Analysis and Prediction of Whether Clients Will Subscribe Bank Term Deposit Through Telemarketing Based on Data of a Portuguese Banking Institution
Project Final Report

I. Problem Description

Our target is to predict if a client would subscribe to a bank term deposit due to telemarketing by using classification machine learning models. Our main task is to predict the consumers' behaviors based on the data we have collected. We believe that it is a good fit to choose the classification method.

Two goals are set for this project: find out the most characteristic features (including client's demographic features and telemarketing features) of the clients who purchased the term deposits, and apply classification models to predict if the client will subscribe to a term deposit.

II. Description of background

People nowadays may think that telemarketing is an outdated marketing method. However, after doing some research, we found that there are many banks and firms still using this so-called old-school advertising method to reach out to their potential customers. Indeed, telemarketing has many advantages that may be underestimated by people and society: It is more interactive compared to some digital marketing nowadays. It also makes marketing results highly quantifiable and significantly more economical (Cruz-Pandy, 2022). Nowadays, although people tend to pay more attention to digital marketing methods compared with traditional marketing methods, old-school methods like telemarketing are still worth analyzing because of their above-mentioned advantages. Especially in recent years, because of the pandemic, telemarketing becomes even more valuable due to remote working (Sihombing et al. 2020). We plan to cut in from the perspective of the banking industry to start this project.

Through previous research, we found that banks generally use telemarketing to promote term deposits, which is a major part of their profits (Chen, 2022) We believe combining the two parts can help us get valuable insights and be able to give some useful suggestions to the banking industry on how to improve the performance of their telemarketing-related business, which would be the main contribution that we try to make and deliver through this project.

III. Description of the Data Set

The dataset (link) showcases a Portuguese banking institution's information on a telemarketing campaign, containing one train data file and one test data file. Each file has 45,211 rows and 18 columns, representing 7 numerical variables and 10 categorical variables. The 7 numerical variables include potential clients' age, average yearly bank balance (the average amount of money held in an account over a 365 days time period), last phone call contact duration (how long has one call been on), number of contacts performed during this campaign for a specific client, days after one client was last contacted, the number of contacts were performed on last marketing campaign, and the day of the month of contact (e.g. 14th December, 14th is the value of the variable). The 10 categorical variables include clients' job type, marital status, education level, contact month, contact type, the outcome of the previous campaign, the outcome of this campaign on that client, and whether a credit is in default, housing loan, or personal loan.

By utilizing the effective information in the data set, we'll get a more comprehensive understanding of customer group characteristics, preferences, and behavioral characteristics.

IV. Description of methods used

According to our goals, we broke the significant question into several small questions. We noticed that most of the categorical variables are socio-demographic, such as job types and marriage status. Therefore, it would be interesting to see whether the clients will differ in subscription performance when they have different variables. Thus, we'd like to analyze if there is a significant subscription difference due to job types, marriage status, etc. On the other side, the numerical variables (for instance, how many times a client had been called before he/she decided to subscribe to the term deposit, and how long the last contact lasted) in the dataset are mainly related to the marketing campaign practice. We would analyze how those variables influence their subscription decisions. By analyzing the previous factors and making predictions, we hope to have a better understanding of the correlation between the factors described above and the clients' subscriptions, and eventually provide some insights on improving the telemarketing campaigns based on our results.

Before running the model, we conducted a correlation analysis to reduce the redundancy of the data set and tried multiple models to test the fits of each variable to the model as well. Eventually, we selected three models for our analysis: logistic regression, random forest

classifier, and LightGBM classifier. The reason we choose the three models is not only because they are classic classification models but also because our data are imbalanced and those models won't be affected significantly by the imbalanced data.

The first model we tried is the logistic regression model. According to Joby's article (2021), "logistic regression models are good for predicting categorical dependent variables, and continuous data can be used as predictor variables." Since all of our variables are continuous or discrete nominal data, we believe that logistic regression would be an appropriate model to predict if the clients will purchase a term deposit using the numerical variables. Both the other two models of our analysis are based on decision trees: the random forest classification and the LightGBM classifier. Random forest classification is a way to combine multiple decision trees to achieve high accuracy and avoid overfitting by averaging the results (Louppe, 2014), and the LightGBM classifier offers a rapid distributed analysis with a lower occupation of storage.

The same variables are used for each model. In addition, we found that when there are too many independent variables, they bring a lot of overfitting. Thus, we tested each variable and eventually sorted out age, account balance, duration, number of current campaign contacts, number of previous campaign contacts, campaign, and days after the last contacts as the independent variables of the model.

V. Analysis and results

Exploratory data analysis (EDA)

Back to the dataset where we conducted the exploratory data analysis, as shown in Fig. 1, the statistical results support our hypothesis that 88.3% of clients rejected the offer. From the result, we can see that our data is imbalanced and we need to make the dataset more balanced to increase the accuracy of our model.

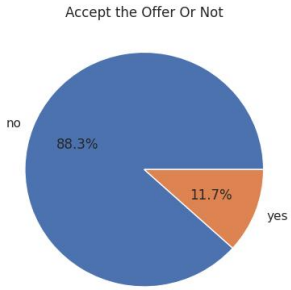

Fig. 1: statistical results of if the consumer would accept the offer

As described in the project ideas section, we analyzed the numerical and categorical variables respectively. During the analysis, we first conducted a correlation analysis and tried to find if there is some correlation between the numerical variables and our target class. Through the analysis, we found that:

- The Lower number of contacts has higher contact duration
- People with a low yearly average balance got more calls
- People under 60s got more calls
- Clients who eventually buy the term deposit tend to have longer call durations during their last contacts

After that, we conducted the analysis based on categorical variables, and the results are shown below in Fig. 2:

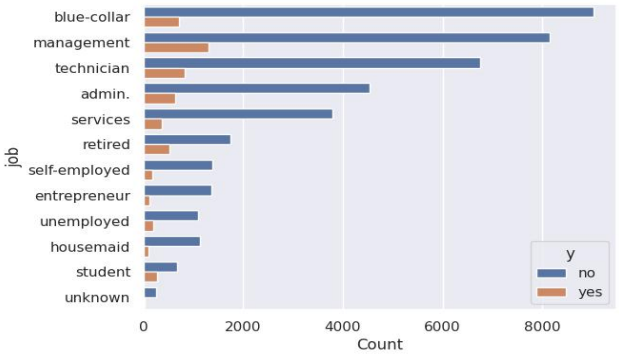

Fig. 2: results of offer acceptance based on categorical variables

Through the categorical variable analyses, the following patterns were found:

- Married clients have the largest number of subscriptions as most calls are made to married clients. However, the single clients' group has the highest percentage of subscriptions.
- People with management-type jobs have the largest number of subscriptions even when most calls are made to people with blue-collar jobs.
- Most calls are made to people who have finished secondary education, but the highest percentage of subscriptions occurs in the group of people who completed tertiary education.
- The groups of clients without housing loans and personal loans have higher percentages of subscriptions than people with loans, though they tend to receive more calls.
- During the campaign, most people were first-time customers of the telemarketing campaign; people who have been involved in the previous campaign and purchased the deposit and also bought the term deposit are more likely to subscribe this time.
- Many people have default credit records.
- Most contacts are made by cellular contacts.

These results provide us guidelines to figure out the major characteristics of people who will purchase the term deposit and the direction to explore more later.

Machine Learning Models

After running all three models, we conducted cross-validation respectively to examine the performance of the model and visualized the result of each model into a confusion matrix shown below. In addition to the separate analysis of categorical and numerical variables, we also conducted analyses with all variables with the three models as a reference to the respective analyses. In the confusion matrix, "0" represents the result of the prediction of people who didn't make the purchase, and "1" represents the prediction of people who accepted the offer. To further check the accuracy of our models we generated two dummies whose responses are 'no' and used our models to predict the result. The result shows our models predict them correctly.

- Numerical variables:

	precision	recall	f1-score	support
0	0.93	0.97	0.95	13206
1	0.64	0.43	0.51	1714
accuracy			0.91	14920
macro avg	0.78	0.70	0.73	14920
weighted avg	0.90	0.91	0.90	14920

	precision	recall	f1-score	support
0	0.90	0.98	0.94	13206
1	0.57	0.17	0.26	1714
accuracy			0.89	14920
macro avg	0.74	0.57	0.60	14920
weighted avg	0.86	0.89	0.86	14920

	precision	recall	f1-score	support
0	0.93	0.97	0.95	13206
1	0.62	0.41	0.49	1714
accuracy			0.90	14920
macro avg	0.77	0.69	0.72	14920
weighted avg	0.89	0.90	0.89	14920

Table 1 (top left): the confusion matrix of random forest

Table 2 (top right): the confusion matrix of logistic regression

Table 3 (bottom left): the confusion matrix of LightGBM

- Categorical variables:

	precision	recall	f1-score	support
0	0.91	0.97	0.94	13206
1	0.56	0.27	0.36	1714
accuracy			0.89	14920
macro avg	0.74	0.62	0.65	14920
weighted avg	0.87	0.89	0.87	14920

	precision	recall	f1-score	support
0	0.89	1.00	0.94	13206
1	0.00	0.00	0.00	1714
accuracy			0.89	14920
macro avg	0.44	0.50	0.47	14920
weighted avg	0.78	0.89	0.83	14920

	precision	recall	f1-score	support
0	0.91	0.99	0.94	13206
1	0.64	0.21	0.31	1714
accuracy			0.90	14920
macro avg	0.77	0.60	0.63	14920
weighted avg	0.88	0.90	0.87	14920

Table 4 (top left): the confusion matrix of random forest

Table 5 (top right): the confusion matrix of logistic regression

Table 6 (bottom left): the confusion matrix of LightGBM

- All variables:

	precision	recall	f1-score	support
0	0.94	0.97	0.95	13206
1	0.69	0.50	0.58	1714
accuracy			0.92	14920
macro avg	0.81	0.74	0.77	14920
weighted avg	0.91	0.92	0.91	14920

	precision	recall	f1-score	support
0	0.90	0.98	0.94	13206
1	0.55	0.20	0.29	1714
accuracy			0.89	14920
macro avg	0.73	0.59	0.62	14920
weighted avg	0.86	0.89	0.87	14920

	precision	recall	f1-score	support
0	0.94	0.96	0.95	13206
1	0.65	0.49	0.56	1714
accuracy			0.91	14920
macro avg	0.79	0.73	0.75	14920
weighted avg	0.90	0.91	0.91	14920

Table 7 (top left): the confusion matrix of random forest

Table 8 (top right): the confusion matrix of logistic regression

Table 9 (bottom left): the confusion matrix of LightGBM

VI. Observation and Conclusion

Generally, according to the results, the three models do not have a significant difference in their performance of predicting purchase behavior between each other as well as with different variables as their f1 score floats in a small interval between 0.94 to 0.95. Overall, the LightGBM classifier performs better than the other two models for both predicting people who accept and don't accept the offer, especially for predicting people who accept the offer; on the other hand, using all variables would have better accuracy in predicting purchase behavior (0.95, 0.94, 0.95) than using categorical (0.94, 0.94, 0.94) or numerical variables (0.94, 0.94, 0.95) separately.

Our project takes telemarketing as the topic and cuts into the banking industry to predict the results of telemarketing activities for specific commodity items in the banks. We paid special attention and analyzed how certain variables would affect people's subscription decisions.

We discussed three models, logistic regression, random forest, and LightGBM classifier in solving the problems we posed, which we think are fit for our topic and the problems we try to solve through comprehensive consideration. Through the preliminary EDA, we got the variables to focus on, the selection of meaningful features and normalization not only improved performance but also reduced processing time while stabilizing the model for this dataset.

Overall, we are satisfied with the results of the project. Due to the length of the project, it may not be in-depth enough, so the analysis of each important variable may be relatively general.

If we have more time for the project, we would further refine and improve our models to improve the accuracy rate. On the other hand, whether these models could be used for other cases would be a great topic for the application of machine learning since the performance of each model varies greatly depending on the subject it is taking (Tékouabou, 2022). Though we added many rows of people who accepted the offers to reduce the imbalance of our dataset and increase the accuracy of predicting people who accept the offer, the results are still not as high as predicting people who don't accept the offer. Therefore, multiple trials with more ways to balance the dataset, greater sample size, and alternative variable sets due to different aspects may be necessary to reach a more complete, general, and significant result.

Ⅶ. References

Cruz-Pandy, A.D.L. (2022) *You should still do telemarketing in 2022 - here's why*, *Wing Assistant*. Available at: https://wingassistant.com/telemarketing-2022/ (Accessed: November 11, 2022).

Sihombing, Ester Hervina, and Nasib Nasib. 2020. The Decision of Choosing Course in the Era of Covid-19 through the Telemarketing Program, Personal Selling and College Image. Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences 3: 2843–50.

Chen, J. (2022) *Time Deposit: Definition, how it's used, rates, and how to invest*, *Investopedia*. Investopedia. Available at: https://www.investopedia.com/terms/t/termdeposit.asp (Accessed: November 11, 2022).

Joby, A (2019) *What is Logistic Regression? Learn When to Use It. Learn Hub.* Available at: https://learn.g2.com/logistic-regression

Louppe, G. (2014) "Understanding Random Forests: From Theory to Practice", Ph.D. Thesis, U. of Liege.

Tékouabou, Stéphane Cédric Koumétio, Ştefan Cristian Gherghina, Hamza Toulni, Pedro Neves Mata, Mário Nuno Mata, and José Moleiro Martins. (2022) "A Machine Learning Framework Towards Bank Telemarketing Prediction." Journal of risk and financial management 15, no. 6: 269.