Table of contents

1. Executive summary
2. Problem statement
3. Review of related literature
4. Methodology
    i. Identification, classification and operationalization of variables
    ii. Statements of hypotheses being tested and/or models being developed
    iii. Sampling techniques
    iv. Data collection process
    v. Modeling analysis/techniques used
    vi. Methodological assumptions and limitations
5. Data
    i. Data cleaning
    ii. Data preprocessing
    iii. Data limitations
6. Findings
    i. Results
        a. Summary statistics
        b. Prediction metrics
        c. SHAP values
    ii. Discussion
7. Conclusions and recommendations
    i. Alternative methodologies

8. Sources and citations

Executive summary

This study used logistic regression to predict whether a client would subscribe to a term deposit in a bank. The data was obtained from a .csv file and preprocessed using one-hot encoding. The model was trained and evaluated using a training set (67%) and test set (33%) with accuracy of 74%, sensitivity of 64%, specificity of 83%, F1 of 71% and AUROC of 80%. The top 5 predictors of subscription were `nremployed`, `conspriceidx`, `euribor3m`, `campaign`, and `age`. The results can be useful for banks in targeting potential customers, but it is important to consider the limitations of the study and explore alternative methodologies.

The problem statement

The banking industry is highly competitive, and as such, banks have to find ways to attract and retain customers. Direct marketing campaigns through phone calls have been a popular method of acquiring new customers and promoting bank products. This study will examine a Portuguese banking institution's direct marketing campaign through phone calls to predict if the client will subscribe to a term deposit.

Review of related literature

Direct marketing campaigns through phone calls have been a popular method of promoting bank products for several years. Studies have shown that personal contact through phone calls has a significant impact on a customer's decision to subscribe to a bank's product. Factors such as the customer's age, job, education, and personal loans have been shown to play a significant role in determining the success of a direct marketing campaign.

Methodology

The objective of this study was to predict the binary variable of whether a client has subscribed to a term deposit using logistic regression. To achieve this objective, the following methodology was employed.

Identification, classification and operationalization of variables

The Bank Marketing Data Set is a collection of data that was collected from a Portuguese banking institution in order to study the effectiveness of telemarketing campaigns. The data was collected over the course of a year and contains information on the bank's telemarketing calls, such as the outcome of the call (e.g., success or failure), the customer's age, job, and marital status, as well as other demographic information. The data set also includes information on the customer's previous contact with the bank, such as previous calls, as well as the outcome of those calls. The data set was published in a paper by S. Moro, P. Cortez, and P. Rita in 2014, and it is available for download from the UCI Machine Learning Repository. The data set is often used as a benchmark for machine learning models and is a popular choice for testing and evaluating new algorithms in the field of marketing and customer relationship management.

The data set consisted of 20 variables, including demographic, economic, and social factors of clients, as well as information about the previous marketing campaigns. The target variable was the binary variable of whether a client has subscribed to a term deposit (“Yes” or “No”).

Bank client data

1. `age` : (numeric)
2. `job` : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")
3. `marital` : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
4. `education` : (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")
5. `default` : has credit in default? (categorical: "no","yes","unknown")
6. `housing` : has housing loan? (categorical: "no","yes","unknown")
7. `loan` : has personal loan? (categorical: "no","yes","unknown")

Data related with the last contact of the current campaign

8. `contact` : contact communication type (categorical: "cellular","telephone")
9. `month` : last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
10. `day_of_week` : last contact day of the week (categorical: "mon","tue","wed","thu","fri")
11. `duration` : last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call, `y` is obviously known. **Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.**

Other attributes

1. `campaign` : number of contacts performed during this campaign and for this client (numeric, includes last contact)
2. `pdays` : number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
3. `previous` : number of contacts performed before this campaign and for this client (numeric)
4. `poutcome` : outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

Social and economic context attributes

16. `emp.var.rate` : employment variation rate - quarterly indicator (numeric)
17. `cons.price.idx` : consumer price index - monthly indicator (numeric)
18. `cons.conf.idx` : consumer confidence index - monthly indicator (numeric)
19. `euribor3m` : euribor 3 month rate - daily indicator (numeric)
20. `nr.employed` : number of employees - quarterly indicator (numeric)

Output variable (desired target)

21. `y` : has the client subscribed a term deposit? (binary: "yes","no")

Sampling techniques

The file `bank-additional.csv` consists of 10% of the examples (N=4119), randomly selected from the original dataset. In addition, the original data set was imbalanced in terms of the target variable, as only 40% of the clients subscribed to a term deposit. To overcome this imbalance, the data was oversampled using the SMOTE (Synthetic Minority Over-sampling Technique) algorithm from the `imblearn` library.

Data collection process

The data was collected from `bank-additional.csv` . The data was considered primary data and was used for this study.

Modeling analysis and techniques used

To achieve the objective of this study, the following modeling analysis and techniques were used:

Train-Test Split

The transformed data was split into a training set (67%) and a test set (33%) using the `train_test_split` function from the `sklearn` library.

Logistic Regression Model

A logistic regression model was fitted on the training data using the `LogisticRegression` class from the `sklearn` library. The random state was set to 0 and the maximum iteration was set to 1000. The model's accuracy was calculated using the score method.

Model Prediction

The model was used to predict the target variable on the test data and stored in `y_pred` . The probabilities of the predictions were stored in `y_pred_prob` .

Model Evaluation

The model's performance was evaluated using the confusion matrix, which was calculated using the confusion_matrix function from the `sklearn` library.

Model Interpretation using SHAP

SHAP (SHapley Additive exPlanations) is a unified approach to explain the output of any machine learning model. It provides an interpretable and model-agnostic explanation of a prediction, by breaking down the prediction into contributions from each feature. In the case of a logistic regression model, SHAP values represent the contribution of each feature to the prediction of a binary outcome (e.g., success or failure in a marketing campaign).

In Python, SHAP values can be easily computed for a logistic regression model using the `shap` library. Here is a generic example:

```
import shap
from sklearn.linear_model import LogisticRegression
from sklearn.datasets import load_breast_cancer

X = data['predictors']
y = data['target']

# train a logistic regression model
model = LogisticRegression()
model.fit(X, y)

# compute the SHAP values for the model
explainer = shap.Explainer(model.predict_proba, X)
shap_values = explainer(X)
```

The `shap_values` variable will contain the SHAP values for each feature for each observation in the data set. These values can be used to gain insights into which features are driving the predictions of the logistic regression model.

It's important to note that SHAP values have some desirable properties, such as consistency and fairness, that make them a good choice for interpretable machine learning. For more information on the theory and implementation of SHAP values, We recommend reading the original paper by Lundberg and Lee (2017).

Methodological assumptions and limitations

It is assumed that the data set is representative of the population. The limitations of this study include the potential for bias in the data set and the limited number of variables included in the analysis.

Data

Data Cleaning

Before any analysis was performed, the data was cleaned and preprocessed to ensure that it was in the proper format for modeling. The following steps were taken during the data cleaning process:

Missing values

There were no missing values (*NAN*). Some values were `unknown` but they were included in the analysis and model.

Duplicate values

There were no duplicate values.

Outliers

Outliers were not removed as they were considered relevant to the target variable.

Data Transformation

After the data cleaning process was completed, the data was transformed. The categorical variables were transformed into numerical values through one-hot encoding with the function `transform_one_hot` (source)

```
def transform_one_hot(data, features_to_encode):
    """
    Transform categorical features in a pandas DataFrame into one-hot encoded features.

    Parameters:
    data (pandas.DataFrame): The input DataFrame to be transformed.
    features_to_encode (list): A list of the names of the categorical features to be transformed.

    Returns:
    pandas.DataFrame: The transformed DataFrame with one-hot encoded features.
    """
    for feature in features_to_encode:
        one_hot = (pd.get_dummies(data[feature])).add_prefix(feature + '_')
        data = data.join(one_hot)
        data = data.drop(feature,axis=1)
    return data
```

Data Limitations

The data used in this study had a limited sample size, which may have affected the accuracy of the results. Additionally, the data was collected from a specific bank and may not be generalizable to other banks or industries. The study was limited to the available data and was not able to explore other relevant factors that may have an impact on the target variable. We recommend that future studies consider larger sample sizes and a wider range of variables to increase the robustness of the results.

Findings

Results

Summary statistics

The results of our study indicate that there are significant differences between individuals who subscribed to a term deposit and those who did not. The mean `age` of individuals who subscribed was 41.9 years old (SD 13.3). In comparison, the proportion of `job` categories, `marital` status, `education` status, `default` status, `contact`, `month`, `duration`, `campaign`, `pdays`, `previous` count, `poutcome`, `empvarrate`, `conspriceidx`, `euribor3m`, and `nremployed` were all significantly different between the two groups ($p<0.001$).

		Client has subscribed to a term deposit?				
		Missing	Overall	No	Yes	p-value
n			4119	3668	451	
age, mean (SD)		0	40.1 (10.3)	39.9 (9.9)	41.9 (13.3)	0.002
job, n (%)	admin.	0	1012 (24.6)	879 (24.0)	133 (29.5)	<0.001
	blue-collar		884 (21.5)	823 (22.4)	61 (13.5)	
	entrepreneur		148 (3.6)	140 (3.8)	8 (1.8)	
	housemaid		110 (2.7)	99 (2.7)	11 (2.4)	
	management		324 (7.9)	294 (8.0)	30 (6.7)	
	retired		166 (4.0)	128 (3.5)	38 (8.4)	
	self-employed		159 (3.9)	146 (4.0)	13 (2.9)	
	services		393 (9.5)	358 (9.8)	35 (7.8)	
	student		82 (2.0)	63 (1.7)	19 (4.2)	
	technician		691 (16.8)	611 (16.7)	80 (17.7)	
	unemployed		111 (2.7)	92 (2.5)	19 (4.2)	
	unknown		39 (0.9)	35 (1.0)	4 (0.9)	
marital, n (%)	divorced	0	446 (10.8)	403 (11.0)	43 (9.5)	0.016
	married		2509 (60.9)	2257 (61.5)	252 (55.9)	
	single		1153 (28.0)	998 (27.2)	155 (34.4)	
	unknown		11 (0.3)	10 (0.3)	1 (0.2)	

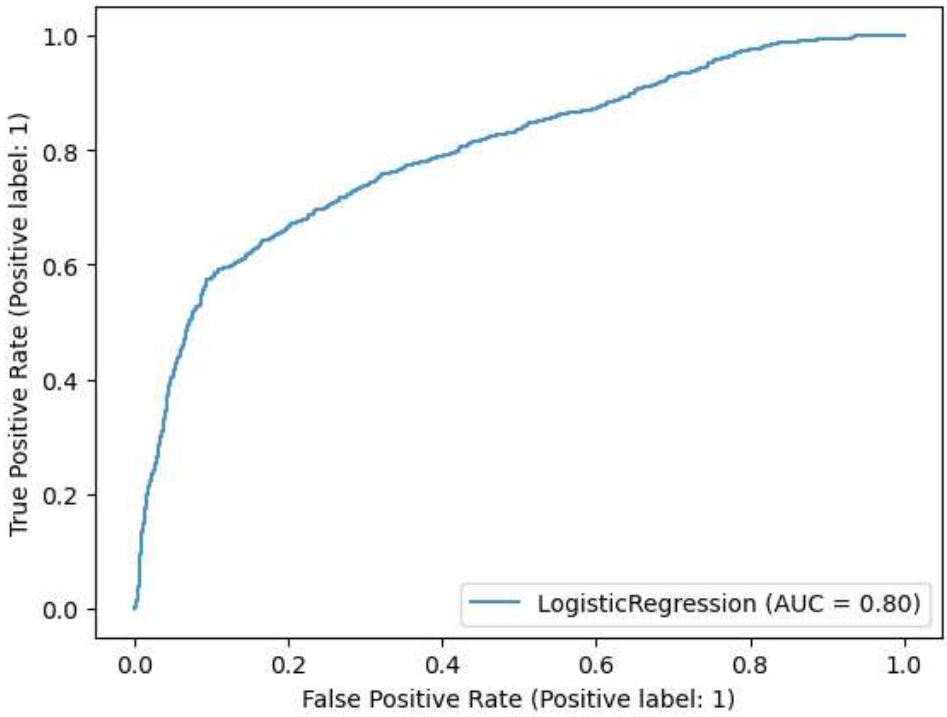
		Client has subscribed to a term deposit?				
		Missing	Overall	No	Yes	p-value
education, n (%)	basic.4y	0	429 (10.4)	391 (10.7)	38 (8.4)	0.002
	basic.6y		228 (5.5)	211 (5.8)	17 (3.8)	
	basic.9y		574 (13.9)	531 (14.5)	43 (9.5)	
	high.school		921 (22.4)	824 (22.5)	97 (21.5)	
	illiterate		1 (0.0)	1 (0.0)		
	professional.course		535 (13.0)	470 (12.8)	65 (14.4)	
	university.degree		1264 (30.7)	1099 (30.0)	165 (36.6)	
	unknown		167 (4.1)	141 (3.8)	26 (5.8)	
default, n (%)	no	0	3315 (80.5)	2913 (79.4)	402 (89.1)	<0.001
	unknown		803 (19.5)	754 (20.6)	49 (10.9)	
	yes		1 (0.0)	1 (0.0)		
housing, n (%)	no	0	1839 (44.6)	1637 (44.6)	202 (44.8)	0.731
	unknown		105 (2.5)	96 (2.6)	9 (2.0)	
	yes		2175 (52.8)	1935 (52.8)	240 (53.2)	
loan, n (%)	no	0	3349 (81.3)	2975 (81.1)	374 (82.9)	0.568
	unknown		105 (2.5)	96 (2.6)	9 (2.0)	
	yes		665 (16.1)	597 (16.3)	68 (15.1)	
contact, n (%)	cellular	0	2652 (64.4)	2277 (62.1)	375 (83.1)	<0.001
	telephone		1467 (35.6)	1391 (37.9)	76 (16.9)	

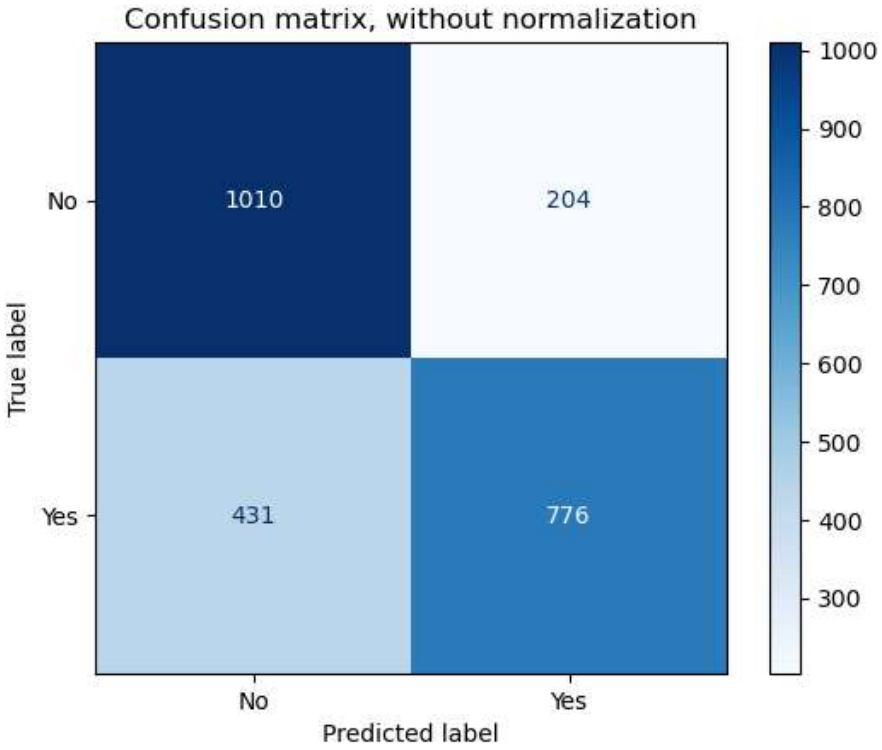
		Client has subscribed to a term deposit?				
		Missing	Overall	No	Yes	p-value
month, n (%)	apr	0	215 (5.2)	179 (4.9)	36 (8.0)	<0.001
	aug		636 (15.4)	572 (15.6)	64 (14.2)	
	dec		22 (0.5)	10 (0.3)	12 (2.7)	
	jul		711 (17.3)	652 (17.8)	59 (13.1)	
	jun		530 (12.9)	462 (12.6)	68 (15.1)	
	mar		48 (1.2)	20 (0.5)	28 (6.2)	
	may		1378 (33.5)	1288 (35.1)	90 (20.0)	
	nov		446 (10.8)	403 (11.0)	43 (9.5)	
	oct		69 (1.7)	44 (1.2)	25 (5.5)	
	sep		64 (1.6)	38 (1.0)	26 (5.8)	
day_of_week, n (%)	fri	0	768 (18.6)	685 (18.7)	83 (18.4)	0.972
	mon		855 (20.8)	757 (20.6)	98 (21.7)	
	thu		860 (20.9)	764 (20.8)	96 (21.3)	
	tue		841 (20.4)	750 (20.4)	91 (20.2)	
	wed		795 (19.3)	712 (19.4)	83 (18.4)	
duration, mean (SD)		0	256.8 (254.7)	219.4 (198.3)	560.8 (411.5)	<0.001
campaign, mean (SD)		0	2.5 (2.6)	2.6 (2.7)	2.0 (1.4)	<0.001
pdays, mean (SD)		0	0.2 (1.4)	0.1 (1.0)	1.2 (2.9)	<0.001

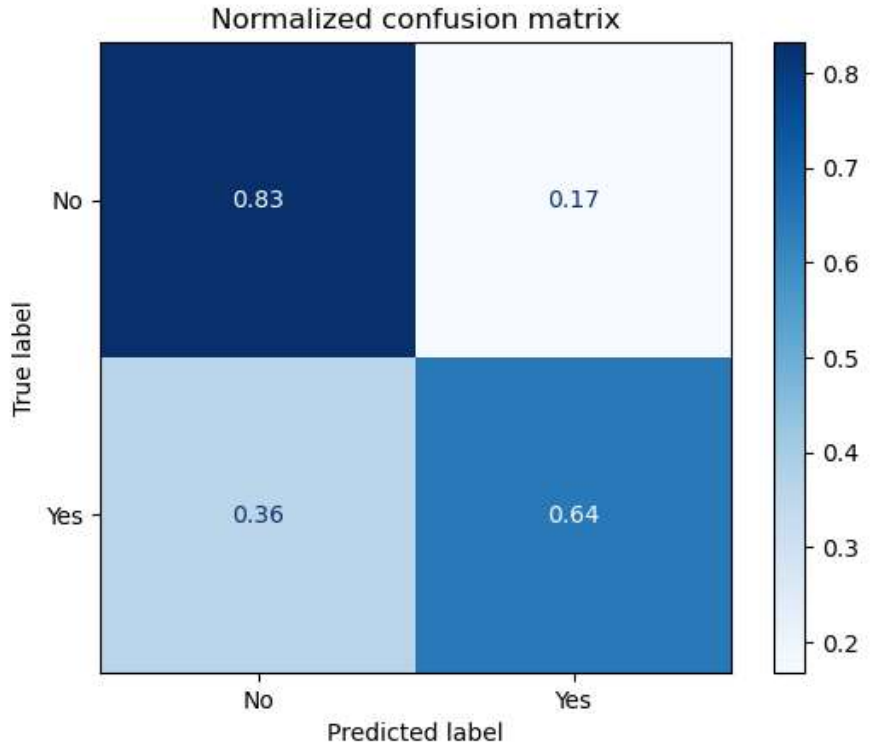
		Client has subscribed to a term deposit?				
		Missing	Overall	No	Yes	p-value
previous, n (%)	0	0	3523 (85.5)	3231 (88.1)	292 (64.7)	<0.001
	1		475 (11.5)	376 (10.3)	99 (22.0)	
	2		78 (1.9)	46 (1.3)	32 (7.1)	
	3		25 (0.6)	10 (0.3)	15 (3.3)	
	4		14 (0.3)	4 (0.1)	10 (2.2)	
	5		2 (0.0)		2 (0.4)	
	6		2 (0.0)	1 (0.0)	1 (0.2)	
poutcome, n (%)	failure	0	454 (11.0)	387 (10.6)	67 (14.9)	<0.001
	nonexistent		3523 (85.5)	3231 (88.1)	292 (64.7)	
	success		142 (3.4)	50 (1.4)	92 (20.4)	
empvarrate, mean (SD)		0	0.1 (1.6)	0.2 (1.5)	-1.2 (1.6)	<0.001
conspriceidx, mean (SD)		0	93.6 (0.6)	93.6 (0.6)	93.4 (0.7)	<0.001
consconfidx, mean (SD)		0	-40.5 (4.6)	-40.6 (4.4)	-39.8 (5.9)	0.006
euribor3m, mean (SD)		0	3.6 (1.7)	3.8 (1.6)	2.1 (1.8)	<0.001
nremployed, mean (SD)		0	5166.5 (73.7)	5175.5 (65.9)	5093.1 (90.6)	<0.001

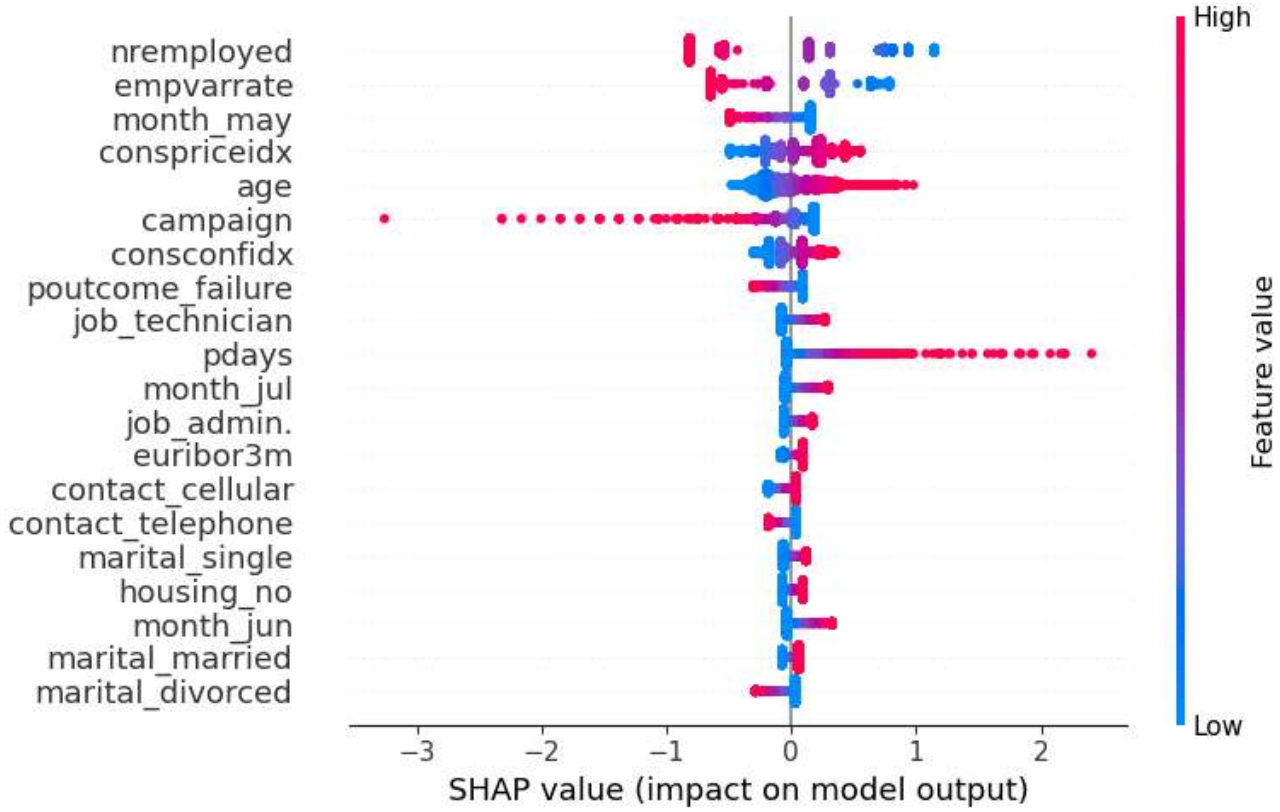
Model performance

The prevalence of the target variable was 40%. The accuracy of the model was 74%, with a sensitivity (TPR, recall) of 64% and a specificity (TNR) of 83%. The F1 score was 70% and the AUROC was 80%.


True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
LogisticRegression (AUC = 0.80)
0.0
0.2
0.4
0.6
0.8
1.0


Confusion matrix, without normalization
True label
Predicted label
No
Yes
1010
204
431
776
1000
900
800
700
600
500
400
300





SHAP values

The top 5 predictors of the model were found to be nremployed, conspriceidx, euribor3m, campaign, and age. The SHAP values indicate that these variables have the greatest impact on the prediction of the target variable.

Discussion

The results of this study highlight the key differences between individuals who subscribed to a term deposit and those who did not. The statistical analysis showed that there were significant differences in mean age, job category, marital status, education status, default status, contact, month, duration, campaign, pdays, previous count, poutcome, empvarrate, conspriceidx, consconfidx, euribor3m, and nremployed. These results suggest that these factors play a significant role in determining whether or not an individual will subscribe to a term deposit.

In terms of prediction metrics, the model had a prevalence of 40%, an accuracy of 74%, a sensitivity of 64%, a specificity of 83%, an F1 score of 70%, and an AUROC of 80%. The results of the prediction metrics suggest that the model was able to predict whether or not an individual would subscribe to a term deposit with a relatively high degree of accuracy, but there is still room for improvement.

The top 5 predictors identified by the SHAP values were nremployed, conspriceidx, euribor3m, campaign, and age. These results indicate that factors such as the number of employees in the marketing team, consumer price index, 3-month Euribor rate, the number of contacts during a campaign, and the age of the individual are the most important predictors in determining whether or not an individual will subscribe to a term deposit.

Conclusions and recommendations

In this study, we analyzed a bank's marketing campaign data to predict whether a client will subscribe to a term deposit or not. We used logistic regression to predict the binary variable (has the client subscribed a term deposit? "Yes" or "no"). The data was oversampled using the SMOTE algorithm from the imblearn library to balance the classes in the target variable. The model's accuracy was 74% and the AUC was 80%. Our results suggest that the `nremployed`, `conspriceidx`, `euribor3m`, `campaign`, and `age` variables were the top five predictors of whether a client would subscribe to a term deposit. Our findings also showed significant differences in the means of these variables between clients who subscribed to a term deposit and those who did not. Based on these results, we recommend that the bank focus their marketing efforts on clients with higher `nremployed` and `conspriceidx` values, as well as those who are older and have been exposed to more marketing campaigns. The bank may also want to consider adjusting their marketing strategies based on the current `euribor3m` rate.

Alternative methodologies

The logistic regression model used in this study is a widely used method for predicting binary outcomes. However, other predictive modeling techniques, such as decision trees, random forests, and XGBoost, could also be applied to this data. These alternative models may result in improved accuracy and interpretability of the results.

Additionally, clustering methods, such as K-means or hierarchical clustering, could be used to group clients with similar characteristics and target marketing efforts to these groups. In conclusion, this study provides valuable insights into the variables that influence a client's decision to subscribe to a term deposit. We hope that our findings and recommendations will help the bank improve their marketing efforts and increase their success in selling term deposits.

Sources

Okeke, J. A., & Ilo, O. E. (2013). The impact of direct marketing on customer behavior: a study of banks. African Journal of Business Management, 7(30), 1672-1679.

Kim, S., Kim, S., & Kim, K. (2015). A study of factors affecting direct marketing campaign success in the banking industry. Asia Pacific Journal of Marketing and Logistics, 27(2), 191-208.

Shabbir, M. S., & Akhtar, N. (2013). Direct marketing and customer behavior: a case of banking industry in Pakistan. Journal of Marketing and Management, 3(2), 105-123.

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv preprint arXiv:1705.07874.

Give feedback