Potential Customers Prediction in Bank Telemarketing

Final Year Project Final Report

Phung Thai Duong

Khuat Duy Bach

A thesis submitted in part fulfilment of the degree of BSc. (Hons.) in Computer Science with the supervision of M.S.E Le Dinh Huynh

Bachelor of Computer Science

Hoa Lac campus - FPT University

21 April 2022

Project Specification

Our team consists of 2 members:

- Phung Thai Duong – HE140170
- Khuat Duy Bach – HE140665

We are both studying in the computer science department at FPT University. With the initial orientation of Assoc. Prof. Phan Duy Hung and with the direct support of M.S.E Le Dinh Huynh, we have done this thesis for over 15 weeks.

When writing this thesis, our most important goal is to learn and apply some specialized knowledge of computer science that we have acquired during our study at FPT University to solve real-life problems. Besides, we also want to explore the implementation process of scientific research and publish a research paper.

After discussing and receiving advice from the mentors, we have chosen the topic of predicting the bank's potential customers based on telemarketing data. Although we still have limited knowledge and experience in research, with the guidance of lecturer Huynh, we can carry out this thesis step by step. After three months of diligent work, our team has conducted a few experiments and obtained favorable results. Our article - Potential customers prediction in Bank TeleMarketing was published and presented at the ICDSA 2022 conference.

During the whole working process, each team member learned new knowledge and applied it to research. We also know the structure and operation of doing a thesis. Fortunately, we had a chance to present our research at an international conference with friends all around the world. Not only learned through the working process we learned from others through the conference. Furthermore, skills such as teamwork, scheduling, division of work for each member, and presentation are also valuable lessons for us.

From a technical perspective, we have learned a lot from imbalanced data preprocessing, features selection, and data encoding methods to get the necessary information for training such as response encoding, one-hot encoding, and Laplace smoothing. The high dimension data visualization is also one of the extreme must-have steps to understand more about how data distribution so that we can choose suitable models for the research aim. Machine learning and data mining models range from simple to complex as expend to their variations, as well as the importance of choosing the appropriate parameters for the data. Metrics to estimate the efficiency of our research also play an important role in how to evaluate the machine learning process.

From a research perspective, we found that the binary classification problem is not only one of the most common problems in life but also one of the banking problems. The selection of features for problem-solving is not an easy task. The search and study of datasets with suitable features for the purpose should be based on the correlation between the features of the data. In the case of poorly correlated data, it may cause misdirections during model training. During our research, we realized that sometimes using simpler training models is more effective in predicting than models with computational complexity. Reference to previous studies is essential to avoid this situation is

a necessary step. Applying the ideas of previous studies has helped guide us in the right direction. We hope this study of ours will contribute to new research in the future as well as expand the effectiveness of previous research.

In the business view, the success of telemarketing can be very low. Although the expense of telecommunications services is not a problem, the cost to pay the caller and the time wasted on non-potential calls are pretty expensive. Customers are not always ready to listen to the calls even though a customer hangs up they might not be interested in the sale product. Therefore, the number of failed calls often takes up the bulk of this work. The prediction of finding potential customers in the telemarketing system is necessary and this is also the most different need of telemarketing compared to other forms of marketing. To solve this real-life business problem, we decide to take research Potential Customers Prediction in Bank Telemarketing.

The process of doing this thesis has brought us a lot of knowledge, which has achieved our learning goals. With the experimental results obtained in this thesis, our team hopes to improve it in the future and bring this problem to reality at a particular credit enterprise.

Acknowledgement

We would like to thank our instructor, Le Dinh Huynh for his patience and time, and for instructing and advising us enthusiastically.

We would like to thank all at my University, FPT, for giving us the best environment to study and grow over the years.

We would like to thank our friends in CS1402, for letting us meet amazing people and learn a lot from them.

We always remember our family's encouragement and support. Thanks to them, we have the will, the energy, and the confidence to pursue our goals.

Abstract

Data mining plays a vital role in the success of direct marketing campaigns by predicting which leads subscribe to a term deposit. This thesis is accomplished to illustrate with practical mining methods that the data is related to a Portuguese banking institution's direct marketing campaigns (phone calls). The algorithms are used: K-Nearest Neighbor, Logistic Regression, Linear Supported Vector Machines, and Extreme Gradient Boosting to classify potential customers for long-term deposits finance products. Response coding is used to vectorize categorical data while solving a machine learning classification problem. Accuracy and AUC scores are key metrics to evaluate performance. We inherited selecting important features from previous research. Our thesis employed a better method by combining response coding techniques with practical algorithms in an unbalanced dataset. The best prediction model achieved 91.07% and 0.9324 of accuracy and AUC score, significantly higher than the prior of 79% and 0.8 respectively.

Keywords: Bank Telemarketing, Data Mining, Response Coding, K-Nearest-Neighbor classifier.

Table of Contents

1 Introduction

In the finance sector, marketing is a tool that helps commercial banks effectively distribute and use the money of individual customers and corporate customers. The goal can be achieved in both indirect marketing (mass marketing) and direct marketing (one-to-one contact) [1]. Direct marketing has shown to yield better results than mass marketing since it provides organizations with better interaction with both current and prospective customers [2].

Telemarketing is a form of direct marketing. In the modern economy, telemarketing still plays a significant role in marketing. With the development of technologies, telemarketing can be taken in face-to-face or formal calls with a low fee. However, with telemarketing campaigns, the success rate of these calls is mostly very low. Because of this, in order to reach the desired number of customers, banks have to waste a large number of calls, operator costs, and call fees. Therefore, reducing the number of failed calls is an important task. From there, the bank can maximize profits. On the other hand, refusing to buy a bank credit package shows that customers do not need it at that time. Minimizing marketing to the wrong audience will prevent customers from receiving meaningless calls.

Our work applies machine learning techniques to classify potential customers through their personal information and favorite. With the customer's personal information, we will predict whether that customer has the potential to use the bank's credit package or not. This personal information is selected for a purpose and is an important factor influencing customers' decisions.

In this thesis, we are focusing on:

- Mining in unbalanced data.
- Encoding category features.
- Model calibrating to acquire the best performance.

The structure of the following parts of this thesis is as follows: Some solutions and the results of other studies was described in section 2. Section 3 provides a generalization about the dataset, followed by the data preprocessing steps. Some machine learning models and their performance are explained in section 4. Conclusion and perspective will present in the last section.

2 Review of Literature

According to research by Ghoddusi et al. (2019) [3], from 2005 to 2018, more than 130 studies were presented that applied machine learning to finance. This study aims to apply machine learning techniques to predict potential customers for selling banking finance products through telemarketing. The data in this study will consist of most related 20 features extracted from 150 the original data from A Portuguese retail bank between 2008 and 2013[4].

In most customer data sets collected by banks, the number of customers who agree to use credit packages usually accounts for a minimal number. As a result, these datasets are imbalanced. This issue is important in model evaluation, as shown in the study by Miguéis et al. (2016) [5] or mentioned in a research paper by Zhang et al. (2015) [6]. Thus, one measure of effectiveness is the AUC, which is independent of the frequency of class or specific false positive/negative data by Martens et al. (2011) [7].

A common approach is to use data mining. According to research by Moro et al. (2014) [4], they used a dataset of 52944 customer calls with 150 attributes obtained by a Portuguese retail bank between 2008 and 2013. With the application of a semi-automatic feature selection to reduce to 22 features and compare the results of 4 data mining models such as Logistic regression, Decision trees, Neural network, and Support vector machine, the best result obtained is AUC = 0.8 with neural network model. A similar study, combining data mining and the Decision tree model of Amponsah and Pabbi (2016) [8], gave very good results with a ROC value of 0.925. Also, in the bank customer classification problem, Kozak and Juszczuk (2018) [9] presented a new model based on the Ant Colony Decision Forest algorithm presented specifically in the study of Boryczka and Kozak (2012) [10] to obtain the results of 0.6436.

In direct marketing datasets, a very common feature that greatly affects the results of learning models is imbalanced data. Ghatasheh et al. (2020) [11] presented an approach to minimize the impact of imbalance using the Meta-Cost Multilayer Perceptron method and the Cost-Sensitive Multilayer Perceptron method, achieving 78.93% and 73.17% results, respectively. Another approach is to solve small problems within a large problem. Research by Moro et al. (2017) [12] in 1915 inbound contacts of total 52944 contacts dataset provides a divide-and-conquer procedure utilizing both the data-based sensitivity analysis for extricating highlight pertinence and master assessment for part the issue of characterizing telemarketing contacts to offer bank deposits products, get the AUC = 0.9247.

For bank telemarketing datasets, highly accurate studies often only deal with a small part of the problem, for example, only predicting with inbound contacts [12] or using a small part of data with few features [8]. The remaining studies using large datasets (over 40,000 records) have not achieved really good results.

3 Dataset and Preprocessing

3.1 Data Description

In our thesis, the data set is very close to the data set in Moro et al. (2014) [4]. This data set was provided by a Portuguese retail bank and publicized on the University of California Irvine (UCI) website for research purposes. Data set was collected between 2008 and 2013, including the negative effect of the global financial crisis. Data set contain 41188 phone contacts with 20 most important features selected from the original data provided by the Portuguese retail bank. Although Selected features have different relative importance levels, these features are a must-have to meet all the business questions that demand a successful telemarketing result. Phone calls took the marketing campaigns. Customers asked if they were interested in Bank financial products (Bank term deposits) or not. **Table 1** shows the description of 20 features in the data used in this thesis.

Table 1. Description of Features.

Feature	Description
age	Numeric
job	Type of job
marital	Matrimony(Categorical)
education	Literacy(Categorical)
default	Is there a credit default? (Categorical)(class: 'yes','no','unknown')
housing	Is there a home loan? (Categorical)(class: 'yes','no','unknown')
loan	Is a personal loan purchasable? (Categorical)(class: 'yes','no','unknown')
contact	Kind of contact communication (Categorical)(class:'cellular','telephone')
month	Last contact month(Categorical)
day_of_week	The week's last contact day (Categorical)
duration	The duration (in seconds) of the last contact(Numeric)
campaign	The total number of contacts made under this campaign and for this customer (Numeric, includes the last contact)
pdays	The number of days since the client was last contacted as part of a prior campaign(Numeric) (class='999' means a client who has never been contacted)
previous	The total number of contacts made prior to this campaign and for this customer (Numeric)

Feature	Description
poutcome	The preceding marketing campaign's result(Categorical)(class= 'failure', 'nonexistent', 'success')
emp.var.rate	Employment variation rate(Numeric, quarterly indicator)
ns.price.idx	Consumer price index(Numeric, monthly indicator)
ons.conf.idx	Consumer confidence index(Numeric, monthly indicator)
euribor3m	Euribor 3 month rate(Numeric, daily indicator)
nr.employed	Number of employees(Numeric, quarterly indicator)

The difference of the dataset compared to the study of Moro et al. (2014) [4] is that we use the feature “duration”. This feature has proven to be an important factor influencing the prediction results [13]. This also makes a lot of sense, given that customers will be more inclined to prolong the call if they are interested in making a bank deposit.

In real life, collected data can be messy and not always ideal for training purposes. Before conducting the machine learning process, it is necessary to analyze the data distribution.

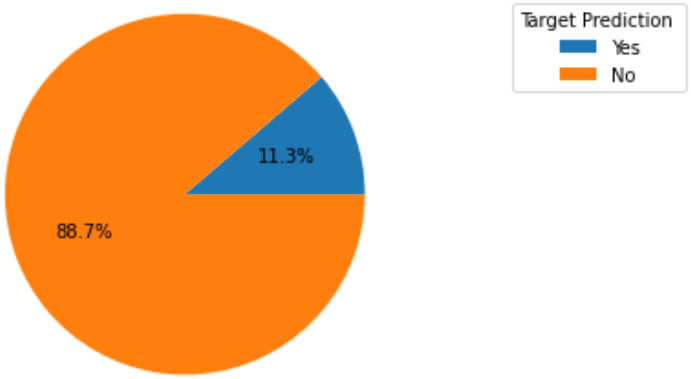

Figure 1. *Target prediction distribution.*

Figure 1 point, the number of rejected calls is about eight times compared to successful calls (88.7% and 11.3% for "no" and "yes" records). Following the imbalance of given data, we decided to use Area Under the Receiver Operating Characteristic score (AUC). In case using accuracy as the metric, machine learning models can produce very high accuracy in training and testing progress. In real-life data, the deployed model can perform very poorly as it tends to correctly predict the label "0" ("naive behavior") more than the label "1". The False positive rate (FPR) and True positive rate (TPR) are used. Only when both TPR and FDR are above the ROC curse's random line can we judge whether our work is efficient or not.

For a more intuitive view of the distribution of data points. We visualized using the T-SNE plot to get a better view of the difference in data distribution(***Figure 2***).

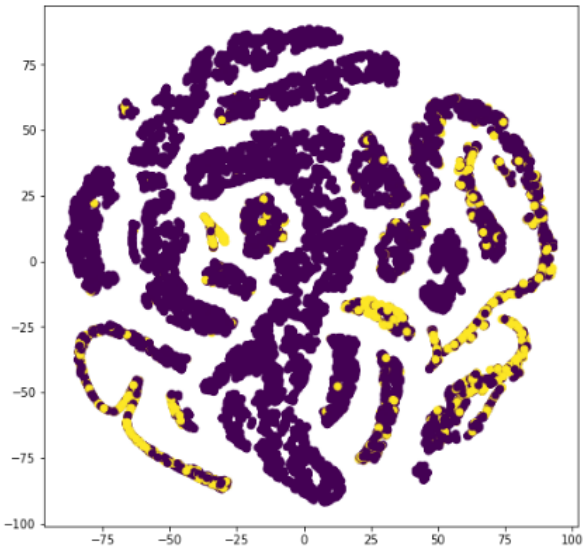

Figure 2a. *Training set.*

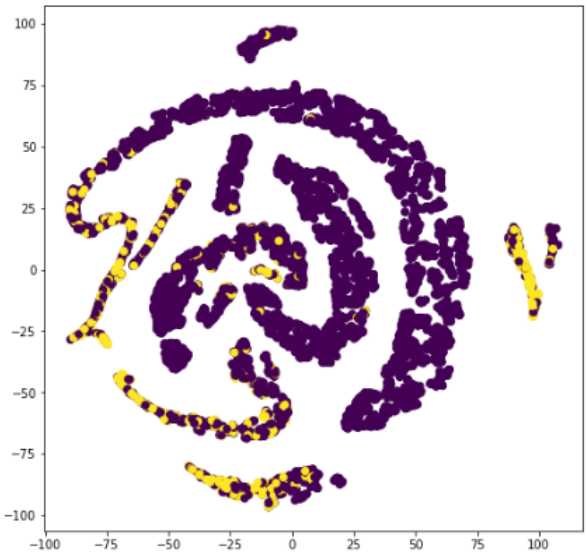

Figure 2b. *Cross-Validation set.*

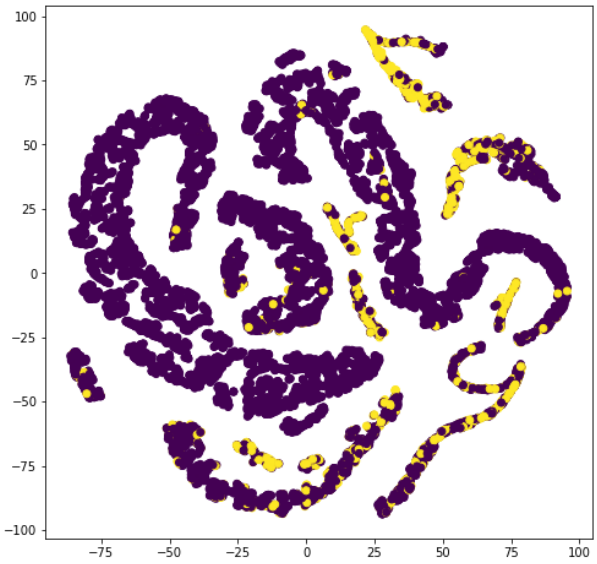

Figure 2c. *Test set.*

Through the observation of T-SNE visualization, even though there seems to be a lot of overlap in the dataset. Based on distribution into specific areas of the yellow ("yes" class) and purple ("no" class) points and the uneven distribution of the data areas of data sets. We can see that the data classification models will be suitable in identifying and extracting the characteristics of the potential of customers.

3.2 Data Correlation

The correlation in the data set also plays an important role in this study. Just in case some features are not related. A correlation with positive results shows the parallel decrease or increase of variables. In contrast, a correlation with negative results shows the increase of one variable but the other decrease. The ***Figure 3*** indicate that Employment variation rate(emp.var.rate), Consumer price index(cons.price.idx), Euribor 3 month rate(euribor3m) and Number of employees(nr.employed) are the most correlation. This is proof the data set meet not only all the major business demand but also meet the requirements of a good data source

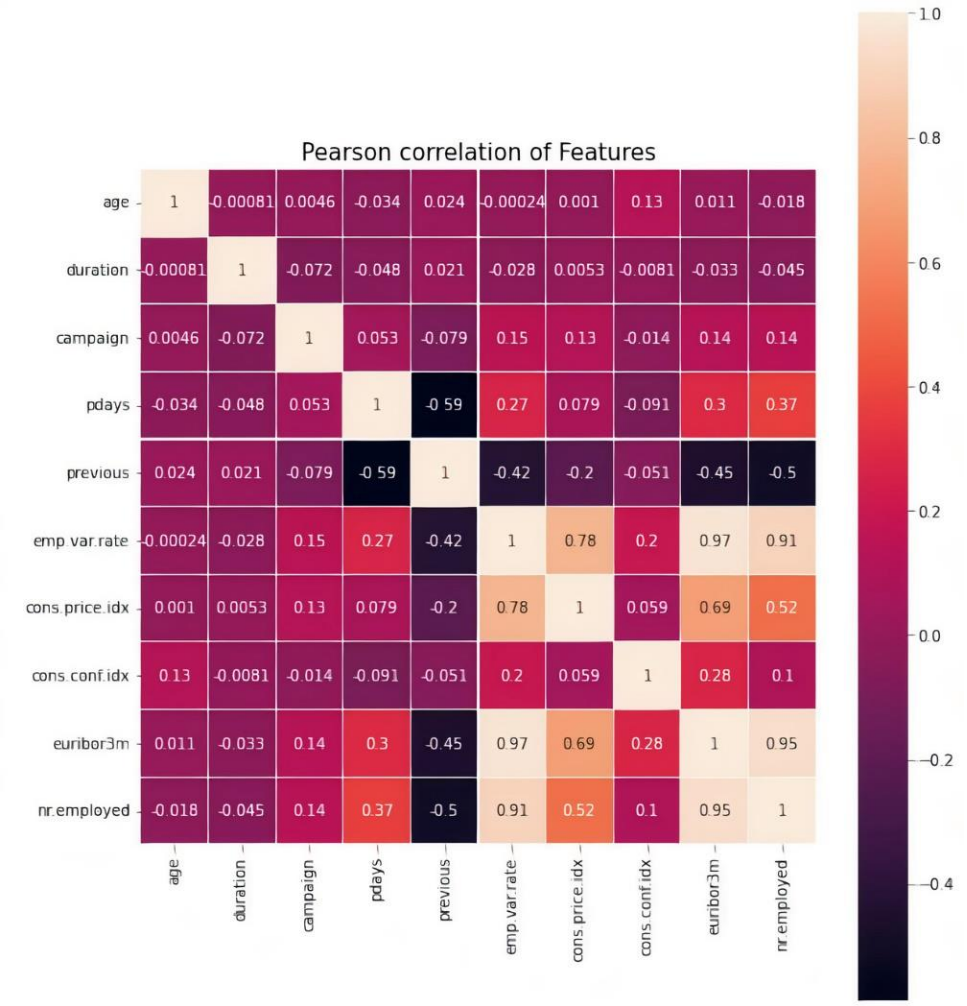

Figure 3. *Data Correlation.*

3.3 Category Data Encoding

This thesis applied models that can only deal with numeric data types. Removing all duplicated and missing values records in the data has been completed. The entire record is reduced by only 21 duplicate items from 41188 to 41176, and there are no missing values. The data set size is still very idealized for the training stage.

Response coding is a technique to represent categorical data. The original idea of the technique is to present a data point belonging to a class of the category. In a case with a K-class classification problem, K-new features will be embedded with the probability calculation of which class data points belong to base on the value of categorical. Laplace smoothing has been shown to not only outperform in text categorization by Zhou et al. (2009) [14], in a study by He and Ding (n.d) [15] improves accuracy when applied to text classifiers such as Naive Bayes. Laplace smoothing is included to avoid zero probability.

To further explain our encoding categorical features method, the response encoding method is based on the probability of a category data points appearing in a class. The formula for this original calculation.

$$P(\text{class=X} \mid \text{category=A}) = \frac{P(\text{category=A} \cap \text{class=X})}{P(\text{category=A})} \quad (1)$$

To avoid zero probability, we apply Laplace smoothing to the previous formula, in our thesis the encoding category feature will be.

$$P(\text{class=X} \mid \text{category=A}) = \frac{P(\text{category=A} \cap \text{class=X}) + \alpha * 10}{P(\text{category=A}) + \alpha * 20} \quad (2)$$

Denote that chosen α in our thesis is 1.

Table 2. Example of encoded data.

	Age	**Duration**	**month_0**	**month_1**	**month**
0	33	335	0.894831	0.105169	Aug
1	51	121	0.894831	0.105169	Aug
2	41	131	0.894831	0.105169	Aug
3	40	339	0.934232	0.065768	May
4	53	79	0.894831	0.105169	Aug

The example of the encoded month feature in **Table 2** shows that there is some change in the training set. The original features month has been replaced with two new features month_1 and month_0. Hence our research is about the binary classification problem, the month_1 is the feature that represents the encode category feature month that shows the probability it's likely to be a potential customer ("yes" class) and month_0 represents the customers not likely interested in a deposit ("no" class).

4 Solutions

4.1 Validation

As noted above, the dataset used in this thesis is highly imbalanced. This is entirely consistent with the actual results of telemarketing at banks. The number of successful calls is minimal in the total number of calls. From this imbalance, conventional assessment tools will no longer maintain accuracy. So, we used AUC scores to evaluate the results of machine learning models.

The AUC score and ROC curve are methods used to evaluate the classification performance. The AUC score (Area Under the Curve) represents the classification level of the model, and the ROC curve represents the probability.

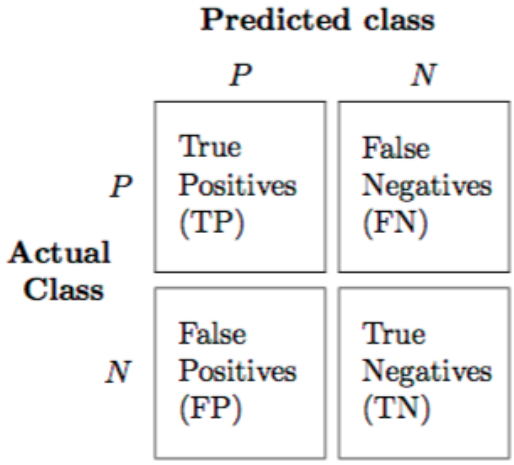

Figure 4. *Evaluation Metrics.*

In the evaluation metrics (***Figure 4***), there are some necessary metrics in calculating the AUC score, which are TP (True Positive), FP (False Positive), TN (True Negative), and FN(False Negative). This is the comparison of results between predictions and actual results.

The metrics used to calculate the AUC score and illustrate the ROC curve are TPR (True Positive Rate) and FPR (False Positive Rate), represented in functions (3) and (4), respectively.

$$TPR\ (Recall) = \frac{TP}{TP + FN} \tag{3}$$

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

4.2 Data mining models

After the data preprocessing steps, the bank telemarketing dataset was removed from the duplicate records and obtained the final size of 41176. With the categorical data encoding, the number of features increased to 30. We have investigated and tested methods in supervised learning and unsupervised learning algorithms to classify labeled data problems. Our team prefers to use an advanced algorithm that parses the information, gains from it, and utilizes those learnings to find significant interest examples. Algorithms can make accurate decisions on their own, but other decisions may require human participation, such as neural networks, which are not being employed at this time. There are four typical and effective classifications used: K-Nearest Neighbors(KNN), Logistic Regression(LR), Linear Support Vector Machines (Linear SVM), and XGBoost(XGB). The performance result will be shown in section 4.3.

In this thesis, a noteworthy problem is data imbalance, with 88.7% being negative and 11.3% positive. In predicting probabilities, predictive models can be overconfident. Furthermore, in the case of this imbalanced dataset, the predictors may likely give preference to the majority class. Because of this, besides using the model performance evaluation tool ROC curve and AUC score, it is necessary to calibrate the probability prediction.

To avoid observable data bias, we divided it into the train, cross-validation, and test sets with the scale of [0.45, 0.22, 0.33] in order. The train set will be learned and calibrated by CalibratedClassifierCV, using the cross-validation set with the 'sigmoid' method similar to Platt's method. Tuning hyperparameters for each chosen algorithm is very important. It helped us understand data better and explain it to the bank business team if needed.

Knn

K-nearest neighbors (KNN) is one of the foremost broadly utilized algorithms not as it were by data scientists but moreover by Artificial intelligent scientists. It's a simple algorithm, also known as a lazy algorithm. KNN is often used to solve classification and regression problems. Based on the efficiency of this algorithm, our thesis applies its classification application as a model for training purposes. KNN algorithm tries to classify classes for the new data by calculating how far the new data point compares to given classes. The test data point will be classified as the same class as the class, which has the closest points within K points the user wants to select. The distance between data points often used in KNN is the Minkowski distance metric.

$$\sum_{1}^{n} |x_i - y_i|^{\frac{1}{p}} \tag{5}$$

Note that:

- n is the number of dimensions
- We get Manhattan distance in case p=1

$$\sum_{1}^{n} |x_i - y_i| \tag{6}$$

- We get Euclidean distance in case p=2

$$\sum_{1}^{n} \sqrt{|x_i - y_i|} \tag{7}$$

In our thesis, we found familiar points in using KNN for classification problems compared to Abdelmoula's credit scoring research [16]. We chose multi-K parameters in our study. We set a set of K parameters that range from 1 to 40 with the step of 2. Since the K start is greater than 5, the data mining models with KNN have returned a better AUC. In our case, 27 is the best value. The difference lies in the data points in our study that are many times larger than the 924 records of Abdelmoula's credit scoring research [16]. The number of K-nearest neighbors also significantly impacts the training stage, and our research shows that using different parameters for K-nearest neighbors can extend the result of data mining models. Choosing the value for k too small will lead to a ratio of high error and Sensing of the data point to bring the local feature. On the other hand, using a large k value can cost a lot in the computation stage. We hope our thesis discovery will make a dedication to other research about binary classification problems in the finance sector.

Logistic Regression

Logistic regression is a prevalent machine learning model for binary classification problems. The input data is calculated through a logistic function, and in this thesis, we use the sigmoid function (function (6)). The output is the probability of the class occurring.

$$f(x) = \frac{1}{1 + \exp -x} \tag{8}$$

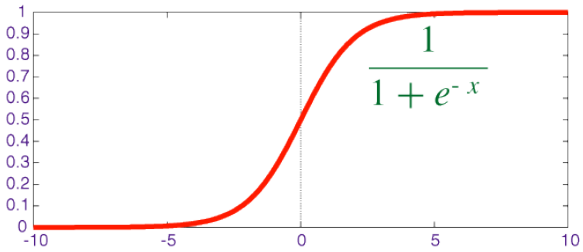

Figure 5. *Sigmoid function graph*

Logistic regression is an easy model to deploy, with good training efficiency. However, with multidimensional datasets, the prediction results can become overfit. Therefore, we choose different Inverse of Regularization parameters(C) from 1e-05 to 1000 to find the appropriate value to avoid overfitting and get the best result. From **Table 3**, we found the model to have the best performance with a C value of 0.001.

Table 3. The result with different C

Inverse of Regularization(C)	**AUC**

1e-05	0.8887
0.0001	0.9244
0.001	**0.929**
0.01	0.9284
0.1	0.9282
1	0.9283
10	0.9282
100	0.9283
1000	0.9282

Linear SVM

One of the models used and achieved good results in this thesis is the Linear SVM. This model performs quite well for high-dimensional data sets. In our works, the linear SVM model is implemented with SGD(Stochastic Gradient Descent) training, and the SGD method plays a classifiers role that is similar to the study of Kabir et al. (2015)[17].

We find that the underlying models become overfit quickly with the large and unbalanced dataset in this study. In SGD classifiers, the Regularization parameter(alpha) is essential in controlling capacity. Because of this, we tried adjusting this parameter between 1e-05 and 1000. In the end, the model got the best results with an alpha value = 0.1 (**Table 4**).

Table 4. The results with different alpha.

Alpha	**AUC**
1e-05	0.5
0.0001	0.5
0.001	0.5
0.01	0.885
0.1	**0.89**

1	0.883
10	0.876
100	0.864
1000	0.865

XGBoost

Extreme Gradient Boosting is constructed from learning models. XGB is known as an ensemble machine learning technique. At a time, trees or base learners are added in order to fit the predictions errors of the previous models. In this study, we tried changing the n_estimators parameter of the XGBClassifier model in the range [10, 50,100, 500, 1000, 2000]. After taking the experiment, we get the best results with the value n_estimators = 1000 (**Table 5**).

Table 5. The results with different n_estimators.

Number of estimators	**AUC**
10	0.8858
50	0.9178
100	0.9212
500	0.9240
1000	**0.9247**
2000	0.923

4.3 Results

Table 1 shows the experimental results compared with the study by Moro et al. [4, 12] - who introduced the first bank telemarketing dataset with four machine learning models: Logistic Regression(LR), Decision Tree(DTs), Support Vector Machine (SVM) and Neural Network (NN).

Table 6. AUC score comparison between authors and Moro et all.

Method	Train AUC	Test AUC	Cross-validation AUC
LR	0.9229	0.9233	0.9287
Linear SVM	0.8853	0.8878	0.8945
KNN	**0.9405**	**0.9295**	**0.9324**
XGBoost	0.9283	0.925	0.9247
LR [4]	0.715		
DTs [4]	0.757		
SVM [4]	0.767		
NN [4]	0.794		

The previous study produced the best result AUC = 0.794 (NN). In research [4, 12], the authors did not use the "duration" feature even though they do concern that "duration" might have a positive change on the result. In this thesis, the authors extend the idea by using the "duration" feature to improve the results. Finally, we received magnificent results compared to previous research.

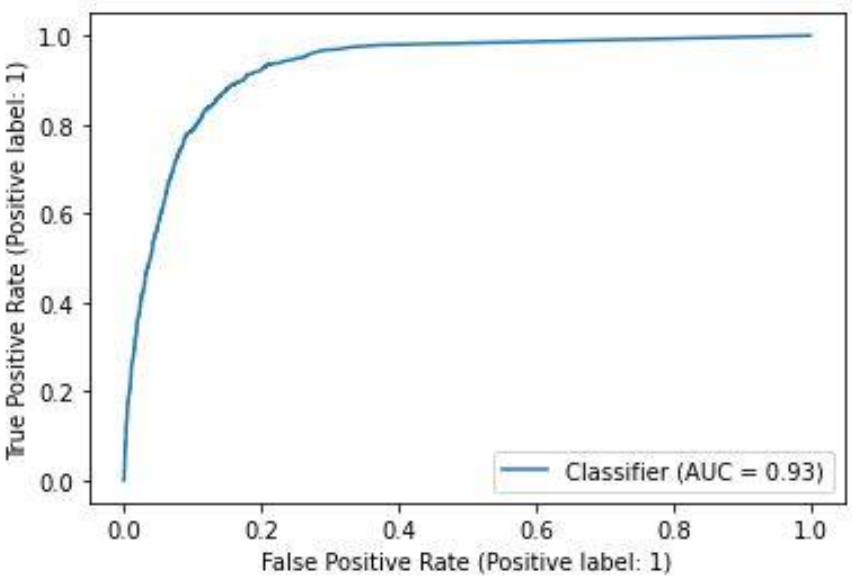

Figure 6. *AUC score of the best experimental result*

Table 7. Confusion matrix of the best experimental result

Actual	**Predicted**	
	success	failure
success	651(4.79%)	838(6.16%)
failure	377(2.77%)	11727(86.28%)

According to the obtained results in **Table 6, Table 7** and ***Figure 6***, the AUC score of the KNN model is the best with train AUC = 0.9405, test AUC = 0.9295, cross-validation AUC = 0.9324. The correct prediction rate of bank telemarketing failure TNR = 0.9688 (Fig. 3.b) is very high. Meanwhile, the correct prediction of success TPR is quite low, only 0.4372. This happens because of an imbalance in the dataset, with the number of failure records being too large (88.7%). We believe the correct prediction of success TPR can be improved if the data sets have more successful records since the TNR is very high with the same approach. However, the overall prediction accuracy rate is still quite high, accounting for 91.07%.

5 Conclusion

Our work demonstrates how machine learning techniques can make an incredible impact on the result of the telemarketing campaign. There are two major steps: data preprocessing and model evaluation. In the first step, cleaning data by removing duplicates records, checking if there were missing values to remove or not, visualizing data to check the imbalance of data set, and applying the response coding technique to encode category features with the help of Laplace smoothing. Moreover, adding a "duration" feature also greatly affects the final result. In the second step, typical efficient algorithms: KNN, LR, Linear SVM, and XGBoost were chosen to determine the best classifier model.

Since the bias of the data set, the Area Under the Receiver Operating Characteristic score is observed to judge the successfulness of research. KNN is the best method with 93% AUC and performance 91.07% accuracy. The experiment results show that the best KNN with k greater than five can be useful for interpreting the business point of view. The thesis can be a good reference for many machine learning problems [18-19].

In the future, we expect to apply this approach to larger datasets to further look at data mining problems. We intend to work with datasets that contain more information on interest values and economic indicators to compare how the state of the economy might affect customers depositing habit behavior. Models and encoding categorical features method still can improve to suit crawling specific information purposes. Through this study, we hope our approach can be extended to new applications in solving binary classification problems in the banking and finance sector as well as the marketing field.

6 References

1. Ling, C., X. & Li, C. (1998) Data Mining for Direct Marketing: Problems and Solutions. In: Proceedings of the International Conference on Knowledge Discovery from Data (KDD 98), New York City, pp.73-79.

2. Elsalamony, H., A. & Elsayad, A., M. (2013) Bank Direct Marketing Based on Neural Network. International Journal of Engineering and Advanced Technology (IJEAT), 2(6), pp.2249 – 8958.

3. Ghoddusi, H., Creamer, G. and Rafizadeh, N., (2019) Machine learning in energy economics and finance: A review. *Energy Economics*, 81, pp.709-727.

4. Moro, S., Cortez, P. and Rita, P., (2014) A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, pp.22-31.

5. Miguéis, V., Camanho, A. and Borges, J., (2017) Predicting direct marketing response in banking: comparison of class imbalance methods. *Service Business*, 11(4), pp.831-849.

6. Zhang, X., Li, X., Feng, Y. and Liu, Z., (2015) The use of ROC and AUC in the validation of objective image fusion evaluation metrics. *Signal Processing*, 115, pp.38-48.

7. Martens, D., Vanthienen, J., Verbeke, W. and Baesens, B., (2011) Performance of classification models from a user perspective. *Decision Support Systems*, 51(4), pp.782-793.

8. Acheampong, A. and Agyepong, K., (2016) Enhancing Direct Marketing using Data Mining: A Case of YAA Asantewaa Rural Bank Ltd. in Ghana. *International Journal of Computer Applications*, 153(7), pp.6-12.

9. Kozak, J. and Juszczuk, P., (2018) The ACDF Algorithm in the Stream Data Analysis for the Bank Telemarketing Campaign. *2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI),*.

10. Boryczka, U. and Kozak, J., (2012) Ant Colony Decision Forest Meta-ensemble. *Computational Collective Intelligence. Technologies and Applications*, pp.473-482.

11. Ghatasheh, N., Faris, H., AlTaharwa, I., Harb, Y. and Harb, A., (2020) Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks. *Applied Sciences*, 10(7), p.2581.

12. Moro, S., Cortez, P. and Rita, P., (2017) A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, 35(3), p.e12253.

13. Moro, S., Laureano, R., Contez, P. (2012) Enhancing bank direct marketing through data mining. In: Proceedings of the Forty-First International Conference of the European Marketing Academy, European Marketing Academy, pp.1–8.

14. Zhou, S., Li, K. and Liu, Y., (2009) Text Categorization Based on Topic Model. *International Journal of Computational Intelligence Systems*, 2(4), pp.398-409.

15. He, F. and Ding, X., n.d. Improving Naive Bayes Text Classifier Using Smoothing Methods. *Lecture Notes in Computer Science*, pp.703-707.

16. Abdelmoula, A. K. (2015) Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. Accounting and Mangement Information System, 14(1), pp.79-106.

17. Kabir, F., Siddique, S., Kotwal, M. and Huda, M., 2015. Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier. *2015 International Conference on Cognitive Computing and Information Processing(CCIP)*,.

18. Luu, N. and Hung, P., (2021) Loan Default Prediction Using Artificial Intelligence for the Borrow – Lend Collaboration. *Lecture Notes in Computer Science*, pp.256-270.

19. Hung, P. and Thinh, T., (2019) Cryptocurrencies Price Index Prediction Using Neural Networks on Bittrex Exchange. *Future Data and Security Engineering*, pp.648-655.