# Telemarketing Decision: Data Mining and Modeling will help bank make decision

Big Data Analytics Project - F2022

CIND820 D1H

SHUXIN ZHU

501148088

Dec 05, 2022

**Table of Contents**

## Abstract

The main topic of this research is about predicting clients who will have a term deposit by telemarketing strategy.A Portuguese bank marketing dataset from 2008 to 2103 is used during the research.

This dataset included 41187 clients and 20 inputs, like the type of job. marital status, education level, etc. Those attributes could be grouped by basic client data, client attributes, economic context attributes and other attributes. The importest predicting model will be between client information(include basic client data and client attributes) and output Y - client decision. In the other word, this research will find out which client will make a term deposit and which attributes may affect them. The second predict is about economic context attributes. For example, based on daily knowledge, the consumer price index ,CPI, will affect short term deposits. This research will mining and modeling data to provide the if the result will be the same for the Portuguese bank marketing dataset.

The original research “A Data-Driven Approach to Predict theSuccess of Bank Telemarketing” (Moro et al., 2014) is using the model to help a Portuguese bank selling long-term deposits and find a semi-automatic feature filter to reduce the ineffective condition. During the research, S. Moro’s terms also compared four differ data mining models to support the research- logistic regression, decision trees (DT), neural network (NN) and support vector machine.

Another popular topic for this database is describing an implementation of a DM project based on the CRISP-DM methodology for Portuguese marketing campaigns(S.Moro et al., 2011). My research will use some code from above research and compare all modeling.

During my research, some tools will be used to prepare, visualize and predict data. For example, python will be used to prepare and model the data. And some plot will be used for visualization. Such as box plots, histograms, decision trees, and Naive Bayes. At the same time, this research illustrated how the data was prepared and conducted predictive models to provide telemarketing strategy recommendations.

For the data preparation, some attributes are objects. Such as, marital status, education, having credit in default, etc. For preparing object attributes, we usually use numeric to mean different answers. For example, clients have different education levels, we could use one to nine to mean “basic.4y” to “unknown”. Meanwhile, some attributes may be deleted since it is duplicate, like “month”- last contact month of year. This attribute is the same as  duration.

Studying this prediction model not only can help banks to understand which clients can use the telemarketing strategy to sell products but also can help bank marketing teams to predict which strategy will be less expensive.

Another topic of the research is how to mining data when the data is being modeled. Some predicted questions may be answered during the research. For example, what is the strongest relationship between variables? And which attributes will be the main consideration? Those kinds of questions will help predict modeling.

## Introduction

Telemarketing is also called the inside or telesales system of marketing and it is a technique applied to conduct direct marketing involving solicits by marketing representatives aiming at customers in purchasing some services or goods. The mode of conducting telemarketing might be possible through preliminary calls, texts, messaging, and email or through web conferences across the world. This type of marketing is regarded as an efficient strategy in business offering immediate response from a client thereby serving as a cost-effective one, especially whenever there is a need to do advertisements through the telephone. This type of marketing helps many companies significantly reduce their employee and infrastructure costs. Many companies have already managed to improve customer service, increase sales, accelerate payments and collect marketing data with the help of telemarketing (Pride and Ferrell).

Another purpose of telemarketing consists in being able to sell products and services to clients in other sales territories: on the local or national level. This type of marketing also helps to follow up with existing clients. Staying in contact with a client enables you to find out more about one's needs and thus improves customer satisfaction, which in turn can have a significant influence on the profitability of the business.

However, telemarketing has several disadvantages that question its effectiveness. Despite the sales value of telemarketing being high, the majority of people associate it with a negative image. Bad experiences of clients with telemarketers create the negative stereotype that makes unwanted calls extremely annoying. Expensive customer lists are often out of date and do not result in many actual sales (Miller, 2004). Another disadvantage of telemarketing is the increased number of frauds. According to Klosek in 2003 people lost approximately $ 40 billion in telemarketing fraud. Thus the Government is annually implementing tougher policies to curb telemarketers Today the connection between telemarketing and customer service is not as obvious as before.

Meanwhile, the telemarketing industry is decreasing. The young generation is not long interested in the call service. All market teams will focus more on the online or retail market. In the other words, the database of telemarketing will be biased for the sample selection in some input data like age, employee, and marriage. After research, we will have the result of affects,  if the biased for the data selection since the telemarketing industry,

Customer service or customer relationship management is a model for managing interactions with current and future clients of the company. It involves using technology to organize sales, marketing, and technical support. Today, “telemarketing” is the word that barely leaves an individual`s lips before one starts cringing. Indeed, telemarketing is a type of customer relationship management, as it serves to increase the client's knowledge about the products and services offered by the company. Despite the ever-present IT technologies, a live phone conversation is much better and faster in resolving objections, answering questions, and getting answers.

The prediction from my research is also about economic context attributes. For example, based on daily knowledge, the consumer price index, CPI, will affect short-term deposits. This is similar to a study that was conducted by Khatibi et al., (2002) in which case perception has been highly majored especially concerning the perception of consumer networks on the quality of service offered. The only difference that would be deduced from the paper on Moro et al., (2014) is that it is a term deposit set while this one is done in almost a minute’s response. However, with due focus on the validity of this study, there will be an alternative or rather definitive demand for the specific data sets contained in dissimilar.

Therefore, this research will focus on mining and modeling data to provide reliability to prove if the result will be the same for the Portuguese bank marketing dataset following an extensive data mining process, particularly on the input variables, socioeconomic attributes, and the predictive model. This data will be as well available for access and additionally important for supplementing the previously gathered data in Moro et al., (2014) either way. My preference would be based on an understanding of the predictive model and how data mining could be done alongside comparing it with the current information.  Base on Moro et al., (2014) research, they already compared four data mining models: Logistic regression, decision trees, neural network and support vector machine. During my research, I also will compare these four data mining models with the CRISP-DM methodology which from research"Using Data Mining for Bank Direct Marketing: An Application of The Crisp-DM Methodology "(S.Moro et al., 2011).

Studying this prediction model not only can help banks to understand which clients can use the telemarketing strategy to sell products but also can help bank marketing teams to predict which strategy will be less expensive. The prediction comes along with the properties that effectively define the model and associated metadata including name and description, processing date for the latest of times, and data filters before it is held for purpose reasons. This reasoning aligns with the similarity in Gabriel and Kanzanjian's strategic implementation of improvement of processes and material structures in organizations. This coupled with implementation as a key state of controlling the strategy process which at times might lack direction and despite being neglected at times, might serve as a core determinant of performance slowly (Galbraith and Kazanjian, 2006).

Meanwhile, based on the research from Moro et al., (2014), the main reason for the Portugues banks to focus on the optimizing targeting for the telemarketing is because that the 2008 financial crisis gave the bank more pressure to increase margin and deduce the expense. Nowadays, the Canadian economy faces the same situation, and the effect of depression will not only affect the banking industry, but also will affect the financial related industry like insurance, cryptocurrency industry, and B2B market. This model may help those industries find the solution.

On the other hand, telemarketing strategies research also will be a good development for the data modeling and mining study. This topic will help data analysis to develop the skill of modeling and analysis strategy.

In summary, once a choice is made on the data mining process, the subsequent steps will include the access, extraction, integration, and preparation for the right data set to be used in data mining to understand the customer needs and demographics before a telemarketing strategy is instituted and implemented. input data needs to be provided in the correct amount, structure, and a suitable format aligned to the modeling level of the algorithms. This perspective in regard describes the generalized thought under which there is a need to express modeling data whilst at the same time describing the major data-cleaning operations that need to be performed. Additionally, this allows for an intensive description of the ways of data exploration before modeling and cleaning up. From a standpoint of a database, data as a body might be termed as being clean but from a data mining perspective, several problems have to be fixed such as missing data. When data misses, the development of adequate consideration of the sufficiency of a study becomes jeopardized. For a data manager, missing information does not remain a central aspect but rather a huge problem for the data miner.

The essence of appropriate data mining will offer an opportunity for efficient telemarketing strategies that have only to come along with strong engagement in the strategy development phase. However, in the data mining process, more reliable data and not missing data needs to be counted as having a positive correlation in the final deduction of the general reverence of the possibilities presented. To strategic telemarketing, data mining is the indispensable due correlation to the conceptual significance of data mining to business analysts. This process entails finding appropriate relationships with data processing, result interpretation, and decision making. Appropriate decision making comes in as a logical attribute that calls for an understanding that it is both a mathematical and scientific instruction in developing and thinking about telemarketing appropriateness and probably activity initiation and in our case, we might need someone who has attributes of Machine learning, neural networks, and automation.

## Methods

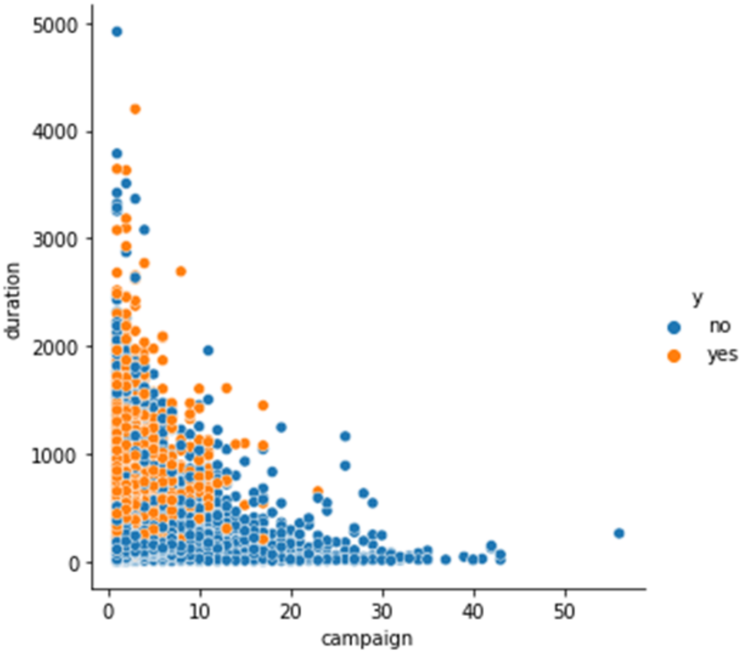
Data mining is the study of data as the object, through a series of steps of data acquisition and analysis, data pre-processing and data modeling to explore the potential information in it. Acquiring the target data is the first step of data mining, and the datasets in this chapter are from the Bank of Portugal telemarketing dataset from the UCI website (http: //archive.ics.uci.edu/ml/datasets). The data analysis allows to obtain the distribution of each feature and the attribute characteristics of different customer segments, which provides auxiliary suggestions for bank telemarketing. Data preprocessing is a preparation for data modeling and different formats of data require the use of different pre-processing methods, especially for unstructured data, which can serve to normalize the data form. This chapter covers two main parts: data acquisition and analysis and data pre-processing.

In industry, it is the data and features that determine the upper limit of machine learning effectiveness, while models and algorithms are only approaching the upper limit infinitely. It is clear that good data is the key to good or bad modeling results. Therefore, when experimenting with data, it is important to be able to find a high quality data set. There are many ways to obtain data, such as using publicly available datasets, using crawlers to get valuable data, collecting data from data trading platforms, using paid APIs to buy data, simulating data experiments, etc. The dataset for this paper is from the open source website UCI and is selected from data related to a marketing campaign conducted by a local banking institution in Portugal. A marketing campaign is the use of telephone calls to one or more telephone contacts to confirm whether a customer will subscribe to a product (bank term deposit). This experimental data has a total of 41,188 items, including 20 features and 1 label, with the classification objective of predicting whether a customer will subscribe to a time deposit service (variable y), corresponding to the task of classification. There are 36,548 items with "no" data label, accounting for 88.7%, and 4,640 items with "yes" data label, accounting for 11.3%.

## Characteristics and meaning of data set

This experimental dataset has a total of 20 features and 1 label. Among the 20 features, half of the variables are category-based and the other half are numerical. All features are considered in terms of both customers and banks, and contain four main blocks, which are: basic customer information, customer financial information, marketing result information and banking industry information. Among them, age, job, marital, education belong to basic information of customers; default, housing, loan belong to financial information of customers; contact, month, day_of_week, duration.

In order to observe the data volume distribution of the two types of labeled data, a scatter plot of the data set with respect to campaign and duration is drawn, as shown in the figure, with the horizontal coordinate being the campaign value of sample xi and the vertical coordinate being the duration value of sample $x_i$. The vertical coordinate is the duration value of sample $x_i$. It is obvious from the data visualization that for the label "y", "no" (i.e., users will not subscribe to time deposit service) is the majority in the dataset; "yes" (i.e., users will subscribe to time deposit service) is the majority. (i.e., users will subscribe to time deposit service) are in the minority, and the dataset is unbalanced.



Scatter plot of bank telemarketing data (in campaign-duration)

Next, the data features are analyzed separately in terms of numerical features and categorical features.

**Numerical characterization**

The experimental dataset has 20 features, 10 of which are numeric: age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, ons.conf.idx, euribor3m and nr.employed characteristics. Numerical characteristics of customers who order time deposit business (i.e. successful customers) are analyzed.

(1) Customer age characteristics are mainly distributed between 31 and 50 years old, presumably because people in this age group need more income to get married, buy a house, pay for their children's education, etc.

(2) The overall call time of the customer duration feature is long, with a median of about 450, indicating that the customer may be interested in the product.

(3) The minimum value of customer CAMPAIGN characteristic is 1 and the maximum value is 23, which indicates that the stronger the willingness to order bank products, the more exchanges customers have in this activity, and customers who have ordered products in this activity have had at least one exchange.

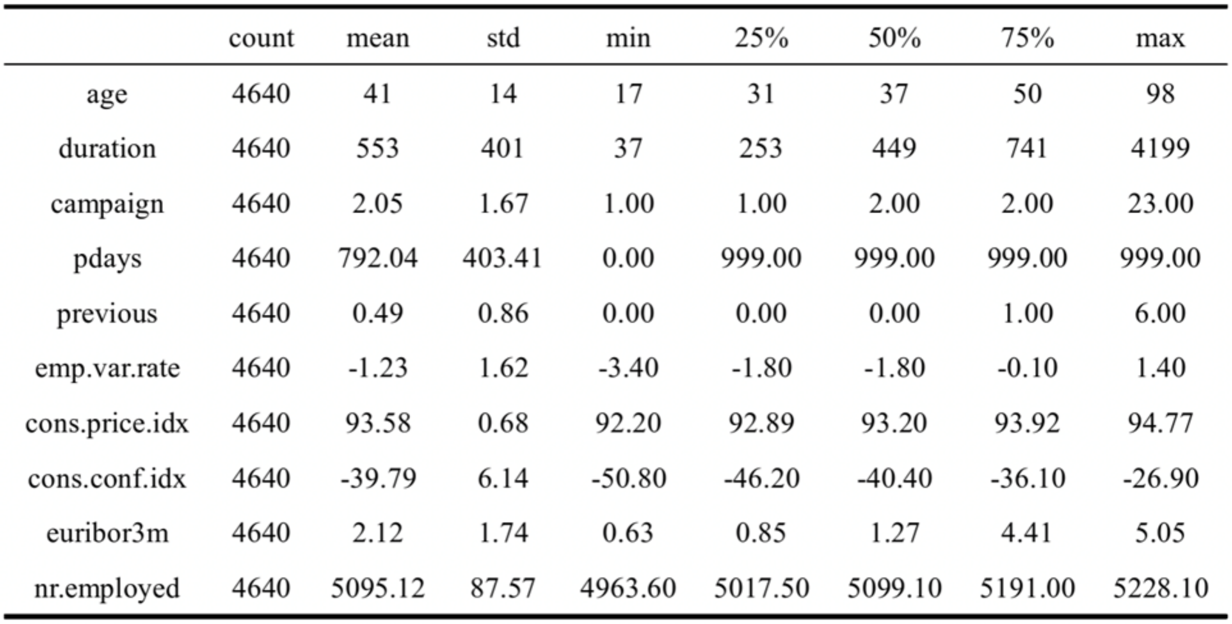
(4) The mean value of customer pdays characteristics is 792.04 and the ¼ quantile is 999, indicating that only a small number of customers have been contacted since the last campaign and most of them are uncontacted. This indicates that regular customers will just take the initiative to order products even if they are not contacted frequently.

(5) The mean value of customer previous characteristics is 0.49, with quartiles 1/4 and 2/4 being 0 and quartile 3/4 being 1.00. This indicates that customers who would have purchased the

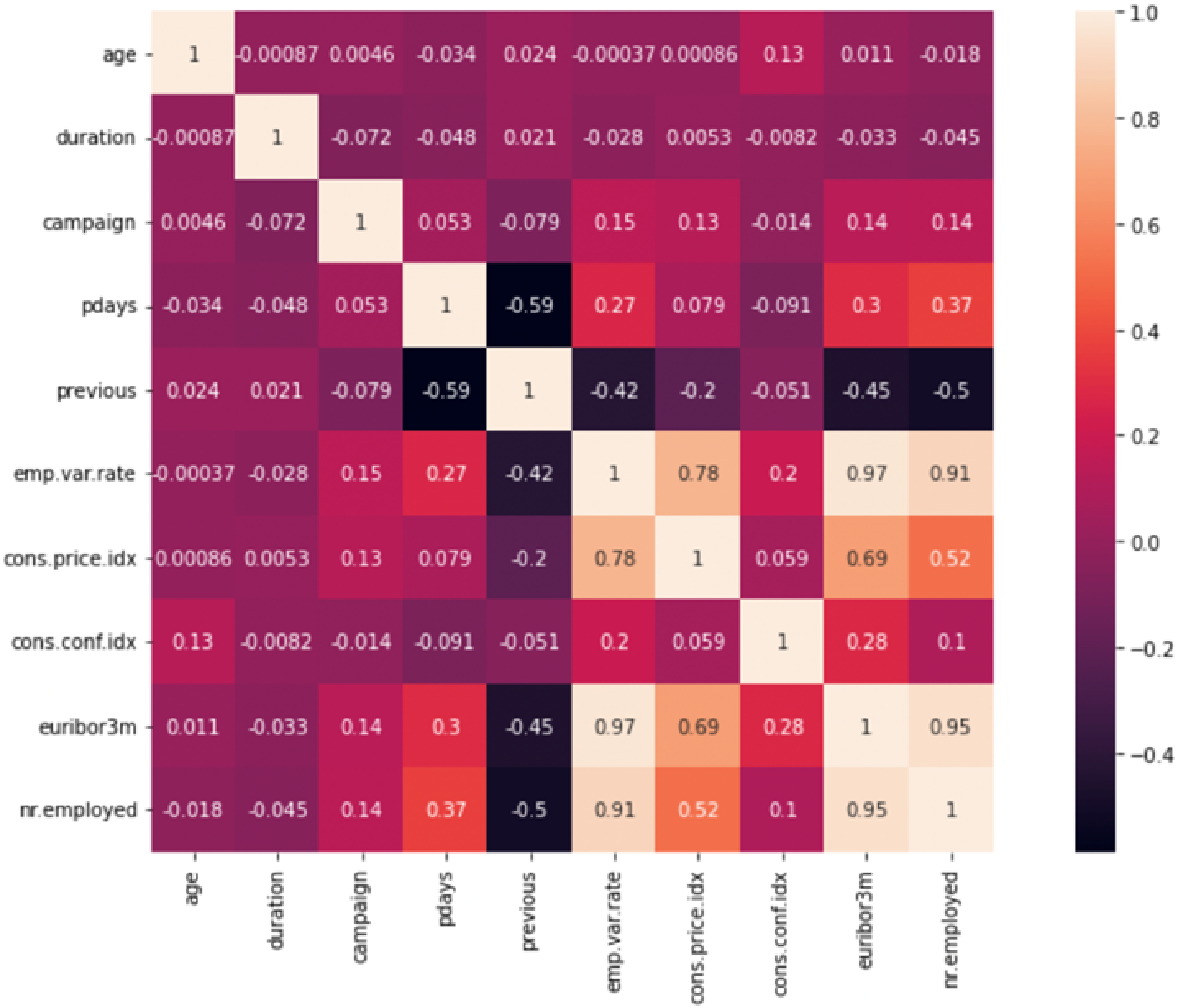
product would have communicated less often before this activity, presumably because they were not interested in purchasing bank products at the time.

(6) The distribution of the five characteristics of customers' emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, and nr.employed is more stable and the extreme differences are small, which is related to the socioeconomic state at that time. Except for the negative values in the emp.var.rate and cons.conf.idx features, the other three features are positive.

Description table of numerical characteristics of customers who order time deposit business



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| age | 4640 | 41 | 14 | 17 | 31 | 37 | 50 | 98 |
| duration | 4640 | 553 | 401 | 37 | 253 | 449 | 741 | 4199 |
| campaign | 4640 | 2.05 | 1.67 | 1.00 | 1.00 | 2.00 | 2.00 | 23.00 |
| pdays | 4640 | 792.04 | 403.41 | 0.00 | 999.00 | 999.00 | 999.00 | 999.00 |
| previous | 4640 | 0.49 | 0.86 | 0.00 | 0.00 | 0.00 | 1.00 | 6.00 |
| emp.var.rate | 4640 | -1.23 | 1.62 | -3.40 | -1.80 | -1.80 | -0.10 | 1.40 |
| cons.price.idx | 4640 | 93.58 | 0.68 | 92.20 | 92.89 | 93.20 | 93.92 | 94.77 |
| cons.conf.idx | 4640 | -39.79 | 6.14 | -50.80 | -46.20 | -40.40 | -36.10 | -26.90 |
| euribor3m | 4640 | 2.12 | 1.74 | 0.63 | 0.85 | 1.27 | 4.41 | 5.05 |
| nr.employed | 4640 | 5095.12 | 87.57 | 4963.60 | 5017.50 | 5099.10 | 5191.00 | 5228.10 |



A correlation coefficient plot is a statistical graph that describes the strength of correlation between two variables. If the correlation coefficient graph is color-coded in addition to the coefficient values, the graph is called the correlation coefficient heat map. The correlation coefficient heat map of numerical variables for customers who order time deposit business (i.e., successful customers) is shown below. The darker the color, the higher the correlation coefficient value between the variables is close to 1. Conversely, the lighter the color, the lower the correlation coefficient value between the variables is close to 0. As can be seen from the following graph, euribor3m and emp.var.rate, euribor3m and nr.employed are positively correlated, with correlation coefficients above 0.9. Therefore, it can be concluded that there is a strong correlation between the euro three-month rate - daily indicator (euribor3m) and the rate of change in employment (emp.var.rate) and the number of employees - quarterly indicator (nr.employed).

**Data preprocessing**

Data preprocessing is the process of checking, deleting or correcting abnormal data. The purpose of data preprocessing is to change the form of the data to fit the model and match the needs of the model. After testing, this paper's bank telemarketing data in this paper are divided into numerical and categorical features, and there are no missing values and duplicate values, so There are no missing values and duplicate values, so there is no need to perform missing value processing and de-duplication processing. In the following, the numerical features processing, categorical features processing, In the following, we will do data pre-processing from four aspects: numerical feature processing, categorical feature processing, data dimensionless processing and unbalanced data processing.

**Category type feature processing**

In real life, not all features are numeric, some features may be category type data. The original input of category-type data is usually in the form of strings, such as the presence or absence of mortgage (yes/no), occupation type (student/entrepreneur/unemployed, etc.). We generally transform the category-type features into numeric features by performing ordinal encoding or dummy variable operations.

Serial number encoding is used to handle data with size relationship between categories, such as education (middle school, high school, college), which can be converted into a number with size relationship (1, 2, 3) by serial number encoding to continue to maintain the size relationship between categories. However, when there is no size relationship between the values of the category characteristics, such as various occupations: student, entrepreneur, manager, other, the ordinal number coding cannot be used and dummy variables should be used.

Also known as dummy variable, dummy variable or nominal variable, dummy variable takes only two kinds of values in the variable: 0 or 1. It is usually used to deal with variables that do not have a size relationship. For category-based variables with n categories, n-1 values are generated. A dummy variable with a reference taking n-1 zeros is the nth category characteristic value of that variable, and a category with a specific meaning is usually selected as the reference. Take the above occupation as an example, there are 4 categories: student, entrepreneur, manager, and other, and set the category of "other" among these 4 occupations as the reference, i.e., the value of each dummy variable is 0. 100, the dummy variable for "entrepreneur" is 010, the dummy variable for "manager" is "001", and the dummy variable for "other "The experimental data in this paper has a total of 10 categorical features, all of which need to be processed and converted into numerical features. The conversion rules are as follows.

The experimental data in this paper have a total of 10 category-based features, all of which need to be processed and converted into numerical features.

The conversion rules are developed as follows.

(1) Since married customers are more willing to order products, the marital features {"married", "single", "divorced ", "'unknown"} correspond to the mapping into {4, 3, 2, 1}.

(2) Since the education feature belongs to ordered variables, the education is coded with ordinal numbers, i.e. {"university.degree", "high.school", "basic.9y", "professional.course", "basic.4y", " basic.6y", "unknown", "illiterate", } corresponding to the mapping into {8, 7

6, 5, 4, 3, 2, 1}.

(3) Since customers without defaulted loans are more willing to order products, the three features of default, housing, and loan {"no", "unknown", "yes "} correspond to mapping into {3, 2, 1}.

(4) Since customers who order products prefer cellular as a contact method, the {"cellular", "telephone"} corresponding to {2, 1} in the contact feature is mapped to {"cellular", "phone"}.

(5) Since customers are more willing to order products if the result of the last activity is good, according to the result of the categorical feature analysis, the poutcome features {"success", "nonexisten", "failure"} corresponding to mapping into {3, 2, 1}.

(6) Remove marital, education, defaul, housing, loan, contact and poutcome remove the features of marital, education, defaul, housing, loan, contact and poutcome, and the three category-based features of job, month and day_of_week are operated by dummy variables.

**Data dimensionless**

The dimensionalization of the data can eliminate the influence of the dimension on the final result, normalize the different features to [0, 1] between [0, 1], which speeds up the convergence of gradient descent and reduces the training time of the model. The dimensionless normalization of data is generally divided into min-max normalization and z-score normalization.

### Min-max normalization

Min-max normalization, also called outlier normalization, is a method of invariant stealing that maps the original data to between [0, 1] through a linear transformation. min-max normalization is given in Eq. However, since min-max normalization is only applicable to traditional accurate small data scenarios, and the min and max values of feature values are very susceptible to outliers, it will lead to results that will be worse and less robust.

$$x^{'} = \frac{x - min}{max - min}$$

where *x* is the original data, min is the minimum value of the sample features, and max is the maximum value of the sample features.

## Telemarketing modeling study

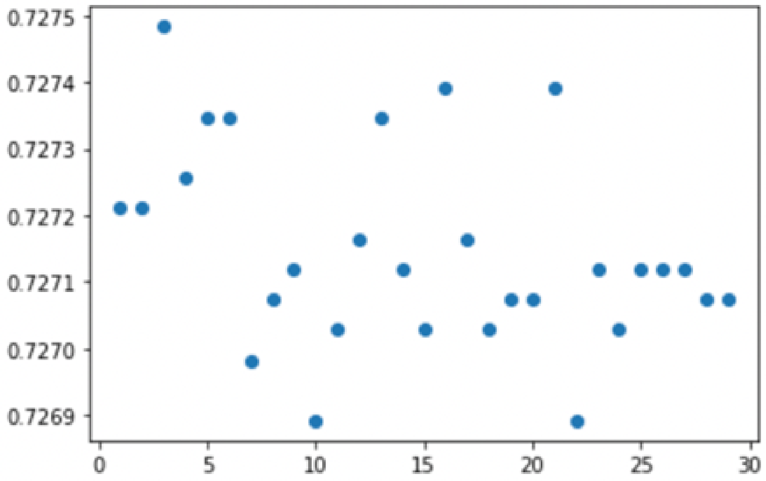
Data modeling is performed using already processed data, and the program used is a python program. In order to maximize the success rate of marketing calls, this chapter first uses the unsampled training set for training, and derives the prediction results of each model and analyzes them; then uses the sampled training set for training to find out the best sampling strategy and the optimal model under this sampling strategy. The optimal model under this sampling strategy is identified. In addition, the test set samples with incorrect predictions of the optimal model are analyzed to classify the prediction results under the optimal sampling strategy into customer segments and provide guidance for bank telemarketing.

This section is about model training using the training set that has not been sampled from the data. Since each model has different learning parameters, we use both learning curve and grid search to find the best parameters, and then substitute the best. The best parameters are then substituted into the model for training, and the trained model is used for test set prediction to obtain the prediction results. The five measures of model prediction The five measures of model prediction results are F1 value, KS value, G_mean value, accuracy accuracy and AUC The optimal classification model is selected by considering the magnitude of the five measures and plotting the ROC curve. The optimal classification model was selected by combining the five measures and plotting the ROC curve.

**Analysis of logistic regression modeling results**

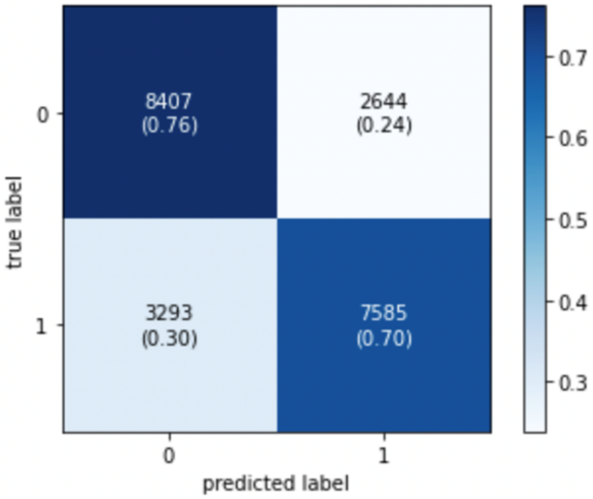
Penalty is a parameter that penalizes the loss function, which is used to limit some of the parameters of the loss function in order to suppress overfitting of the model. "l1" refers to the L1 regularization, which can make the value of some parameters become 0 and generate a sparse weight matrix for feature selection and The "l1" refers to the L1 regularization, which can make some parameters become 0 and generate a sparse weight matrix for feature selection and overfitting prevention; "l2" refers to the L2 regularization, which can make the parameters' values as close to 0 as possible instead of 0 and is used to prevent overfitting. The default value of penalty in the model is "l2", i.e., L2 regularization. max_iter is the number of iterations of the model, which can be set artificially, or the default value of 20, i.e., the model will stop training after 20 iterations.

In this paper, the learning curve is used to learn the values of the penalty and max_iter parameters. The learning curve is shown below, and the vertical coordinate of the learning curve takes the accuracy value because the training data is unbalanced at this time. The horizontal coordinate is the number of iterations parameter max_iter, which is an integer value between [1, 20], and the vertical coordinate is the zhun que l value, i.e., as the number of iterations max_iter.



Regression Accuracy

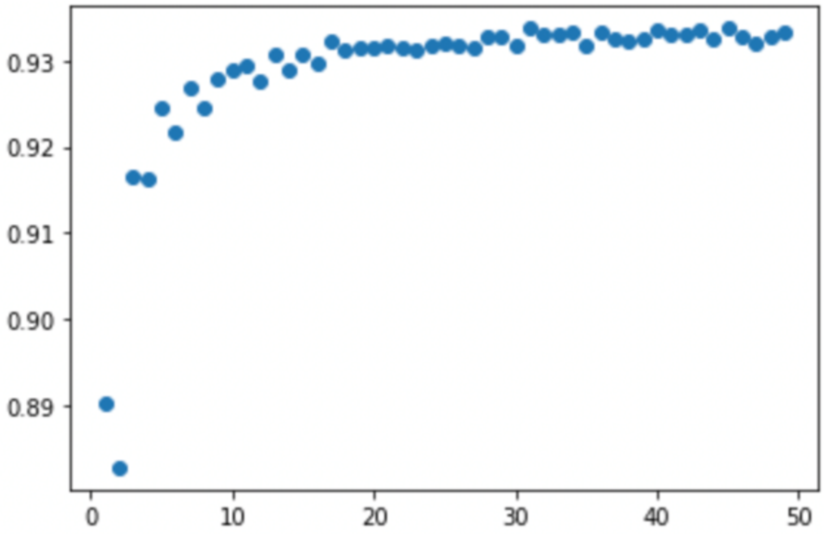
The selected parameters are substituted into the model for training and then the test set prediction is performed, and the confusion matrix of the test set is obtained as shown in the table.

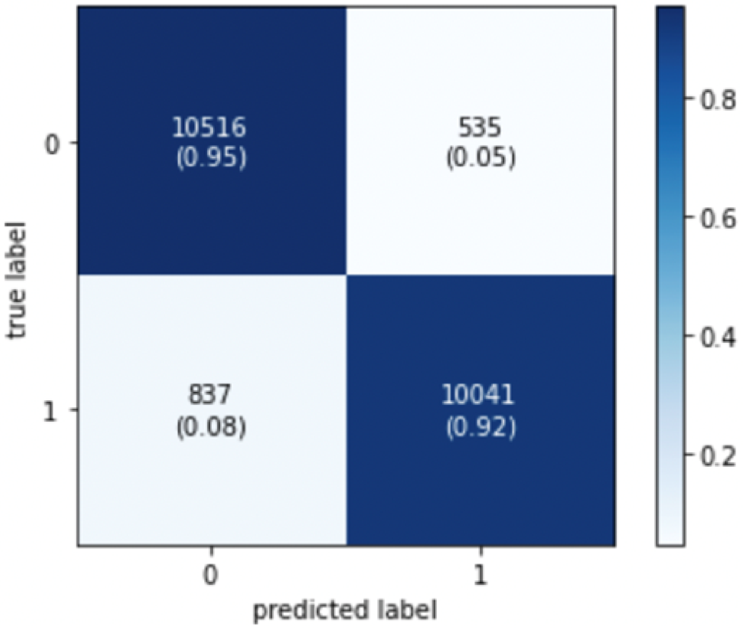


The accuracy of the final result is 0.729111-Precision: 0.498、Recall score: 0.514.

**The random forest modeling regression**

The random forest modeling regression is shown below.





Precision: 0.498, Recall score: 0.514.

### K- Nearest Neighbor regression algorithm modeling

KNN (K- Nearest Neighbor) method, originally proposed by Cover and Hart in 1968, is one of the simplest machine learning algorithms and belongs to the classification algorithm of supervised learning. The idea of the algorithm is simple and intuitive: the classification problem: if a sample belongs to a class if most of the K most similar (i.e., most neighboring) samples in the feature space belong to that class, then the sample also belongs to that class.

The KNN formula is shown below：

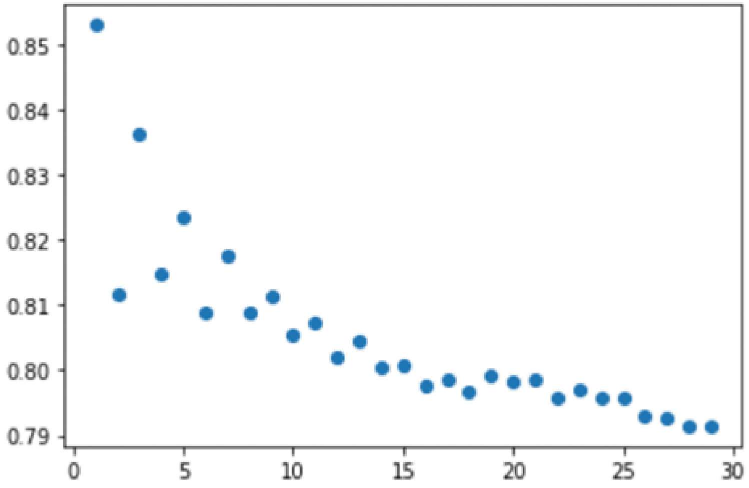
$$d(x,y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} :$$

KNN is a non-parametric, inert algorithmic model. What is non-parametric and what is inert? Non-parametric does not mean that the algorithm does not require parameters, but it means that the model does not make any assumptions about the data, as opposed to linear regression (which we will always assume is a straight line). This means that the structure of the model built by KNN is determined by the data, which is more in line with reality, after all, in reality the situation often does not match the theoretical assumptions.

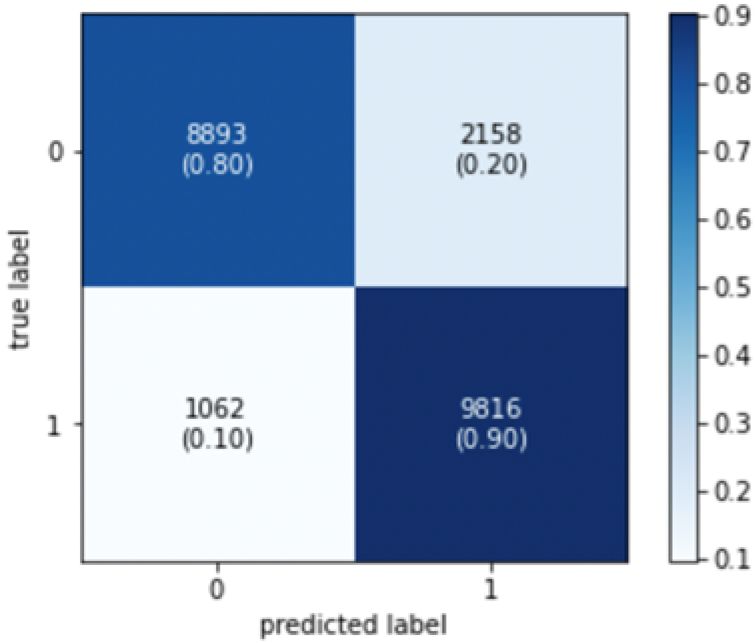
What does inertia mean again? Consider that the same classification algorithm, logistic regression requires a lot of training of the data before finally getting an algorithmic model. The KNN algorithm, on the other hand, does not need to, it does not have an explicit process of training the data, or the process is very fast.

An algorithmic model will be obtained. The KNN algorithm does not need to, it does not have an explicit process of training data, or the process is very fast.

To summarize, the KNN model is violently calculating the distance between the predicted values and the data in the model, and then classifying by the nearest data.



The final result: 0.852795，The confusion matrix is shown below：



Precision: 0.893，Recall score: 0.805

### Result Understanding

An article about machine learning from Marco Santos mentions that recall would be the next way to detect if an apple was poison- we need to focus on how many poison apples may have rather than being too concerned with missing. Same for this study, Precision, means when the bank knows the group of people who will buy the product, who will be the TRUE buyer. Recall means that the modeling will be used to predict how many people will be successfully buying the production in the group of clients. Precision is an easy way to make a decision for modeling. Compared to the precision and recall, banks should choose KNN for the final decision.

## Conclusion

With the rapid development of the Internet and the advent of the era of big data, a variety of industrial data in a substantial increase, how to quickly and accurately analyze data, extract useful information and value realization, is the theme of the era of big data. Telemarketing as a means of marketing, the amount of data is generally large and highly prone to imbalance, traditional data analysis methods have limited capacity and poor processing results. For this reason, this paper proposes telemarketing based on data mining: on the one hand, to improve the data imbalance phenomenon from the data level, on the other hand, to improve the model prediction effect from the algorithm level. The main work done in this paper is as follows.

(1) In-depth analysis of the current situation of telemarketing at home and abroad and the transformation of marketing methods. By comparing traditional bank telemarketing

with bank telemarketing in the era of big data, it is concluded that bank telemarketing based on data mining can help banks achieve digital transformation, improve marketing success rate and acquire potential customer characteristics.

(2) Telemarketing examples were analyzed. There are 41,188 experimental data in this paper, and by analyzing each feature of the data, the following characteristics of successful customers are obtained: generally young and middle-aged people aged 31 to 50; with high school education or above; more stable jobs, such as technicians, administrators and other occupations; married people with stable marital status; no bad records, such as defaulted loans; light economic burden, no loans and mortgage and other economic pressure; prefer cellular this contact. When conducting bank telemarketing, these target customer characteristics can be supplemented to increase the success rate of telemarketing.

(3) Classification models in data mining are investigated, comparing the effects of unsampled and sampled training sets on model prediction results. In this paper, the principles of logistic regression, random forest, and KNN algorithms are introduced and the models are applied. After sampling the training set, KNN is the best sampling strategy.

Telemarketing as a traditional marketing method, although the impact of the Internet, still has certain advantages. The data mining-based bank telemarketing proposed in this paper can help banks achieve accurate marketing and improve. This paper proposes a data mining-based bank telemarketing that can help banks achieve accurate marketing and increase the success rate of marketing. Due to the time, the research in this paper is not perfect and there are shortcomings, mainly in the following aspects. It is mainly reflected in the following aspects.

(1) Lack of data set. The lack of data set here does not refer to the amount of data, but to the lack of real customer data set of bank telemarketing. The experimental data in this paper comes from the public data set on UCI, which has fewer features and is older, and the conclusions drawn may not be applicable to the current bank telemarketing. If domestic banks can provide real customer datasets, they can definitely provide useful suggestions for bank telemarketing by means of data mining.

(2) Improvement of data processing method. Due to the low dimensionality of the experimental dataset features, this paper only processed the existing features. In the future, new features can be added through feature creation, and then important features can be filtered from the new features for data modeling.

(3) Handling unstructured data with new techniques. Bank telemarketing is to use telephone as a medium for voice If conditions allow, speech technology can be used to extract key information directly from the audio, such as If the conditions allow, we can use speech technology to extract key information directly from the audio, such as the tone and mood of the customer, to improve the success rate of bank telemarketing.

All code is include in https://github.com/LilithZz/CIND820

## References

S. Moro, P. Cortez and P. Rita(2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31

S. Moro and R. M.S. Laureano(2011). Using Data Mining for Bank Direct Marketing: An Application of The Crisp-DM Methodology.

Becker, D. & Becker, P.B (1994). Customer Service and the Telephone. *Business Skills Express.*

Galbraith, J. R. & Kazanjian, R. K. (2006). Strategy Implementation, Structure Systems and process, St 12 Paul, MN: West Publishing.

Klosek, J. (2003). The legal guide to e-business. *Westport*, CT: Praeger

Marshall, J.J. & Vredenburg, H. (1991). The roles of outside and inside sales representatives: Conflict or cooperation? *Journal of Direct Marketing*. Volume 5, Issue 4.

Miller, J.B. (2004). Mainstream Marketing Services. Federal Trade Commission: Resources and Legal Analysis. Retrieved from: ReclaimDemocracy.org.

Zajas, J.R. & Church, O.D. (1997). Applying Telecommunications and Technology from a Global Business Perspective.

Santos, M. (2020, May 18). Precision or Recall: Which Should You Use? The Differences among Precision, Recall, Accuracy, and F1 Score. Towards Data Science.

https://towardsdatascience.com/explaining-precision-vs-recall-to-everyone-295d4848edaf