

# Optimizing Bank Product Campaign through Propensity Modeling: A Case Study of a Portuguese Bank

Chen Bo(A0251236M), Chen Zhihong(A0251133W), Guo Jingjing(A0251535J), Huang Hongjie(A0262724E),  
Lu Jiahong(A0262661E), Luo Weifeng(A0262774W), Zheng Xi(A0251476B)  
National University of Singapore

March 31, 2023

## ***Abstract***

This project aims to evaluate and optimize the efficiency of a term deposit campaign in the banking industry through precision marketing, using customer data from a Portuguese bank. The data will be split into training and testing datasets, and various machine learning algorithms (such as Logistic Regression, Decision Tree, Random Forest, Neural Network, and Support Vector Machine) will be used to build propensity models that predict whether customers will make term deposits after being invited to the campaign. To handle a class imbalance in the dataset, the weight adjusting method and Synthetic Minority Over-sampling Technique will be used. The efficiency of the model will be evaluated based on the Area Under Curve (AUC) score. After comparison, the Logistic model was selected as the final model with an accuracy score of approximately 0.82.

## ***Key Words***

Propensity Model, Logistic Regression, Decision Tree, Random Forest, Neural Network, Support Vector Machine (SVM), Synthetic Minority Over-sampling Technique (SMOTE), Area Under Curve (AUC)

## **I. INTRODUCTION**

The torrent of the era of mobile Internet and big data technology is engulfing traditional finance institutions, urging financial institutions to continuously increase investment in technology and innovate product design and marketing model to respond to customers' increasingly personalized, changeable and sophisticated financial needs. Relatively speaking, the extensive traditional marketing model is not only bad in efficiency, but it also leads to the loss of customers. Currently, the marketing model of the banking industry is gradually transforming from creative marketing, direct marketing to precision marketing, mainly reflected in: Data-driven instead of creative priority in marketing planning; bold attempts in marketing delivery Internet marketing methods; innovative customer-centric, precise marketing with banking characteristics pin mode.

Several machine learning algorithms have been applied in customized marketing prediction like Logistic Regression (LR), Decision Tree (DT), Neural Network (NN), and Support Vector Machine (SVM), etc. However, not all the algorithms mentioned above may give a satisfying prediction outcome and it seems unreasonable to intuitively choose the best one to apply without multiple testing. Specifically, LR and DT have the benefit of fitting models that are easy for people to understand while also making excellent forecasts in classification tasks. When compared to DT or traditional statistical modeling like LR, NN and SVM are more flexible with no priori restriction, showing learning capabilities ranging from linear to complicated nonlinear mappings. Based on this, NN and SVM tend to provide correct predictions, but the resulting models are challenging for people to understand. Consequently, several algorithms that are potential useful should be tested before choosing the best one and the classification performance may differ among different problem contexts, comparison outcome of several studies in different situations provide strong evidence. For example, better prediction result is gained from SVM when predicting the bank customer churning (He et al., 2014) and bank risk assessment (Chen, 2019), similar

performances by using NN and SVM were obtained in satellite image analysis and prediction of car prices (Cortez, 2010), whereas DT outperformed NN and SVM for bankruptcy prediction although more rule nodes than desired were observed (Olson et al., 2012).

This project will focus on customer behavior on term deposits which is a crucial source of income for banks through interest margin, and try to evaluate and optimize how to improve efficiency of a term deposit campaign. To implement this objective, the customer data of a Portuguese bank will be used and split into two parts for training and testing respectively. Training dataset will be used to extract key features, make classification for target customer identification, and build propensity models by using different machine learning algorithms to predict whether customers will make term deposits after being invited into a precision marketing campaign. The test dataset will sequentially be used to evaluate the model efficiencies and then make comparisons and optimization. The remaining components are as follows: a second section forming a literature review about the evolution of bank marketing and applied algorithms; a third section introducing the prediction method including data and model; a fourth section detailing the model outcome and efficiency comparison; a fifth section outlining further discussion; a sixth section illustrating weakness and next step that can develop, and a final section briefing the conclusion.

## **II. LITERATURE REVIEW**

### *2.1 The Evolve of Bank Marketing*

As was introduced by Bank of China (2017), the marketing in banking has experienced several stages.

The stage of creative marketing (1960s-70s).

Marketing activities emphasize the creativity and artistry of advertising, and do not pay attention to logic and analysis, so as not to restrict marketing creativity. Advertising masters represented by Bill Bernbach, the founder of Hengmei Advertising Company, have created many advertisements that are both artistic and novel. During this period, the banking industry began to hire advertising agencies and promotion experts to study marketing plans, and to attract customers by printing advertisements on gifts such as umbrellas.

The direct marketing stage (1980s-90s).

The development of the media and advertising industry and the abundance of promotional methods have made it easier to attract customers. Marketing has begun to focus on maintaining customer relationships, emphasizing market segmentation and positioning, and at the same time began to use data as a guide to predict customers based on purchase time, frequency, and amount. Behavior. During this period, the banking industry formed a marketing concept focusing on creating a warm service environment and friendly staff attitudes, focusing on clarifying product positioning and segmenting customer groups to form differentiated advantages.

Stage of precision marketing (from the 21st century to the present).

With the introduction of the concept of precision marketing, the importance of data in marketing promotion has been further highlighted, and data is widely used in the whole process of insight into customer behavior, correlation analysis and marketing interaction. During this period, in marketing management, through the construction of systems and marketing platforms, the bank realized the digital marketing process reconstruction of marketing activities including pre-analysis, planning, monitoring during the event, and post-event statistics and summarization.

### *2.2 The Concept of Enterprise Marketing and Precision Marketing*

Li (2018) stated that marketing strategy refers to the process of identifying the actual needs of customers and utilizing information on their purchasing behavior, market expectations, and other relevant data to plan and organize production and business activities. It aims to understand and analyze various market conditions to accurately select and seize opportunities that meet consumer needs and maximize the company's long-term growth and success. Precision marketing is a type of

marketing strategy that emphasizes precise targeting of customers using modern information technology to establish personalized communication systems. It has become a crucial aspect of network marketing and is considered a core concept in this field. In summary, precision marketing is an approach that focuses on accurately understanding and meeting the needs of customers while minimizing costs and maximizing profits for the company.

### *2.3 Propensity Modeling Techniques to Predict Customers' Behaviors*

Propensity modeling is a set of statistical techniques that aim to predict the likelihood of customer behavior based on their past actions and characteristics, basically, it can help businesses improve their marketing campaigns, target customers more effectively, make better business decisions, and reduce customer churn. In fact, many studies have applied propensity models to predict customers' behaviors as well as the brought results presented high accuracy. However, as various techniques are widely used in propensity modeling such as logistic regression, decision trees, random forests, neural networks, support vector machine, etc. and each has pros and cons, it is hard to choose the best model among them and it depends on the data and problem as well (Moro et al., 2014).

#### *1. Logistic Regression*

Logistic regression is one of the most common techniques for propensity modeling. It is a type of classification model that estimates the probability of an event occurring based on a set of independent variables. For example, logistic regression can be used to predict the probability of a customer buying a product based on their age, gender, income, etc. Logistic regression has the advantages of being interpretable, fast, and accurate on simple data. However, it also has some limitations such as needing a simple, clean data set, working best with linearly separable data, and having a difficult mathematical explanation of the result. Jill (2011) points that logistic regression could be a primer for research that needs quantity analysis to assess model.

#### *2. Decision Tree*

Decision trees are another technique for propensity modeling. They are graphical models that split the data into branches based on certain criteria or rules. For example, a decision tree can be used to predict the probability of a customer buying a product based on whether they are male or female, whether they have children or not, whether they have a high or low income, etc. Decision trees have the advantages of being easy to understand and visualize, able to handle complex and nonlinear relationships, and able to deal with missing values and outliers. However, they also have some drawbacks such as being prone to overfitting, time-consuming, sensitive to small changes in the data, and having high variance. Olson et al. (2012) compared decision tree, neural network, and support vector machine for the prediction of bankruptcy, and the accuracy generated by decision tree was found to be higher than those from other algorithms. They also showed how adjusting the minimum support level can reduce the number of rules and increase the comprehensibility of decision trees, which is a common problem.

#### *3. Random Forest*

Random forest is an ensemble model that consists of many decision trees that vote on the final outcome. Random forest is robust and adaptable that can handle complex and nonlinear relationships, deal with missing values and outliers, and provide feature importance measures. However, random forest is also cost-intensive since the number of trees can be huge, it may be costly to forecast using this model, and it is under low control to information except for the number of trees and their depth, which is prone to overfitting, limited control over their customers' behavior forecasting models, and less interpretable than logistic regression. From a recent research, Carlos et al. (2019) provide additional diversity for the decision trees by means of the use of imprecise probabilities and they found that random forest model provide better results than only use the random forest with multivariate decision trees.

#### *4. Neural Network*

Neural networks are another technique for propensity modeling. They are complex models that consist of multiple layers

of interconnected nodes that mimic the structure and function of biological neurons. Neural networks can learn from the data and adjust their weights and biases accordingly. Neural networks can handle complex and nonlinear relationships, deal with high-dimensional data, and achieve high accuracy and generalization. However, neural networks also have some challenges such as being difficult to interpret and explain, requiring large amounts of data and computational resources, and being sensitive to hyperparameters and initialization. Bahari et al. (2015) proposes an efficient CRM-data mining framework for the prediction of customer behavior and two classification models, Naïve Bayes and Neural Networks, are applied to a dataset of credit card customers. After comparing their accuracy, the Neural Networks have a higher accuracy than Naïve Bayes. Ładyżyński et al. (2019) presents a novel approach for direct marketing campaigns in retail banking using deep learning and random forests. They applied two classification models, deep neural networks and random forests, to a dataset of bank customers and compares their performance, both models achieve high accuracy and precision, but deep neural networks have a slight edge over random forests.

### 5. Support Vector Machine

Support vector machines are another technique for propensity modeling. They are models that try to find the optimal hyperplane that separates the data into two classes with maximum margin. Support vector machines can handle linearly separable data as well as nonlinearly separable data by using kernel functions that transform the data into higher-dimensional spaces. Support vector machines have the advantages of being robust to outliers, having low variance and high accuracy, and being able to handle sparse data. However, they also have some disadvantages such as being difficult to interpret and tune, requiring long training time and memory space, and being sensitive to noise and imbalanced data. Chen (2019) provides a unique KYC data-based method to improve the detection of default risk and suspicious behavior. With a dataset of bank customers and branches, higher accuracy and precision were found by using support vector machines than decision trees although decision trees have an advantage over support vector machines in terms of interpretability and computational efficiency.

SVM is a supervised machine learning algorithm. After the introduction of kernel method, SVM has a wider application, which can be used as both linear classifier and nonlinear classifier. Kernel functions used in SVM include multilayer perceptron (MLP), Gaussian RBF Network, direct/inverse multiquadric, etc. In this study, SVM algorithm with Gaussian RBF Network is mainly adopted.

SVM is optimized by solving dual Lagrange problem, and the decision function is

$$\hat{y} = \text{sign} \left( \sum_{i=1}^m \sum_{j=1}^m \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} K(x^{(i)}, x^{(j)}) + b \right)$$

Where  $\alpha_i$  is Lagrange multipliers for each of the sample constraints,  $K(x^{(i)}, x^{(j)})$  is kernel function.

Two learning strategies are used for SVM, maximizing geometric spacing between data sets and separated hyperplanes, and minimizing hinge loss functions. After substituting constraints, the final decision function is (Sanchez, 2003)

$$\begin{aligned} \min W(\vec{\alpha}) &= - \sum_{i=1}^m \alpha^{(i)} + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j K(\vec{x}^{(i)}, \vec{x}^{(j)}), \\ \sum_{i=1}^m \alpha_i y_i &= 0, \\ 0 \leq \alpha_i &\leq C \quad \forall i = 1, \dots, m. \end{aligned}$$

In this case, kernel function is Gaussian RBF Network, the expression is

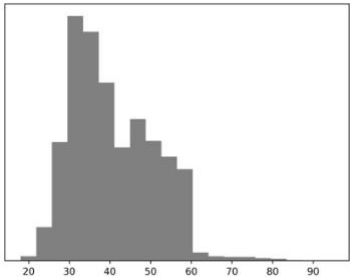
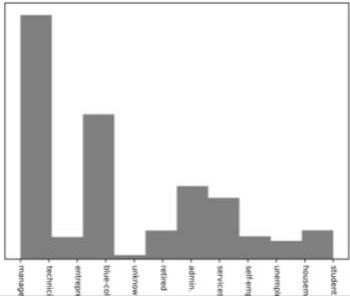
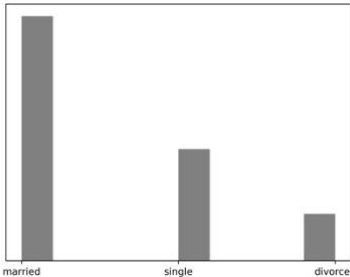
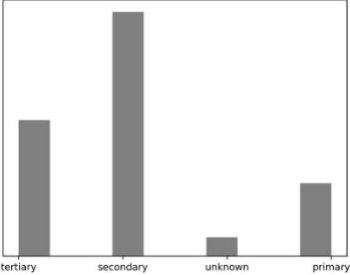
$$\exp(-\|\vec{x} - \vec{y}\|^2)$$

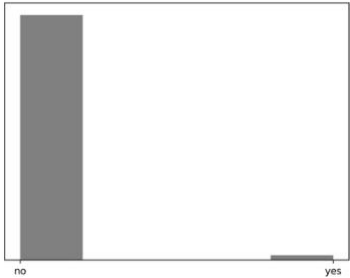
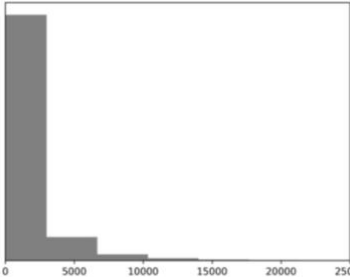
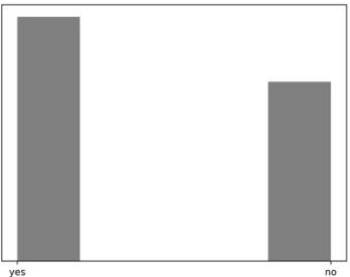
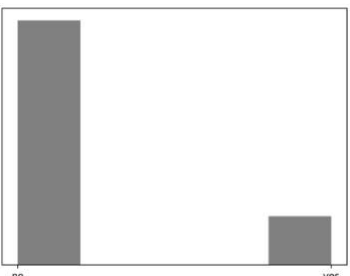
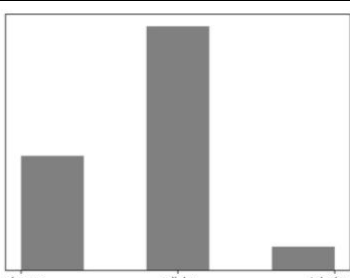
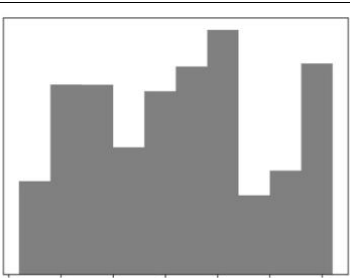
### III. METHODOLOGY

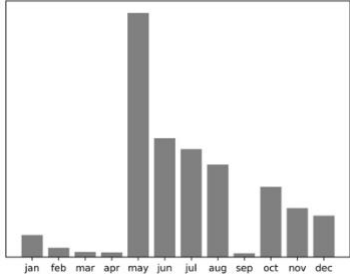
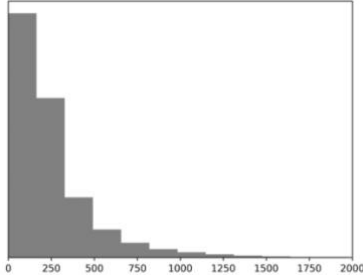
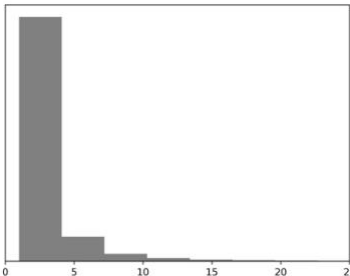
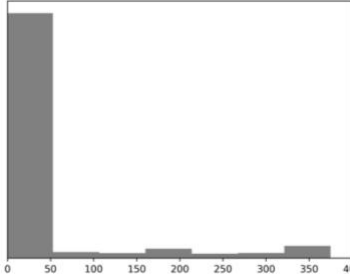
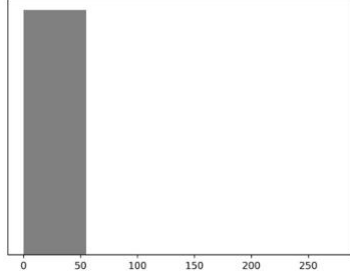
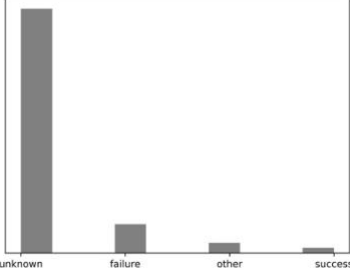
#### 3.1 Data Description


The project aims to predict if the clients of Portuguese banking institution will subscribe to a term deposit during the marketing campaign. The data records the 17 metrics of consumer information through phone call, including 6 numeric variables and 11 categorical variables.

Table 1. Features Of Bank Customers

	<i>Variables</i>	<i>Type</i>	<i>Description</i>	<i>Visualization</i>
1	age	Numeric	Age of customers	
2	job	Categorical	Job type	
3	marital	Categorical	Marital status	
4	education	Categorical	Education status	

5	default	Categorical	Has credit in default Y/N	
6	balance	Numeric	Account balance	
7	housing	Categorical	Has housing loan Y/N	
8	loan	Categorical	Has personal loan Y/N	
9	contact	Categorical	Contact communication type	
10	day	Categorical	Last contact day	

11	month	Categorical	Last contact month	 <table><caption>Data for Last contact month</caption><thead><tr><th>Month</th><th>Frequency</th></tr></thead><tbody><tr><td>jan</td><td>10</td></tr><tr><td>feb</td><td>5</td></tr><tr><td>mar</td><td>5</td></tr><tr><td>apr</td><td>5</td></tr><tr><td>may</td><td>150</td></tr><tr><td>jun</td><td>100</td></tr><tr><td>jul</td><td>90</td></tr><tr><td>aug</td><td>80</td></tr><tr><td>sep</td><td>5</td></tr><tr><td>oct</td><td>40</td></tr><tr><td>nov</td><td>25</td></tr><tr><td>dec</td><td>25</td></tr></tbody></table>	Month	Frequency	jan	10	feb	5	mar	5	apr	5	may	150	jun	100	jul	90	aug	80	sep	5	oct	40	nov	25	dec	25
Month	Frequency																													
jan	10																													
feb	5																													
mar	5																													
apr	5																													
may	150																													
jun	100																													
jul	90																													
aug	80																													
sep	5																													
oct	40																													
nov	25																													
dec	25																													
12	duration	Numeric	Last contact duration/seconds	 <table><caption>Data for Last contact duration/seconds</caption><thead><tr><th>Duration Range (seconds)</th><th>Frequency</th></tr></thead><tbody><tr><td>0-250</td><td>150</td></tr><tr><td>250-500</td><td>100</td></tr><tr><td>500-750</td><td>40</td></tr><tr><td>750-1000</td><td>20</td></tr><tr><td>1000-1250</td><td>10</td></tr><tr><td>1250-1500</td><td>5</td></tr><tr><td>1500-1750</td><td>2</td></tr><tr><td>1750-2000</td><td>1</td></tr></tbody></table>	Duration Range (seconds)	Frequency	0-250	150	250-500	100	500-750	40	750-1000	20	1000-1250	10	1250-1500	5	1500-1750	2	1750-2000	1								
Duration Range (seconds)	Frequency																													
0-250	150																													
250-500	100																													
500-750	40																													
750-1000	20																													
1000-1250	10																													
1250-1500	5																													
1500-1750	2																													
1750-2000	1																													
13	campaign	Numeric	# of contacts during campaign per client	 <table><caption>Data for # of contacts during campaign per client</caption><thead><tr><th>Contacts Range</th><th>Frequency</th></tr></thead><tbody><tr><td>0-5</td><td>150</td></tr><tr><td>5-10</td><td>20</td></tr><tr><td>10-15</td><td>10</td></tr><tr><td>15-20</td><td>5</td></tr><tr><td>20-25</td><td>2</td></tr></tbody></table>	Contacts Range	Frequency	0-5	150	5-10	20	10-15	10	15-20	5	20-25	2														
Contacts Range	Frequency																													
0-5	150																													
5-10	20																													
10-15	10																													
15-20	5																													
20-25	2																													
14	pdays	Numeric	# of passed days from last contact	 <table><caption>Data for # of passed days from last contact</caption><thead><tr><th>Days Range</th><th>Frequency</th></tr></thead><tbody><tr><td>0-50</td><td>150</td></tr><tr><td>50-100</td><td>5</td></tr><tr><td>100-150</td><td>2</td></tr><tr><td>150-200</td><td>10</td></tr><tr><td>200-250</td><td>5</td></tr><tr><td>250-300</td><td>2</td></tr><tr><td>300-350</td><td>10</td></tr><tr><td>350-400</td><td>10</td></tr></tbody></table>	Days Range	Frequency	0-50	150	50-100	5	100-150	2	150-200	10	200-250	5	250-300	2	300-350	10	350-400	10								
Days Range	Frequency																													
0-50	150																													
50-100	5																													
100-150	2																													
150-200	10																													
200-250	5																													
250-300	2																													
300-350	10																													
350-400	10																													
15	previous	Numeric	# of contacts before campaign per client	 <table><caption>Data for # of contacts before campaign per client</caption><thead><tr><th>Contacts Range</th><th>Frequency</th></tr></thead><tbody><tr><td>0-50</td><td>150</td></tr><tr><td>50-100</td><td>5</td></tr><tr><td>100-150</td><td>2</td></tr><tr><td>150-200</td><td>1</td></tr><tr><td>200-250</td><td>1</td></tr></tbody></table>	Contacts Range	Frequency	0-50	150	50-100	5	100-150	2	150-200	1	200-250	1														
Contacts Range	Frequency																													
0-50	150																													
50-100	5																													
100-150	2																													
150-200	1																													
200-250	1																													
16	poutcome	Categorical	Outcome of last campaign	 <table><caption>Data for Outcome of last campaign</caption><thead><tr><th>Outcome</th><th>Frequency</th></tr></thead><tbody><tr><td>unknown</td><td>150</td></tr><tr><td>failure</td><td>20</td></tr><tr><td>other</td><td>10</td></tr><tr><td>success</td><td>5</td></tr></tbody></table>	Outcome	Frequency	unknown	150	failure	20	other	10	success	5																
Outcome	Frequency																													
unknown	150																													
failure	20																													
other	10																													
success	5																													

17	term_deposit	Categorical	If clients subscribe term deposit Y/N	
----	--------------	-------------	---------------------------------------	---

### 3.2 Modelling

#### 1. Data Cleaning Process

We removed the columns "Day" and "Month" because they represent the last contact date and last contact month of the year, respectively. We retained "Last contact duration" in the model to ensure that there is at least one feature related to the contact. We then discovered that the "Outcome of the previous marketing campaign" column had 84.6% unknown data, which prompted us to remove the column from the dataset.

The outcome of term deposit for this dataset implies that 90.71% of customers did not choose to purchase the term deposit product, while less than 10% did. This extreme imbalance in the data led us to adopt different methods for data balancing in the modeling process.

#### 2. Data Balancing

In our model construction, we have employed two methods to balance the imbalanced data. The first approach is by setting the '*class\_weight*' parameter in *sklearn* to balance the class weights. This function automatically handles the imbalanced data problem by adjusting the weight of each class based on its frequency, giving more weight to the minority class samples and thus focusing the model more on correctly classifying the minority class. The second method we adopted is SMOTE (Synthetic Minority Over-sampling Technique). The SMOTE algorithm randomly selects a minority class sample and then selects a random sample from its *k* nearest neighbors. A new synthetic sample is generated on the line between these two samples, and the new sample is added to the dataset to increase the number of minority class samples (Elreedy et al., 2019). By using the two different balancing data methods mentioned above, we obtained the following eleven models.

#### 3. Model Specifics

Table 2. Outcomes Of Modelling

	<i>Classification Method</i>	<i>Balancing Method</i>	<i>Classification Accuracy</i>	<i>AUC</i>	<i>Precision (yes)</i>	<i>Recall (yes)</i>	<i>f1-score(yes)</i>	<i>f1-score(no)</i>
1	Logistic Regression	class_weight="Balanced"	0.82	<b>0.80</b>	0.32	<b>0.78</b>	<b>0.45</b>	0.89
2	Logistic Regression	SMOTE	0.90	0.66	0.48	0.36	<b>0.41</b>	0.95
3	Decision Tree	class_weight="Balanced"	0.88	0.65	0.37	0.38	0.36	0.93
4	Decision Tree	SMOTE	0.87	0.66	0.33	0.38	0.36	0.93



5	Random Forest (n=10)	class_weight= "Balanced"	0.91	0.58	0.57	0.18	0.29	0.95
6	Random Forest (n=50)	class_weight= "Balanced"	0.91	0.59	0.58	0.20	0.29	0.95
7	Random Forest (n=10)	SMOTE	0.91	0.63	0.51	0.29	0.37	0.95
8	Random Forest (n=50)	SMOTE	0.91	0.64	0.52	0.31	0.39	0.95
9	SVM	class_weight= "Balanced"	0.81	<b>0.84</b>	0.27	<b>0.68</b>	<b>0.41</b>	0.88
10	SVM	SMOTE	0.80	<b>0.84</b>	<b>0.79</b>	<b>0.69</b>	<b>0.40</b>	0.88
11	Neural Network	/	0.91	<b>0.85</b>	0.56	0.11	0.26	0.95

After comparing the results in the table above, we found that although the Random Forest and Neural Network models have the highest accuracy, they perform poorly in predicting customers who will purchase the term deposit. AUC is short for "Area Under the ROC Curve", which calculates the total area under the entire ROC curve in two dimensions, from (0,0) to (1,1), using integral calculus. AUC is a comprehensive indicator of performance for all classification thresholds. A higher AUC value indicates a better model performance ("Classification: ROC Curve and AUC", 2023). By comparing the AUC values of these 11 models, we found that model 1, 9, 10, and 11 have the highest AUC score around 0.8.

As our business goal is to identify more customers who are likely to purchase the product and to conduct personalized marketing, therefore, the primary focus should be on maximizing the true positives (i.e., correctly identifying customers who will make a term deposit, also known as recall). Thus, recall of the "yes" group should be prioritized over precision in this case. We need to rely on the recall score of the "yes" group to select the optimal model. Model 11, the Neural Network model has the lowest recall score, indicating that it cannot detect the target customers well enough, hence we drop model 11. Model 1 is the logistic model with the weight adjusted balance method, and it has a recall score of 0.78, that means the has a high ability to correctly identify positive cases. Model 9 and model 10 are two SVM models, they all have recall scores around 0.7, which also indicates they have good ability to correctly identify targeted customers.

Therefore, we chose model 1 the logistic regression with weight adjusted balance method as the best model, and two SVM models are the second and third best model.

#### 4. Result

##### Logistic Regression

The final equation of model 1 is as below:

$$\log\left(\frac{P}{1-P}\right) = -0.5251 - 0.00222X_{age} + \dots - 1.0478X_{contactUnknown}$$

Table 3. Parameters of Logistic Model

<i>Parameter</i>	<i>Coefficient</i>	<i>Parameter</i>	<i>Coefficient</i>	<i>Parameter</i>	<i>Coefficient</i>
age	-0.0022	balance	3.60e-05	duration	<b>0.0057</b>
pdays	<b>0.0009</b>	previous	<b>0.0896</b>	job_admin	<b>0.1110</b>
job_entrepreneur	-0.0849	job_housemaid	-0.0755	job_management	-0.1530
job_self-employed	-0.0701	job_services	-0.1794	job_student	<b>0.1497</b>
job_blue-collar	-0.4678	job_retired	<b>0.3976</b>	job_technician	-0.1241
job_unemployed	-0.0215	job_unknown	-0.0067	marital_divorced	<b>0.0228</b>
marital_married	-0.3765	education_tertiary	-0.0204	housing_no	<b>0.2382</b>
marital_single	-0.1710	education_primary	-0.2898	education_secondary	-0.2011
education_unknown	-0.0134	default_no	-0.4404	default_yes	-0.0844
housing_yes	-0.7630	loan_no	<b>0.0461</b>	loan_yes	-0.5709
contact_telephone	<b>0.2561</b>	contact_unknown	-1.0478	campaign	-0.1328
contact_cellular	<b>0.2669</b>	intercept	-0.5251		

#### IV. DISCUSSION

In the basis of the density graph presenting above, most of customers had predicted probabilities of less than 50%, which reveals the conversion rate of the marketing campaign is unexpectedly low. If the purpose of the campaign is to attract new subscribers, then the low conversion rate indicates the marketing campaigns are moderately failed due to several reasons, for example, the campaign may be reaching the audience who is not interested in the financial products being offered, or the campaign itself is not compelling enough, etc.

From the resulting chart of logistic regression, the balance variable is intensely correlated to the interests of subscribing the deposit products. It is reasonable to assume that consumers with high balance or low balance have apparently distinct depositing interests. Another highly correlated feature is the job type where administration employees, students, and retirees have positive correlation to the depositing interests. These three roles of job are secure positions compared to other positions; consumers prefer to purchase secure financial products such as the terms of deposit. Additionally, a longer campaign duration and clear customer contact records can improve the likelihood of customers purchasing the product.

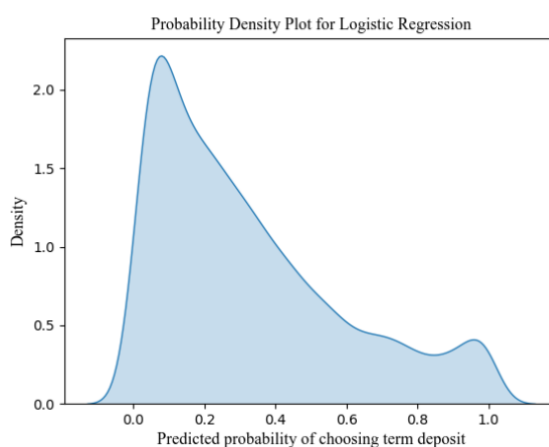


Figure 1. Probability Density Plot For Logistic Regression

To consider the marital status, divorced consumers hold the strongest correlation of the deposit interests, and the coefficient of single consumers is higher than consumers who have married. One possible reason for this result might be that married families need financial products at low risks and their propensity is on the products recommended by their trust portfolio managers instead of subscribing term of deposit through phone call campaigns. A large portion of the single consumers are supposed to have propensity on deposits into accounts, such that they are more inclined to subscribe to the terms of deposit than married consumers.

To summarize the findings above, this marketing campaign needs improvement to gain subscribers. Balance is one of the key factors to be considered when promoting the campaigns. Administration employees, students and retirees have high propensity in terms of deposit when targeting audience by job type. Divorced cohorts have the highest interest in subscribing to this campaign than other marital statuses.

## **V. WEAKNESS AND NEXT STEPS**

### *5.1 Weakness*

1. The accuracy rate of the model is only around 0.85, with the highest accuracy rate being only 0.9. While this is relatively high, it still leaves room for improvement. One potential reason for this could be that the model is not capturing all of the relevant factors that influence customer behavior. For example, the dataset used in this study only includes 16 features, and there are likely many other variables that could be relevant, such as demographic factors, psychographic factors, and environmental factors.
2. The choice of model is limited to the current dataset, with the random forest model performing the best among all selected models. However, as more categorical data is added with increasing data, the accuracy of the random forest model may decline.
3. The low accuracy of the "yes" class is also a weakness of the current model. This could be due to the imbalanced nature of the dataset, with far more examples of the "no" class than the "yes" class. To address this, alternative methods for balancing the dataset could be explored.
4. Finally, it is worth noting that the accuracy of the model may be impacted by external factors that are difficult to control for, such as changes in market conditions or shifts in customer preferences. As a result, it is important to periodically evaluate and update the model to ensure that it remains effective over time.

### *5.2 Next Steps*

1. As noted, one potential area for improvement is to conduct an importance analysis of the existing features to determine which ones are most predictive of customer behavior. This could involve using techniques such as feature selection or principal component analysis to identify the most relevant variables. Additionally, new features could be added to the dataset based on insights from domain experts or additional data sources.
2. Another way to improve the model is to optimize the algorithm parameters. For example, the SVM model's parameters C and gamma could be tuned using grid search or other optimization techniques to improve the accuracy of the model.
3. To address the issue of imbalanced data, several approaches could be explored. One option is to use more advanced resampling techniques such as ADASYN (Adaptive Synthetic Sampling) to generate synthetic data points for the minority class.
4. Finally, it may be useful to explore alternative algorithms that are better suited to handling larger and more complex datasets. For example, deep learning techniques such as convolutional neural networks could be applied to the data to capture more complex interactions between variables.

## VI. CONCLUSION

To sum up, the era of mobile Internet and big data technology has changed the traditional marketing model, prompting financial institutions to increase investment in technology, innovate product design and marketing strategies, in order to meet the increasingly personalized and complex financial needs of customers. In this context, precision marketing and machine learning technologies have emerged as effective tools to improve the efficiency of marketing campaigns and enhance customer engagement.

Several machine learning algorithms have been applied to customize marketing forecasts, including logistic regression, decision trees, neural networks, and support vector machines. To choose the best algorithm for a given problem context, one should test multiple algorithms before choosing the most efficient one. In this paper, the SVM algorithm is selected as the final model, and the accuracy score is about 0.81.

In this study, after cleaning and preprocessing the data, the project team built 11 models using logistic regression, SVM, random forest, and neural network algorithms along with different data balancing techniques. In this paper, evaluated the models using accuracy, AUC, precision and recall metrics. Based on observation, while random forest and neural network models had the highest accuracy, the models performed poorly at identifying customers who would buy a product. Model 1, the weight-adjusted balanced logistic regression model, had the highest recall in the Yes category and was the best model for identifying potential customers. Despite the success of the model, there were some limitations and weaknesses including the limited number of features, possible drop in accuracy as more categorical data is added, the imbalanced nature of the dataset, and the potential influence of external factors. To address these weaknesses, the paper recommend several next steps, such as performing importance analysis on existing features, optimizing algorithm parameters, exploring advanced resampling techniques, and exploring alternative algorithms such as deep learning. Although the model has some limitations, it provides a promising basis for identifying customers who are likely to purchase term deposits and for personalized marketing.

Overall, the paper highlights the importance of adopting emerging technologies and data-driven approaches to optimize marketing strategies in the banking industry. The project's findings could help banks refine their marketing strategies to better target customers, improve customer engagement and improve overall performance.

## REFERENCES

- [1] Lin Qingpeng. Precision Marketing Strategy Research based on Big Data Mining.
- [2] Dong Jing, Shen Qian, Innovation and Enlightenment of Bank Marketing Model.
- [3] Chen, T.-H. (2020). Do you know your customer? Bank risk assessment based on machine learning. *Applied Soft Computing*, 86, 105779.
- [4] Moro, S., Cortez, P. & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*. 62, 22-31.
- [5] He, B., Shi, Y., Wang, Q., & Zhao, X. (2019). Prediction of customer attrition of commercial banks based on SVM model. *Journal of Systems Engineering and Electronics*, 30(4), 783-791.
- [6] Ładyżyński, P., Zbikowski, K., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*, 134, 28-35.
- [7] Bahari, F., & Elayidom, S. M. (2015). An efficient CRM-data mining framework for the prediction of customer behavior. *Procedia Computer Science*, 46, 725-731.
- [8] Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. In 10th Industrial Conference on Data Mining (ICDM 2010) (pp. 572-583). University of Minho.
- [9] Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, 52(2), 464-473.
- [10] Stoltzfus JC. (2011). Logistic regression: a brief primer. *Acad Emerg Med*:1099-104.
- [11] Dina Elreedy & Amir F. Atiya. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*.
- [12] Sanchez, A. VD. "Advanced Support Vector Machines and Kernel Methods." *Neurocomputing (Amsterdam)*, vol. 55, no. 1, 2003, pp. 5-20.
- [13] Classification: ROC Curve and AUC. (n.d.). Google Developers. <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>