Machine Learning For Profit Based Investing
Predictive Modelling (2024 P3A)

Koen van Esterik
kd.vanesterik@student.han.nl

7th of July 2025

1. Introduction

For every commercial company, there is an obligation to make a profit in order to make investments, achieve growth, and maintain continuity [1]. This responsibility is essential for the company towards employees as well as shareholders. For shareholders, this consists of value creation, thereby increasing the return on their investments.

This also applies to investment companies, which primarily invest in other businesses to achieve value creation and returns. The success of an investment company is determined by its ability to make the correct investment decisions in order to maximize shareholder profit.

In this context, the investment company BlackRock has initiated research into possibly acquiring a Portuguese bank. In theory, this investment could yield a good return for BlackRock's shareholders. In addition to regular financial activities, the Portuguese bank conducts telemarketing campaigns to sell financial products. BlackRock seeks to explore whether it can create shareholder value with its knowledge and expertise in optimizing these telemarketing campaigns.

The most extensive telemarketing operation consists of campaigns to sell subscriptions for bank deposits. The telemarketers use data from existing customers to make outbound calls. This yields mixed results because it is difficult to predict how a prospect will respond to a potentially intrusive phone call. Many of these calls are recorded in a dataset, indicating whether the prospect wants to subscribe to a bank deposit.

The recorded dataset serves as input for the research. The research focuses on a selection procedure for telemarketing prospects to be called. This procedure is based on a machine learning predictive model. This predictive model is trained on data from existing customers and then predicts, when given new data, whether a prospect wants to subscribe to a bank deposit. These predictions will answer

the question of how much the effectiveness of telemarketing campaigns can be improved.

The research can be summarized as:

- Increase conversion rates of telemarketing campaigns,
- By developing a selection procedure based on a predictive model,
- That performs better than calling all telemarketing prospects,
- In order to achieve profit maximization for the shareholders of BlackRock.

The predictive model is based on binary classification because we want to predict whether a prospect will subscribe to a bank deposit: yes or no. This information is available in the previously mentioned dataset and will serve to train and test the model.

Binary classification is a supervised learning method used to categorize data into one of two possible outcomes. When evaluating the performance of the binary classification model, the following terms [2] are generally used:

- True Positive (TP): The model correctly predicts a positive outcome.
- False Negative (FN): The model incorrectly predicts a negative outcome.
- False Positive (FP): The model incorrectly predicts a positive outcome.
- True Negative (TN): The model correctly predicts a negative outcome.
- Thresholds: The determination of where model predictions should be attributed in terms of TP, FN, FP, and TN.

Different standard metrics are calculated using these terms — e.g., accuracy, precision, recall, etc. These metrics are generally applied to evaluate a classification model. In this research, we propose a novel metric: a metric that calculates the maximum profit yielded by the model based on the previously mentioned terms (TP, FN, FP, TN, and thresholds). We call this metric the **Maximum Profit** (MP) metric.

First, we determine the predictions for each threshold based on the probabilities produced by the classification model:

$$\vec{y_{pred}} = \sum_{i=1}^{\vec{t}} \begin{cases} 1 & \text{if } \vec{y_{probs}} \geq t_i \\ 0 & \text{otherwise} \end{cases}$$

where:

- $\vec{y_{pred}}$, vector with predictions for all threshold values;
- $\vec{y_{probs}}$, vector with probabilities;
- $\vec{t}$, vector with all threshold values;

Next, we input the predictions into a confusion matrix to determine the true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) for all thresholds:

$$\vec{tps}, \vec{fns}, \vec{fps}, \vec{tns} = \sum_{i=1}^{\vec{t}} \text{confusion\_matrix}(\vec{y_{true}}, \vec{y_{pred}})$$

where:

- $\vec{y_{pred}}$, vector containing predictions for all threshold values;
- $\vec{y_{true}}$, vector containing actual values;
- $\vec{t}$, vector containing all threshold values;
- $\vec{tps}$, vector containing TPs for all threshold values;
- $\vec{fps}$, vector containing FNs for all threshold values;
- $\vec{fps}$, vector containing FPs for all threshold values;
- $\vec{fps}$, vector containing TNs for all threshold values;

Finally, we calculate and determine the maximum profit derived from all threshold profit values:

$$p = \max\left(r * \vec{tps} - c * (\vec{tps} + \vec{fps})\right)$$

where:

- p, scalar with the maximum profit;
- r, scalar with the profit per successful call;
- c, scalar with the cost per call;
- $\vec{tps}$, vector with TPs for all threshold values;
- $\vec{fps}$, vector with FPs for all threshold values;

We introduce the MP metric to create a connection between the technical aspect of the research and the business case. This is based on the idea that standard metrics often do not provide sufficient insights regarding the strategic decisions [3] that BlackRock typically has to make.

The MP metric will be instrumental in evaluating all possible classification models and will select a final model. The selected classification model will then be used to perform calculations and make a comparison with:

1. The current telemarketing procedure where all prospects are called.
2. The proposed procedure where the classification model makes a preselection of prospects.

This analysis will provide BlackRock with an answer to whether the potential for improvement in telemarketing campaigns is sufficient to justify BlackRock's investment in the Portuguese bank from a shareholder's perspective.

2. Methodology

This research can be conducted with any predictive data analysis tool setup. We have used the following system configuration:

- Python 3.11
- PDM
- Jupyter
- Pandas
- NumPy
- Scikit-Learn

The description in the repository [4] for this research indicates how to set up this study. This ensures that the research itself and the results can be validated.

2.1. Dataset

The dataset [5] for the research is described as follows:

- Time series data
- 41,000+ instances
- 20 features
    - 10 numerical
    - 10 categorical
- Target with binary values: yes/no
- No missing values

The data were collected from May 2008 to November 2010. The list below shows the structure of the dataset:

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
10   duration        41188 non-null  int64
11   campaign        41188 non-null  int64
12   pdays           41188 non-null  int64
13   previous        41188 non-null  int64
14   poutcome        41188 non-null  object
15   emp.var.rate    41188 non-null  float64
16   cons.price.idx  41188 non-null  float64
17   cons.conf.idx   41188 non-null  float64
18   euribor3m       41188 non-null  float64
19   nr.employed     41188 non-null  float64
20   y               41188 non-null  object
```

This overview serves as a reference for various descriptions in this document.

2.2. Data Cleaning

Apart from the fact that there are no missing values in the dataset, another task is required for data cleaning. The dataset description [5] indicates that the feature *duration* must be removed. This is to prevent questionable predictions.

In our opinion, this is related to possible data leakage during the training of the predictive model. The idea is that a prediction of a call cannot be made in advance if the *duration* of the same call has not yet been recorded.

Therefore, this *duration* feature must be removed.

2.3. Feature Engineering

The dataset is structured as a time series. However, a concrete timestamp is missing from the data. By using the period determination from the data description — along with the values in the *month* feature, the year for each instance can be determined. This engineered *year* feature will produce multiple insights in terms of data analysis.

2.4. Preprocessing

A number of preprocessing tasks need to be performed before model training can take place.

1. Transform the target feature from *yes/no* to numerical values. [6]
2. Convert all categorical values to numerical values. [7]
3. Scale all numerical values to a range from zero to one. [8]

These steps are important for developing a well-performing model.

2.5 Train Test Split

A regular train-test split [9] should be executed:

- shuffle
- no stratify
- 20% test size

Even though the dataset is structured as time series data, we believe using some sort of rolling window approach would produce a suboptimal predictive model. This is because the model should discover patterns in terms of prospect profiles that can be determined independently of each other. Therefore, we propose a regular split.

2.5. Model Shortlist

The models eligible for evaluation need to meet the following conditions:

- The model must have a probability function (to be able to use the MP metric).
- The model must not take longer than one minute to train a batch of data.

The models that meet these conditions are:

- AdaBoost [10]
- Gradient Boosting [11]
- K-Nearest Neighbors [12]
- Logistic Regression [13]
- Random Forest [14]
- XGBoost [15]

There may be more models that satisfy the above conditions, but for this research, we limit ourselves to this shortlist.

2.6. Procedure

The procedure for conducting the research consists of the following steps:

1. Load the dataset.
2. Apply the described data cleaning.
3. Apply the described transformations.
4. Split the dataset into a training set and a testing set.
5. Hyper-tune all models based on the MP metric as a score.
6. Select the best-performing parameters for each model.
7. Model all models through cross-validation.
8. Predict the probabilities of all models through cross-validation.
9. Calculate the maximum profit for all models.
10. Evaluate the maximum profit for all models.
11. Select the model with the highest maximum profit.
12. Use the predictions of the selected model to compare the current telemarketing procedure with the proposed telemarketing procedure.

Perhaps you have now earned yourself a cup of coffee.

3. Exploratory Data Analysis

Before we conducted the research procedure described, we also examined the dataset itself. During this examination, we discovered a number of peculiarities that might be important for the performance of the final predictive model.

3.1. Data Imbalance

The created *year* feature described earlier allows for the data to be categorized by year. This categorization reveals the following two noteworthy points.
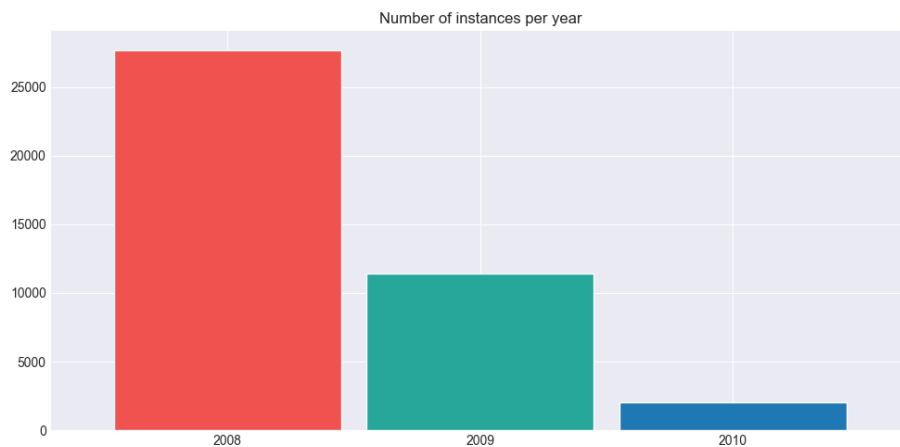

Figure 1: Number of instances per year

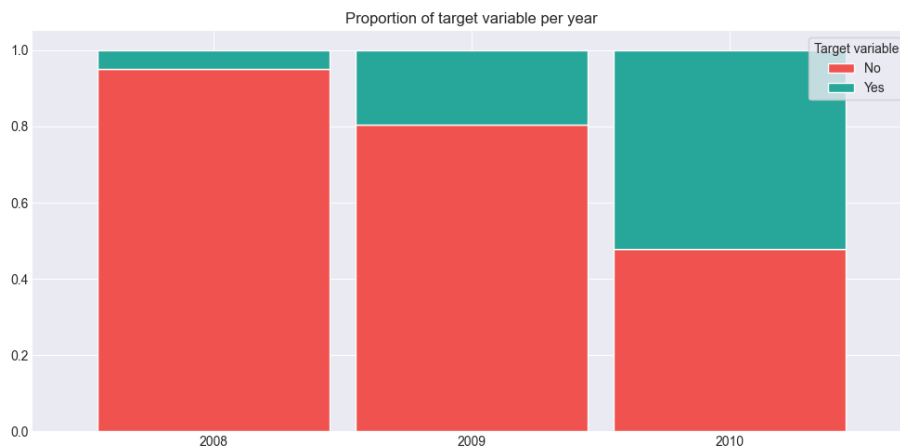The bulk of the data is concentrated in the year 2008, as shown in Figure 1.

Figure 2: Proportion of target variable per year

Additionally, the distribution of the target variable varies significantly by year, as shown in Figure 2.

We do not know why this imbalance is present in the data. One speculation is that there was a financial crisis in 2008, and therefore, there were more negative

responses that year than in other years to the questions of the telemarketers. However, we cannot verify this as we are not in contact with the owners of the dataset.

3.2. Approached vs Not-Approached

The *pdays* feature indicates the number of days since the last contact within the ongoing telemarketing campaign, according to the data description [5]. The value *999* is an exception to this rule, as it indicates that the prospect has not yet been contacted.

This mix of numerical and categorical values within a single feature can pose a problem when training a predictive model, as the model cannot make this contextual distinction.

Figure 3 shows that the distribution of approached versus non-approached prospects in the year 2008 is marginal.
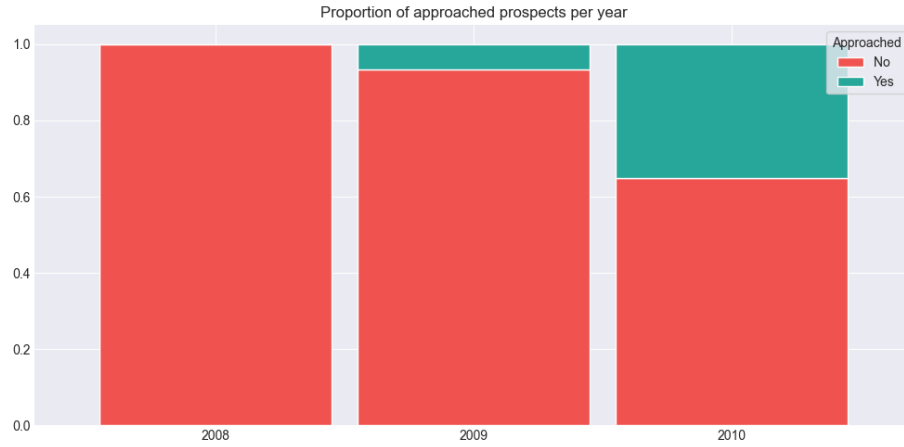

Figure 3: Proportion of approached prospects per year

Moreover, instances classified as *not-approached* do contain information suggesting that the prospect has indeed been contacted. This raises the suspicion that there are errors in the dataset. Unfortunately, this cannot be verified.

4. Results

We used cross-validation to initially compare the shortlist of classification models with one another. This cross-validation utilized the training set and also provided input for the proposed MP metric. The metric requires a ratio of revenue to cost, which we derived from the training set. This is because we want to

perform cross-validation based on a realistic proportion that the training set could provide.

Setting	Value
Cost Per Call	100.00
Revenue Per Success	400.00

The above ratios yielded the following results:

Model	Optimal Threshold	Profit	Profit Margin
AdaBoost	0.25	328,300	43.02%
Gradient Boosting	0.24	354,100	45.89%
K-Nearest Neighbors	0.21	248,300	42.06%
Logistic Regression	0.23	351,500	45.16%
Random Forest	0.2	**362,500**	**42.73%**
XGBoost	0.24	203,800	22.53%

The MP plots in Figure 4 should be interpreted as follows: at a threshold of 0, all prospects are called, whereas at a threshold of 1, no prospects are called.

With these results, it is clear that the Random Forest model yielded the most profit. This is achieved with the following hyperparameters:

```
{
  'max_depth': 10,
  'min_samples_leaf': 3,
  'min_samples_split': 10,
  'n_estimators': 100
}
```

These hyperparameters produced the profit curve plot shown in Figure 5.

It is important that the probabilities predicted by the model are correctly calibrated. This is because the MP metric uses these probabilities to calculate the profit per threshold. As demonstrated in Figure 6, the Random Forest model is reasonably calibrated. This means that the predicted probabilities correspond to the actual probabilities of the positive class.

We also want to evaluate whether the selected model is underfitting or overfitting. We do this using the same MP metric but based on normalized values calculated by the metric. This is because absolute profit figures are not suitable for calculating a learning curve, as these values are not in the same unit or scale, which prevents a fair comparison. Figure 5 illustrates the learning curves of the selected model.
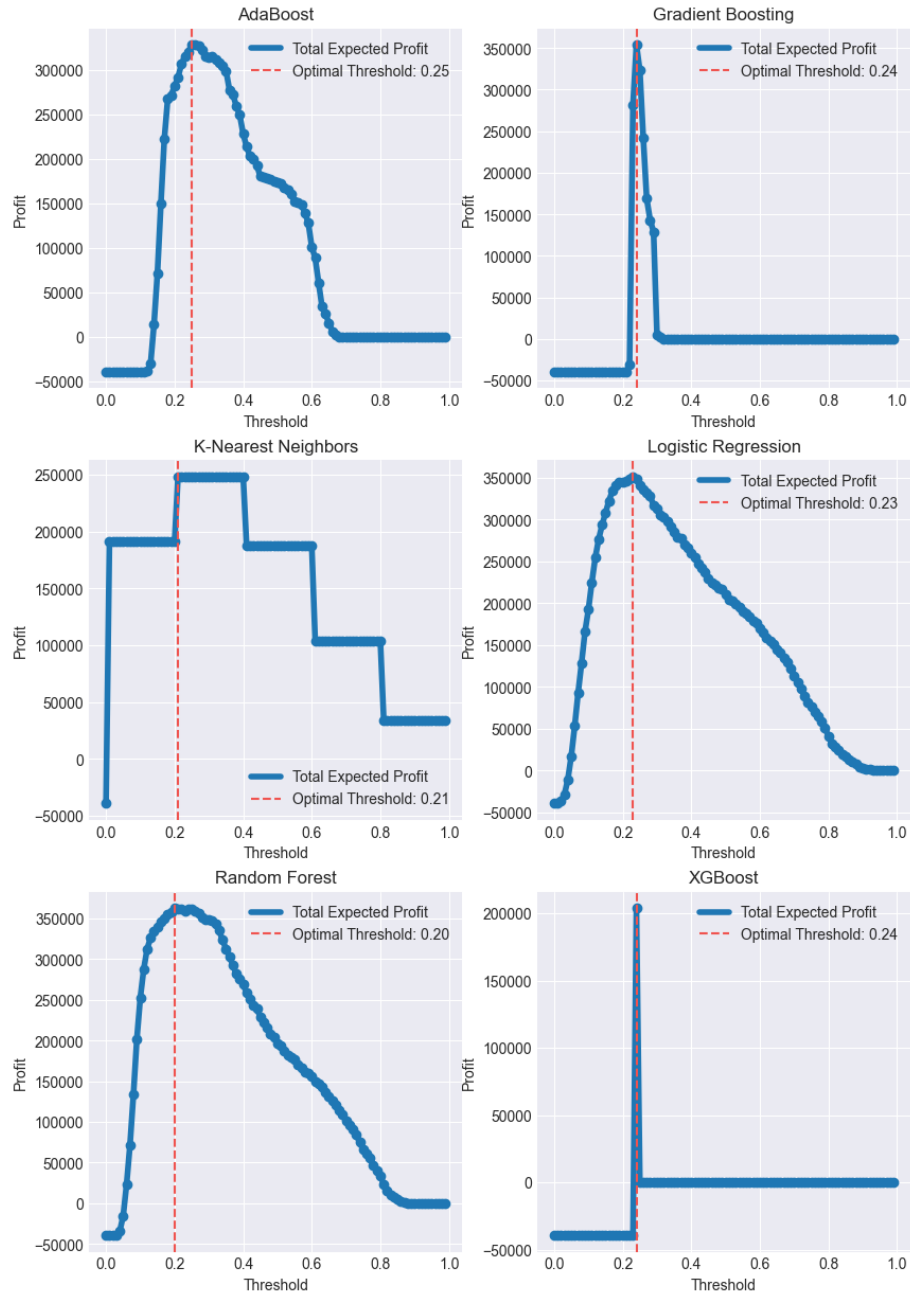
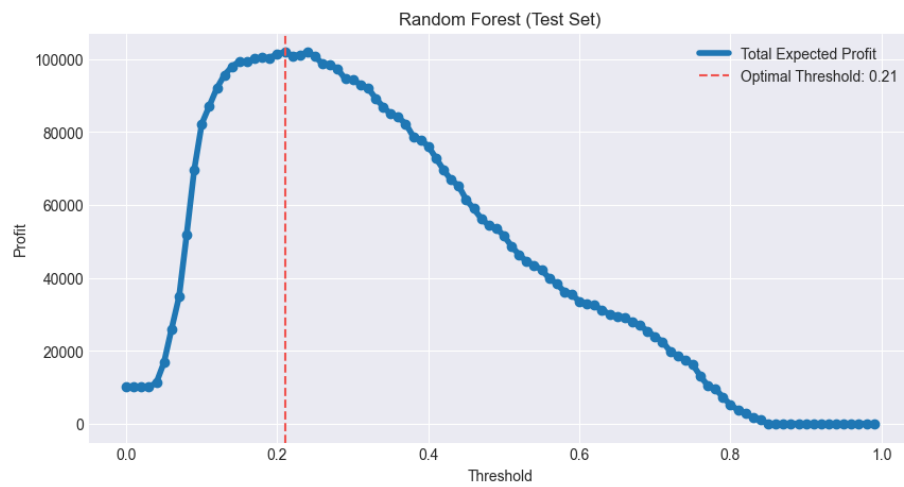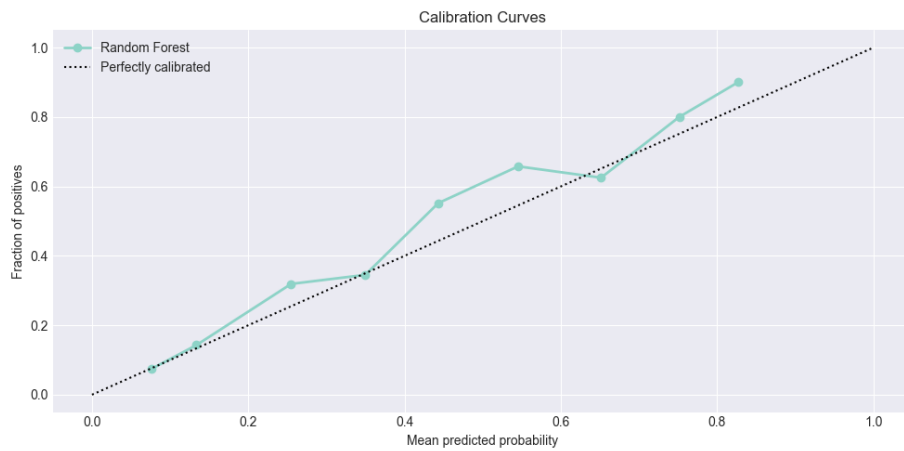Figure 4: Model Selection

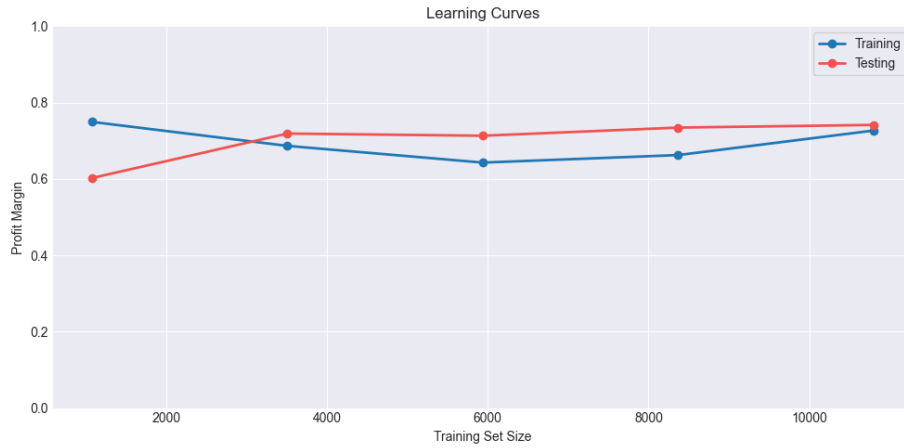Figure 5: Model Evaluation

Figure 6: Model Calibration

Figure 7: Learning Curves

A note regarding the learning curves plot is that the MP metric is suggesting a slightly underfitted model. This could be due to the novelty of the MP metric. This should be adjusted in follow-up research, to fully calibrate the metric when used as a scoring function.

Done a lot ... perhaps it's time for a little dance.

5. Discussion

During the investigation, several findings were made that may require further research.

- The dataset is relatively heavily concentrated in the year 2008 in terms of instances.
- The dataset is imbalanced with respect to the target feature, particularly in the year 2008.
- The *pdays* feature has a mix of numerical and categorical values.
- The dataset is relatively old, considering this research takes place in 2025.

Contact with the owners of the dataset is required, in order to resolve the described issues in potential follow-up research. This to gain more domain knowledge and, based on that, make more informed decisions regarding its implementation.

6. Conclusion

In conclusion, we analyze the effectiveness of the current and proposed telemarketing procedures and evaluate their impact on profitability.

The revenue-cost ratio for the calculations is as follows:

Setting	Value
Cost Per Call	100.00
Revenue Per Success	400.00

Which gives the following results:

Procedure	Profit
Call All Prospects	10,000
Call Preselected Propspects	102,000

When evaluating these results, we can conclude that there are significant opportunities for optimization within the current telemarketing procedure. This leads us to the conclusion that BlackRock's investment in the Portuguese bank is justified from the shareholders' perspective.

As an old Dutch expression goes: "Gas op die lolly!"

References

[1]   "*Principles of Corporate Finance*," *Wikipedia*. Feb. 04, 2024. Accessed: Jun. 23, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Principles_of_Corporate_Finance&oldid=1203323000

[2]   "Evaluation of binary classifiers," *Wikipedia*. May 25, 2025. Accessed: Jul. 06, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Evaluation_of_binary_classifiers&oldid=1292184015

[3]   "Data Science for Business: What you need to know about Data Mining and Data Analytic Thinking." Accessed: Jul. 05, 2025. [Online]. Available: https://fosterprovost.com/publication/data-science-for-business-what-you-need-to-know-about-data-mining-and-data-analytic-thinking/

[4]   K. van Esterik, "Vanesterik/mads-telemarketing-assignment." Jul. 04, 2025. Accessed: Jul. 05, 2025. [Online]. Available: https://github.com/vanesterik/mads-telemarketing-assignment

[5]   S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, Jun. 2014, doi: 10.1016/j.dss.2014.03.001.

[6]   "LabelEncoder." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[7]   "OneHotEncoder." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[8] "MinMaxScaler." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[9] "Train_test_split." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.model_selection.train_test_split.html

[10] "AdaBoostClassifier." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[11] "GradientBoostingClassifier." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[12] "KNeighborsClassifier." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[13] "LogisticRegression." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[14] "RandomForestClassifier." Accessed: Jul. 06, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[15] "XGBClassifier." Accessed: Jul. 06, 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/xgbclassifier/