Bank Telemarketing Report

Steven Lio

01/20/2022

DSCI 542 Lab2

- GitHub Repo: **https://github.com/stevenlio88/Bank_Telemarketing_Report.git**

Audience personas

- Walter Mitty is a Marketing Campaign Manager for a bank who manage telemarketing campaign. He has a background in finance and marketing and only recently heard about data science and how machine learning algorithms can help provide actionable insights from data and wants to know if the past telemarketing results collected from customers can help him to develop better telemarketing strategies and optimize budget spending as well as campaign success rate.

Report:

Bank Telemarketing Report

Executive Summary:

We detailed the process of building a Logistic Regression Model and show that a out-of-box model with existing customer information can be good at predict the probability of a customer who will subscribe to the new campaign when contacted by the bank through telephone. The simple model used 20 attributes and have a accuracy of 91% and can recall 65% of all the customer who actually subscribed to previous campaign and similar performance on newly unseen data. From studying the model, we now know the best time of the year to run campaign should align with consumer price index which we can use from historical data. Tuesday and Wednesday are the best time for calls as it shows most successful rate are calls from those days. We can further improve the model and develop more flexible and effective telemarketing strategy by using other information from customer such as other demographic information, current product status and customer values. Then combine with the probability we predicted for each customer we can maximize our campaign success rate by going after the high value and high potential customers based on their customer segments in order to get them on board on tailored products designed to specifically to target them.

Introduction

In this report we build a Logistic Regression Model that can predict the probability of a customer who will subscribe to the new campaign when contacted by the bank through a telephone. There are two main benefits of creating this successful prediction model, one is that now the marketing team can now know which customer are likely to subscribe to the new campaign and prioritize to contact them through telemarketing which will improve the campaign success rate as oppose to having cold calling every customer. The second benefit is that using regression model we can also gain insights in how different variables can influence the probability of a customer who will subscribe to the new campaign hence the marketing team can develop more efficient marketing strategies to target different customer group to optimize campaign budget usage.

Logistic Regression model has been proven as one of the powerful machine algorithm which specialize in binary classification tasks as well as providing probability estimate on a success see Moro, S., Cortez, P. & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems. 62, 22-31. For the problem we try to answer, the probability of success will be a customer will subscribe to the new campaign when contacted by the bank over the phone.

Methods

In this section we discuss the methodology and model building process.

Data

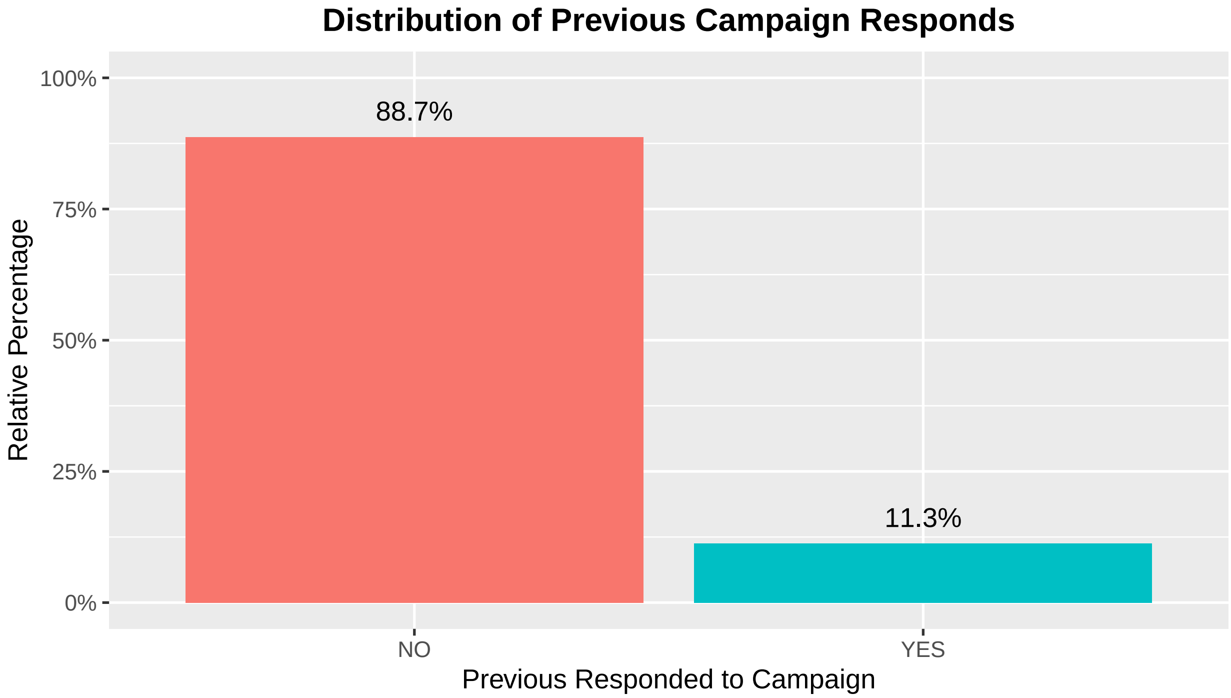
The data used in this report is provided by UCI's Bank Marketing Data Set which contains customers information and attributes regarding their previous contact with the bank and the output of the last contact. There are a total of 20 features for a given customer and his/her response to the previous outcome of the telemarketing call. In order to train the Logistic Regression, we will reserve 20% of the data as test case to validate our model performance hence 80% of the data is used as training data.

Table 1: Attribute from Banking Data.

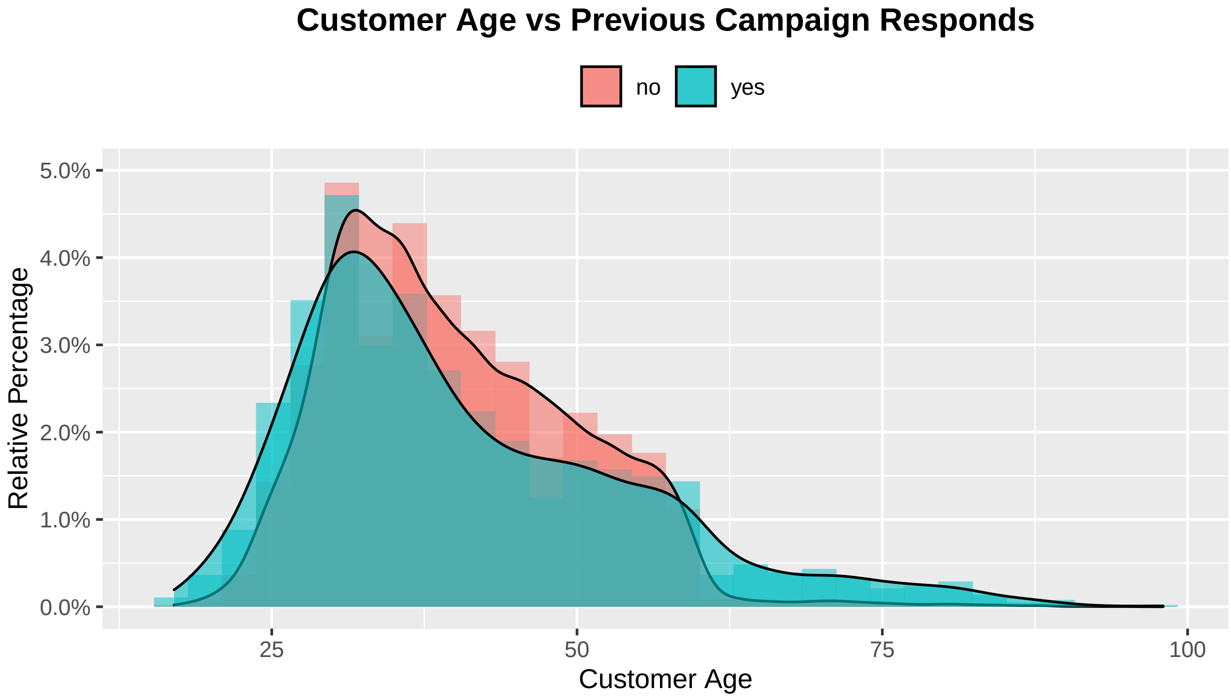
Attribute	Type	Data Type	Description
age	Input	Numeric	Age
job	Input	Categorical	Occupation
marital	Input	Categorical	Marital status
education	Input	Categorical	Education Level
default	Input	Categorical	Has credit in default?
housing	Input	Categorical	Has housing loan?
loan	Input	Categorical	Has personal loan?
contact	Input	Categorical	Contact communication type
month	Input	Categorical	Last contact month of year
day_of_week	Input	Categorical	Last contact day of the week
duration	Input	Numeric	Last contact duration, in seconds
campaign	Input	Numeric	Number of contacts performed during this campaign and for this client
pdays	Input	Numeric	Number of days that passed by after the client was last contacted from a previous
previous	Input	Numeric	Number of contacts performed before this campaign and for this client
poutcome	Input	Categorical	Outcome of the previous marketing campaign
emp.var.rate	Input	Numeric	Employment variation rate
cons.price.idx	Input	Numeric	Consumer price index
cons.conf.idx	Input	Numeric	Consumer confidence index
euribor3m	Input	Numeric	Euribor 3 month rate - daily indicator
nr.employed	Input	Numeric	Number of employees - quarterly indicator
y	Target	Binary	Has the client subscribed a term deposit?

Exploratory Data Analysis

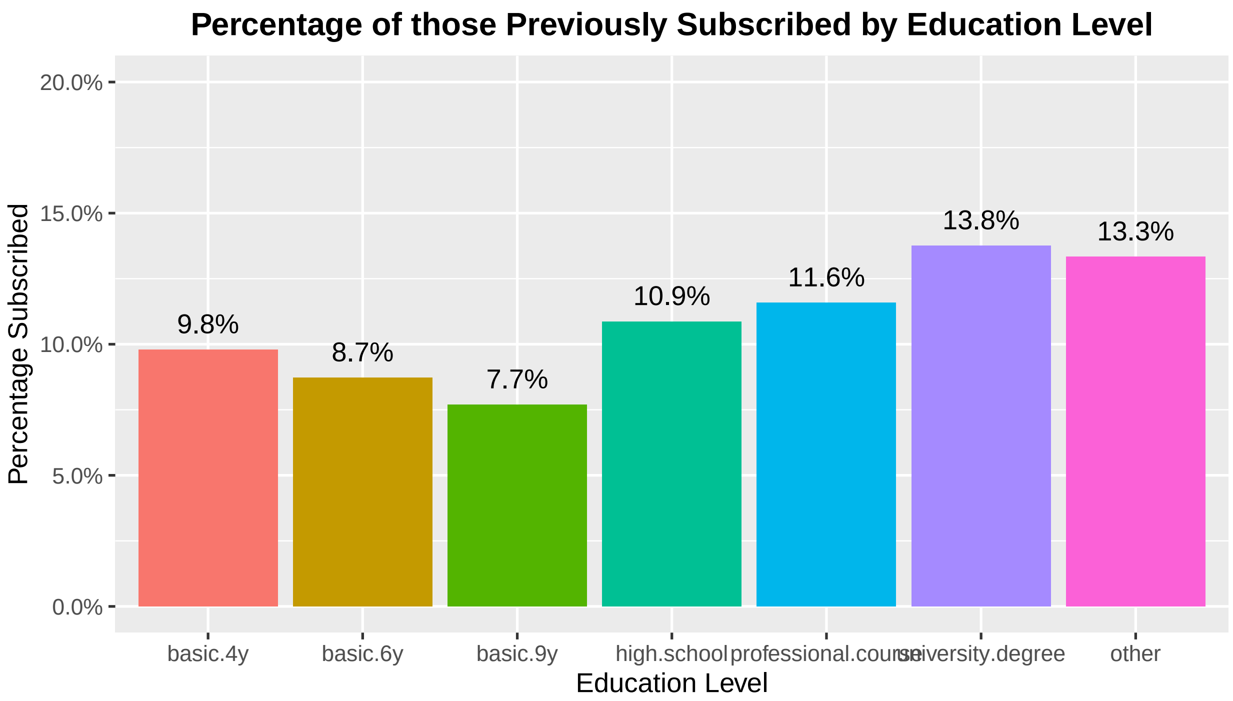
In this section we are going to explore few relationship between the variables and see how they are related to the customer's response to telemarketing.



We can see the previous telemarketing campaigns on average has 11.3% subscription rate.



It is not apparent that customer's age are distinctly different on how they will react to telemarket campaign.



Customer's education level seems to may have contribute to the customer's decision on subscribing to the new campaign when contact through the phone. 13.8% of the customer who have a university degree are more likely to subscribe in previous campaign. Note the average subscription rate was 11.3%.

Model Building

We build a Logistic Regression to predict the probability and the binary outcomes for customer who - `yes`: subscribed to previous campaign after contact - `no`: not subscribed to previous campaign after contact

All 20 variables will be used in building the model, an partial output from the model is shown below:

```
## # A tibble: 54 x 5
##    term              estimate std.error statistic p.value
##    <chr>                <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)      -210.       42.8      -4.91    0
##  2 age                -0.001     0.003    -0.221   0.825
##  3 jobself-employed   -0.196     0.186    -1.06    0.291
##  4 jobstudent          0.113     0.179     0.63    0.529
##  5 jobhousemaid       -0.101     0.209    -0.482   0.63
##  6 jobservices        -0.128     0.16     -0.802   0.422
##  7 jobadmin.          -0.009     0.144    -0.065   0.948
##  8 jobblue-collar     -0.265     0.152    -1.74    0.081
##  9 jobtechnician      -0.039     0.15     -0.263   0.793
## 10 jobmanagement      -0.092     0.162    -0.565   0.572
## # ... with 44 more rows
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik    AIC    BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl>  <dbl>  <dbl>    <dbl>       <int> <int>
## 1        23203.   32949 -6798. 13703. 14148.   13597.       32897 32950
```

```
## [1] 0.9125645
```

```
## [1] 0.6753794
```

The model performance will be assessed on `accuracy` (% of outcome are predicted by the model) and `recall` (% of the customer who actually subscribed are predicted correctly by the model).

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

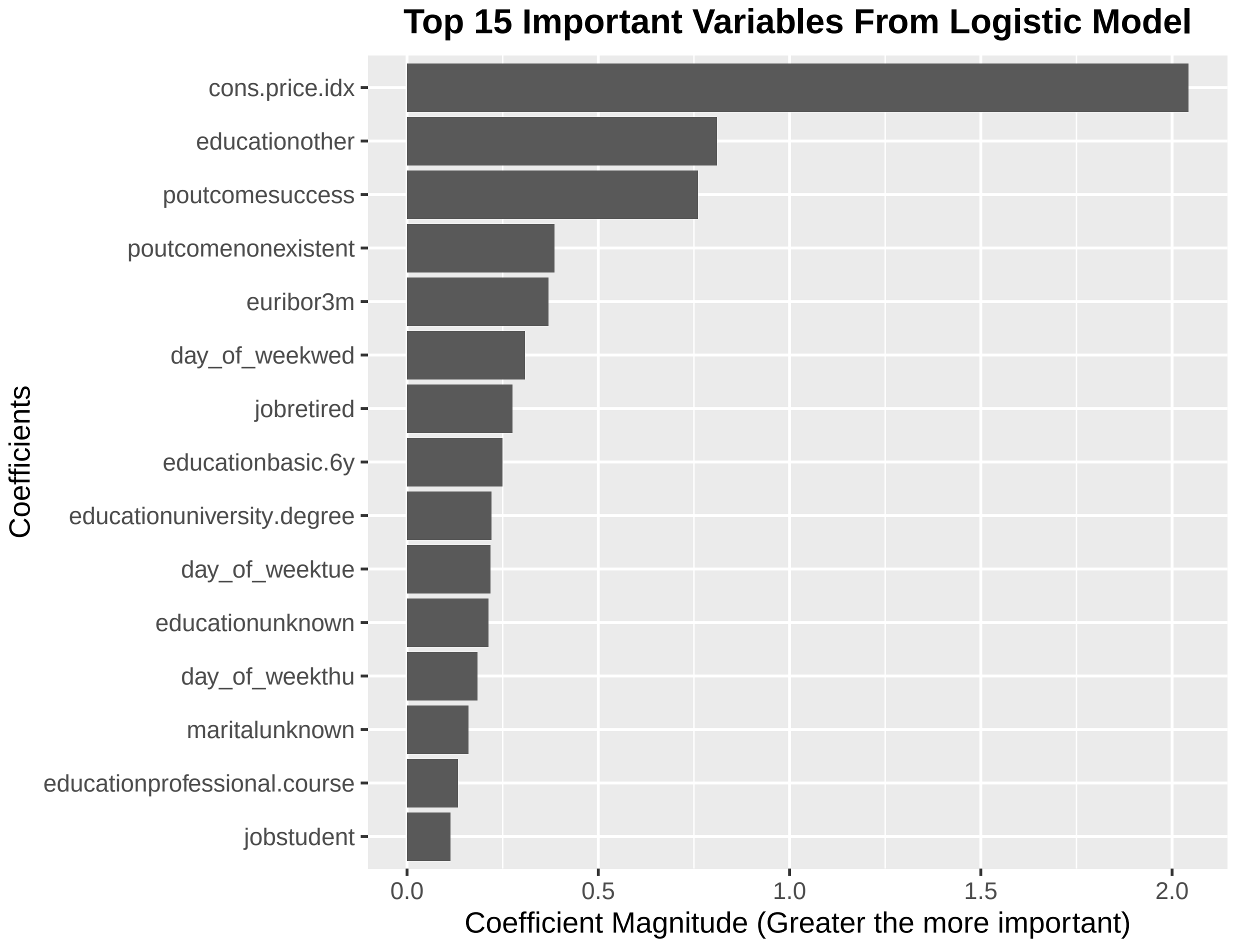
```
## [1] 0.9099296
```

```
## [1] 0.6465927
```

Results

The base Logistic Regression model above has an accuracy of 91.1% and a recall rate of 67%. Hence the model is successfully classify customers 91.1% of the time for their actual decision from previous telemarketing campaign, the model is also successfully recalled 65% of those who actually subscribed previously. The model also have a accuracy of 91.0% and recalled 65% from new data that is not used in training the model.

This shows that a simple robust Logistic Regression can be successful to predict the output of customer's decision from a telemarketing campaign and perform just as well on new data. Hence we can use this model to calculate the probability of the customer who will subscribe to the next telemarketing and we can create a list of highly potential customers and those who are least likely. Further more we can combine the probability along with other customer values attribute such as customer tenure, account values and their number of financial products purchased. Then we can further be more efficient at gaining high value and high potential customer to maximize the campaign effectiveness.



The graph above showed the top 15 variables (order by coefficient magnitude) that are important and significant to the model in the prediction. We can see that `cons.price.idx -` the quarterly consumer price index rank as the top variable in predicting the customer's decision. Note that the second variable is `poutcomesuccess` which is the outcome from previous marketing campaign which is not necessary available for all new customers.

Also The timing variables `day_of_weekwed`, `day_of_weektue` hinting that campaigns ran on Tuesday and Wednesday seems to have more successful with customer's subscription to a telemarketing campaign hence prioritize on increase the number of calls during these two days would improve overall success rate.

`Education` as shown are also one of the top variable in predicting customer's decision at varies degree. These variables can help us to create customer segments along and develop more tailored products/campaigns for different customer segments. But it may requires further analysis to have a more meaningful conclusion.

Conclusion

Overall, a straight out-of-box Logistic Regression model with the variables collected can already be successful at predicting the outcome of customer's decision in the telemarketing campaign. From the model we can now predict the probability of a customer who will subscribe from the telemarketing campaign as well as calls on Tuesday and Wednesday are more likely to be successful than the other days. Also best campaign run time should be check against the quarterly consumer price index to gauge customer's sensitivity to consumer prices to pick the time of the year for the best successful campaign.

Further work can be done to assess other information we could be used in the model to further refine the model performance. Some variables can be used are `gender`, `ethnicity`, `caller's gender`, `preferred language`, `Previous Campaign Type` etc. We can also create customer segments along with customer values so that we can maximize the potential profits from telemarketing campaign, with the probability predicted from the model we can be most efficient and develop effective and flexible telemarketing strategies and optimize campaign budget and ROI.

References

Moro, Sérgio, Paulo Cortez, and Paulo Rita. 2014. "A Data-Driven Approach to Predict the Success of Bank Telemarketing." *Decision Support Systems* 62: 22–31.