香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

ACT6241 Machine Learning and Business Analytics

Project Report

Team Member

Name	Student ID
杜雨欣	222021088
袁晨	222021059
左昕叶	222021090
白含岭	222021054
杨煜文	222021067

1. Introduction

1.1 Background and motivation

Telemarketing, as a convenient, fast and economic marketing communication tool, has been widely used in enterprises and consumers. The past decade has seen that more than 60% of our residents use telephone inquiries and consulting services, and more than 20% of residents use telephone reservations and consumption. This means that consumers are not only concerned about the basic functions of the products, but also the added value of the products is an important factor to be considered. Therefore, it is essential to consider how to use telemarketing strategies to maximize business revenue.

1.2 Objective

The research objective of this paper is to explore the distribution of customer deposits by analyzing their age, job, education, marital, housing and loan separately through a machine learning approach. An intelligent decision system is also constructed to predict the optimal outcome of maximizing the bank's revenue from term deposit business using direct telemarketing under different budget constraints. Thus, it helps banks to precisely market customers with certain characteristics to maximize deposit under different budget constraints. This helps decision makers identify and prioritize customers with certain features in marketing activities and select the group of customers to be contacted, thus helping banks to improve the rational use of their resources.

1.3 Key Assumptions

We set a fixed budget X and help banks achieve maximum revenue within this budget. The way how banks generate revenue through absorbing deposit is to reinvest or re-lend the money. Thus, larger deposit can leverage larger revenue or profit.

Customers will be sorted based on these two metrics. The cost is determined by the number of phone marketing. It would be equal to Call duration * call cost * labor cost. We will elaborate on the assumptions regarding customer value later.

1.3.1. Age

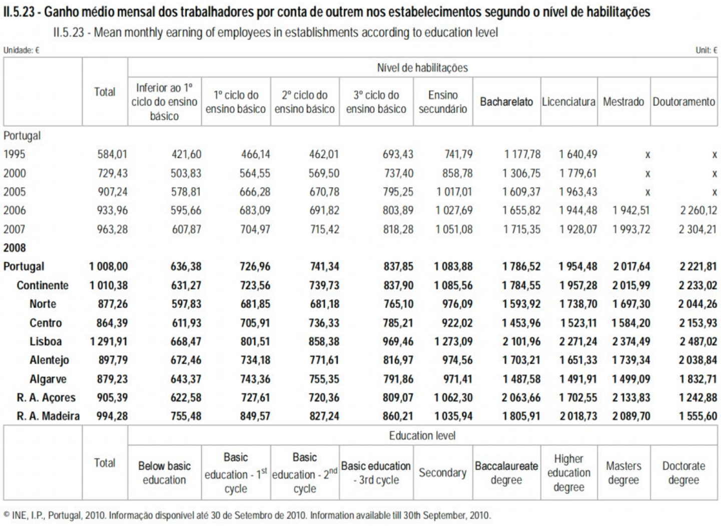
Based on the Life-Cycle Hypothesis of Savings, the savings curve reaches its peak in middle age and is lower during teenager and retirement. For minors, we assume that such customers could not generate revenue. Most banks require parents to participate in the management of minors' bank accounts. Marketing targeting minors is often ineffective. For customers aged 18 to 60, we assume that their savings will increase with age. Since people's income levels increase with years of work, while the burden on families gradually decreases. For individual aged 60 and above, their savings usually gradually decrease due to a sudden drop in income and health expenses.

1.3.2. Job

Taking into account the diversity and stability of income sources and the necessary expenses, we assign weight to the three types of employment status. From the occupation perspectives, blue collar, maid, and service industries are highly mobile and easily replaceable professions. The high cost of living and lower income will leave them with lower savings. Management and entrepreneurs typically have higher income levels. At the same time, they have good financial literacy. They will have more sources of income. They would also increase the savings rate for potential investment opportunities. Especially for entrepreneurs, they are more inclined to retain deposits to cope with potential risks in their operations. We assign higher weight to their potential deposits. Technicians are also given relatively high weights due to their high income. Students lack income and have low savings awareness, therefore assigning lower values.

1.3.3 Education

According to Statistics Portugal, there is statistical evidence that education level has some impact on people's earning.



II.5.23 - Ganho médio mensal dos trabalhadores por conta de outrem nos estabelecimentos segundo o nível de habilitações

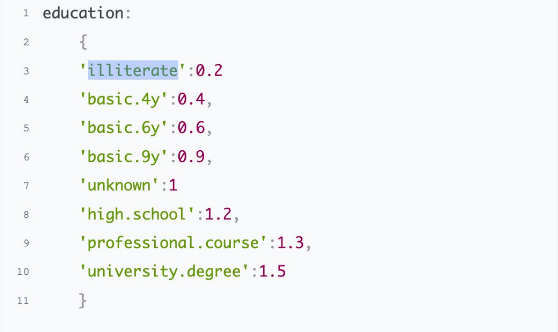
II.5.23 - Mean monthly earning of employees in establishments according to education level

Unidade: € Unit: €

	Total	Nível de habilitações								
		Inferior ao 1º ciclo do ensino básico	1º ciclo do ensino básico	2º ciclo do ensino básico	3º ciclo do ensino básico	Ensino secundário	Bacharelato	Licenciatura	Mestrado	Doutoramento
Portugal										
1995	584,01	421,60	466,14	462,01	693,43	741,79	1 177,78	1 640,49	x	x
2000	729,43	503,83	564,55	569,50	737,40	858,78	1 306,75	1 779,61	x	x
2005	907,24	578,81	666,28	670,78	795,25	1 017,01	1 609,37	1 963,43	x	x
2006	933,96	595,66	683,09	691,82	803,89	1 027,69	1 655,82	1 944,48	1 942,51	2 260,12
2007	963,28	607,87	704,97	715,42	818,28	1 051,08	1 715,35	1 928,07	1 993,72	2 304,21
2008										
Portugal	**1 008,00**	**636,38**	**726,96**	**741,34**	**837,85**	**1 083,88**	**1 786,52**	**1 954,48**	**2 017,64**	**2 221,81**
Continente	**1 010,38**	**631,27**	**723,56**	**739,73**	**837,90**	**1 085,56**	**1 784,55**	**1 957,28**	**2 015,99**	**2 233,02**
Norte	**877,26**	**597,83**	**681,85**	**681,18**	**765,10**	**976,09**	**1 593,92**	**1 738,70**	**1 697,30**	**2 044,26**
Centro	**864,39**	**611,93**	**705,91**	**736,33**	**785,21**	**922,02**	**1 453,96**	**1 523,11**	**1 584,20**	**2 153,93**
Lisboa	**1 291,91**	**668,47**	**801,51**	**858,38**	**969,46**	**1 273,09**	**2 101,96**	**2 271,24**	**2 374,49**	**2 487,02**
Alentejo	**897,79**	**672,46**	**734,18**	**771,61**	**816,97**	**974,56**	**1 703,21**	**1 651,33**	**1 739,34**	**2 038,84**
Algarve	**879,23**	**643,37**	**743,36**	**755,35**	**791,86**	**971,41**	**1 487,58**	**1 491,91**	**1 499,09**	**1 832,71**
R. A. Açores	**905,39**	**622,58**	**727,61**	**720,36**	**809,07**	**1 062,30**	**2 063,66**	**1 702,55**	**2 133,83**	**1 242,88**
R. A. Madeira	**994,28**	**755,48**	**849,57**	**827,24**	**860,21**	**1 035,94**	**1 805,91**	**2 018,73**	**2 089,70**	**1 555,60**
		Education level								
	Total	Below basic education	Basic education - 1st cycle	Basic education - 2nd cycle	Basic education - 3rd cycle	Secondary	Baccalaureate degree	Higher education degree	Masters degree	Doctorate degree

© INE, I.P., Portugal, 2010. Informação disponível até 30 de Setembro de 2010. Information available till 30th September, 2010.

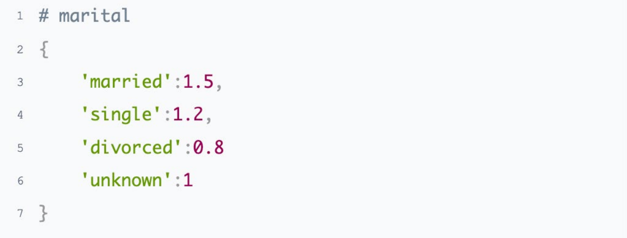
Higher education is correlated with higher income for individuals, and thus higher deposit savable. We assume customers with unknown education level to be 1, and higher level with higher index:



```
education:
    {
    'illiterate':0.2
    'basic.4y':0.4,
    'basic.6y':0.6,
    'basic.9y':0.9,
    'unknown':1
    'high.school':1.2,
    'professional.course':1.3,
    'university.degree':1.5
    }
```

1.3.4Marital

According to common sense, the marriage of a customer will affect his deposit level for marriage's effect on a customer's property level. Married customer, with two salary earners, would have more money to deposit. While single customer, he may also deposit most of his money in the bank because there's not so many conditions for him to spend money. But a divorced customer, the divorce may spend him a lot, so he may not have a lot of money to deposit. So, we made the following assumptions:



```
# marital
{
    'married':1.5,
    'single':1.2,
    'divorced':0.8
    'unknown':1
}
```

1.3.5 Loan (Binary)

Customers with active loans may be inclined to maintain lower deposit amounts in their bank accounts, as they are likely to allocate a portion of their income towards loan repayment. Furthermore, during the 2008-2010 period, the global financial crisis had a significant impact on the banking industry, potentially leading to increased financial stress for borrowers. As a result, customers with loans might have been more cautious in their spending habits and focused on repaying their debts, leaving less money available for deposits.

1.3.6 Housing (Binary)

Homeownership typically represents a significant financial commitment and can influence a customer's deposit behavior. Customers who own houses may have larger deposit amounts due to a sense of financial stability associated with property ownership. However, considering the 2008-2010 period's real-world context, the housing market faced a downturn due to the financial crisis. This could have led to decreased home values and negative equity for some homeowners, affecting their overall financial health. Consequently, homeowners during this time might have had less disposable income to contribute to deposits, while non-homeowners might have been more cautious about investing in property and opted to increase their savings instead.

2. Dataset preparation

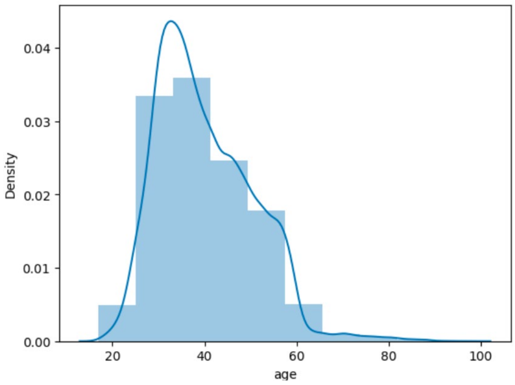
2.1 Data-type and Definition of Variables

```
 0    age              26246 non-null   int64
  1   job              26246 non-null   object
  2   marital          26246 non-null   object
  3   education        26246 non-null   object
  4   default          26246 non-null   object
  5   housing          26246 non-null   object
  6   loan             26246 non-null   object
  7   contact          26246 non-null   object
  8   month            26246 non-null   object
  9   day_of_week      26246 non-null   object
  10  campaign         26246 non-null   int64
  11  pdays            26246 non-null   int64
  12  previous         26246 non-null   int64
  13  poutcome         26246 non-null   object
  14  emp.var.rate     26246 non-null   float64
  15  cons.price.idx   26246 non-null   float64
  16  cons.conf.idx    26246 non-null   float64
  17  euribor3m        26246 non-null   float64
  18  nr.employed      26246 non-null   float64
```

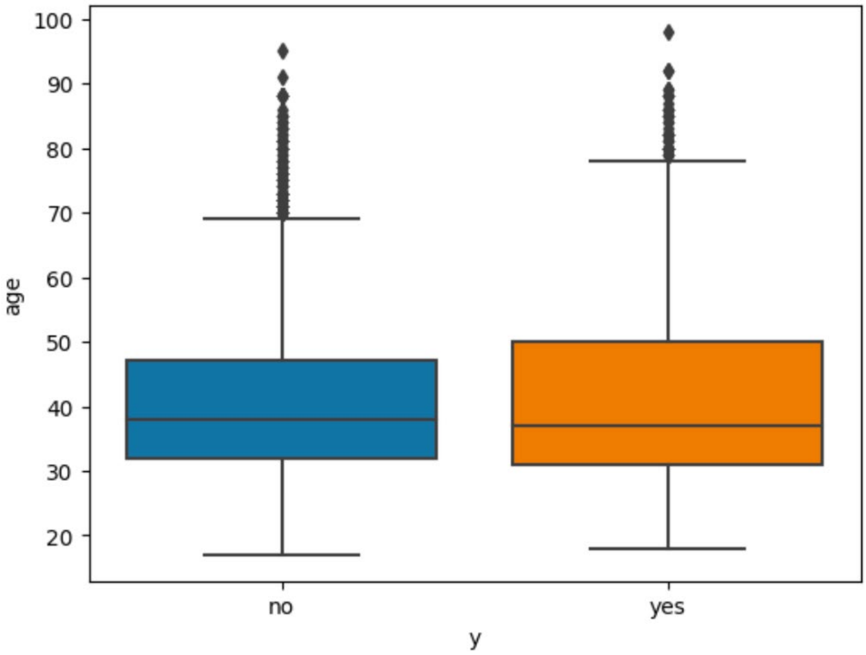
There are a total of 19 variables，between them, job、marital、education、default、housing、loan、contact、month、day_of_week、poutcome are non-numerical variableDescription and Preprocess of Variables that require special treatment.

2.2 Description of Variables

2.2.1Age distribution among customers

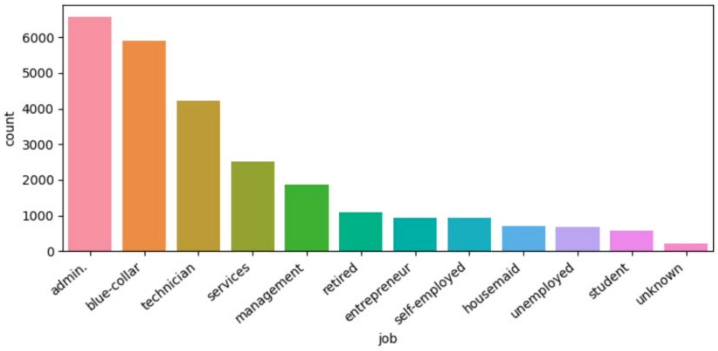


Age ranges from 17 to 98, and the average age is 40.2 years old.

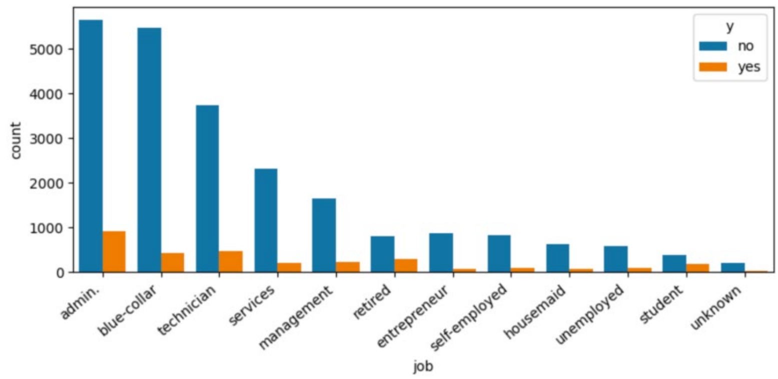


Analyzing the relationship between age and marketing success, the older you are, the easier it is to be successfully marketed. Successful customers are mainly between 30 and 50 years old, and the marketing success rate of customers around 50 years old is very high.

2.2.2 Career analysis

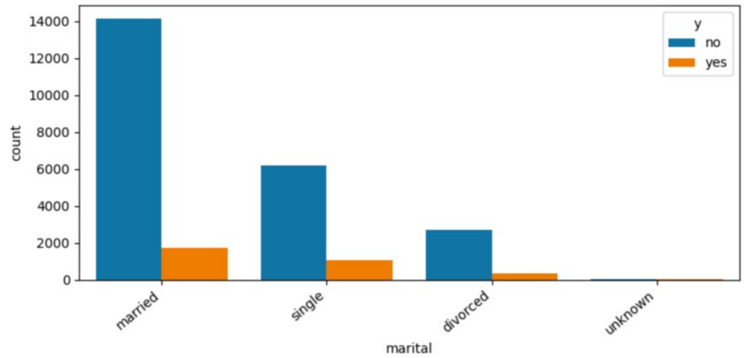


The largest number of occupations is amin, followed by blue-collar workers and technology practitioners.



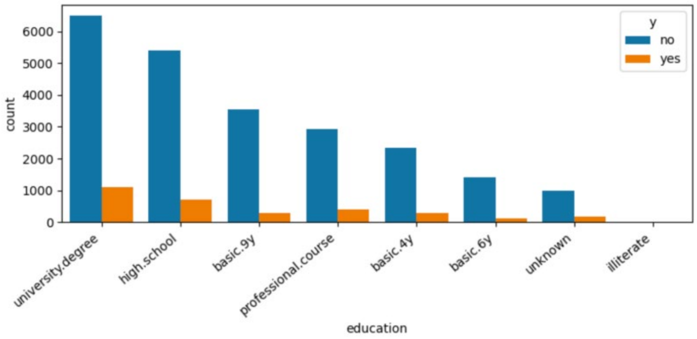
Analyzing the success rate of its marketing, it can be seen that administrators, managers and scientific and technological personnel are more likely to succeed in marketing.

2.2.3 The relationship between marriage and marketing



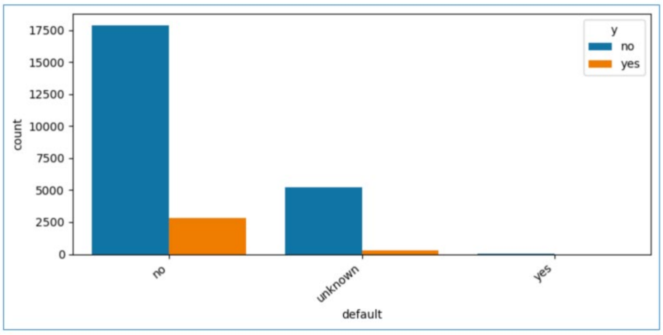
Seems marital status has no impact on people's decision.

2.2.4 Education level



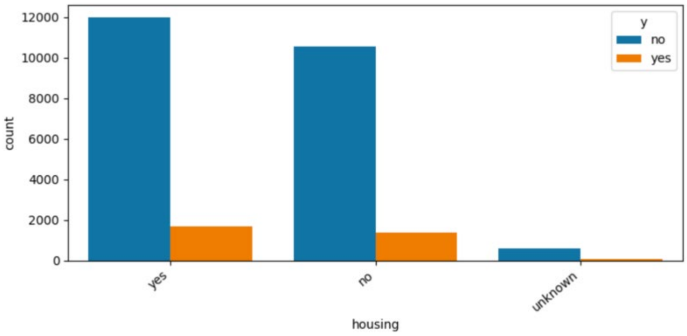
The higher the education level, the easier it is to succeed.

2.2.5 Credit in default



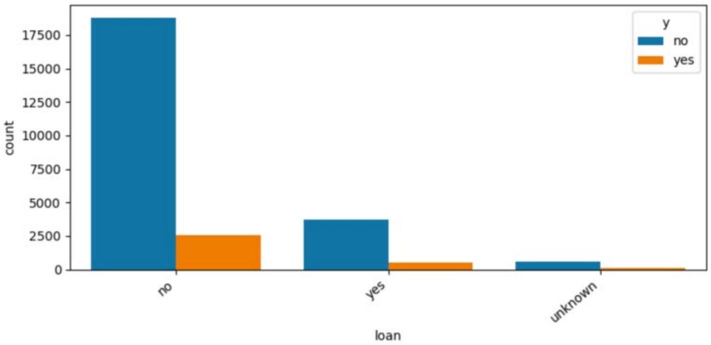
Pretty much customers' default history are unknown (almost 1/3). In addition, customers without a record of credit violation are more likely to succeed in marketing.

2.2.6 Housing



There are more customers who have housing loan than customers who don't. The impact of housing loan is not clear here. But in guts call, will people having housing loan have less probability to take long-term loan? We need to explore further.

2.2.7 Loan



Most of the customers don't have personal loan. Again, will people having personal loan have less probability to take long-term loan? We need to explore further.

2.3 Preprocess of Variables

Step1: The training data is divided into 9:1, of which 9 are used as training data and 10% is used as verification data to verify which model is the best.

Step 2: One-hot encoding of category data.

Onehot Coding: Monothermal Coding Is One-Hot Coding, Also Known As One Effective Coding. The Method Is To Use N-bit State registers To Encode N States. Each State Is Separately By Its Independent register, And At Any Time, Only One Of Them Is Valid. All non-numerical features are encoded by onehot, which can process category features, unlike numerical coding, which will be affected by the size of the value.

The characteristics of one-hot coding are:

['job', 'marital', 'education',
'default', 'housing', 'loan',
'contact', 'month', 'day_of_week','poutcome']

3. Model

3.1 Decomposition of business problem

To decompose the problem, as is discussed in Section 1, we need to know how banks generate profits through term deposit. Generally, banks reinvest or re-lend the deposit. Given a fixed budget, larger deposit can leverage larger profit. To simplify the problem, we assume the marketing and operating budget is fixed. Costs include costs per customer * total customers reached and fixed costs. To achieve higher deposit within budget, we use direct marketing to prioritize the valuable customers. Customer value is estimated as the expected deposit amount that he or she may contribute, i.e., predicted probability multiplied by predicted amounts. We predict the probability of subscription of each customer, and make assumptions regarding the potential deposit amount based on customer attributes, e.g., social status and financial situation. While we have no intention to quantify the predicted deposit amount in this study, we believe that banks can adjust the expected deposit amount based on our assumptions.

3.2.1 XgBoost

XGBoost (eXtreme Gradient Boosting) is a popular machine learning algorithm that is widely used for both regression and classification tasks. It is an ensemble method that combines multiple weak models (decision trees) to create a stronger model. The algorithm works by iteratively training decision trees on the residuals of the previous tree, thereby gradually reducing the error of the model.

What's more, XGBoost has undergone deep consideration in system optimization and machine learning principles. It is not an exaggeration to say that the scalability, portability, and accuracy provided by XGBoost have pushed the upper limit of machine learning computation constraints. The system runs more than ten times faster on a single machine than the popular solutions at that time, and can even handle data at the billion-level in a distributed system.

XGBoost had following advantages:

(1) Easy to use: Compared to other machine learning libraries, users can easily use XGBoost and achieve fairly good results.

(2) Efficient and scalable: It is fast and effective when processing large datasets and does not require high hardware resources such as memory.

(3) Strong robustness: Compared to deep learning models, it can achieve similar results without requiring fine-tuning.

(4) XGBoost internally implements boosted tree models, which can automatically handle missing values.

XGBoost also has following disadvantages:

(1) In comparison to deep learning models, it cannot model spatiotemporal data

and cannot effectively capture high-dimensional data such as images, speech, and text.

(2) When having massive training data and can find a suitable deep learning model, deep learning can achieve much higher accuracy than XGBoost.

3.2.2 SVM(Support Vector Machine）

We choose the SVM model to classify the data to predict the customer's response of telemarketing. Support Vector Machine is a supervised learning algorithm which applies kernel functions. It classifies for determining the hyperplane which corresponds to the maximum margin between two categories (distance between hyperplane and nearest data points). Support vectors are critical points closest to the hyperplane.

SVM has following advantages:

(1) SVM is a novel small-sample learning method with a solid theoretical foundation. It basically does not involve probability measures and the law of large numbers, so it is different from existing statistical methods. It avoids the traditional process from induction to deduction, and achieves efficient "transductive inference" from training samples to forecast samples, which greatly simplifies the usual problems of classification and regression.

(2) The final decision function of SVM is determined by only a small number of support vectors, and the complexity of the computation depends on the number of support vectors rather than the dimensionality of the sample space, which in a sense avoids the "dimensionality disaster".

(3) The SVM algorithm is simple, using only a few support vectors to determine the final result, which can help us capture key samples and also eliminate a large number of redundant samples.

SVM also has following disadvantages:

(1) SVM algorithm is difficult to apply to large scale training samples

Since SVM solves the support vector with the help of quadratic programming, and solving quadratic programming will involve the computation of a matrix of order m (m is the number of samples), the storage and computation of this matrix will consume a lot of machine memory and computing time when the number of m is large. The main improvements for the above problem are J. Platt's SMO algorithm, T. Joachims' SVM, C.J.C. Burges et al.'s PCGC, Xue-Gong Zhang's CSVM, and O.L. Mangasarian et al.'s SOR algorithm

(2) Difficulties in solving multi-classification problems with SVM

The classical support vector machine algorithms only give algorithms for two-class classification, while in practical applications of data mining, multi-class classification problems are generally to be solved. It can be solved by the combination of multiple two-class support vector machines. There are mainly one-to-many combinatorial models, one-to-one combinatorial models and SVM decision trees; then it is solved by constructing combinations of multiple classifiers. The main principle is to overcome the inherent shortcomings of SVM and combine the advantages of other algorithms to solve the classification accuracy of multi-class problems. For example, it is combined with rough set theory to form a combined classifier for multi-class problems with complementary advantages.

3.2.3 Logistic Regression

Logistic regression is a discriminative model that are commonly used in classification. It is suitable for determining the relationship between a binary dependent and multiple other continuous dependent variables. At the same time, there are many regularization methods for logistic regression. This increases the penalty term, which helps to reduce the complexity of the model and prevent overfitting. And compared with Naïve Bayes and decision tree, logistic regression would learn the importance of individual features through its weight parameters. It is a flexible model to handle nonlinear relationship. Therefore, we choose this model.

The advantage of logistic regression also includes: easy to use and understand, less likely to occur multicollinearity, low computational complexity and fast speed. It's also easy for us to observe probability.

This model also has some disadvantages: 1) Underfitting will result in a decrease in accuracy, 2) Unable to handle a large number of multi-class features or variables well.

4. Performance Evaluation

A brief decomposition of the business problem: Maximize the deposit amounts to leverage higher principal for reinvestment. Making assumptions regarding savable income and predicting subscription probability to allocate budget to high-value customers.

Simply using accuracy or F-measure to evaluate the models is not enough. To align with our objective (maximizing profit), we use cost-benefit matrix to calculate expected profit per customer for each model. We assume the average deposit would be € 10,000. Net rate of return is estimated as 5-year rolling return of S&P portfolio[1] subtracted by deposit interest rate in Portugal [2](both are (weighted average rates for the period May 2008- Oct 2010). Costs per customer are largely variable telephone costs and labor costs, which can be trivial. However, the benefit of retaining a customer and reinvesting the deposit is substantially high. Thus, the opportunity cost of losing a customer is also high. For this reason, we should tolerate false positives but try best to minimize false negatives.

Estimated Rate of Return (Net) = 7%

	Actual Positive	Actual Negative
Predicted Positive	Benefit = 10000*7% = 700	Cost = 3
Predicted Negative	Cost = 700	Benefit = 0

	Accuracy	**F-measure**	**FPR**	**FNR**	**Expected Profit**
XGB	0.8438	0.477	0.1232	**0.4006**	**309.87**
SVM	0.8853	0.2782	0.0203	0.8141	49.78
LR	**0.8956**	**0.3687**	**0.0182**	0.7436	106.35

The results show that XGBoost has the highest expected profit and the lowest false

[1] Data source for return of S&P 500: https://xueqiu.com/8431147782/238855474

[2] Data source for deposit interest rate in Portugal: https://d.qianzhan.com/xdata/details/4e80e1deba8be017.html

negative rate (FNR). Although Logistic regression has the highest accuracy, F-measure and False positive rate, it has low expected profit due to the punishment caused by the high FNR. Thus, XGBoost is the most suitable model for our business scenario. Actually, XGBoost also has the highest AUC score.

5.Conclusion

In our research, we analyzed the distribution of customer deposits based on age, occupation, education, marital status, housing, and loans using visualization techniques. We developed an intelligent decision support system to predict the optimal outcome of using telephone marketing to maximize the bank's revenue from term deposits within a limited budget. We compared the performance of three different machine learning models - XGBoost, SVM, and Logistic Regression - in accurately predicting the success rate (y) of telephone marketing for different customers, with XGBoost performing the best. We then ranked customers based on their average deposit amount that telephone marketing could bring to the bank and assigned priority to customers accordingly. This allowed us to contact customers in order of priority, within a fixed number of calls that could be made under a limited budget, to maximize revenue and profits. This decision support system helps decision-makers identify and prioritize customers with certain characteristics in marketing activities and select the target customer groups, enabling the bank to maximize its resources and achieve business growth.

Reference

[1]Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

[2]Tian, T., Gao, J., Li, X., Zhao, B., Wang, S. and Liu, J., 2020. A novel machine learning model based on XGBoost for improving the diagnosis of prostate cancer. Journal of cancer research and clinical oncology, 146(2), pp.391-401.

[3]Zheng, Y., Zhang, X., Chen, C., Wang, Y. and Zhu, H., 2021. Feature selection and multi-class classification of plant diseases based on XGBoost algorithm. Journal of ambient intelligence and humanized computing, 12(1), pp.563-574.

[4]Wang, Q., Zhao, X., Gao, F. and Yan, L., 2019. A hybrid feature selection method based on XGBoost and particle swarm optimization for credit scoring. Expert systems with applications, 119, pp.42-53.