Customer Analytics

MARK5827

Group Research Project

Tutor Name: ☐ Hauke
☑ Amy

Group: 9099.B

Students Names(IDs):

L Jing	z524***
M Jain	z525***
R Korupalli	z523***
Y Dong	z530***
S Jiang	z528***

Date submitted: 7 Nov 2020

1. Executive Summary

This search project is aimed to develop a model that can be used for bank X to target prospects who will buy term deposits in a future campaign. A few stages are applied in the studying process: define the research goal, data analysis and variable selection, model creation and fine-tuning, model estimation and interpretation and final conclusions.

From the prediction some conclusions can be derived: customers with a success or others previous campaign results are very likely to trigger a purchase in the campaign. Moreover, Customers who are married with high level education, without housing loan and private loan are more likely to have a positive response. Customers contacted by cellular and telephone with more than 1 min contact duration are more likely interested in the campaign and have a higher probability to make a positive response.

By using targeted strategy only 2290 out of all 4521 customers are targeted and the cost is $16030. The profit for the bank is going to increase to $74510 from total revenue $90540.

2. Research Background and Objectives

In the past, Bank X always sold products to its customers through rather broad sales campaigns, in which customers were cold-called by call-center personnel. In retrospect, the bank found those campaigns rarely paid-off. The costs for such campaigns were huge, however, only a small number of customers for which the campaign triggered a purchase. And one of the most important reasons is that these campaigns were untargeted.

Therefore, The bank decided to adopt a more targeted approach in its future campaigns. Based on existing clients' data and behaviour from previous campaign, one of the research goal is to provide an approach from existing customers to identify the prospects, who

have a high probability of enrolling in the term deposits campaign, another is to determine what percentage of the customers should be contacted based on the price of agency call cost to avoid a loss from the campaign.

A short outlook on the structure of the report is provided as below:

● Executive summary, background and objectives

● Data analysis and variable selection

● Methodological approach, model creation and fine-tuning

● Model estimation and interpretation

● Implications for Campaign and final conclusions

3. Data analysis and variable selection

A rich dataset bank campaign is provided, which contains information about a random set of existing bank clients, details about the contacts made in the last campaign and other previous campaign attributes, so the data is regarded as valid and reliable. Out of 4521 observations the number of positive response is 521. The dataset attributes is as follows:

6 numeric variables:

Age	***Balance***	***Duration***	***Campaign***	***Pdays***	***Previous***

4 categorial variables measured by “yes” and ”no”, which are transferred into numbers 1 for “yes” and 0 for “no” accordingly:

Housing	***Loan***	***Default***	***Response***

6 categorial variables with more than two groups are transferred into multiple dummy variables using one hot encoding:

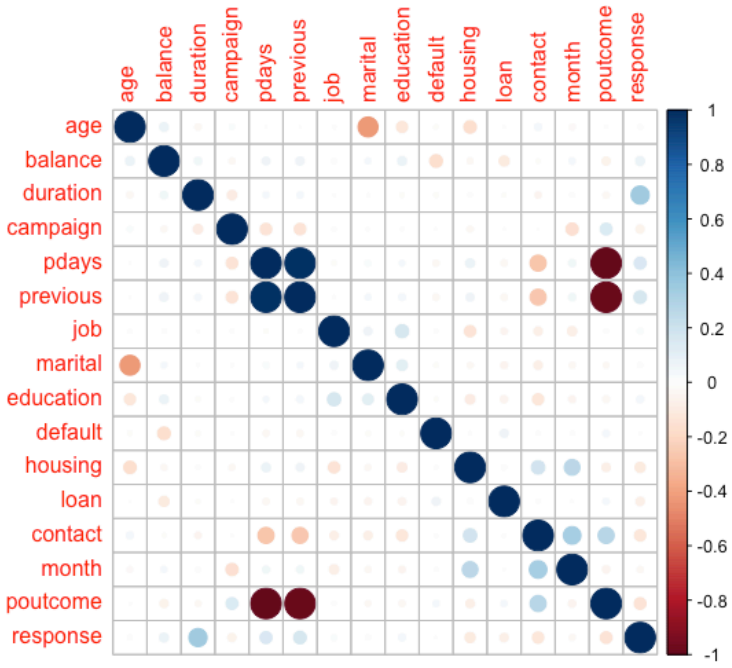
Job	***Marital***	***Education***	***Contact***	***Month***	***Poutcome***

Data analysis and variable selection:

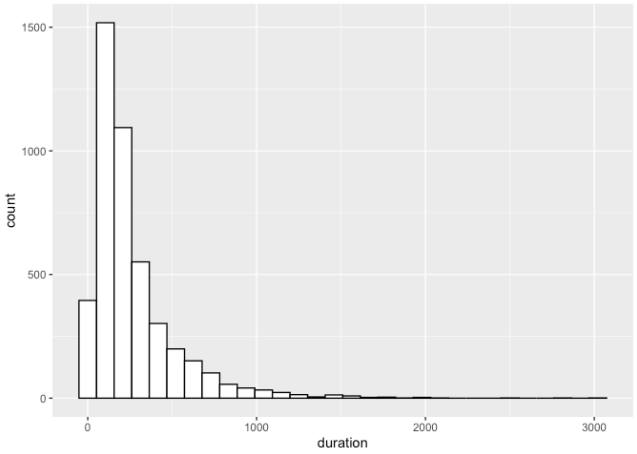
An overall view of correlation between variables is applied before variable transformation as shown below and some viewpoints can be observed:

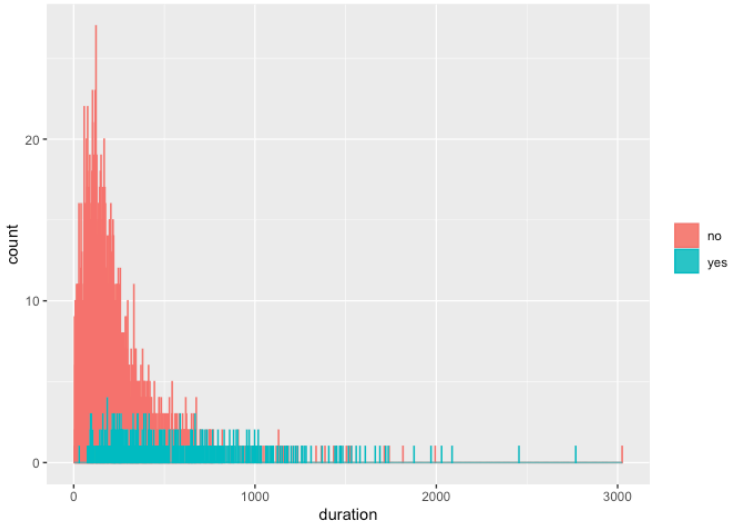
- Correlation between response and other variables: Duration has very strong correlation with response; pdays, previous, housing, contact and poutcome have a strong relationship with response; balance, campaign, loan, education have some relation with response; and for age, job, marital, default and month, the relation between response and these variables is barely observed.

- The correlation between pdays and previous, poutcome and pdays, poutcome and previous is close to 1, therefore, only one variable of each pair can be applied in the model because of too much multicollinearity. More attention is needed for age and marital as well due to the same reason.



Duration is the time duration in seconds about the last contact made in the most current campaign. Through studying and comparing with the duration of positive response, the minimum duration is 30 seconds for positive response, and the probability of positive response is pumped with duration over 60 seconds as the distribution showing, which means people had last contact time over 1 minute with a high probability to trigger a purchase behaviour.

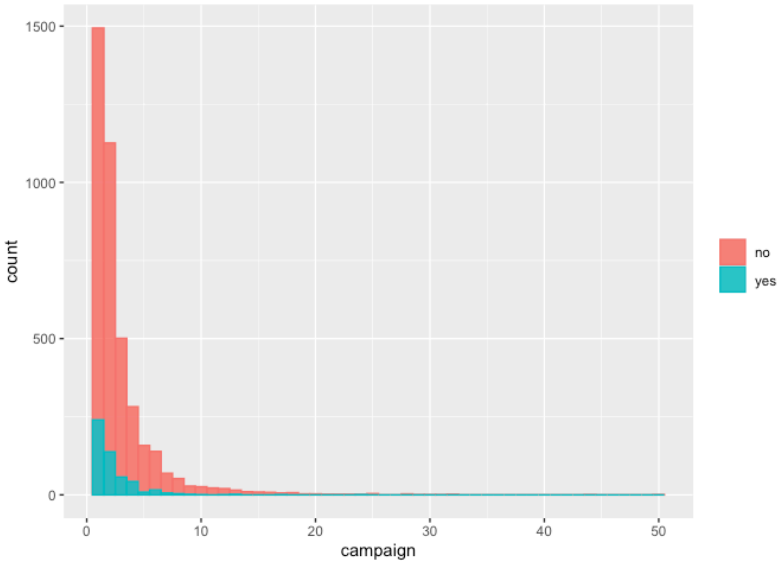




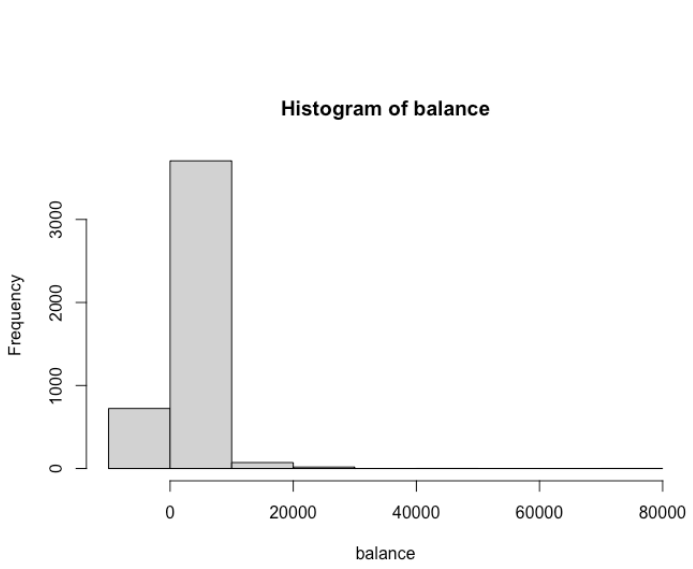
Poutcome is the outcome of the previous marketing campaign. From statistic, there are 83 positive response Out of 129 poucome.success category, which means that more than 64% customers with poutcome.success have had triggered an actually purchase. Similarly, poutcome.failure has positive response with a probability of 19%. Therefore, poutcome is valuable and has a high priority to be kept in the model implementation.

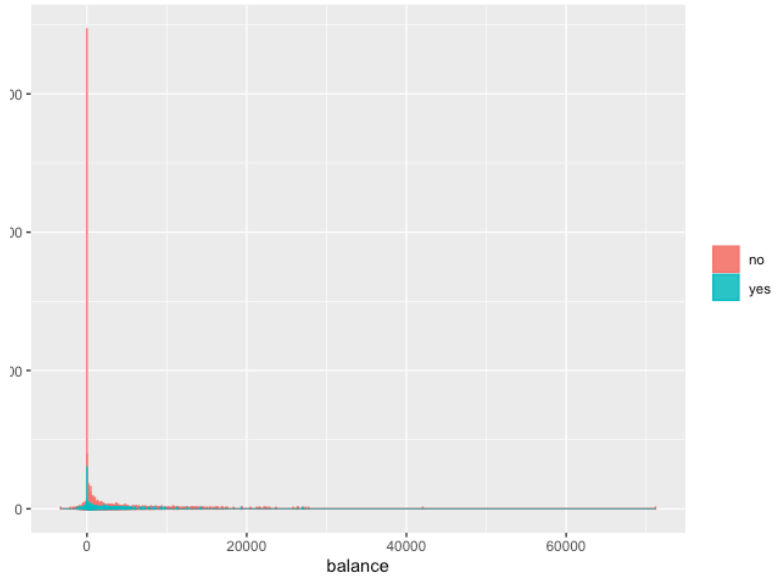
Since the multicollinearity between pdays and previous is too high, **pdays** and **previous** will be considered to be removed with a high priority.

Campaign is the number of contacts performed during the campaign and for the client. The positive response of campaign is mainly between [1, 5] comparing with [0, 50] overall. So campaign variable is optional.



Balance: Many negative values can be observed in balance variable. With further data digging, positive response has an extremely high probability when balance is 0 and positive response is very obvious even the balance is negative, which doesn't make sense. Since from a common sense some money will be held by the bank when the customer triggered a purchase, the bank could not successfully hold money for forthcoming purchase when the balance is negative, it is especially true in the next term deposits campaign. Therefore, balance variable is very suspicious and of limit referential value during the modelling.

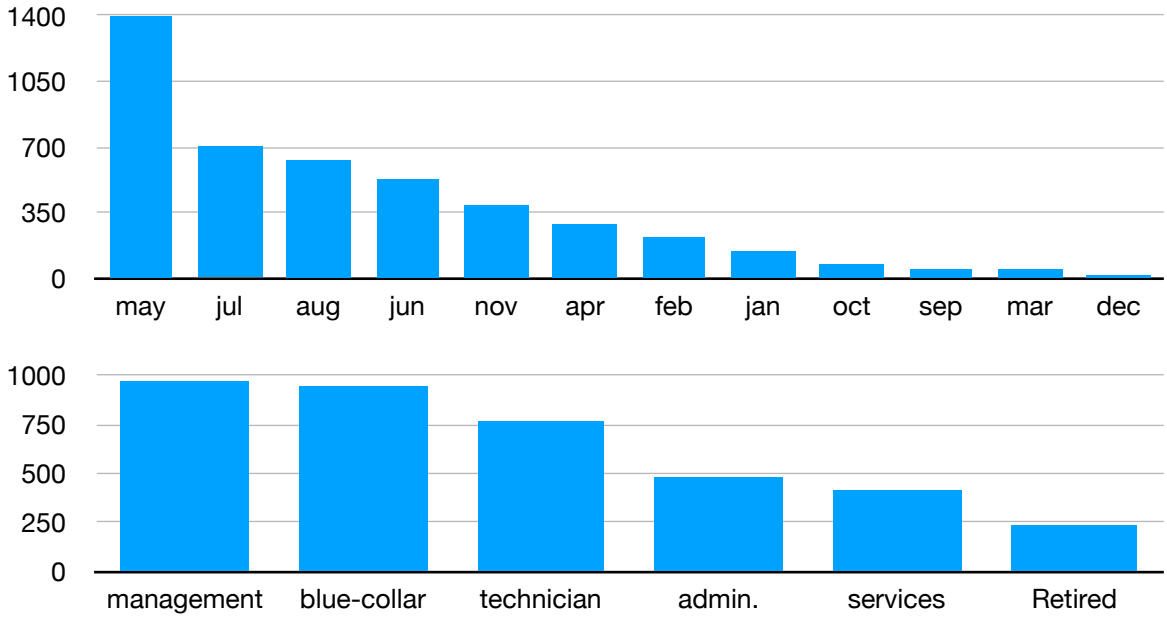




Marital and **Age**: There is not much valuable information got based on the data provided, Marital is three-category variable which is more applicable than numeric variable logistic regression model, and marital and age have a high multicollinearity, therefore, both of them optional, but marital has a higher priority than age during modelling.

Education, housing, loan, contact and default: All of them have two or three categories, which are applicable for logistic regression model, therefore, all of them will be kept or discarded based on their performance during model fine-tuning.

Month and job: both of them have Many categories, partial of them might be kept if they have good performance during model fine-tuning, however, months or job types with big data proportion will be regarded having a higher priority than others since we want to keep as much information as possible.



In summary, the priority of variables selection during model fine-tuning will be as follows:

Variables kept with high priority	Variable optional with high priority	Variable optional with low priority	Suspicious variable	Variable discarded
Poutcome	*Housing*	*Job*	*Balance*	*Pdays*
Duration	*Loan*	*Month*		*Previous*
	Contact	*Age*		
	Campaign			
	Education			
	Merital			
	Default			

4. Methodological Approach, model creation and fine-tuning

Since the campaign response is non-normally distributed outcomes “yes” or ”no”, so **logistic regression model** is selected to do the prediction.

Before applying into model, all categorial variables measured by “yes” and ”no” (including response variable) are transferred from “yes” and “no” into 1 and 0 accordingly. Other categorial variables with more than two groups are transferred into multiple dummy variables using **one hot encoding** approach. Only N-1 dummy variables can be chosen to apply into the model when one variable is transferred into N dummy variables.

Almost all variables are applied to create the first model, then variables are removed from model one by one based on their performance during the **model fine-tuning** process until the best one established. **Log() function** is applied to reduce the influence of numeric values to the model , for example duration and campaign. Besides this, to keep the model as simple as possible during the model fine-tuning process, the model is used as fewer variables as possible.

5. Model evaluation and interpretation

Model evaluation:

- ANOVA tests: the variance in the set of data explained by the model is significantly greater than unexplained variance, and the test is significant 2.2e-16($p<0.05$).
- Hosmer Lemeshow test: it tests whether the predicted values and the actual values are significantly different, and the model results is 0.2363 insignificant(>0.05).
- Nagelkerke R-Square: the Nagelkerke R-Square should be between 0 to 1, the higher there better, and the value of the final model is 0.4090424.
- summary(model): in the summary(model) all model parameters are shown as significant ($p<0.05$).
- vif(model): all values of VIFs are within 2, which are good.

In summary, the final model performs good and meets all main requirements of evaluation.

Following variables are included in the final model :

As mentioned in data analysing part, poutcome.success and outcome.other indicates the results of last campaign, which are meaningful and valuable form both common sense and statistical perspective as shown. Duration variable indicates that customers with more than 1 minute contact call seemed be interested in the campaign and had a high probability of positive response.

poutcome.success	*contact.cellular*	*marital.married*	*month.aug*
poutcome.other	*contact.telephone*	*education.tertiary*	*month.nov*
duration		*housing*	*month.jul*
		loan	*month.may*

From a common sense, whether the customer is married, with a high level education, having hosing loan or personal loan or not will actually affect their finance status and purchase behaviours.

And customers with a cellular or telephone contact method are more accessible by call-centre channel for the bank's campaign, and the customers only might trigger a purchase behaviour once they've been reached. For month variable indicates the last contact month of year, which to some extent indicates the campaign running month.

6. Implications for Campaign

The model can help address bank X's untargeted and barely paid-off issue in campaigns from a statistical perspective. In this case the cutoff was 7/180, i.e. 0.039. Any customer below the cutoff level will not be contacted and all the customers above the threshold level of 0.039 will be known as prospects. The confusion matrix is as follows, and the accuracy is about 0.6008, the sensitivity is 0.5533 and specificity is 0.9655.

	Reference	
Prediction	**0**	**1**
0	2213	18
1	1787	503

The detailed comparison between untargeted campaign and targeted campaign using the model is as follows: Only 50.6%(2290/4521) customers will be contacted in targeted campaign.

	Untargeted campaign	Targeted campaign
Number of customers contacted	4521	2290
Costs per customer contact	$7.00	$7.00
Number of positive response	521	503
Worth of each positive response	$180.00	$180.00
Predicted(cost)	$31647.00	$16030.00
Predicted(revenue)	$93780.00	$90540.00
predicted(Total profit)	$62133.00	$74510.00

Costs info:

Untargeted campaign: 4521*7 = $31647

Targeted campaign: 2290*7 = $16030

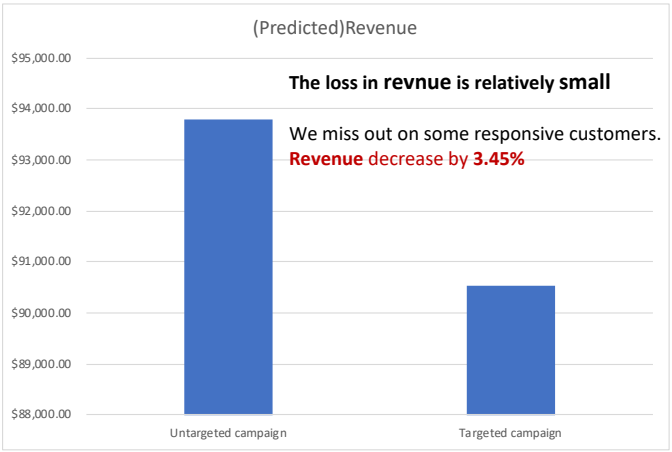
Cost reduction: (16030-31647)/31647 =-49.3%

Revenue info:

Untargeted campaign: 521*180 = $93780

Targeted campaign: 503*180 = $90540

Revenue reduction: (90540 - 93780)/93780 = -3.45%

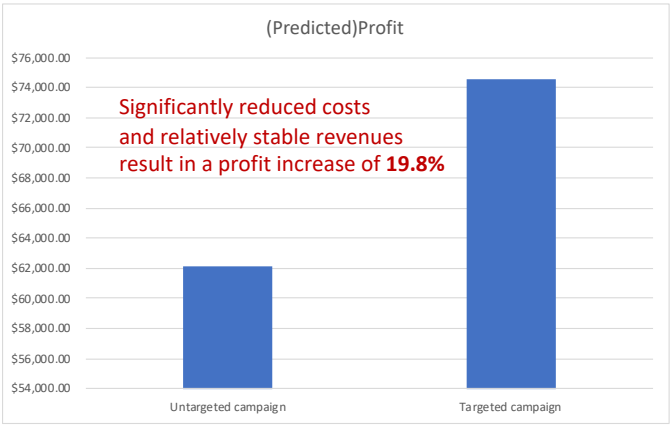


Profit info:

Untargeted campaign: 93780 - 31647 = $62133

Targeted campaign: 90540 - 16030 = $74510

Profit increase: (74510 - 62133)/62133 = 19.8%



In summary, the total cost of the campaign will be almost halved and reduced more than 49% using targeted strategy. Though the total revenue might be slightly reduced about 3.45%, the profit of the campaign would be increased about 20%. Therefore, the model is useful and can help bank X make more profits in the coming campaigns.

7. Conclusion

From this research, based on the current data set provided the overall accuracy for the final model is 60.08%. The specificity is more than 96.5% which means that the model can predict more than 96.5% positive response correctly. The sensitivity is about 55.3%, which means that this model can correctly predict about 55% negative response. Therefore, by using this model more than 96.5% prospects and about 55% customers with negative responses will be identified, which means that prospect customers will be conserved, and at the same time the valueless customers will be filtered out efficiently, therefore, from a marketing perspective the bank can directly focus on prospect customers and save a lot of costs by discarding the customers with negative response.

From a management perspective, below recommendations are provided:

- Customers are very likely to trigger a purchase if the outcome of the previous marketing campaign is success or other.
- Customers with the last contact made in Nov, May, Jul and Aug are more likely to buy the product.
- Customers who are married with high level education, without housing loan and private loan are more likely to have a positive response.
- Customers contacted by cellular and telephone with more than 1 min contact duration are more likely interested in the campaign and have a higher probability to make a positive response.

Reference

Kumar, V., & Reinartz, W. J. (2006). Customer relationship management: A databased approach. Hoboken: Wiley. Chapter 11.
Grigsby, M. (2015). Marketing analytics: A practical guide to real marketing science. Kogan Page Publishers. Chapter 5.

Chapman, C., & Feit, E. M. (2015). R for marketing research and analytics. New York, NY: Springer. Chapter 9.2.

Nussbaumer Knaflic, Cole (2015)., Storytelling with data : a data visualization guide for business professionals. Hoboken, New Jersey: Wiley