weak predictive power (`default`, whether the client has credit in default, was 21% missing missing. `pdays`, the number of days that passed by after the client was last contacted from a previous campaign, was 96% missing). We also removed missing observations of features that were important and had with minor missingness (`education`, `housing`, `loan`, `job`, and `marital`); Since the missingness is low, we will not lose much information by dropping these rows. Lastly, we removed the `duration` column since this attribute highly affects the output target (e.g., if `duration`=0 then `y`="no") and was not recommended for use with predictive models per the data notes.

Since logistic regression requires categorical variables to be encoded numerically, we performed feature engineering on to map the ordinal categorical variable 'education' to numeric values corresponding to the number of years of schooling a client received. We also transformed data using one-hot encoding to nominal categorical variables.

We standardized the numerical columns to ensure logistic regression performed well across features on the same scale. To address multicollinearity, we applied principal component analysis (PCA) to reduce the feature space and remove redundant predictors. The PCA results show that the first principal component (PC1) is dominated by campaign (1.36), indicating that the number of contacts during the campaign is the primary driver of this dimension. In contrast, the second principal component (PC2) is shaped by broader social and economic variables—`emp.var.rate`, `cons.price.idx`, `euribor3m`, and `nr.employed`—which all load strongly (0.43–0.45). While campaign contributes negatively to PC2 (-0.28), its influence here is weaker compared to these macroeconomic features.

In our data, it was observed that the vast majority of cases resulted with the client not subscribing (outcome = no). Because of that, we have an inherent imbalance in our data. To handle this, we've considered over-sampling the minority class, under-sampling the majority class, assigning a larger penalty for the minority class, and generating synthetic samples. Assigning a larger penalty for the minority class involves hyperparameter tuning which would can only be done during model development, which can be used later. To immediately handle the imbalance we would need to employ a sampling technique. Generating synthetic samples would risk introducing noise to our data and over-sampling the minority class would risk overfitting. As our dataset has a rich number of samples (n=41,188), we elected to under-sample the majority class as we were able to retain a sufficient amount of data (n=8,516) and did not have to risk overfitting or generating noise.

Business insight recommendations

Based on our exploratory data analysis, we present the following recommendations for Portuguese banking institution to increase the number of clients subscribing to term deposits. Our first recommendation is to increase the total number of bank clients. Our second recommendation is to monitor daily, monthly, and quarterly economic indicators, and focus spending resources during economic times of improved likelihood of a client subscriptions. This will save money on marketing efforts when uncontrollable and strong predictor variables such as `cons.conf.idx`, `emp.var.rate`, `cons.price.idx`, `euribor3m` are unfavorable. WoE and IV analysis showed that these uncontrollable economic factors had strong potential to predict whether clients are willing to subscribe and by saving resources during this time, the bank can redeploy these resources when opportunity is the highest.

Our third recommendation is to focus marketing efforts on bank clients who fit a profile of having a higher likelihood of subscribing by hitting client specific criteria. Namely, those clients who did not `default`, those who subscribed during the previous marketing campaign (`poutcome`), and those who work admin jobs, are retired, or are students (`job`). We also recommend contacting clients in the December, March, October, and September months, since these months had a rate of subscription. We recommend that the bank prioritize contacting those who can be reached by cell phone. Our WoE and IV analysis shows that clients who fall within the 31-55 `age` range have a lower likelihood of subscribing and should be de-prioritized.

References

Bhalla, Deepanshu (2015). *Weight of Evidence (WOE) and Information Value (IV) Explained.* https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html#what_is_weight_of_evidence. Accessed: Sunday Sept. 28th.

Moro, S., P. Rita, and P. Cortez (2014). *Bank Marketing.* UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K306.