Predicting the success of bank telemarketing

1 Introduction

argeting the right set of customers is of high importance in any marketing scheme to avoid wasting unnecessary resources on customers unlikely to purchase, while not missing out on customers that would purchase. This is especially true within telemarketing, where the cost of reaching a customer is significantly higher than for a simple digital impression. It is cost-effective and time-saving for the company, since it eliminates the cost and the effort to contact the clients who have a low chance of buying or subscribing to a product.

In this project, the data[1][2] used are from the direct marketing campaigns (phone calls) of a Portuguese banking institution from May 2008 till November 2010, seeking to market term deposits to new customers. It has information about the bank branches, and the customers that were contacted by the bank, and, importantly, which of the customers subscribed. As such, for the purposes of this project we will be taking on this Portuguese Bank as our client, with the goal to build a classification model to predict if the customer will subscribe to a term deposit or not to help them maximize the profitability of their marketing campaign. We apply different ML techniques Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) and choose the best model that has the highest area under the receiver operating characteristic curve (ROC-AUC) score. Then we tune the model to maximize the profitability for the bank.

2 The Data

The data set has 41188 rows (number of clients), and 21 attributes. The columns have some client info, e.g., age, job, marital status, education, housing, loan, etc., plus some social and economic context attributes of the bank branch. The outcome is binary, i.e., if the client subscribes the outcome is 1, and if the client doesn’t subscribe it is 0. The data is imbalanced. Only 11.26% of the data has positive outcomes.

[1] “Bank Marketing Data Set,” UCI Machine Learning Repository

[2] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.

2.1 Adding the ‘year’ column

The clients (rows) are sorted by the date of contact but the dataset is missing the year column. I found the starting index and ending index for each month so I could assign the year column to each entry. It turns out that in the first year, 2008, we only have contacts in months of May-Aug and Oct-Dec. In 2009, the contacts were from March till December. In 2010, contacts were made from March till November.

2.2 Missing data

There are no null values in the dataset but some columns have ‘unknown’ values. The number of ‘unknown’ values in the following columns are: **‘jobs’:** 330, **‘marital status’:** 80, **‘education’:** 1731, **‘default’:** 8597, **‘housing’ and ‘loan’:** each have 990 unknown values (the “unknown” rows are the same).

3 EDA

3.1 Correlated variables and column selection

- The heat map shows that 'euribor3m' and 'emp.var.rate' are highly correlated so we drop one of them ('emp.var.rate').
- **‘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. *This column is dropped in the EDA section.*
- **‘pdays’:** the number of days passed by after the last call. ‘999’ means that the client was not previously contacted and this case needs special handling. 39673 cases of ‘999’ are present. I replaced the ‘999’ value with '-1'. After this replacement the range of this variable is [-1, 27].
- For every 'Cons.price.idx' there is a unique 'cons.conf.idx'. So it is safe to drop one of these columns. *The 'cons.conf.idx' column is dropped in the EDA section.*
- For every 'Emp.var.rate' there is a unique ‘nr.employed'. So it is safe to drop one of these columns. *The 'nr.employed' column is dropped in the EDA section*

3.2 Exploring the outcome by different attributes

3.2.1 Outcome by year

Fig 3.1 clearly shows that the proportion of positive outcomes has increased year by year. However, the number of contacts in 2008 is much higher than 2009 and 2010 as it is seen in the count plot Fig 3.2. The number of negative outcomes significantly decreased year by year. This shows either the bank has adopted a better model for targeting the customers, or this is the result of economic recession during the 2008-2010 period which has affected the subscription trend.

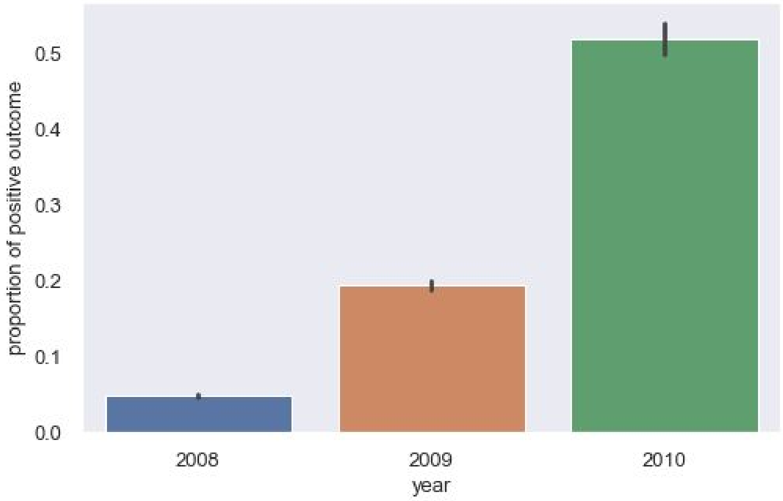


Fig 3.1
A barplot showing the proportion of positive outcomes by year.

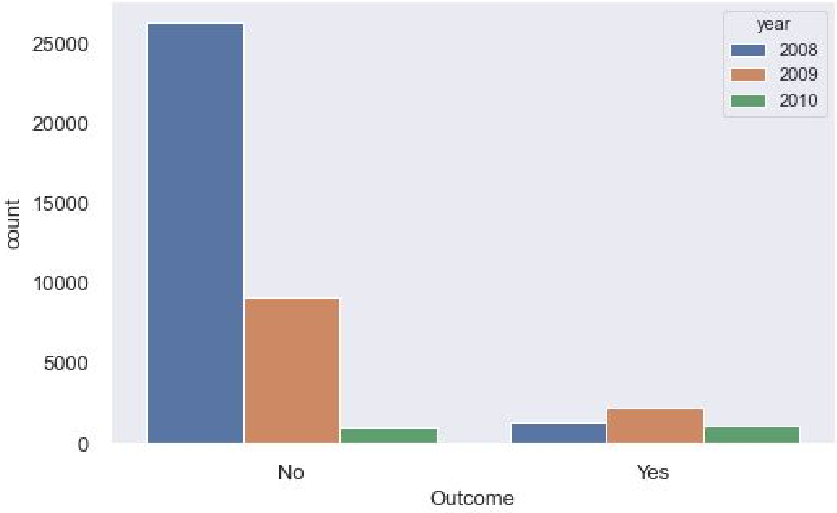


Fig 3.2
A countplot showing the number of positive and negative outcomes by year.

3.2.2 Outcome by year and month

Fig 3.3 shows that in 2008 the proportion of positive outcomes in October (the purple bar) was much higher than other months. In 2009, May and April had the lowest proportion of positive outcomes and in 2010 all proportions for different months got closer to each other. This shows that the trend has changed significantly from 2008 to 2010. This data is during the economic crisis which could have affected the outcome trend significantly. Some months are missing for example year 2008 does not have any contacts for months of Jan-Apr and Sep. Year 2009 has contacts for March-December months, and year 2010 has March-November.

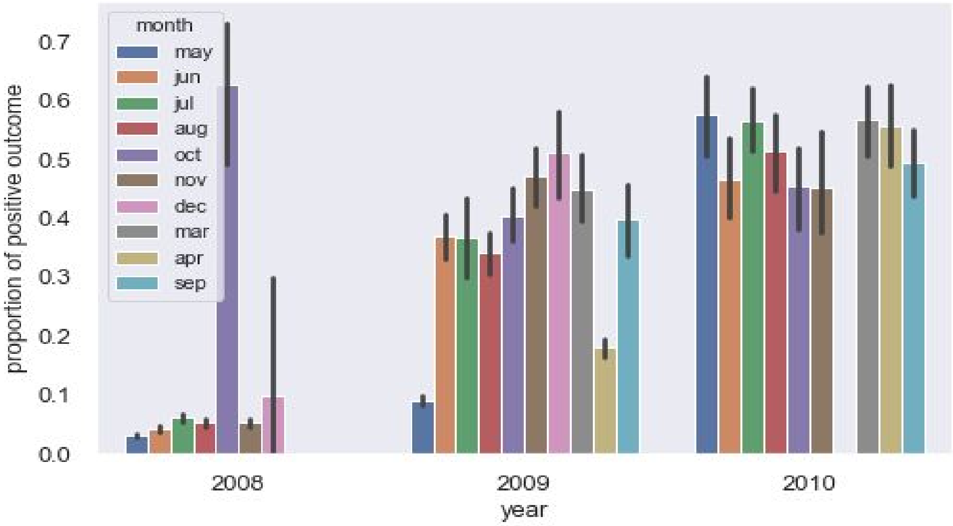


Fig 3.3
A barplot showing the proportion of positive outcome by year and month.

3.2.3 Outcome by job

In Fig 3.4 we can see that students and retirees have the highest proportions of positive outcomes. The p-value for the ‘job’ column is 3.268412e-199, based on significance level of α=5%, this column is significant.

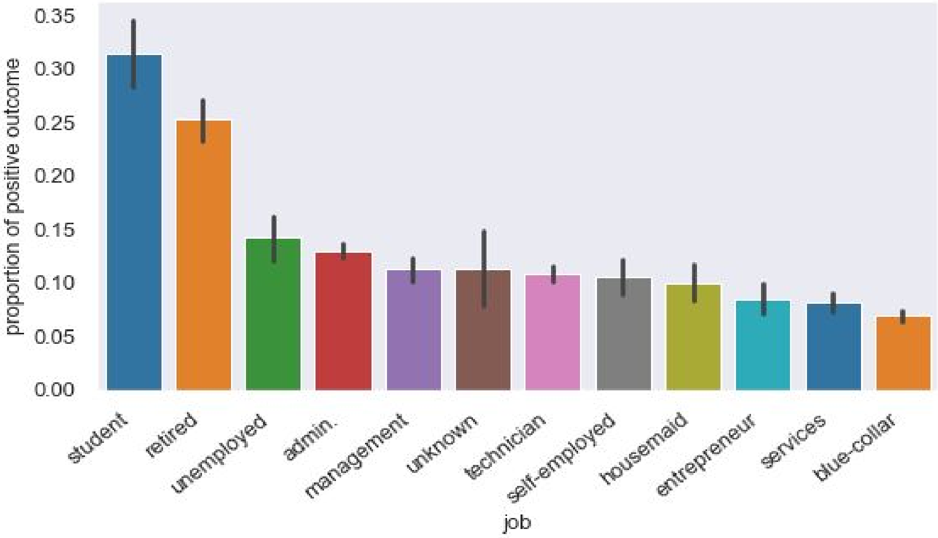


Fig 3.4
A barplot showing the proportion of positive outcome by job.

3.2.4 Outcome by education

The illiterate had a higher proportion of positive outcome but the confidence interval for this group is so big and overlaps with other groups. The p-value for the ‘education’ column is 3.746768e-38, based on significance level of ɑ=5%, this column is significant.

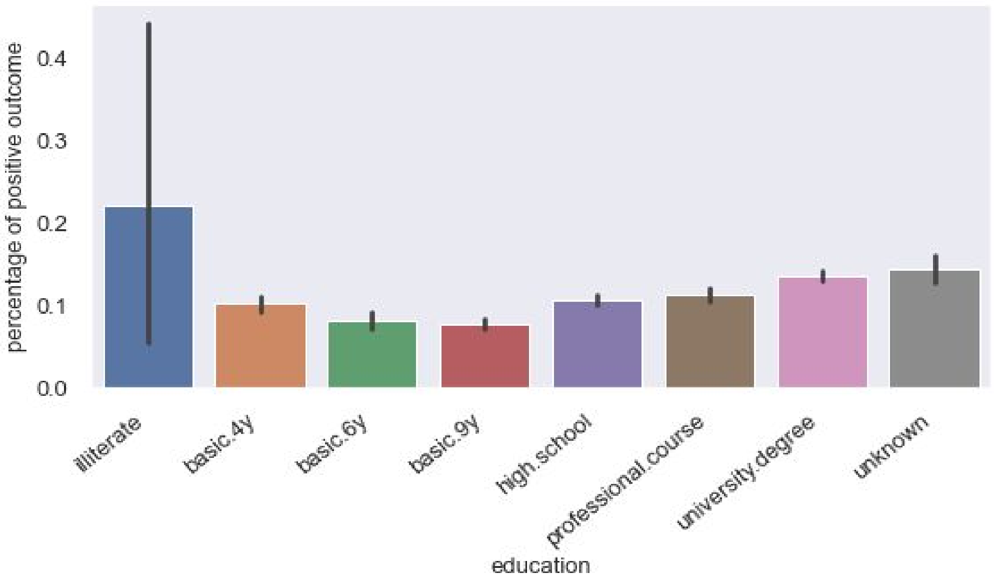


Fig 3.5
A barplot showing the proportion of positive outcome by education.

3.2.5 Outcome by age

Customers younger than 25 and customers older than 60 have a bigger proportion of positive outcomes, Fig 3.6. This is consistent with the result from the ‘Outcome by job’ section above which shows students and retirees have a high proportion of positive outcomes. The p-value for the ‘age’ column is 3.716987e-299, based on significance level of ɑ=5%, this column is significant.

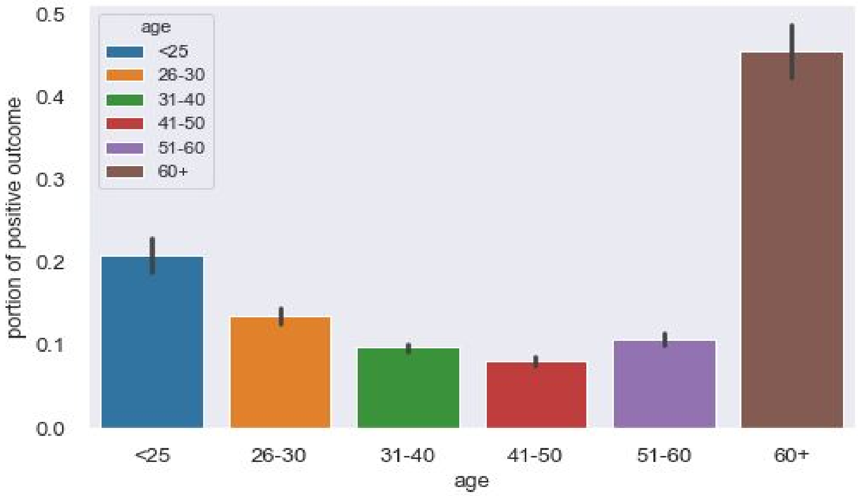


Fig 3.6
A barplot showing the proportion of positive outcome by age group.

4 Model selection

In order to find the best model that fits our data, we apply different classification methods; RF, LR, KNN, and DT. The ROC-AUC score for these models are listed in Table 4.1. The best model with the highest score is RF, so we choose this model for tuning and further investigation.

Table 4.1 ROC-AUC score for different classification models applied to our data set.

Model	ROC-AUC score
Random Forest	0.8027
Logistic Regression	0.7945
KNN	0.7000
Decision Tree	0.4924

5 Model tuning

The ultimate goal of any marketing strategy is to increase profitability for the client. We tune our best model, RF, by maximizing profitability and find a new threshold for this model.

5.1 Profitability

We define the profitability function, which takes the cost of each call from the bank to customers and the revenue generated for the bank if the customer subscribes as inputs, and calculates the profit made by the bank by this model. Since we don't have access to exact values of these inputs, in order to calculate the profit we need to estimate these values. We estimate the cost of the call to each client is $5 and the average loan amount is 5k, they offer 5% back but loan out to someone else for 7%, which results in a 2% differential = $100 profit on a term deposit. These numbers are estimates and may vary for each bank. The exact values can easily be adjusted by the client in the code, with appropriate inputs for the cost and average revenue for each term deposit.

5.2 Threshold tuning

By graphing profitability vs threshold for the RF model, it becomes apparent that maximum profitability is achieved by threshold=0.271, where overall profitability of $87,925. This result is based on the cost=$5 and average-revenue=$100 inputs for the profitability function. To put this in context, we have 12,353 rows in our test set, of which 1,424 purchased term deposits (fig 5.2), so our max profitability, if we only called the clients that would purchase and no one else, would be 1,424*$95 = ~$135k. For a baseline model, if we called everyone the cost would be 12,353*$5 = $61,765, and the profitability would then be 1,424*100 - $61,765 = $80,635. This shows that our model increases the profitability of a telemarketing campaign by ~9% over simply calling everyone on the list.

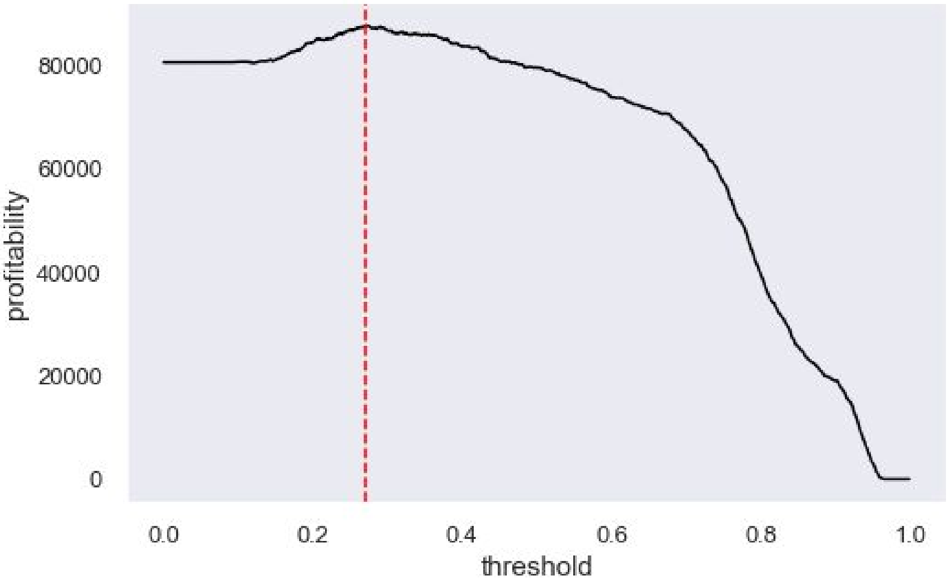
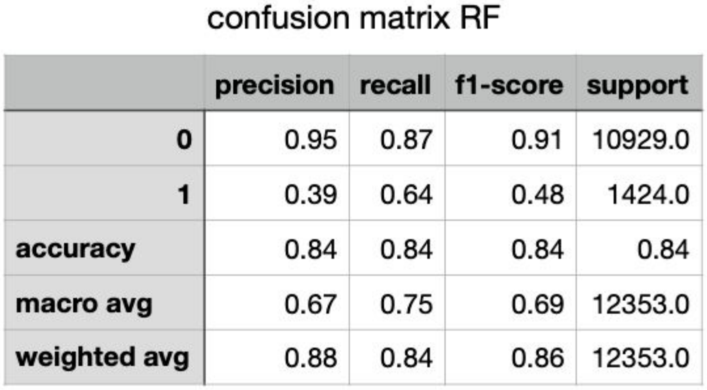


Fig 5.1

Profitability vs threshold for RF model.



confusion matrix RF

	precision	recall	f1-score	support
0	0.95	0.87	0.91	10929.0
1	0.39	0.64	0.48	1424.0
accuracy	0.84	0.84	0.84	0.84
macro avg	0.67	0.75	0.69	12353.0
weighted avg	0.88	0.84	0.86	12353.0

Fig 5.2

Confusion matrix for RF model.

6 Takeaway and future research

In the end, in this case we found we could build a model that could increase profitability of a telemarketing campaign by ~9% compared to the simple strategy of calling everyone on the list. The cost of marketing and revenue for each company is different. The profitability function that we defined in this project takes these variables as an input and returns the profitability. Therefore, it makes it easy to generalize the model to different clients.

The most important feature (fig 6.1) is the 3 months interest rate of Euro Interbank Offer Rate (EURIBOR). This result is consistent with the finding in S. Moro, et. al.[3] that is based on a bigger data set from the same banking institution. The second important feature is the Consumer Price Index (CPI) and the third one is the number of days that passed by after the client was last contacted from a previous campaign (pdays).

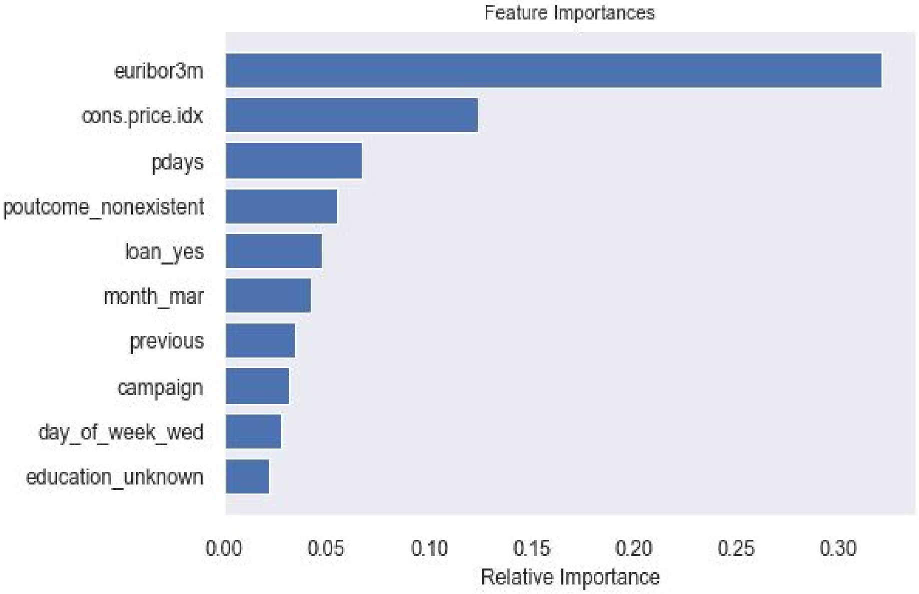


Fig 6.1
Feature importances based on Random Forest model

The data is taken during the economic recession between 2008-2010. As it is clear from the EDA sections, outcome by year and outcome by year and month, the subscription trend has dramatically changed during these years. It would be interesting to find the best model for each year from 2008 to 2010. Then analyze how the recession has affected the subscription. This will give us an insight about how another recession might affect the subscription trends.

[3] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.