Faculté
des sciences économiques
et de gestion
Université de Strasbourg

Data Visualization Essay
(Take-home Project)

By

MAHAMAT IBRAHIM ISSA GUIRE

&

CHENYIN WU

Academic Year 2021-2022

INTRODUCTION

As the context of this essay, we are going to provide a description of the take-home project we did, to make our points clear. This take-home project is based on 3 databases containing information on users of a museum card association. We have completed a series of tasks from data cleaning, data analysis to sample clustering, model prediction, etc. During this period, data visualization has always been our most important tool. Therefore, this essay will combine data visualization graphs to introduce our project in detail. And at the end of it, we will also attach the R code we wrote to answer the questions.

PART 1 : DATA PREPARATION

TASK 1 : Describe the most interesting variables by plotting distributions, correlations, and co-occurrence.

From Figure 1 we can see that, among all users, female users account for more than male users in all age groups. And from the perspective of distribution, users are generally older, the number of users born before 1965 is obviously more than the number of users born after 1965. What we need to pay attention to is an interesting point, that the number of users born in 1950 is staggering, which is not in line with the general distribution logic, but the reason for this phenomenon is still unclear. It may be an error in the original data, or a certain age-specific promotion has been carried out, etc.

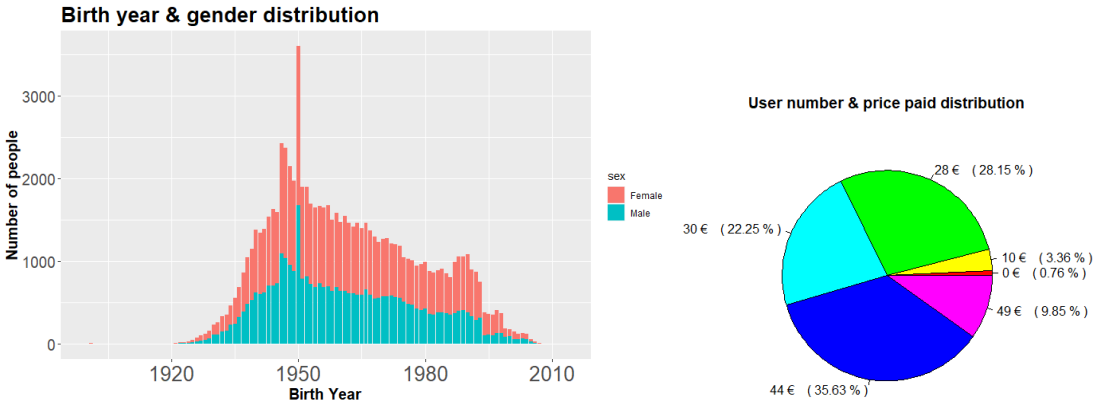

Figure 1 & 2

Figure 2 is the distribution of the user number based on the price they paid. We can see that, except for a very small portion of users who spent very little to buy this card (0 € or 10 €), and about 10 percent of users spent the most expensive price for this card, which is 49 €, the vast majority of users paid one of the 3 prices : 28 €, 33 € or 44 €. According to other variables, we know that prices are largely affected by the type of price reduction.

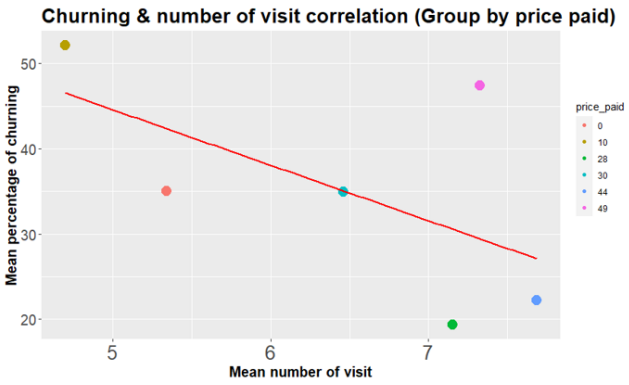

Figure 3

Figure 3 shows our initial understanding of the possible correlations in the data. We categorized each customer according to the price they paid when purchasing this card, and then we want to know the correlation between the average number of uses of each group of users (how many times they went to the museum during the validity period of the card) and the average percentage of churning according to the grouping. In general, we can draw a preliminary conclusion that the higher the price paid for this card, the higher the average number of visits for the users, and the lower the average percentage of churning. But of course, this is only our preliminary understanding of the data and does not have complete accuracy.

TASK 2 : Is there some problem with variables? Are there any specific problems we should take care about?

When we were completing the first task, we had discovered some problems with variables, so we must effectively clean up these three databases.

To clean the database called “data1”, we should firstly check if the variable called “codcliente” in this database has any duplicate values. This variable is the identity of the user so it should not be duplicated. As the second step, we also need to make sure that there are no missing values in “si2014” (binary variable for churning) & “abb13” (variable for the starting date of the card in 2013). What we need to pay attention to is that, for the variables called “ultimo_ing.x” (date of last visit) and “abb14” (renewal date in 2014 if renewed), it is reasonable to have some missing values, because the user may never use the card after buying it, or he may not update his card in 2014.

To clean the database called “in13”, we just need to check if there are any missing values for the whole database. Since this database containing mainly the records of visiting the museum for each of users, we do not need to delete the duplicate values in each variable. After confirming that there are no missing values in this database and no duplicate records in which all variables are identical, we can say that this database has been cleaned up.

To clean the database called “an13”, we need to check if the “codcliente” (identity of user) variable has duplicate values, to delete non-logical values in

the variable “data_nascita” (birth year) which are null or later than 2013. After that, we still need to delete non-logical values in variable “cap” (local area code) which are composed of letters. And finally, we should check if there are missing values for the whole database, the answer is yes so we need to go further to treat them in our third task.

TASK 3 : Analyze the pattern of missing values. Is there any variable we should drop from the analysis?

In order to better understand the existence of missing values in these three databases, we used tools specifically designed to check for them.

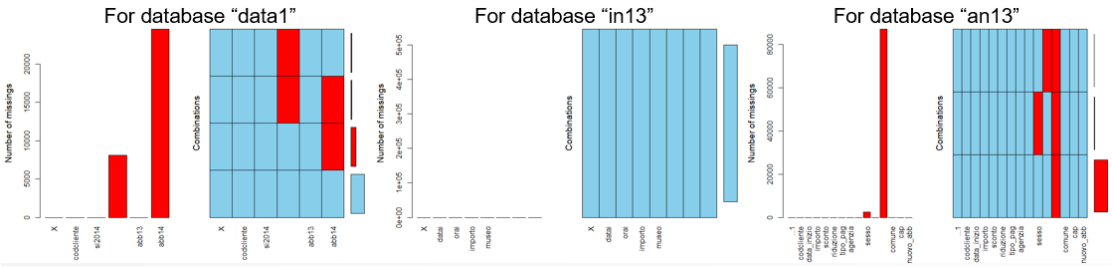

Figure 4, 5 & 6

For missing values in database called “data1”, we have to delete these rows where “si2014” (binary variable for churning) equals to 1, and at the same time “abb14” (renewal date in 2014 if renewed) is missing values. Because the meaning of these rows is not logical : it means that their card was updated in 2014 but there is no renewal date in 2014.

For missing values in database called “an13”, we should delete the rows where “sesso” (gender of user) are missing values. We also noticed that the whole variable “professione” (employment status of user) is empty, so it is better to drop it from the analysis.

TASK 4 : Can we cluster the observations? Is there a cluster with most churners?

For clustering the observations, we introduced five variables to create the clustering model. The first variable we used is called “Churning” (binary variable, 0 for churners and 1 for non-churners), the second variable is called “Number of visits” (number of uses of the card), the third variable is “Gender” (binary variable, 0 for female and 1 for male), the fourth variable we used is called “Birth year” (the year the user was born), and the final variable is the price paid to get this card. We clustered the observations on a 10 by 10 matrix. Figures 7, 8 & 9 show separately the number of observations in each cluster, the composition of elements distinguishing each cluster from other clusters, and the property plot according to the codes plot.

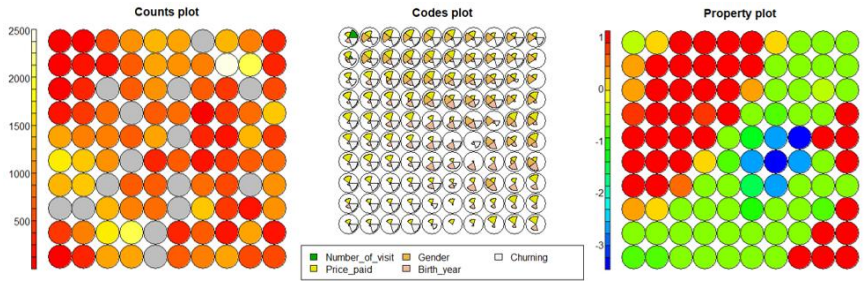

Figure 7, 8 & 9

If we look at the five figures below, we can then get a general impression that, most of the churners have a lower number of visit and a higher price paid for this card. According to the gender plot and the birth year plot, we can say that there are more women than men in churners, and the average age of churners is younger than the average age of overall users, which means that young people have a higher probability to be a churner.

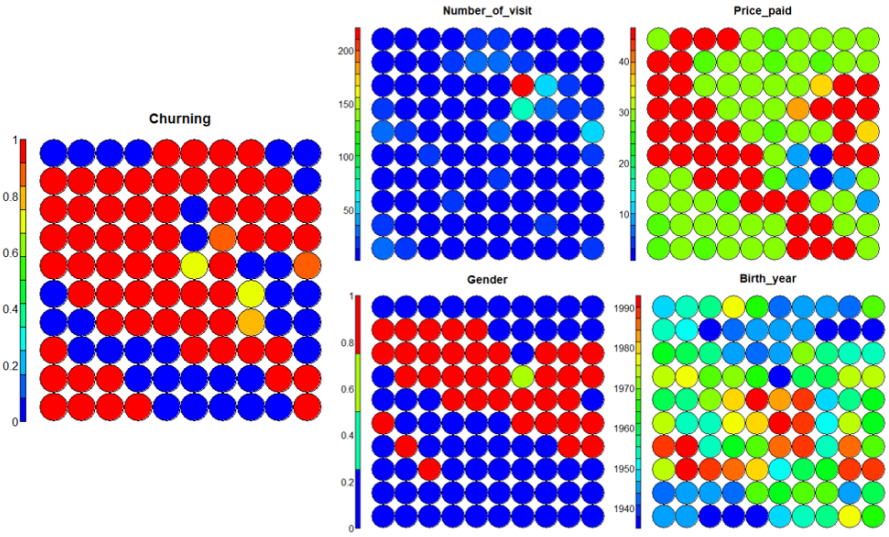

Figure 10, 11, 12, 13 & 14

TASK 5 : Draw some geographical maps of the distribution of card holders, percentage of churners, and average revenues.

We have created 3 geographical maps for the region of northwest Italy according to the number of card holders for each area, the average percentage of churners for each area and the average revenue per user for each area on the point of view of the card association (in a specific area, how much revenue each user of this card association can bring to the card association on average).

Figure 15 tells us that the distribution of card holders is mainly concentrated around three cities and their surrounding areas: Torino, Cuneo and Aosta. Figure 16 shows that, in areas with enough big number of card holders, the distribution of churner percentage is similar. And these areas with extreme percentages, are areas with not many users, so maybe it doesn't have enough meaning.

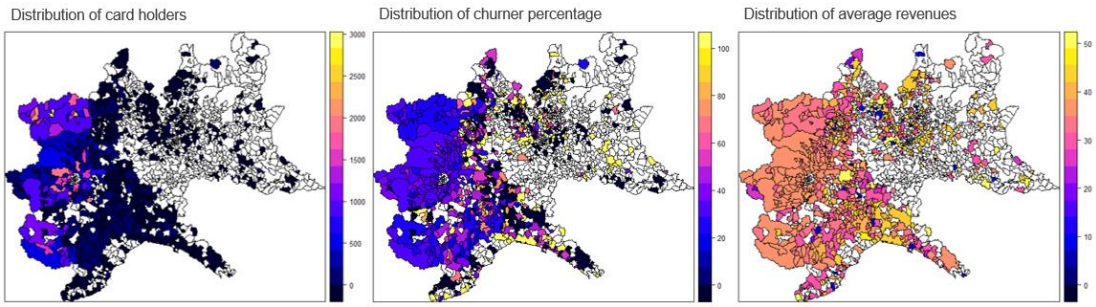

Figure 15, 16 & 17

Figure 17 gives us a picture that, in areas with enough big number of card holders, the distribution of average revenues around the city of Aosta is slightly lower than other areas, we can guess that in order to expand the number of users, perhaps the card association has carried out relatively more promotional activities around this city, resulting in a decrease in average revenues.

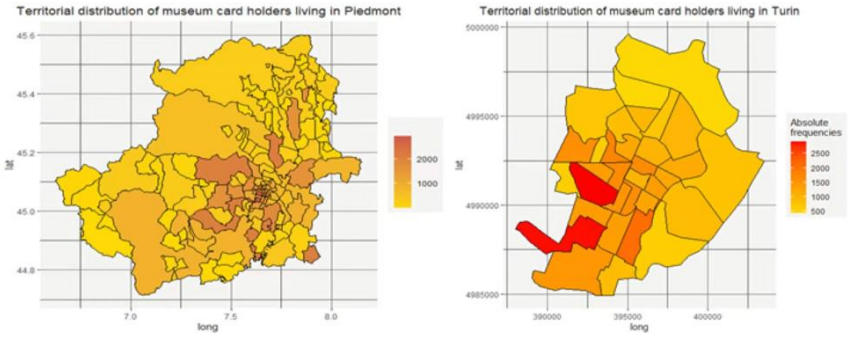

Figure 18 & 19

Figures 18 and 19 tell us the territorial distribution of card holders in Piedmont area and in Turin. It can be seen that card holders are mainly concentrated in Turin and its surrounding areas, and there are two areas with the most card holders in the southwest part of Turin. Probably because there are a large number of companies in these two areas, the number of residents is large there.

PART 2 : CHURN & MARKETING CAMPAIGN

TASK 1 : Is there an impact of age and gender on the probability of churning? Identify a model to answer the question.

For answering this question, we decided to use the Logit Model to measure the impact of age and gender on the probability of churning. As the result, we can see from the Figures 20 and 21, that the age for the card holders in 2014 has a negative impact on the probability of churning. It means that, in general, the older the card holder is, the lower the percentage of churning they have. At the same time, being a male (Gender = 1) has a positive impact on the probability of churning comparing with being a female (Gender = 0). In another word, a

male card holder has a higher percentage of churning than a female when they are in the same age.

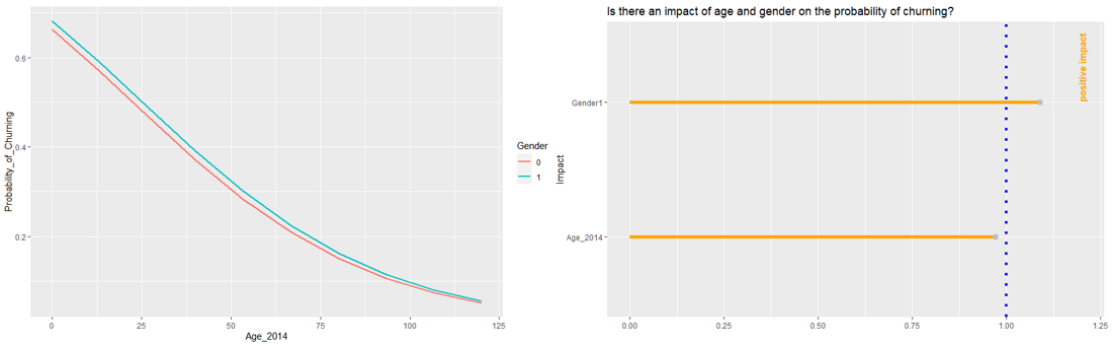

Figure 20 & 21

TASK 2 : Which models could we use to predict churners? Show the ROC curves and the distribution of predicted probability for them.

For creating models to predict churners, we introduced 8 variables in total. They are Churning (1 for churners and 0 for non-churners), Price paid for the card, Type of reduction, Gender (M for male and F for female), Age of 2014 (2014 minus the birth year), Number of visits for each card holder, Amount the client should pay if they don't have the card and finally Type of payment. And we chose Logit Model, Recursive Partition Model and Conditional Tree Model to make predictions. Figure 22 shows us the ROC curves for these 3 models according to their classification performance, then we can get a preliminary judgment, that is, Logit Model and Conditional Tree Model may have better performance.

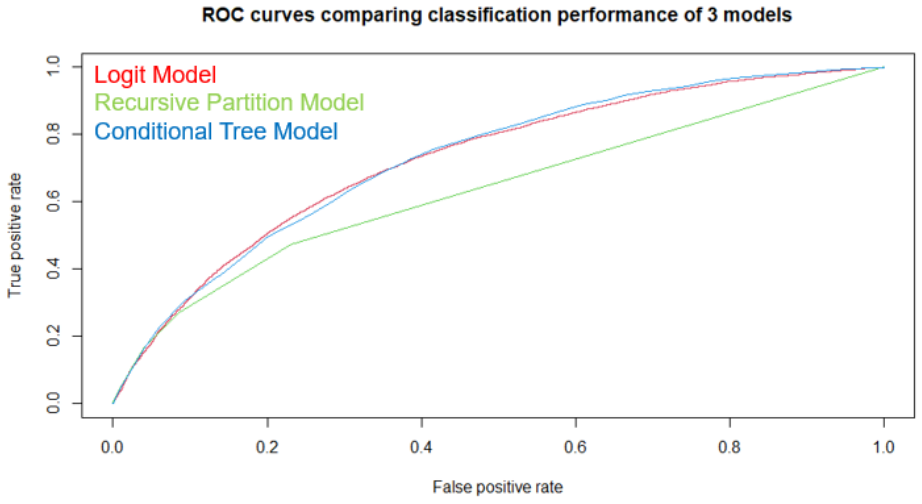

Figure 22

And Figures 23, 24 and 25 are the distributions of predicted probability on the test set for each model, grouping by the true result for churners (in red curve) and non-churners (in blue curve). As the distribution feature of each model, we can say that Logit Model has a distribution which is basically continuous, and

relatively uniform, Recursive Partition Model only has several probabilities which can be predicted, and Conditional Tree Model gives a probability for each possible outcome of the conditional tree.

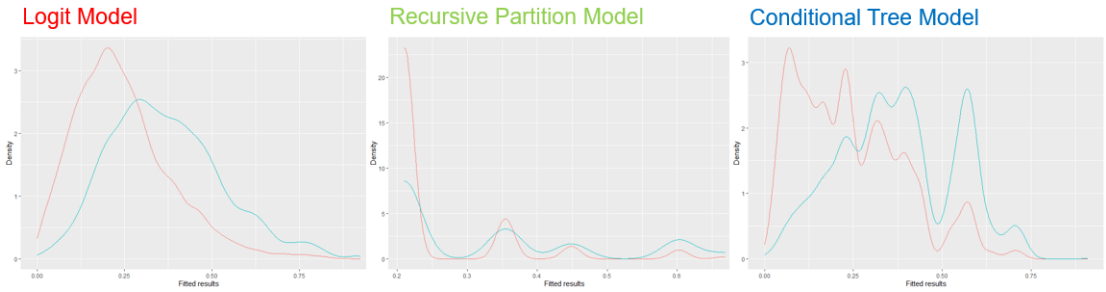

Figure 23, 24 & 25

TASK 3 : Which model are we going to use? Compare the expected profits for each model with the real profit curve.

The following table shows the prediction accuracy of each of the three models. From this we can conclude that the Conditional Tree Model has an accuracy of 74.29%, which is the best among the three models.

Logit Model	**0**	**1**	**RP Model**	**0**	**1**	**CT Model**	**0**	**1**
0	13919	769	**0**	14097	591	**0**	13509	1179
1	4506	1061	**1**	4626	941	**1**	4028	1539
Rate	73.96%		**Rate**	74.24%		**Rate**	74.29%	

Table 1

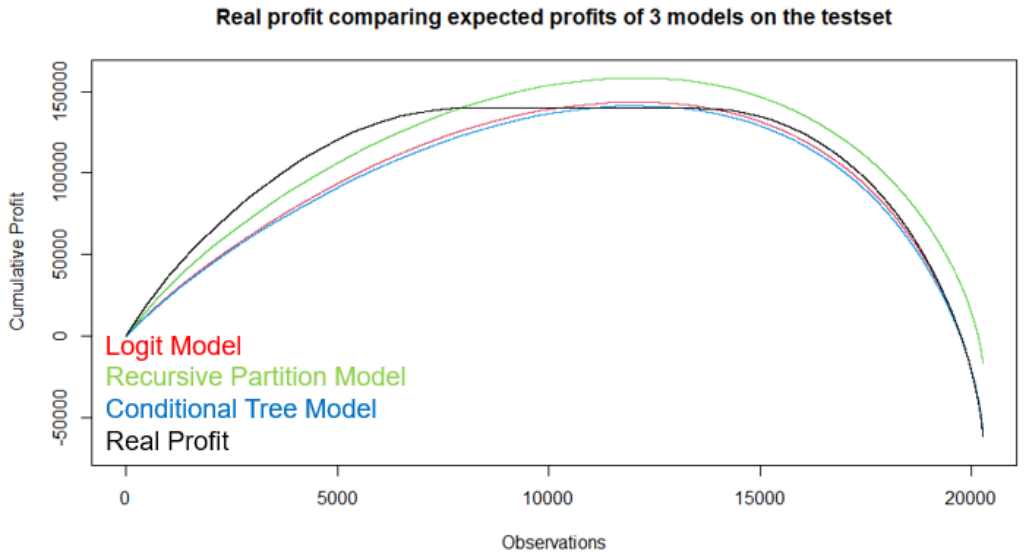

Figure 26

After this we calculated the expected profits of 3 models and the real profit of the test set. It should be noted here that in order to simplify the calculation, we have simplified the cost of the card association for each time of the museum visit to 5 euros, but in the real case, this cost is often less than 5 euros.

Expected profit : Pr * 0 + (1 - Pr) * [Price paid for the card – (5 euro * Number of visits)]

Real profit : (1 – Churning) * [Price paid for the card – (5 euro * Number of visits)]

We then ordered the consumers on the test set according on their expected profits for the card association. From Figure 26, we can see that the curve that fits the real profit best is the expected profit of the Conditional Tree Model. Combined with the above comparison, we finally come to the conclusion that we will use the Conditional Tree Model to make further predictions because it has the relatively best performance among them.

TASK 4 : Given that we contact consumers by expected profits, what is the percentage of consumers we are going to contact?

Next, we used the Conditional Tree Model to predict the expected profit with marketing campaign. We simulate the profits in 2014, assuming that consumers behavior is the same and draw the expected profit curves with different values of ALPHA.

Expected profit with marketing campaign : [Pr * 0.9 * 0 + (1 – Pr * 0.9) * profits] - alpha

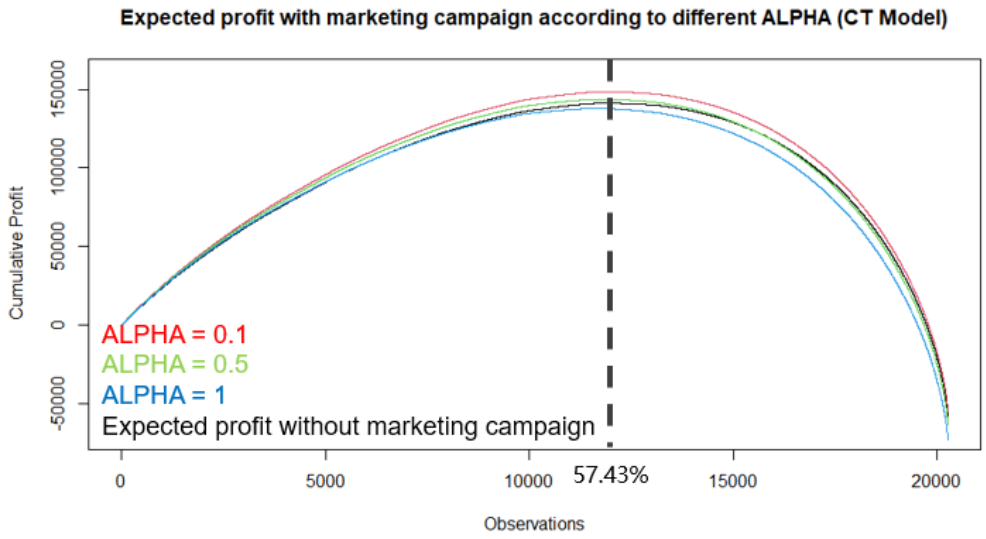

Figure 27

From Table 2, we can know that under our prediction, we should contact the top 57.4% consumers according to their expected profit, for maximizing the expected profit with marketing campaign. When ALPHA is 0.1, our expected profit change will increase by 5.19%. When ALPHA is 0.5, then it will increase by 1.89%. When ALPHA is 1, we should no longer carry out any marketing campaign because the expected profit change will decrease by 2.23%.

CT Model	ALPHA = 0.1	ALPHA = 0.5	ALPHA = 1
Percentage to contact to maximize profit	57.43%	57.43%	57.42%
Expected profit change	+5.19%	+1.89%	-2.23%

Table 2