Applying Data Science Methods to Analyze the Success of a Bank Marketing Campaign

Sumeet Ranjan
Dept. of Mechanical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
190040120@iitb.ac.in

Kunind Sahu
Dept. of Metallurgical Engg. & Materials Sci.
Indian Institute of Technology Bombay
Mumbai, Maharashtra
kunind@iitb.ac.in

Ritwik Gupta
Dept. of Mechanical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
ritwikgupta@iitb.ac.in

***Abstract*—In recent times, banks have been facing countless challenges, especially when loan activities are concerned. To be able to provide borrowers with the funds needed, banks must compete for deposits. Banks also face ongoing pressure by shareholders, both public and private, to achieve earnings and growth projections. Regulators place added pressure on banks to manage the various categories of risk [1]. It is imperative that banks rework their marketing strategy and base it on data backed facts. Banking is a very constrained industry where marketing the schemes properly is of utmost importance. In the following paper we've analysed the marketing campaign of Banco de Portugal during and after the Great Recession to increase client subscription to term deposits and have compared the performances of various Machine Learning Models to accurately predict the success of the campaign for a particular client.**

I. Introduction

Marketing is one of the primary components of business management and commerce [2]. Marketing is very important because it helps businesses distinguish themselves in a Monopolistic Competition. Firms must advertise their brands for consumers to recognize it. If they succeed, they will earn positive (and more) profits [3]. A subset of Classical Marketing is Direct Marketing. Direct Marketing is a type of marketing where we pick out intended target clients out of a large population of potential clients such that they meet a certain criteria. There are many aspects of Direct Marketing such as Telemarketing, Mobile Phone Marketing and E-Mail Marketing etc. However there are many challenges faced by direct marketing especially being labelled as spam or junk [4], leading to low response rates. How do banks come into play here? The most significant method via which banks make money is charging interest on the capital it lends out to customers. Banks makes money by charging a higher interest rate when it lends the depositors' money to borrowers than the interest rate offered to the depositors [5].

However, banking is a very competitive industry due to the entry of various entities such as fintech startups, insurance agencies, credit card companies and cryptocurrencies. Banks also need to compete with each other for consumer deposits. So, banks need to use their scarce resources, especially their time, in marketing their various schemes to the right sample of the population, to improve their already low success rate, which, essentially is, the definition of Direct Marketing. The computational problem of selecting a set of favourable clients to target during a marketing campaign is strongly NP-Hard and it is highly unlikely that a constant factor approximation algorithm can be proposed to solve this problem [6]. Hence, we have turned to Data Science to approach this problem to determine the drivers of success of the campaign for a particular client and use those drivers to determine the success of the campaign for a potential client The main idea behind the analysis is to explore the data and use Descriptive Analysis & Hypothesis Testing techniques to determine the features which are correlated with the success of the advertisement campaign

II. Related Work

- Chaitra Hegde, Aakash Kaku and Neelang Parghi first try to tackle this problem by applying Data Science methods to the dataset. They applied Exploratory Data Analysis Techniques to the dataset to fill up the missing values to the best of their ability to simulate the real world. They also used machine learning models such as Logistic Regression, Decision Trees, Random Forest Classifier & Adaptive Boosting (AdaBoost) to predict which clients will subscribe to a term deposit. Instead of finding out trends during an Exploratory/ Descriptive Analysis of the data, they used AdaBoost as a black box to determine the importance of each feature in predicting the success of the campaign. Also, instead of F1 Score which is often and the most widely used evaluation metric for imbalanced classes, they used AUROC as an evaluation metric for their models [7] and did not attempt to explore the avenue of synthetically correcting the class imbalance present in the data. The achieved the best mean AUROC Score of 0.8036 on their test data. Their ingenious way of feature engineering to simulate real world as closely as possible will be implemented in this project.
- The work of Sandy Wu, Andy Hsu, Wei-Zhu Chen, Samantha Chien has taken a similar approach in tackling the problem, they have done a basic Exploratory Data Analysis and data cleaning and spent much more time in the Predictive Analysis pipeline of the problem. In contrast with the previous example, this paper has used the SMOTE algorithm to synthetically correct the severe class imbalance [8] present in the dataset and trained the

machine learning models such as Naïve Bayes, Logistic Regression, Decision Tree and Random Forest on both, the imbalanced and the SMOTE corrected dataset, though they still used the AUROC metric to evaluate their models on both, the imbalanced dataset and the SMOTE corrected dataset [9]. Their usage of oversampling the positive examples to develop a more robust model will be used in this project.

- Moro et al. attempted to solve a significantly bigger problem as compared to the previous two works on this topic and analyzed a large set of features (= 150) and implemented a semi-automatic feature selection to select a reduced set of 22 features. They trained the data on 4 Machine Learning Models namely Logistic Regression, Decision Trees, Neural Networks and Support Vector Machines. They evaluated model performance using two metrics - AUROC (Area under the Receiver Operating Characteristic Curve) & ALIFT (Area under the LIFT curve). For them, the Neural Network model showed the best results out of the 4 models and had AUC = 0.8 & ALIFT = 0.7 which allowed them to reach out to 79% of the potential clients. They used two knowledge extraction methods namely Sensitivity Analysis and Decision Trees to extract the list of relative importance of every feature and to establish the credibility of their models. However, in my opinion their approach as a whole tended to rely on black box approaches where the reader is unable to understand the nuances of how the analysis came to such a conclusion, and also their approach did not explore any potential oversampling techniques to deal with the class imbalance present in the data and also did away with any kind of Descriptive / Exploratory Data Analysis in favour of the semi-automatic feature selection tool to maintain the focus of the analysis on the Machine Learning Models themselves [10].

III. Dataset & Feature Engineering

The dataset contains data collected by a Portuguese retail bank consisting of 41188 instances each described by 20 attributes. The dataset is highly imbalanced as only 4640 (11.3%) of the records being attributed with success. Each record included client personal information (e.g., age, education), the client's relationship with banks in the past (e.g., loan, default), whether a campaign had previously made contact with them (e.g., days since last contact), and records of various social and economic context attributes (e.g., consumer confidence index) when the client was contacted [10], [11].

A thorough exploratory and descriptive data analysis supported by statistical hypothesis testing successfully reduced the number of features to be used in the model down to 14. Several categorical features like the ones with values stored as 'unknown' have been filled with a random probability weighted choice of other meaningful categories in that feature. Also, for some highly correlated features, such as *job, education*, the unknown values for *job* have been filled by first checking out the *education* and finding out the most suitable job for that education category using a crosstab to determine relative probabilities of jobs for each education categories and the same has been done for the feature *education*. numerical variables (such as *pdays*) with very high variances have been suitably simplified, binned (to convert them into categorical variables) and framed to reduce both the complexity of our model and prevent overfitting on the training set. The models have been trained twice, once with the original dataset (D1) and a second time with a modified version of our dataset (D2), where, to counter the class imbalance in the target variable, we've implemented randomised oversampling by duplicating the records with target variable equal to the minority variable attributed to success (i.e., y=1 indicating clients have subscribed to a term deposit account) 7 times as the imbalance was in the ratio of roughly 1:8.

IV. Computational Environment & Data Mining Models

A. Goal Definition

1) Data Analysis and Interpretation: The goal of our Data Analysis is split into 3 parts: Exploratory Data Analysis, Descriptive Data Analysis and Hypothesis Testing.

Exploratory Data Analysis: We take stock of our data, find out the missing values, determine which variables are categorical, which variables are numerical; are there any null values in the data, how is the distribution of each numerical feature, is there any correlation between numerical variables and what is the relative composition of each categorical variable.

Descriptive Data Analysis: We implemented sophisticated data processing techniques using the `pandas` library of `Python` to make stronger conclusions about our data. We tried to get to know: how each variable was linked to our desired output, what trends we could see between each variable and the output, for categorical features if any category of a feature stood out in affecting the target variable. We tried to analyse and isolate as to when and whom should the bank's marketing campaign target to extract the most out of their campaign.

Hypothesis Testing: This part of the data analysis pipeline was done to be statistically sure of the conclusions we made during the Descriptive Analysis of our data. We used a χ^2 Contingency Tables Test and a Cramer's V Test to determine any correlation between categorical columns and the target variable which is also categorical.

2) Data Mining Model Implementation: **Precision** and **recall** are two important metrics on the basis of which we intend to evaluate our models. In this context, precision of the model will provide us with the fraction of people who would respond positively to our campaign, from the set of people that the model has considered to respond favourably. Recall, on the other hand, will give us the fraction of people our model would have missed had it only considered the "favourable set" of people.

$$Precision\,(P) = \frac{TP}{TP + FP} \tag{1}$$

$$Recall\,(R) = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ Score = \frac{2PR}{P + R} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$False\ Positive\ Rate\,(FPR) = \frac{FP}{FP + TN} \tag{5}$$

Where TP is the number of True Positives (no. of positives predicted correctly), TN is the number of True Negatives. The importance of either of these metrics could not be understated in this scenario, so we have used an **F1** score, which provides a balance between both precision and recall, along with Accuracy. Another benefit of the F1 score is that it is a better measure to use when there are a large number of True Negatives (TN) in our sample, which is exactly the case.

We've trained our model on two types of datasets- our original dataset **(D1)** and the oversampled version of our original dataset **(D2)**. Due to the severe class imbalance in **D1**, we've used F1 score as an evaluation metric over accuracy and AUROC.

AUROC is the Area under the Receiver Operating Characteristic Curve. ROC is the plot of FPR vs R at different thresholds. AUROC and Accuracy are very poor evaluation metrics when imbalanced classes are considered. AUROC is unaffected by the skew in the data and it is suggested that AUROC tends to mask poor performing models & gives us a specious value which is due to high False Positives rather than True Positives [12].

Accuracy is not suitable because the impact of the least represented, but more important examples, is reduced when compared to that of the majority class. For instance, if we consider a problem where only 1% of the examples belong to the minority class, an high accuracy of 99% is achievable by predicting the majority class for all examples. Yet, all minority class examples, the rare and more interesting cases for the user, are misclassified. This is worthless when the goal is the identification of the rare cases [13].

For our Dataset D2, we have included Accuracy and F1 Score as the model evaluation metrics. We've yet again excluded AUROC because it is fundamentally incoherent in terms of misclassification costs: the AUC uses different misclassification cost distributions for different models, basically, using AUC is equivalent to using different metrics to evaluate different classification rules [14], and since we would like to compare different models on their results, we reject AUROC as an evaluation metric.

Hence, finally, in any case, the goal of our models is to maximise the F1 Score and the Accuracy of Classification.

B. Computational Environment

In this project we have used four binary classification models as implemented in the `scikit-learn` library in `Python`:

- Logistic Classifier
- Neural Network Classifier
- Random Forests Classifier
- Support Vector Machine (SVM) Classifier

The entire dataset has been normalised before proceeding with these models.

C. Data Mining Models

Logistic Regression is probably the most commonly used supervised learning classification method and the results obtained here from the Logistic Classifier have been used as a benchmark for more complex models. We have used the *saga* solver, an extension of the stochastic average gradient descent approach, which uses a random sample of previous gradient values and allows for L1 regularization. A 4-fold cross-validation procedure has been used to improve the estimated performance of the machine learning model and reduce the standard error associated with the results.

For the **Neural Network Classifier**, we have implemented the multilayer perceptron architecture with a single hidden layer (H). This H is a hyperparameter which decides the complexity of the architecture. We have used *Adam* optimising algorithm combines the advantages of two extensions of the classic stochastic gradient descent - RMSPRop and AdaGrad. The biggest advantage of Adam over other conventional optimisation techniques is that it converges faster by countering the oscillation problem faced by other optimisation techniques during gradient descent, by using adaptive learning rate and "momentum", a weighted average of the previous updates to the weights, while updating during backpropagation.

The **Random Forest Classifier** consists of an ensemble of decision trees (DT) where each DT is a branching structure that represents a set of rules, distinguishing values in a hierarchical form. Depth of each DT and the number of such DTs to be used in our ensemble were the hyperparameters which were selected after running a gridsearch on the training set for each ensemble,

The **Support Vector Machine Classifier** which uses a Kernel trick was also used in the hope that the F1 Score could further be improved. The SVM transforms the input into a higher dimensional vector space using the kernel trick and attempts to find the maximum margin decision boundary which separates the positive and the negative examples, the hyperparameters tuned for the SVM were the Regularisation Parameter (C) and the nature of the Kernel Function. SVMs are prone to overfitting due to large dimension of the parameters after the Kernel transform and do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (in the `scikit-learn` library).

V. Evaluation

A. Data Analysis and Interpretation

1) Descriptive Data Analysis: Preliminary exploratory and descriptive data analysis helps us in dropping redundant features. It was found that a few features such as *default* ,*day-of-week*, *default*, *education*, *duration*, *loan*, *marital* etc don't give any meaningful information, and removing them from

being fed to our models might help us prevent overfitting on the training set. Several meaningful correlations of favourable outcomes with respect to age, job, education, mode of communication and attributes related to the previous campaign were found. It would be more efficient to target clients younger than 32 or older than 56 years of age; high-school educated or higher. Retirees and students responded very favorably, and people with white-collar jobs were more likely than those with blue-collar jobs to subscribe. People who were followed up with a few times, and who had been contacted in an earlier campaign responded favourably too.

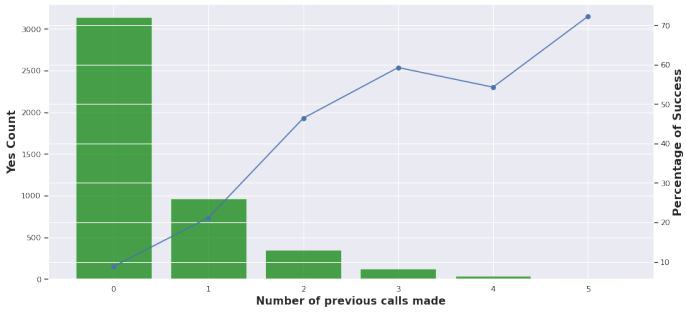

Fig. 1. Plot of percentage possibility of success against the number of calls

2) Hypothesis Testing: **Cramer's V** test tells us that a few features like *pdays* (days since last contact), *month* of contact and *campaign* (number of times the client was contacted) are the most important, and others like *marital* and *education*, the least. The χ^2 **Contingency test** results agree with our earlier analysis that the variables *loan* and *housing* are not important. We have

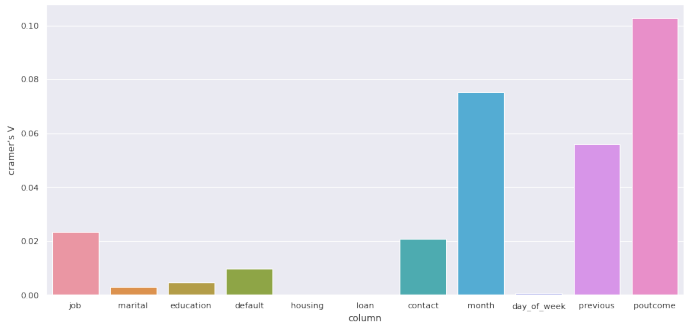

Fig. 2. Cramer's V test showing importance of each feature

B. Data Mining Model Implementation

We have dropped six features going into the model implementations which are *marital*, *education*, *default*, *loan*, *duration* and *day-of-week*.

1) Hyperparameter Tuning:: For **original dataset D1**:

- Logistic Regression: C = 10
- Neural Network: Hidden Layer Size = (75, 75) & Learning Rate (α) = 0.05
- Random Forest Classifier: Number of Trees = 150 , Maximum Depth of Each Tree = 19
- Linear SVM : C = 5

For **Minority Class Oversampled Dataset D2**:

- Logistic Regression: C = 0.02
- Neural Network: Hidden Layer Size = (125, 125) & Learning Rate (α) = 0.05
- Random Forest Classifier: Number of Trees = 50 , Maximum Depth of Each Tree = 19

2) Test Data Results: For dataset **D1**, the classifier models underwent Hyperparameter Tuning (evaluated on the F1 Score) and were subsequently tested on the test dataset. The accuracy, F1 score, precision and recall for each model on the test data has been shown in Table 1.

TABLE I
UNBALANCED DATASET

Model	Evaluation Metrics			
	F1 Score	**Accuracy**	**Recall**	**Precision**
Logistic Regression	0.328	89.76%	0.218	0.662
Neural Network	0.374	89.78%	0.267	0.627
Random Forest	0.396	89.55%	0.299	0.586
Support Vector Machine	0.077	83.90%	0.059	0.113

For dataset **D2**, we dropped the SVM classifier because of extremely poor performance. The accuracy, F1 score, precision and recall for each model on the test data has been shown in Table 2.

TABLE II
BALANCED DATASET

Model	Evaluation Metrics			
	F1 Score	**Accuracy**	**Recall**	**Precision**
Logistic Regression	0.712	74.27%	0.631	0.816
Neural Network	0.742	76.58%	0.672	0.830
Random Forest	0.894	89.41%	0.891	0.898

VI. DISCUSSIONS

From the results above, we can make some really insightful conclusions:

- The marketing team should try to target retirees and clients aged 56 and older, since our results have shown that they are very likely to subscribe to a term deposit account.
- Students who have completed at least high-school level of education, aged 26 or less respond favourably.
- People working blue-collar jobs are less likely to subscribe and this demographic is more likely to have completed high-school or university degree education, which is all the more reason to avoid them.
- No more than 3 calls should be made during the campaign as it has been shown that after 3 calls during the campaign, there is a sharp decrease in the mean subscription rate which goes even below 10%

- Those clients who had previously subscribed to a term deposit account must be targeted because they are much more likely to subscribe to a term deposit account again.
- Following up and calling people who have subscribed earlier helps. Calling with a 1 day gap gives a mean success of 30% and calling after 2+ days gives us a mean success rate of approximately 60+%, but since it is a bin it is recommended to call after a gap of 2 days.
- For most of the banks which are engaged in a constant competition the model with the highest F1 Score will be better suited to test if a future targeted client would subscribe to a term deposit account or not. For the models we have tested so far, the Random Forest Classifier Model for both Datasets **D1** & **D2** gave us the best and highest F1 Scores and Accuracy.
- The SVM performed the worst of all models and it was nowhere near the performance of any of the other three classifiers. That is why we decided to exclude it while training **D2**.

VII. Future Work

For a future analysis on a similar subject, it would be beneficial to include more features about the personal data of the client, though including a lot of personal data would not be a problem for analysis or the model, but may rather pose several ethical questions and may also lead to an increase in divide between the rich and the poor clients.

Also an inclusion of much smoother numerical data and some more features related to the economics and general well being of the populace which will help us make more informed analysis and prediction of the success of our campaign.

Additonally, some other sophisticated methods to deal with the class imbalance present in the data can be tried out.

References

[1] Mishler, Lon; Cole, Robert E. (1995). Consumer and Business Credit Management. Homewood: Irwin. pp. 128–29. ISBN 978-0-256-13948-8.

[2] Drucker, Peter (1954). The Practice of Management. New York: Harper & Row. p. 32.

[3] Kulihala, Hisashi,"Advertising Costs, Monopolistic Competition, and International Trade". Economic Journal of Hokkaido University, Vol. 29, pp. 86

[4] What is a "Whitelist" and why do I want to work with a "Whitelisted"

[5] Bowman, Cynthia. How Do Banks Make Money?

[6] Fabrice Talla Nobibon, Roel Leus, Frits C.R Spieksma, "Optimization models for targeted offers in direct marketing: exact and heuristic algorithms". European Journal of Operational Research, vol. 210(3), 2011, pp. 670–683

[7] Chaitra Hegde, Aakash Kaku, Neelang Parghi. Predicting Bank Marketing Campaign Success using Machine Learning, 2017.

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Vol. 16 (2002), pp. 321 – 357.

[9] Sandy Wu, Andy Hsu, Wei-Zhu Chen, Samantha Chien. Predicting Customer Purchase to Improve Bank Marketing Effectiveness.

[10] [Moro et al., 2014] 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

[11] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[12] L. A. Jeni, J. F. Cohn and F. De La Torre, "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics," 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, 2013, pp. 245-251, doi: 10.1109/ACII.2013.47.

[13] Paula Branco, Luis Torgo, Rita Ribeiro. A Survey of Predictive Modelling under Imbalanced Distributions. arXiv, 1505.01658, 2015.

[14] Hand, D.J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn 77, 103–123 (2009). https://doi.org/10.1007/s10994-009-5119-5

Appendix

The Python Notebook containing the entire code for the above analysis is present in **this** GitHub link. In this notebook, we've done a very detailed analysis of all the parameters along with explanations, conclusions and supporting graphs to help grasp a notion of the analysis.