Predicting the Success of Bank Telemarketing Using Machine Learning

Atheer Albarqi, asa387@drexel.edu
Emily Wang, ew552@drexel.edu
Xi Chen, xc98@drexel.edu
College of Computing and Informatics
Drexel University
Philadelphia, PA

***Abstract*—In this study, we examine telemarketing practices for promoting long-term bank deposits to potential bank customers. A number of pre-processing steps are applied to the dataset including log transformation, removal of correlated features, and oversampling of the minority class. Two machine learning algorithms, Decision Tree (DT) and Logistic Regression (LR), are employed to determine the subscription of long-term bank deposits. DT model yields an F1 score of 0.90 for the positive class and LR yields an F1 score 0f 0.73. DT is the better model for predicting the potential customers who have an interest in long-term deposits through telemarketing. In future marketing campaigns, the results of the machine learning model could be used by managers to prioritize and select which customers to contact next.**

***Index Terms*—Decision Tree, Logistic Regression, Resilient Distributed Dataset (RDD), Telemarketing.**

I. Introduction

Companies often conduct marketing campaigns to attract new customers and grow their business. Direct marketing is used by businesses to reach certain categories of customers to meet specific goals. One of the most widely used campaign channel is the telephone, which is called telemarketing. Notably, many financial services providers adopted telemarketing strategy to reach out to new customers and to provide better services to existing customers, and to meet their specific needs. This project focuses on telemarketing phone calls to sell long-term deposits. Within a campaign, the human agents execute phone calls to a list of clients to sell the deposit (outbound), or, if the client calls the contact center for any other reason, they are asked to subscribe to the deposit (inbound).

Technology can allow companies to rethink marketing through the analysis of data to build longer and tighter relations in line with company needs. Machine learning techniques can therefore be used to support managerial decision making.

In this project, we will develop classification algorithm that can automatically predict the results of a phone call to sell long term deposits by using a machine learning approach. This study is inspired by the research paper by Moro et al., where the researchers used a similar dataset to determine the success of bank telemarketing in selling bank long-term deposits [1]. Another inspiration is the project by Palaniappan et al., whose aim is to help banks increase their accuracy of customer profiling through classification as well as to identify a high probability group of customers to subscribe to a long-term deposit [2].

The results from the machine learning model would be valuable for managers in making decisions on prioritizing and selecting the next customers to be contacted during future marketing campaigns. For instance, the probability of success of the current campaign can help the business decide how many customers to contact in the next campaign, making the campaigns more effective and reducing the time and costs of such campaigns.

II. Dataset

The dataset is published by UCI Machine Learning Repository [3]. It contains real data related with direct marketing campaigns (phone calls) of a Portuguese banking institution from May 2008 to June 2013. Often, more than one contact to the same client was required to assess whether the product (bank term deposit) would be subscribed. There are 41,188 examples in the dataset and 21 attributes. The input variables include demographic information about the clients, information related to the last contact of the current campaign (how and when the phone call was made), and social and economic context attributes. The output variable is whether or not the client has subscribed to a term deposit.

III. Methodology

The machine learning task is a binary classification which predicts whether the client will subscribe to the term deposit. However, our mission in analyzing this dataset is to not only build an effective predictive machine learning model, but also uncover interesting patterns in clients and the features' relationship with the success of the campaign. Therefore, we will conduct exploratory data analysis to show the relationships between different attributes. Then, we will preprocess the data and prepare the feature for the ML models. Lastly, we will split the data to train and validate the models. The methodology of the analysis and preprocessing steps is shown in Fig. 1.

A. Data Preprocessing

The data will go through a number of pre-processing steps. First, from the exploratory data analysis, we see that some of the features (age, campaign, previous) are highly skewed

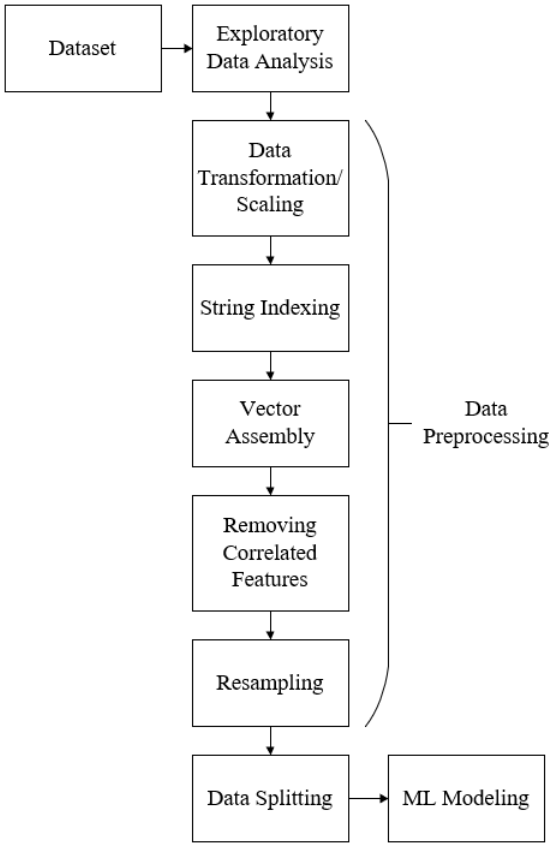

Fig. 1. Data processing flow overview.

to the right. Highly skewed data can make our statistical analysis invalid [4]. Log transformation will be applied to these features to make the data as "normal" as possible and, thus, increase the validity of the associated statistical analyses. In the string indexing stage, the remaining categorical variables such as "education", "marital", and "job" are converted to integer labels.

All features are then assembled into a vector in the vector assembly stage. From the feature vectors, the similarities/correlations between different features will be calculated. Classical machine learning models such as logistic regression or random forest become unstable in the presence of high feature correlations [5]. Therefore, the features with a correlation score of more than 0.85 are removed. The dataset is also highly imbalanced, with the target column containing 89% "no" and 11% "yes". In such cases, standard classifiers tend to be overwhelmed by the majority class and ignore the minority class, and the performance drops significantly [6]. To rectify this issue, we will over-sample the minority class by using SMOTE (Synthetic Minority Over-sampling TEchnique). The preprocessed data is then randomly split into 80-20 train-test distribution.

B. Machine Learning Modeling

In this work, we test two binary classification machine learning models as implemented in the PySpark ml or mlib packages: logistic regression (LR) and decision trees (DT).

LR is a benchmark model for classification. Logistic regression uses a logistic function to calculate the probability of an observation belonging to a particular class. LR is a popular choice (e.g., in credit scoring). Due to the linear combination of its independent variables (x), the model is straightforward to interpret. Yet, the model cannot model adequately complex nonlinear relationships [1].

DT is a branching structure that represents a set of rules in a hierarchical tree form. This representation can be translated into a set of if-then rules that are easy to understand. The advantages of a decision tree as compared to logistic regression include its ability to allow for training models on large datasets in addition to quantitative and qualitative input variables [7]. Several parameters can be adjusted in the DecisionTreeClassifer to improve performance and prevent overfitting. In Decision Trees, parameters are very crucial because overfitting is common. For instance, the max depth represents the maximum number of tree layers that will have, which in our case is 30. We find that increasing the max depth leads to better performance.

C. Tools

We use Apache Spark (PySpark) as our primary tool. Two machine learning libraries, mllib and ml, are provided in Spark. The operation of mllib is based on RDD (resilient distributed dataset), while ml is based on DataFrame, which is a mainstream machine learning library. The ml package includes three main abstract classes: Transformer, Estimator, and Pipeline.

Transformer classes transform data by appending a new column to the DataFrame. At a high level, when deriving from the Transformer abstract class, each new Transformer class needs to implement the .transform() method. This method requires passing a DataFrame to be transformed, which is usually the first and only mandatory parameter.

The pipeline in pyspark ML is used to represent the end-to-end process from transformation to evaluation (with a series of different stages), which can perform necessary data processing (transformation) on some input raw data (in the form of DataFrame), and finally Evaluate the model.

A pipeline can be thought of as consisting of a series of different stages. When the fit method is executed on a Pipeline object, all stages are executed in the order specified in the stage parameter; the stage parameter is a list of transformer and evaluator objects. The fit method of the pipeline object executes the transform method of each transformer and the fit method of all evaluators.

IV. Exploratory Data Analysis

There are many variables that determine whether a marketing campaign will be successful or not. Today, these variables are know as the marketing mix theory, which is also referred to as the 4 Ps, i.e. product, price, promotion "communication", and place "distribution" [8]. These marketing mix elements are the four key decision areas that must be managed to work

together in a single marketing plan to satisfy customer needs better than the competition and allow the firm to make a reasonable profit [9].

Not all four of the elements above apply the our dataset. Therefore, for the exploratory data analysis, we will focus on two elements from a modified version of the marketing mix, Population and Place. We will seek out to answer the following exploratory questions:

1) What is the demographic distribution among the clients (age, job, marital status, education)?
2) Based on this dataset, what type of customer will be more likely to acquire the product? This will help the company develop more targeted marketing campaigns in the future.
3) Which distribution channel is the most effective (telephone, cellular)?

To answer the motivating question (1), Table. I shows the demographic distribution among clients. The mean age is 40, and the minimum, median, and maximum ages are 17, 38, 98, respectively. The fact that the maximum age is almost three times the the median age shows that the distribution of age is heavily skewed to the right. Therefore, logarithmic transformation will need to be applied to the age column to normalize the distribution of the feature. From the distribution of the rest of the demographic variables, we can see that a lot of clients have administrative jobs, are married, and have a university degree.

TABLE I
CLIENT DEMOGRAPHIC DISTRIBUTION

Demographic Attribute	Type	Value
Age (years)	Mean	40
	Median	38
	Min	17
	Max	98
Occupation (%)	Admin	25.3
	Blue-collar	22.5
	Technician	16.4
	Services	9.6
	Management	7.1
	Retired	5.2
	Entrepreneur	3.5
	Self-employed	3.5
	Others	7.2
	Unknown	0.8
Marital Status (%)	Married	60.5
	Single	28.1
	Divorced	11.2
	Unknown	0.2
Education (%)	University Degree	29.5
	High School	23.1
	Basic 9-Year	14.7
	Professional Course	12.7
	Basic 6-Year or lower	15.7
	Unknown	4.2

To answer the motivating question (2), Table. II shows the top 5 demographics who have subscribed to term deposits. The mean age is 40, and the minimum, median, and maximum ages are 17, 38, 98, respectively. The fact that the maximum age is almost three times the the median age shows that the distribution of age is heavily skewed to the right. Therefore, logarithmic transformation will need to be applied to the age column to normalize the distribution of the feature. From the distribution of the rest of the demographic variables, we can see that a lot of clients have administrative jobs, are married, and have a university degree. Figure. 2 shows the age distribution among clients who did and did not subscribe to term deposits. Clients over the age of 60 seem to be more likely to subscribe to term deposits than younger clients.

TABLE II
DEMOGRAPHICS WITH THE HIGHEST SUBSCRIPTIONS

Demographics			Subscriptions
Job	**Marital Status**	**Education**	
Admin	Married	University Degree	386
Admin	Single	University Degree	360
Technician	Married	Professional Course	196
Admin	Married	High School	175
Management	Married	University Degree	169

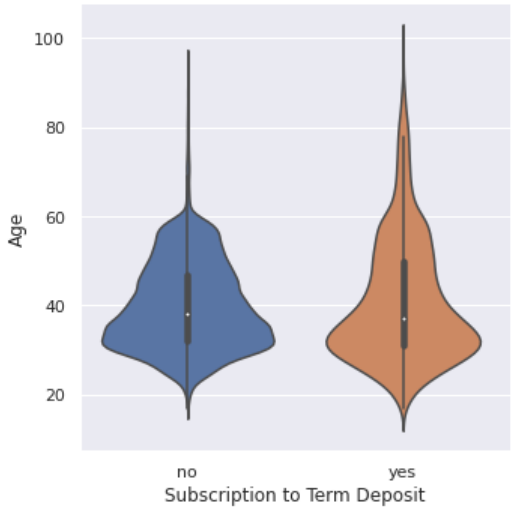

Fig. 2. Clients over the age of 60 are more likely to subscribe to term deposits than younger clients.

To answer the motivating question (3), Table. III shows the success rates of the two different contact methods. Cellular phone has a success rate of 14.7% whereas telephone has a success rate of 5.2%, which makes cellular phone the more effective channel of communication.

TABLE III
SUCCESS RATES OF TWO CHANNELS OF COMMUNICATION

Contact Method	Result	Count	Success Rate
Cellular	No	22291	14.7%
	Yes	3853	
Telephone	No	14257	5.2%
	Yes	787	

V. Experiments and Results

From the preprocessed dataset, we use machine learning models to predict the results of a phone call to sell long term deposits. The result is presented in Table. IV, which shows different evaluation metrics like precision, recall, F1 score, support is called classification report. Precision measures the percentage of classified positives that were actually positive. Recall measures the percentage of correctly identified true positives or not. F-measure is a weighted harmonic mean of precision and recall. Accuracy is the amount of correctly identified classifications over the total number of classifications. These five types of measurements are calculated to each classifier and compared to determined which classifier performs the best with the given dataset. The equations are shown below.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

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

The area under an ROC (Receiver Operating Characteristic) curve is a measure of the usefulness of a test in general, where a greater area means a more useful test. The areas under ROC curves are used to compare the usefulness of tests. In addition, we have used Logistic Regression Training Summary to provides a summary for a Logistic Regression Model and summarize the model over the training set. The area under ROC is shown in Fig. 3.

Fig. 4 visualizes the precision-recall (PR) curves to compare the performance of the models with Recall on the x-axis and Precision on the y-axis. We can clearly see here that DT is the better performing model.

While using a classification problem, we need to use metrics to check efficient machine learning models. Among the two models, DT yields better results; it has outperformed LR on every metric. Since the class distribution of the dataset was originally as imbalanced with 89% majority class and 11% minority class, we also saw that resampling has significantly improved the F1 score for our evaluation.

TABLE IV
Classification Model Performance on Train Dataset

Evolution Metrics	Logistic Regression	Decision Tree
Accuracy	0.75	0.90
Precision	0.74	0.89
Recall	0.72	0.90
F1-Score	0.73	0.90
Area Under ROC	0.79	0.94

The results of this study can inform banking institutions about the chance of any client subscribing to the term deposit.

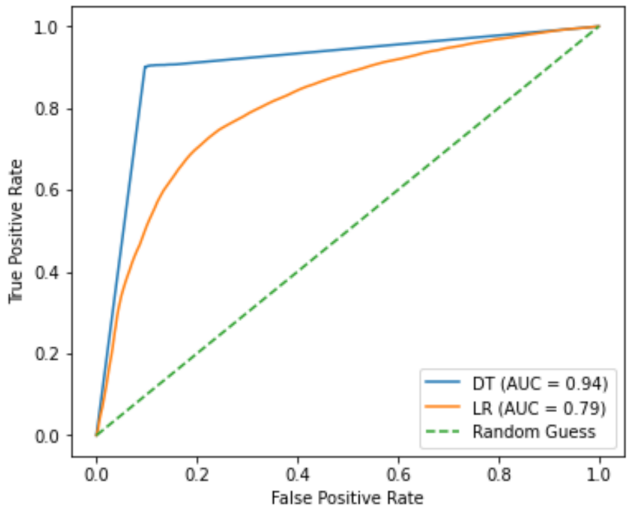

Fig. 3. ROC Curve for the models.

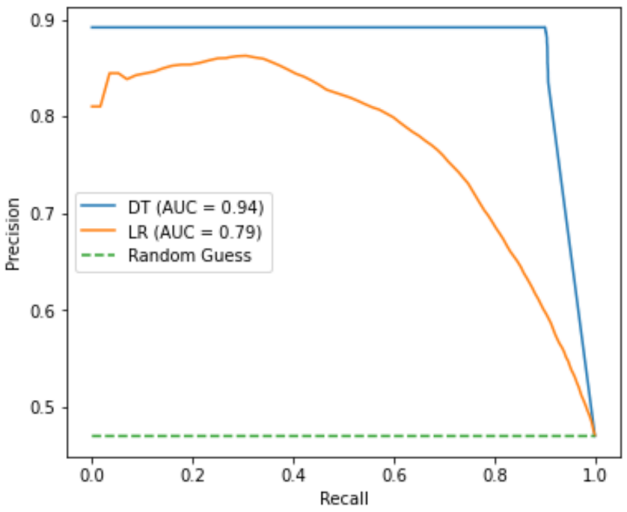

Fig. 4. Precision-Recall Curve for the models.

This ability can help them better target the clients and allocate resources appropriately. For example, a manager could choose to spend more resources on customers with high probabilities of subscribing to ensure the outcome and decrease the amount of resources spent on low-likelihood clients.

VI. Conclusion and Future Works

In this paper, we present two machine learning models, Logistic regression and decision Tree to predict automatically the results of a phone call to sell long-term deposits using PySpark. A number of pre-processing steps were applied to the dataset including log transformation, removal of correlated features, and oversampling of the minority class. Both models performed well (AUC greater than 0.7). The results confirm that the DT model provides an F1 score of 0.90 compared to the F1 score of 0.73 for LR, which makes DT the better model for predicting the potential customers who have an interest in long-term deposits through telemarketing. It may

be possible for service providers to use the results to develop more targeted telemarketing marketing campaigns that will ultimately increase banks' earning capacity.

For future work, the prediction accuracy of decision trees can be improved by Ensemble methods, such as Random Forest and Gradient-Boosted Tree. The performance of the logistic regression model can also be improved by one-hot encoding of the categorical labels.

REFERENCES

[1] S. Moro, P. Cortez and P. Rita, "A Data-Driven Approach to Predict the Success of Bank Telemarketing," Decision Support Systems, Elsevier, vol. 62, pp. 22-31, Jun. 2014

[2] S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, "Customer profiling using classification approach for Bank Telemarketing," JOIV : International Journal on Informatics Visualization, vol. 1, no. 4-2, pp. 214–217, 2017.

[3] D. Dua and C. Graff, "UCI Machine Learning Repository," UCI Machine Learning Repository: Bank Marketing Data Set, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/bank+marketing. [Accessed: 28-Feb-2022].

[4] C. Feng, H. Wang, N. Lu, T. Chen, H. He, Y. Lu, and X. M. Tu, "Log-transformation and its implications for data analysis," Shanghai Arch Psychiatry, vol. 26, no. 2, pp. 105–109, Apr. 2014.

[5] L. Tolo̧si and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," Bioinformatics, vol. 27, no. 14, pp. 1986-1994, Jul. 2011.

[6] S. H. Javaheri, M. M. Sepehri, and B. Teimourpour, "Chapter 6 - Response Modeling in Direct Marketing: A Data Mining-Based Approach for Target Selection," in Data mining applications with R, Y. Zhao and Y. Cen, Eds. Waltham, MA: Academic Press, 2014, pp. 153–180.

[7] K. Kirasich, T. Smith, and B Sadler, "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Science Review: vol. 1, no. 3, 2018.

[8] E. J. McCarthy, J. F. Grashof, and W. D. Perreault, Basic marketing: A managerial approach. Homewood, IL: Irwin, 1971.

[9] M. Zineldin and S. Philipson, "Kotler and Borden are not dead: Myth of Relationship Marketing and truth of the 4Ps," Journal of Consumer Marketing, vol. 24, no. 4, pp. 229–241, Jul. 2007.