FPT Education

FPT UNIVERSITY

Potential Customers Prediction in Bank Telemarketing

CST491_G6

Our Team Members

- Phùng Thái Dương – HE140170
- Khuất Duy Bách – HE140665

A thesis submitted in partial fulfillment of the degree of BSc. In Computer Science with the supervision of M.S.E Lê Đình Huynh

Our Goal

- Learning some techniques in Computer Science
- Applying some specialized knowledge of computer science that we have acquired during our study at FPT University to solve real-life problems.
- Exploring the implementation process of scientific research and publish a research paper

TABLE OF CONTENTS

INTRODUCTION

- 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. [1,2]

- To reduce resources for calling no potential customers. Our study applies machine learning techniques to classify potential customers through their personal information and favorite.

We consider the responses of customers as a binary classification problem. Two classes are “yes,” which denotes the customers have interest in a term deposit, and “no” which represents the customers who do not have interest in a term deposit.

In this paper, we are focusing on:

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

Literal Review

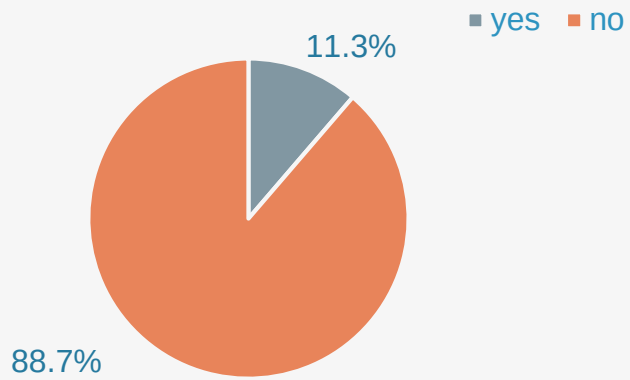
- According to research by Ghoddusi et al. (2019), from 2005 to 2018, more than 130 studies were presented that applied machine learning to finance.
- Research by Moro et al. (2014), they used a dataset of 52944 customer calls with four machine learning models: LR, DTs, NN, SVM get the best AUC = 0.8
- A similar study, combining data mining and the Decision tree model of Amponsah and Pabbi (2016), gave very good results with a AUC value of 0.925
- Ghatasheh et al. (2020) presented an approach to minimize the impact of imbalanced using the Meta-Cost Multilayer Perceptron method and the Cost-Sensitive Multilayer Perceptron method, achieving 78.93% and 73.17% results

DATASET AND PREPROCESSING

This data set was provided by a Portuguese retail bank and publicized on the University of California Irvine (UCI) website for research purposes. The data set was collected between 2008 and 2013, including the negative effect of the global financial crisis.

Data Description

- Data set contain 41188 phone contacts with 20 most important features selected from the original data provided by the Portuguese retail bank.



- 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.

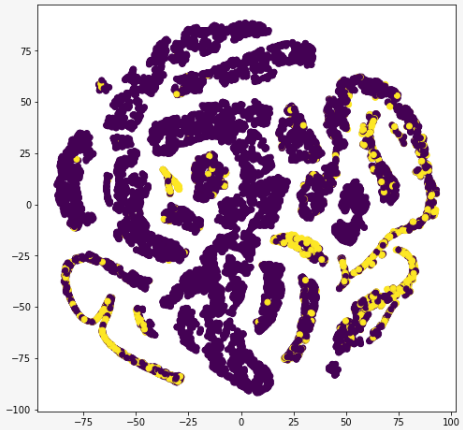
Features Description

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)

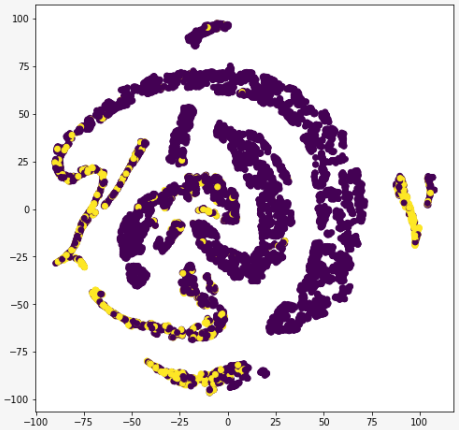
day_of_week:	The week's last contact day (Categorical)
duration	The duration(in seconds) of the last contact(Numeric)
campaign:	The total number of contacts made under this campaign and for this customer(Numeric, includes the last contact)
pdays	The number of days since the client was last contacted 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)
poutcome	The preceding marketing campaign's result(Categorical)(class= 'failure', 'nonexistent', 'success')
emp.var.rate	Employment variation rate(Numeric, quarterly indicator)
cons.price.idx	Consumer price index(Numeric, monthly indicator)
cons.conf.idx	Consumer confidence index(Numeric, monthly indicator)
euribor3m	Euribor 3 month rate(Numeric, daily indicator)
nr.employed	Number of employees(Numeric, quarterly indicator)

Data Visualization with T-SNE

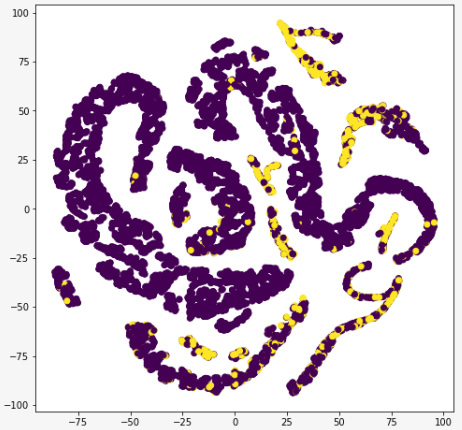
- For a more intuitive view of the distribution of data points. We visualized using the T-SNE plot to get a better view of the difference in data distribution.



a) Training set



b) Cross-Validation set

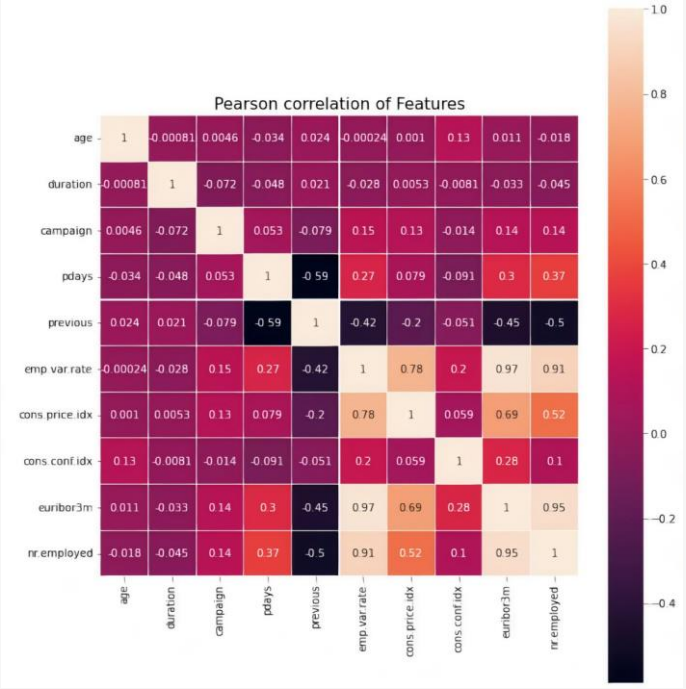


c) Test set

Data Correlation

The correlation in the data set also plays an important role in this research. 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
- This is proof the data set meet not only all the major business demand but also meet the requirements of a good data source



Category Data Encoding

- 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.

Respond encoding

- The formula for this original calculation.

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

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

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

- Denote that chosen α in our thesis is 1.

Data encoded example

	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 table 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).

Experimental and Result

KNN

K-Nearest Neighbors

LR

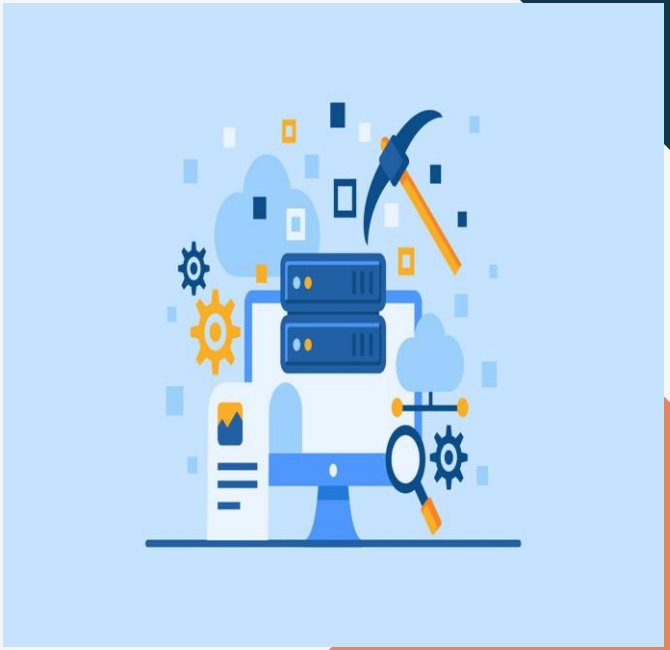
Logistic Regression(LR)

Linear SVM

Linear Support Vector Machines

XGBoost

Extreme Gradient Boosting

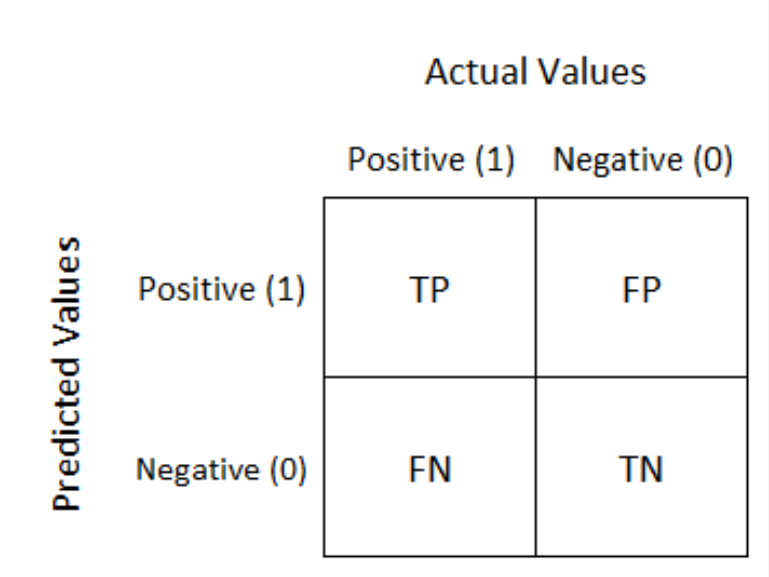


Evaluation models

- The AUC score and ROC curve are methods used to evaluate the classification performance.

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

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



Logistic Regression

- Logistic regression is a very popular 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

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

- We choose different Inverse of Regularization parameters(C) from 1e-05 to 1000 to find the appropriate value to avoid overfit and get the best result.

- Finally, We get the best result with C = 0.001

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

- 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.
- In SGD classifiers, the Regularization parameter(alpha) is quite important in controlling capacity. We tried adjusting this parameter between 1e-05 and 1000.
- Finally, We get the best result with alpha = 0.1

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

- XGBoost is efficient for structured or tabular datasets on classification problem and regression prediction modeling
- 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].
- Finally, We get the best result with n_estimators = 1000

The results with different n_estimators

Number of estimators	AUC
10	0.8858
50	0.9178
100	0.9212
500	0.924
1000	**0.9247**
2000	0.923

KNN Model

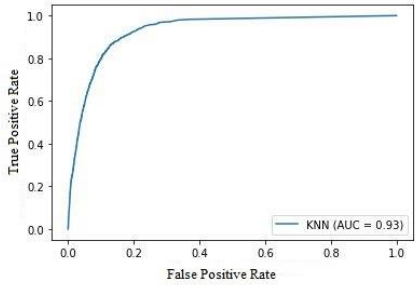
- K-nearest Neighbors (KNN) is one of the foremost broadly utilized algorithms not as it were by data scientists but moreover by Artificial intelligent scientists.
- KNN algorithm tries to classify classes for the new data by calculating how far the new data point compares to given classes.
- We chose multi-K parameters in our research. We set a set of K parameters that has a range from 1 to 29 with the step of 2.
- Finally, With K >= 5, the result becomes better. We get the best result with K = 27.

Result

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		

- Table Shows the experimental results compared with the study by Sergio Moro - 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).
- In this research, the authors extend the idea using the “duration” feature to improve the research results. Finally, we received magnificent results compared to previous research.

Result



(a)

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

(b)

a. AUC score and b. Confusion matrix

- 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
- The correct prediction of success TPR only about 0.4372

Conclusion

- 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.
- 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.



THANK YOU FOR LISTENING

Do you have any questions?