Improve Telemarketing Efficiency in Banking Industry by Using Machine Learning Methods

Yixiang Huang

Yuchen Zeng

Kshitiz Parashar

Executive Summary

This report investigates the inefficiency problem of traditional telemarketing methods in the bank industry, and try to solve this problem using machine learning techniques. Logistic regression, KNN, Random Forest, and XGBoosting are used to predict potential subscribers. The Random Forest model achieves the best predicting result. The results indicate that machine learning-based method is more useful in guiding how to optimize the marketing activity rather than predicting the potential subscribers. According to these models, the report gives five recommendations regarding marketing intensity, customer segmentation, seasonality, channel choosing, and remarketing activity. By following these recommendations, banks can reduce marketing costs while increasing revenue by capturing more potential subscribers.

Problem Background

Telemarketing is a common marketing tool, especially in banking institutions. All customers are required to provide some basic information like job, income, financial experience, investment history, and other demographic information when opening a bank account and need to update these anytime when increasing their credit limit, or purchasing other financial products and services. Therefore, compared to other industries, banks have more comprehensive and accurate information about their customers. Telemarketing is a direct marketing method, and most telemarketers have some knowledge of the target customer before calling, so it has become one of the most common ways for banks to promote their new financial products. Of course, a customer’s willingness to subscribe to a term deposit is also influenced by the overall economic indicators in addition to a personal financial situation. Banks often need to take into account the multiple economic indicators when promoting their products, including bank interest rates, currency exchange rates, stock fluctuation, and other general economic indicators.

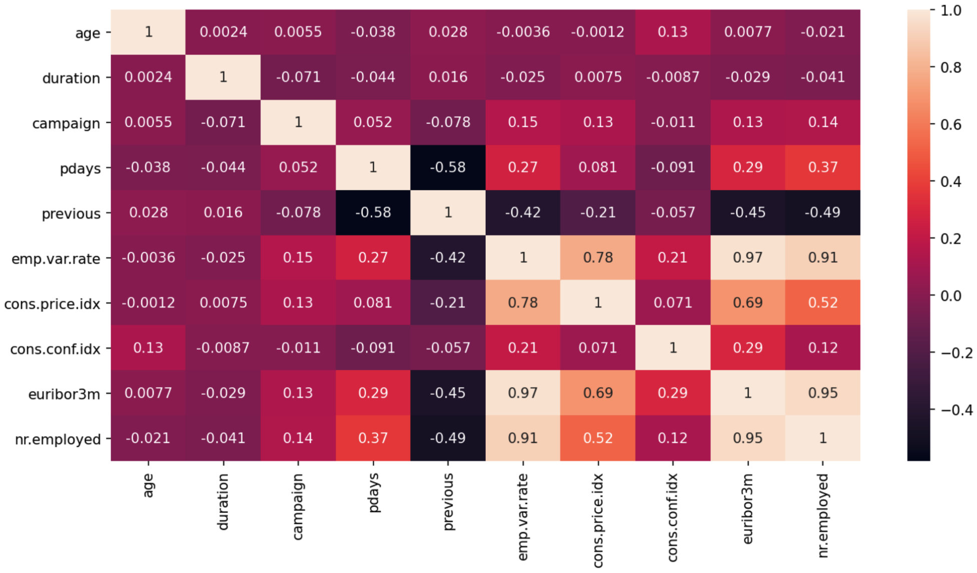
Traditional Approaches and Pain Points

In order to attract more customers to subscribe to financial products, banks often make regular telephone calls to promote them. Telemarketers explained the features of the product in detail like the short-term rate, long-term return, stability, and risk. Before calling, the telemarketer needs to go through the basic information of customers briefly, such as the account's transaction history, credit status, investment behavior, and risk assessment. All telemarketers are expected to have a good understanding of the products they are promoting and are required to clarify customers' questions quickly. In most cases, it is rare that a single call can make a deal, this is why there are multiple telephone promotions regularly. To improve the probability of successful marketing in the next call, banks need to record the customer's feedback, call duration, customer's interest, and the reasons for their hesitation. However, it's undoubtful that telemarketing has been used less frequently in recent years. On the one hand, with the advent of the digital age, marketing tools have become more diverse. The emergence of new marketing platforms such as the internet and social media has allowed banking institutions to discover new marketing channels. On the other hand, people are less interested in answering the phone from strangers, preferring to learn and shop for financial products through their own exploration. For banks, the cost of operating a full telemarketing department is quite high but the return is not appealing as expected. Many banks prefer to outsource this work, which can reduce operational costs compared with doing it by themselves

EDA and Data Preprocessing

The data comes from a bank in Portugal. Its telemarketing department made multiple contacts with existing customers in 2012 to promote a term deposit and recorded whether the customer subscribed to the term deposit in the end. In this project, we firstly did exploratory data analysis to investigate the correlation between variables and their distribution.

According to the heatmap, the correlations among euribor3m(euribor 3 month rate), np. employed(number of employees), and e,p.var.rate(employment variation rate) are higher than 0.91. In order to mitigate the interrelationship and avoid multicollinearity of regression, we need to drop two of these. Employment Variation Rate is positively correlated with the number of employees. The increase or decrease of employee number will result in the change of employment variation rate. Based on domain research, the employment market is strongly related to the currency rate, so we can just keep "euribor 3 month rate" and drop the other two features to avoid multicollinearity in our mode.



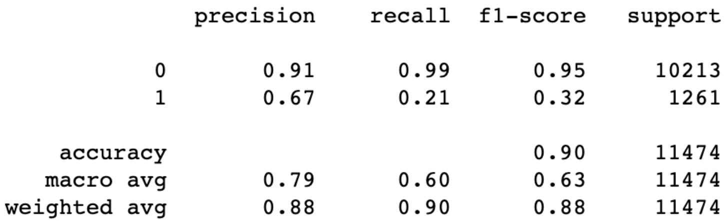
In addition, we dropped several variables in the following studies due two various reasons. Firstly, duration is not known before a call is performed. In other words, values of this feature were recorded after knowing the subscription status (targeted variable). Thus, this input should only be included for modeling because of its bias. Thus, we exclude it from our analysis. Secondly, for 'default' feature, there are 30485 'no' and only 3'yes', the other 7757 values are missing. This feature is highly skewed, so we drop it.

Finally, to deal with missing values, we calculated the proportion of missing values for all columns. It turns out that the percentage of missing value for 'job','marital','housing' and 'loan' are quite small, so we drop these missing values directly from the dataset.

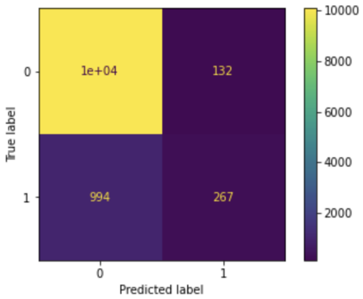
Machine-Learning Based Approaches

(1) Logistic regression

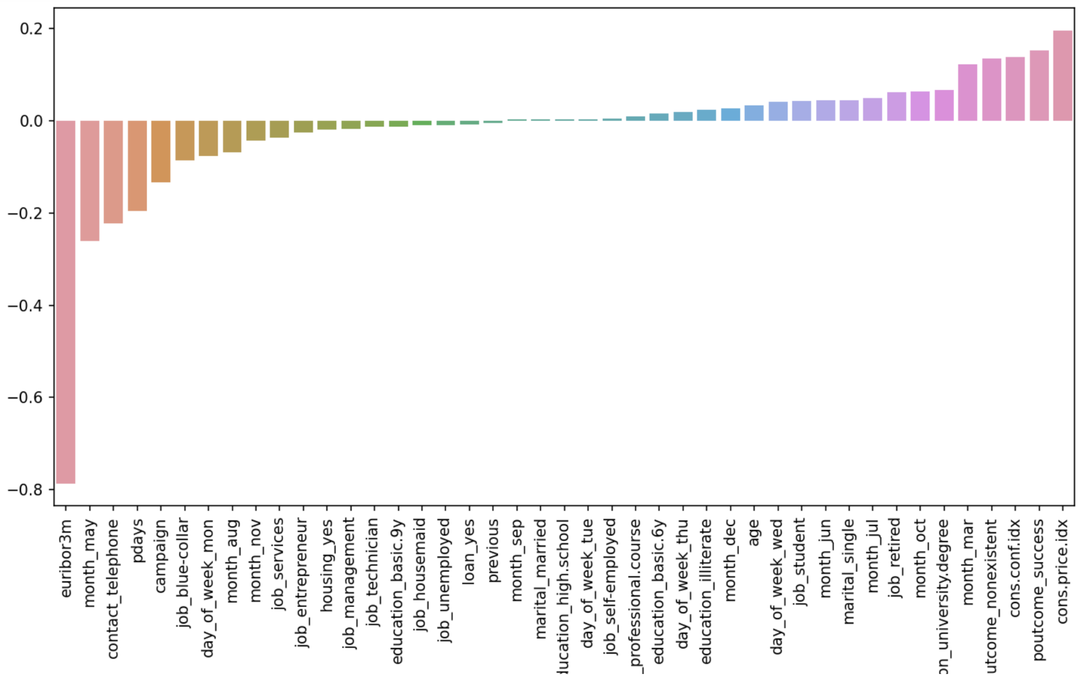
We used cross-validation at 10 folds to evaluate the model performance. To avoid overfitting and multicollinearity, we use ridge regression(L2 regulation) to shrinkage coefficients and penalize the insignificant predictors.



```
              precision    recall  f1-score   support

           0       0.91      0.99      0.95     10213
           1       0.67      0.21      0.32      1261

    accuracy                           0.90     11474
   macro avg       0.79      0.60      0.63     11474
weighted avg       0.88      0.90      0.88     11474
```



Both precision and recall are high in prediction 0 while low in 1, which means this model can predict well for people who subscribed to the term deposit. It can just correctly predict 21% of all subscribed customers and the precision rate is only 67% among this 21%. Because the dataset is very imbalanced, the number of customers who choose ‘No’ is 8 times larger than the number of “Yes”, we decide to use F1 score as criteria to measure the model’s accuracy. The F1 score of logistic regression is only 0.3216867469879518.

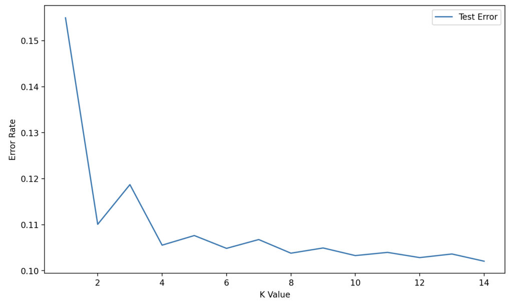


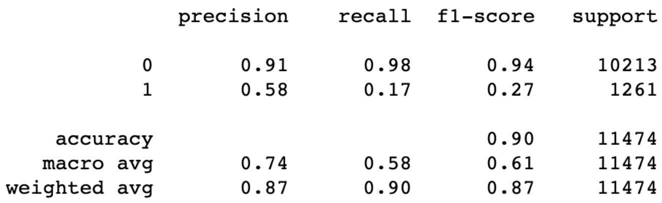
We further investigated the coefficient for each variable. Variables with higher coefficients like ‘cons.price.idx’, ‘poutcome_success’, ‘cons.conf.idx’ and etc on the right will increase odds of subscribing term deposit (y = 1). In the contrast, Variables with smaller coefficients like ‘job_services’, ‘month_nov’, ‘month_aug’ and etc on the left will increase odds of not subscribing term deposit (y = 0). For example, the coefficient of ‘cons.price.idx’ is 0.195990. e^β = e^0.195990 = 1.2165 will be the odds ratio that associates ‘cons.price.idx’ to the successful subscription. In other words, an increase of 1 unit in ‘cons.price.idx’ multiplies the odds of subscribing term deposit by 1.2165. Contract, the coefficient of ‘job_service’ is -0.036987. e^β = e^-0.036987 = 0.96368. We can say a customer who worked in the service industry has 3.632%(1 - 0.06368) fewer odds of subscribing term deposit.

(2) KNN

For the K-nearest neighbors model, we iterated K from 1 to 15 and find the optimal K value with a low error rate(1- accuracy score). When K is 6, the error rate is very small at about

0.08, so we use K = 6 to predict and got the F1 score 0.2678. The accuracy is worse than logistic regression.





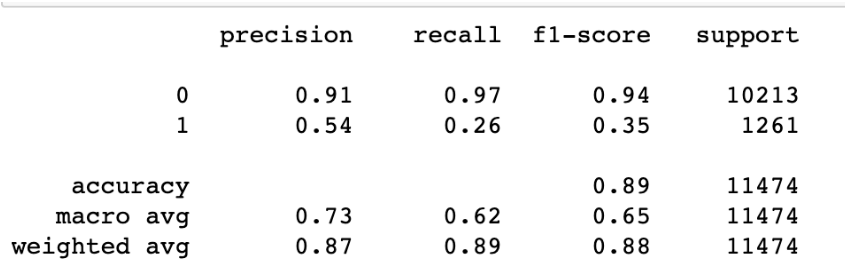
```
              precision    recall  f1-score   support

           0       0.91      0.98      0.94     10213
           1       0.58      0.17      0.27      1261

    accuracy                           0.90     11474
   macro avg       0.74      0.58      0.61     11474
weighted avg       0.87      0.90      0.87     11474
```

(3) Random Forest

Random forest is an ensemble of decision tree algorithms. It is a combination of bootstrap and aggregation of multiple decision trees. To maximize the model's accuracy, we developed different combinations of hyperparameters from the number of trees in the forest, the number of features to consider when looking for the best split and whether to use out-of-bag samples to estimate the generalization score. Because random forest used bootstrap to sample the observations for each decision tree, the out-of-bag sample is not necessarily used to estimate the model.

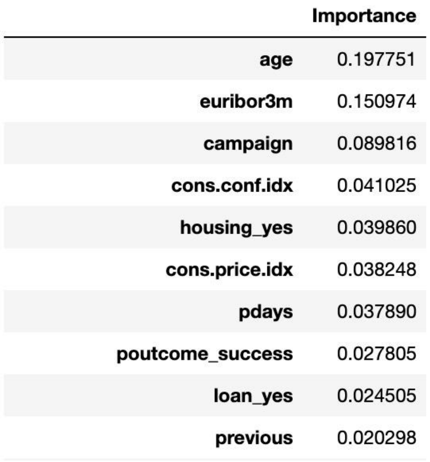
Hyperparameter	Value Lists
Number of individual trees	[64, 100, 128, 200]
Maximum features used in each tree	[2, 3, 4]
Bootstrap option (with or without replacement)	[True, False]
OOB score	[True, False]

The cross validation shows that, the combination of 100 trees, 4 maximum features, bootstrap with replacements and without out of bag score gives the best out-of-sample model performance. Thus, we decide to choose this parameter combination in our final Random Forest model.



```
              precision    recall  f1-score   support

           0       0.91      0.97      0.94     10213
           1       0.54      0.26      0.35      1261

    accuracy                           0.89     11474
   macro avg       0.73      0.62      0.65     11474
weighted avg       0.87      0.89      0.88     11474
```

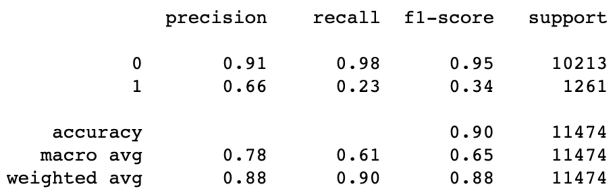
The F1 score of Random Forest is 0.3495978552278821, which is higher than KNN and logistic regression. The model told us the importance of all features. The following table shows the most important ten features.



	Importance
age	0.197751
euribor3m	0.150974
campaign	0.089816
cons.conf.idx	0.041025
housing_yes	0.039860
cons.price.idx	0.038248
pdays	0.037890
poutcome_success	0.027805
loan_yes	0.024505
previous	0.020298

(4) XGBoosting

We did the hyperparameter tuning of XGBoosting from the maximum depth per tree: a deeper tree might increase the performance, but also the complexity and chances to overfit; and the number of trees in our ensemble. F1 score for XGBoosting is 0.3448275862068966, lower than that of Random Forest.



```
              precision    recall  f1-score   support

           0       0.91      0.98      0.95     10213
           1       0.66      0.23      0.34      1261

    accuracy                           0.90     11474
   macro avg       0.78      0.61      0.65     11474
weighted avg       0.88      0.90      0.88     11474
```

Recommendations and Business Value

(1) Model Comparision

Model	Accuracy	F1-score	Recall
Logistic regression	0.90	0.32	0.21
KNN	0.90	0.27	0.17
Random Forest	0.89	0.35	0.26
XGBoosting	0.90	0.34	0.23

In this study, the dataset is highly imbalanced with eight times as many “No subscribe” as ‘Subscribe’. In this case, we can not use accuracy to evaluate model performance (the high accuracy is due to data’s imbalancement). Thus, we use F1-score and recall together to evaluate model performance. The recall is important in this situation because we care more about how accurately we can predict the ‘Subscribe’ case, and recall indicates the percentage of correctly predicted ‘Subscribe’. According to F1-score and recall, the Random Forest gives the best performance.

Although the models do not perform very well in predicting ‘Subscribe’, these machine learning models can provide the bank useful information on how to conduct marketing activities better. In the following section, we will give several recommendations according to these models.

(2) Recommendation for telemarketing

Recommendation 1: Adjust marketing intensity according to the economic cycles

According to the coefficients of logistic regression, we know consumer price index and consumer confidence index will add odds of subscribing. For telemarketing, it suggests that customers' willingness to subscribe is affected by the social-economic indicators. Banks can make more marketing calls when the socio-economic indicators show increasing trends.

Recommendation 2: Segment old customers according to previous marketing outcome

In addition to social-economic indicators, the outcome of the previous marketing campaign can also affect the marketing result. As long as the customers didn't refuse the former marketing campaign, telemarketers should keep contacting them. This group of customers is either loyal to the bank, keen on its various financial products, or may also be easily impressed by telemarketing. This requires the marketing department to make a priority list of customers and pay more attention to these customers.

Recommendation 3: Arrange marketing activities according to seasonal patterns

If the customers had been contacted in March, October, July, and June, the odds of successful subscribing will increase. This is means that some seasonal factors also influence telemarketing. However, we didn't have more data to dive deeper into the seasonal effects. For banks, they need to review and compared the different marketing campaigns held in different months and find the potential correlation.

Recommendation 4: Choose the most effective reaching channel

Compared with telemarketing via cellular phone, telephone marketing has a smaller successful probability. As the same before, we didn't have more related information to compare these two groups of customers. It's possible that cell phone users tend to talk longer with telemarketer (to validate this assumption, the bank need extra data or experiment to rule out possible confounding effects). The bank can further consider more channels, and conduct a series of AB tests to find the most effective reaching channels.

Recommendation 5: Remarketing to increase marketing effect

The more days since the last marketing campaign, the less likely to subscribe. It tells banks that regular marketing campaigns are very important. They should not lose contact with customers too long. Finding a suitable interval for telemarketing to make sure to connect with customers regularly.

Summary and Conclusions

In conclusion, Logistic Regression, KNN, Random Forest, and XGBoosting are not very useful in predicting potential subscribers. However, in these business settings, predicting power is less important than model interpretability. The cost of one extra marketing call is not significant, therefore, conducting more calls to cover more possible subscribers is a better choice than sticking to the model predicted results and missing some subscribers. What's more important in this case is how we can interpret the models to optimize bank's marketing activity.

In our case, we dig into the logistic regression model and give 5 recommendations regarding marketing intensity, customer segmentation, seasonality, channel choosing, and remarketing activity. By following these suggestions, banks can improve their marketing efficiency.