Data Mining for Business Analytics — Assignment 5
Carlos Fernandez — Teddy Le Nguyen

1. **Estimate the impact of the campaign. You may incorporate several measures of impact (I expect you to determine these on your own). Explain in detail how you got your estimation. You may provide an Excel file to complement your answer.**

   Given the past marketing campaign where SGB sent out 25% steakhouse discounts to subscribed customers, I considered estimating the impact of the campaign in the following ways.

   a. Our first approach is based on the expected value framework.

	Purchase (P)	Not Purchase (NP)
Targeted (T)	V(x)- cost of targeting (C)	-cost of targeting (C)
Not targeted (NT)	V(x)	0

Since we were only given data about the customer's previous transactions at 100 different places prior to the marketing campaign, I proceeded to use these variables to estimate the value of the customers. As I proceeded with the evaluation, I made a few assumptions:

1. The 100 places are related to meat:

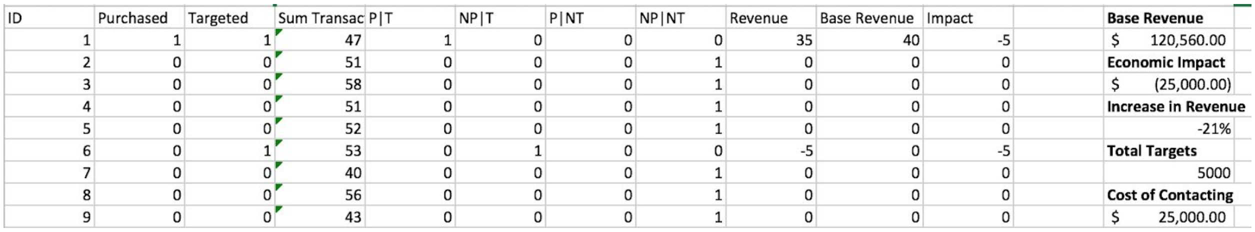
   As the transactions were used to decide who to target with the steakhouse discount, I assumed that these places are related to meat dining, consumption, or purchase (grocery stores, butcher shops, certain restaurants, etc).

2. Each person spends around $40 at the steakhouse:

   In reality, this can very much differ, the price of a dish at steakhouse can differ based on the type of cut or type of the restaurant. On average, a steak dinner (in the US) would cost around $40; therefore, V(x) = $40

3. The only cost is the cost of targeting (C) :

   Since the offer is not at the bank's expense and rather, it could be a product of a partnership between SGB and the Steakhouse (i.e. the discount acts as an advertisement for the steakhouse and if the customer dines at the steakhouse, they would transact with the bank). Therefore, I assumed that the only cost to the bank comes from targeting the customer, which is $5.



ID	Purchased	Targeted	Sum Transac	P|T	NP|T	P|NT	NP|NT	Revenue	Base Revenue	Impact		**Base Revenue**
1	1	1	47	1	0	0	0	35	40	-5		$ 120,560.00
2	0	0	51	0	0	0	1	0	0	0		**Economic Impact**
3	0	0	58	0	0	0	1	0	0	0		$ (25,000.00)
4	0	0	51	0	0	0	1	0	0	0		**Increase in Revenue**
5	0	0	52	0	0	0	1	0	0	0		-21%
6	0	1	53	0	1	0	0	-5	0	-5		**Total Targets**
7	0	0	40	0	0	0	1	0	0	0		5000
8	0	0	56	0	0	0	1	0	0	0		**Cost of Contacting**
9	0	0	43	0	0	0	1	0	0	0		$ 25,000.00

The previous campaign has a negative economic impact of $25,000 and a decrease in 21% of revenue. This is due to the fact that the revenue generated from the campaign is generally less than without the campaign. The revenue for each client depends on whether they were P|T, NP|T, P|NT, or NP|NT with their respective value in the chart above. Although the targeting cost is only $5/targeting decision, the downfall of this marketing campaign could be due to the fact that they only targeted those who had a high probability of purchasing, due to the fact that they were interested in meat consumption. However, the campaign did not take into account the fact that those customers who were interested in meat, would go to the restaurant themselves, without the need for the bank to expend on discount offers.

b. Our second approach is to understand the impact of the campaign via causal inference. The discount caused a client to purchase if it changed their decision from not purchasing to purchasing. Such an effect is calculated by the formula:

Y1 - Y0 where Y1 is the decision to purchase after being targeted

Y0 is the decision to purchase without being targeted

The causal inference must hold the following assumptions:

1. Experimentation takes the value of 1 and 0
2. There are no confounding variables, and the two groups are comparable
3. SUTVA (“no interaction”) where each treatment affects one and only one individual

While the first assumption is met, the third assumption may not, though it is unlikely to make a difference. There are confound variables in the two groups as the dataset only provides the transaction history of each customer while there obviously are other hidden variables that could make the two groups inherently different, thus, not comparable.

The ideal method of matching would be Propensity Score Matching where customers in the control and treatment groups are matched based on the score P(T|X). After the model calculates each customer’s P(T|X) score, it would match customers from each group with the same scores together, calculating the value of Y1 - Y0, thus, the effect of the campaign.

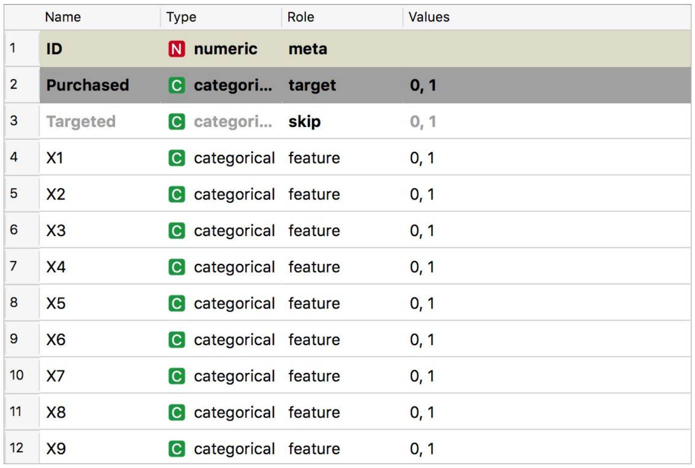
2. **Propose how to improve targeting in future campaigns using this data and predictive modeling. Describe your proposal in detail. You may provide an Excel file to complement your answer (e.g., to show the decisions made by a predictive model).**

We must decide who to target with a 25% steakhouse discount offer with the objective of maximizing the expected revenue for the campaign.

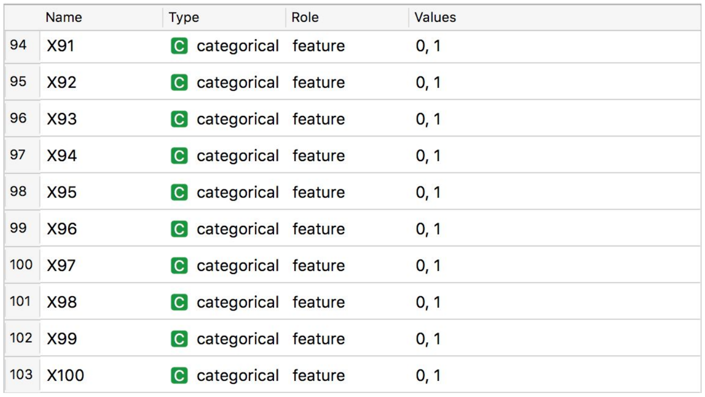
To calculate the expected revenue, we need:
1. A customer’s annual value, V(x)
2. A customer’s probability of purchase, p(P|x, NT)

We could then use historical data to build a predictive model to sort customers according to how likely they are to purchase prior to the discount, p(P|x, NT).

An instance is a customer of SGB, identified by their ID (meta attribute). The target variable is the probability that the customer goes to the steakhouse. The features are whether or not they did a transaction at location i, denoted by variable X with i from 1 to 100, as categorical variables. We should not include Targeted (categorical) because these were the decisions made by the previous campaign, which may cause leakage in our prediction model.

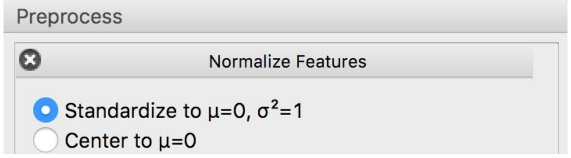


	Name	Type	Role	Values
1	**ID**	N **numeric**	**meta**	
2	**Purchased**	C **categori...**	**target**	**0, 1**
3	Targeted	C categori...	**skip**	0, 1
4	X1	C categorical	feature	0, 1
5	X2	C categorical	feature	0, 1
6	X3	C categorical	feature	0, 1
7	X4	C categorical	feature	0, 1
8	X5	C categorical	feature	0, 1
9	X6	C categorical	feature	0, 1
10	X7	C categorical	feature	0, 1
11	X8	C categorical	feature	0, 1
12	X9	C categorical	feature	0, 1



	Name	Type	Role	Values
94	X91	C categorical	feature	0, 1
95	X92	C categorical	feature	0, 1
96	X93	C categorical	feature	0, 1
97	X94	C categorical	feature	0, 1
98	X95	C categorical	feature	0, 1
99	X96	C categorical	feature	0, 1
100	X97	C categorical	feature	0, 1
101	X98	C categorical	feature	0, 1
102	X99	C categorical	feature	0, 1
103	X100	C categorical	feature	0, 1

I preprocessed the data, standardizing the numeric features with a mean of 0 and var of 1.



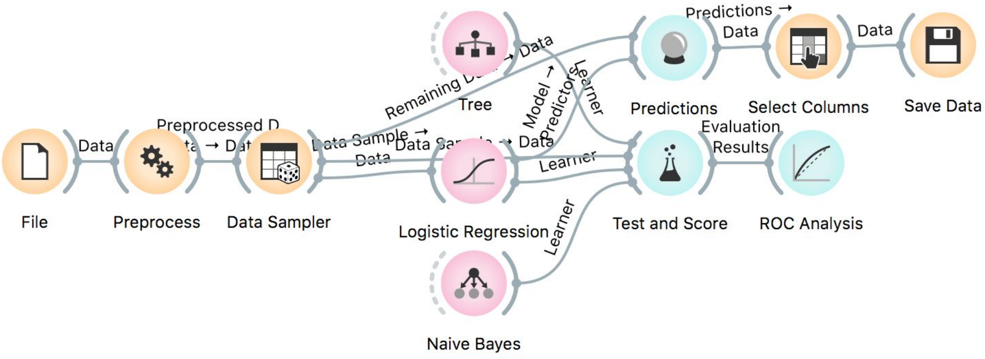
Preprocess

Normalize Features

● Standardize to $\mu=0$, $\sigma^2=1$

○ Center to $\mu=0$

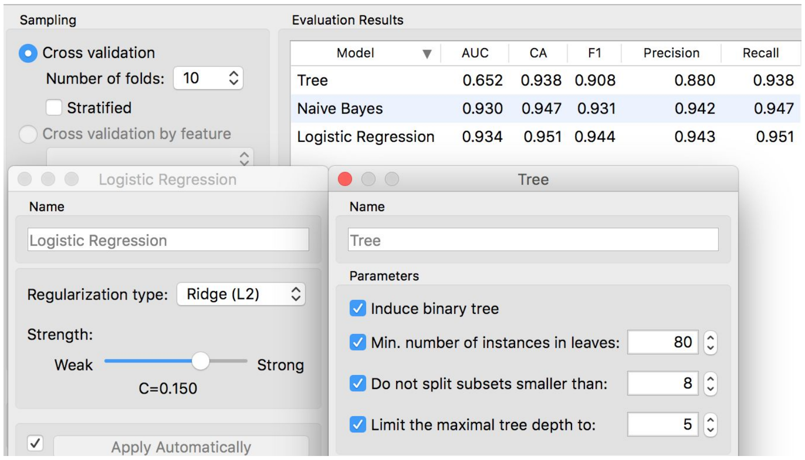
I considered decision trees, logistic regression, and Naive Bayes as our potential models since this is a classification problem.



I sampled 20% of the dataset to use for training the models, since it is a quite large dataset, and the other 80% to test the model. I selected replicable sampling to maintain consistent answers. I used cross-validation with 10 folds to compare models with AUC.

I adjusted in the tree model the minimum number of instances in leaves doubling starting from 2 to 180. I increased the maximal tree depth by 1 incrementally. I found the optimal number was 80 instances and tree depth of 5, yielding the AUC value of 0.652.

After trying both Ridge and Lasso regularization and a complexity constant of 100, 10, .015, and .005, the max AUC of the logistic regression model of 0.934 ultimately outperforms the other models at C=0.15 and Ridge regularization.



Sampling

● Cross validation

Number of folds: 10

□ Stratified

○ Cross validation by feature

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.652	0.938	0.908	0.880	0.938
Naive Bayes	0.930	0.947	0.931	0.942	0.947
Logistic Regression	0.934	0.951	0.944	0.943	0.951

Logistic Regression

Name: Logistic Regression

Regularization type: Ridge (L2)

Strength: Weak — Strong

C=0.150

☑ Apply Automatically

Tree

Name: Tree

Parameters

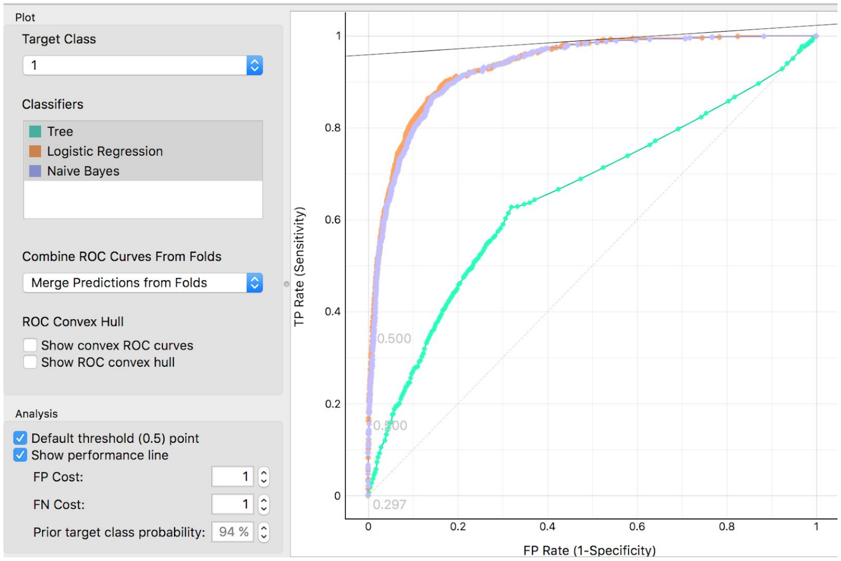
☑ Induce binary tree

☑ Min. number of instances in leaves: 80

☑ Do not split subsets smaller than: 8

☑ Limit the maximal tree depth to: 5

In the ROC analysis, we can see for the most part the logistic regression model and Naive Beyes closely follow each other while the tree model line significantly lower. Using the iso performance line, we can see the Logistic Regression again outperforms the other models where it is tangent (with the assumption that FN Cost = FT Cost). However, even if the assumption does not hold, the LR model seems to outperform the others in other areas of the ROC chart.



Plot

Target Class: 1

Classifiers

🟩 Tree

🟧 Logistic Regression

🟪 Naive Bayes

Combine ROC Curves From Folds: Merge Predictions from Folds

ROC Convex Hull

□ Show convex ROC curves

□ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

FP Cost: 1

FN Cost: 1

Prior target class probability: 94 %

I, therefore, decided to proceed with the Logistic Regression model. The model performs quite well with an AUC of 0.935. In addition, I used the expected value framework mentioned in question 1 to evaluate the model. I proceeded with the same assumptions, namely:

1. The 100 places are related to meat purchase and consumption, therefore, they are relevant to our business decision.
2. Each person spends around $40 at the steakhouse: V(x) = $40
3. The only cost is the cost of targeting (C): C=$5

We want to target a customer with the offer when the value of contacting them EVt(x) is greater than the value of not contacting them, EVnt(x).

$$EVt(x) - EVnt(x) = [p(P|x,T)*(V(x)-C) + (1-p(P|x,T))*-C] - [p(P|x,NT)*V(x) - (1-p(P|x,NT))*0]$$

$$EVt(x) - EVnt(x)= p(P|x,T)*(V(x)-C) + (1-p(P|x,T))*-C - p(P|x,NT)*V(x)$$

We want to contact if EVt(x) > EVnt(x):

$p(P|x, T)*(V(x)-C) + (1-p(P|x, T))*(-C) > (p(P|x, NT))*(V(x)) + (1-p(P|x, NT))*(0)$
$p(P|x, T)*(V(x)-C) + (1-p(P|x, T))*(-C) > (p(P|x, NT))*(V(x))$

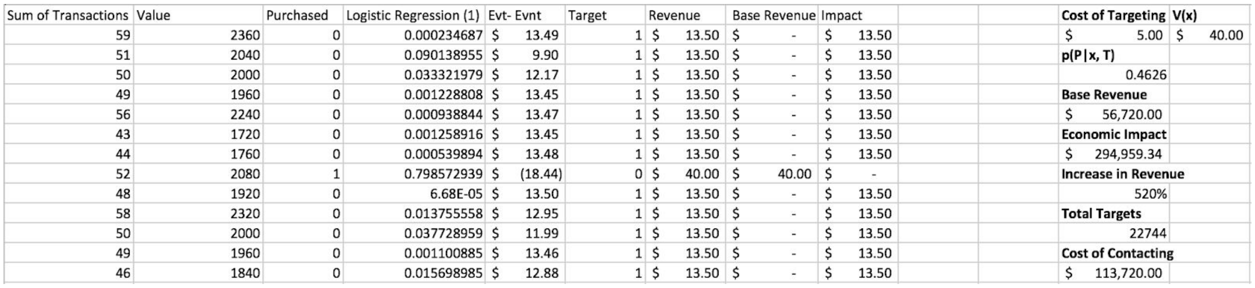
Rearrange the left side to represent the effect of the offer*expected revenue generation:

$p(P|x, T)*(V(x) - p(P|x, T)*C + p(P|x, T))*C - C > (p(P|x, NT))*(V(x))$
$[(p(P|x, T)-(p(P|x, NT))]*(V(x)) - C > 0$

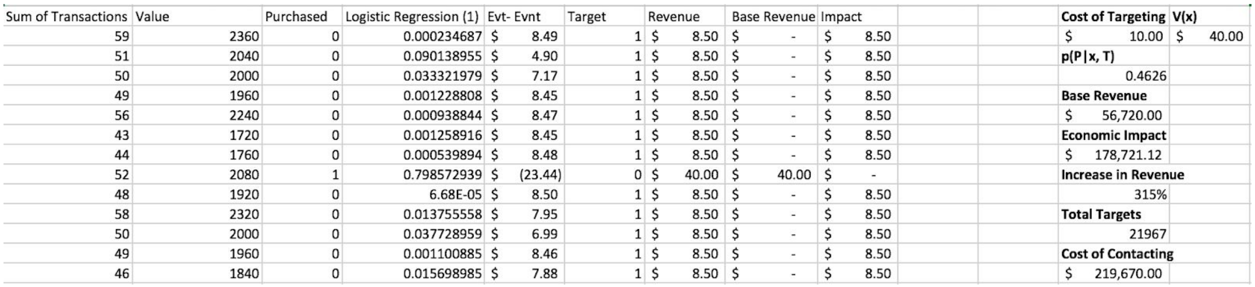
We can obtain values of p(P|x, NT) and customer values from the prediction model. Although the model does not predict p(P|x, T), we can acquire such data from the previous campaign. I calculated the average value of p(P|x, T) from the previous campaign by dividing the sum of P|x, T over the total number of customers targeted, yielding a p(P|x, T) of 0.4626. After exporting data from the Prediction widget, the model suggests the following targeting decisions.

3. **Estimate by how much your proposal could improve things compared to the current targeting procedure. Explain in detail how you got your estimation. You may provide an Excel file to complement your answer.**

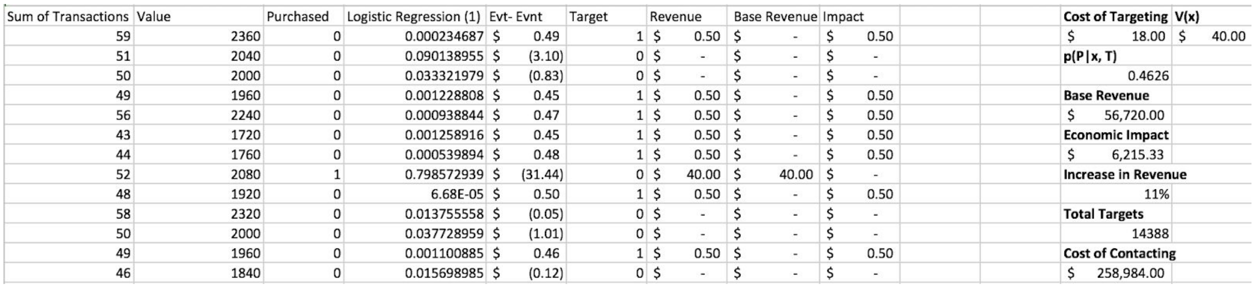
According to the model's predictions on the test data of 40,000 customers, SGB would send the discount offer to 22744 customers whose expected benefit from the offer is greater than the targeting cost of $5. The total economic impact would be $294,959, resulting in a 520% increase in revenue. The estimated cost of targeting of $5 may underestimate the real cost; therefore, I also measured the economic impact given higher targeting costs.



Sum of Transactions	Value	Purchased	Logistic Regression (1)	Evt- Evnt	Target	Revenue	Base Revenue	Impact			Cost of Targeting	V(x)
59	2360	0	0.000234687	$ 13.49	1	$ 13.50	$ -	$ 13.50			$ 5.00	$ 40.00
51	2040	0	0.090138955	$ 9.90	1	$ 13.50	$ -	$ 13.50			**p(P|x, T)**	
50	2000	0	0.033321979	$ 12.17	1	$ 13.50	$ -	$ 13.50			0.4626	
49	1960	0	0.001228808	$ 13.45	1	$ 13.50	$ -	$ 13.50			**Base Revenue**	
56	2240	0	0.000938844	$ 13.47	1	$ 13.50	$ -	$ 13.50			$ 56,720.00	
43	1720	0	0.001258916	$ 13.45	1	$ 13.50	$ -	$ 13.50			**Economic Impact**	
44	1760	0	0.000539894	$ 13.48	1	$ 13.50	$ -	$ 13.50			$ 294,959.34	
52	2080	1	0.798572939	$ (18.44)	0	$ 40.00	$ 40.00	$ -			**Increase in Revenue**	
48	1920	0	6.68E-05	$ 13.50	1	$ 13.50	$ -	$ 13.50			520%	
58	2320	0	0.013755558	$ 12.95	1	$ 13.50	$ -	$ 13.50			**Total Targets**	
50	2000	0	0.037728959	$ 11.99	1	$ 13.50	$ -	$ 13.50			22744	
49	1960	0	0.001100885	$ 13.46	1	$ 13.50	$ -	$ 13.50			**Cost of Contacting**	
46	1840	0	0.015698985	$ 12.88	1	$ 13.50	$ -	$ 13.50			$ 113,720.00	



Sum of Transactions	Value	Purchased	Logistic Regression (1)	Evt- Evnt	Target	Revenue	Base Revenue	Impact			Cost of Targeting	V(x)
59	2360	0	0.000234687	$ 8.49	1	$ 8.50	$ -	$ 8.50			$ 10.00	$ 40.00
51	2040	0	0.090138955	$ 4.90	1	$ 8.50	$ -	$ 8.50			**p(P|x, T)**	
50	2000	0	0.033321979	$ 7.17	1	$ 8.50	$ -	$ 8.50			0.4626	
49	1960	0	0.001228808	$ 8.45	1	$ 8.50	$ -	$ 8.50			**Base Revenue**	
56	2240	0	0.000938844	$ 8.47	1	$ 8.50	$ -	$ 8.50			$ 56,720.00	
43	1720	0	0.001258916	$ 8.45	1	$ 8.50	$ -	$ 8.50			**Economic Impact**	
44	1760	0	0.000539894	$ 8.48	1	$ 8.50	$ -	$ 8.50			$ 178,721.12	
52	2080	1	0.798572939	$ (23.44)	0	$ 40.00	$ 40.00	$ -			**Increase in Revenue**	
48	1920	0	6.68E-05	$ 8.50	1	$ 8.50	$ -	$ 8.50			315%	
58	2320	0	0.013755558	$ 7.95	1	$ 8.50	$ -	$ 8.50			**Total Targets**	
50	2000	0	0.037728959	$ 6.99	1	$ 8.50	$ -	$ 8.50			21967	
49	1960	0	0.001100885	$ 8.46	1	$ 8.50	$ -	$ 8.50			**Cost of Contacting**	
46	1840	0	0.015698985	$ 7.88	1	$ 8.50	$ -	$ 8.50			$ 219,670.00	



Sum of Transactions	Value	Purchased	Logistic Regression (1)	Evt- Evnt	Target	Revenue	Base Revenue	Impact			Cost of Targeting	V(x)
59	2360	0	0.000234687	$ 0.49	1	$ 0.50	$ -	$ 0.50			$ 18.00	$ 40.00
51	2040	0	0.090138955	$ (3.10)	0	$ -	$ -	$ -			**p(P|x, T)**	
50	2000	0	0.033321979	$ (0.83)	0	$ -	$ -	$ -			0.4626	
49	1960	0	0.001228808	$ 0.45	1	$ 0.50	$ -	$ 0.50			**Base Revenue**	
56	2240	0	0.000938844	$ 0.47	1	$ 0.50	$ -	$ 0.50			$ 56,720.00	
43	1720	0	0.001258916	$ 0.45	1	$ 0.50	$ -	$ 0.50			**Economic Impact**	
44	1760	0	0.000539894	$ 0.48	1	$ 0.50	$ -	$ 0.50			$ 6,215.33	
52	2080	1	0.798572939	$ (31.44)	0	$ 40.00	$ 40.00	$ -			**Increase in Revenue**	
48	1920	0	6.68E-05	$ 0.50	1	$ 0.50	$ -	$ 0.50			11%	
58	2320	0	0.013755558	$ (0.05)	0	$ -	$ -	$ -			**Total Targets**	
50	2000	0	0.037728959	$ (1.01)	0	$ -	$ -	$ -			14388	
49	1960	0	0.001100885	$ 0.46	1	$ 0.50	$ -	$ 0.50			**Cost of Contacting**	
46	1840	0	0.015698985	$ (0.12)	0	$ -	$ -	$ -			$ 258,984.00	

In increasing the targeting cost, we can see that the model still proves to be effective in increasing the company's revenue, until the breakeven point of around $20/per target. However, such cost per customer seems to be unrealistic, and quite high compared to an average targeting cost. Therefore, we perceive that the model is effective in deciding which customer to target in order to increase profit.

Another thing to note is that the model differs from the previous campaign. While the model suggests that we send the discount to those who did not go to the steakhouse during the previous marketing campaign, the previous campaign sent offers mainly only to those who did but resulted in a much less significant economic impact.

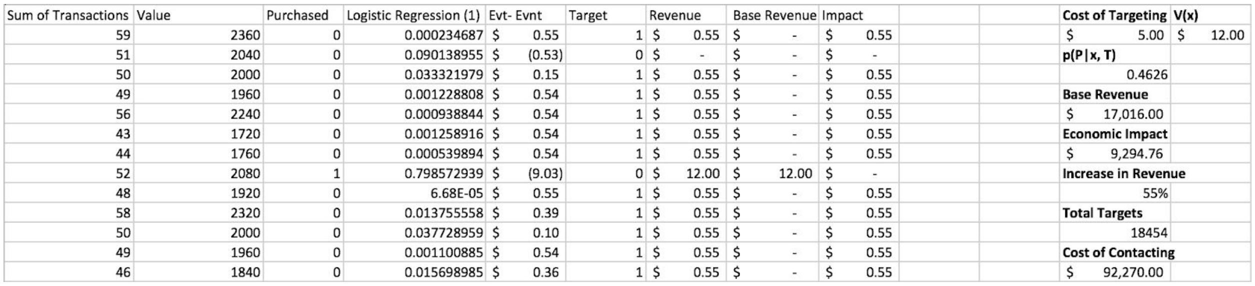
4. **Mention any potential limitations of your estimates and your proposal. Suggest ways in which these limitations may be addressed, including pros and cons of implementing those suggestions.**

1. Given that the model was based on the expected value framework, I used the average probability that a customer would purchase if targeted, calculated by dividing the sum of P|x, T over the total number of customers targeted, to estimate p(P|x,T) for this model. Though this might be a more realistic method than assuming perfect offer p(P|x,T) = 1 (which we can see from the data that it is not the case), if the population in which the model is tested is different than the population of the previous campaign, we may not able to use this number at all.

If that were the case, we may have to assume the perfect offer on the new population while assessing the model's sensitivity to that assumption by observing the change in economic impact revenue increase in changing the probability.

Another alternative would be to examine the highest probability of P|x,NT that would allow for any p(P|x,NT) and still be profitable.

2. I estimated the value of a customer's purchase if they were to go to the steakhouse. Since we do not know anything about the steakhouse, the estimated value can easily be an underestimate or overestimate. However, this can easily be solved once provided with more details about the restaurant. Given that all the other assumptions hold, the model would still provide decisions that are profitable. For example, if the purchase were to be $12 (which is quite cheap in a steakhouse), the model would provide the following economic impact.



Sum of Transactions	Value	Purchased	Logistic Regression (1)	Evt- Evnt	Target	Revenue	Base Revenue	Impact			Cost of Targeting	V(x)
59	2360	0	0.000234687	$ 0.55	1	$ 0.55	$ -	$ 0.55			$ 5.00	$ 12.00
51	2040	0	0.090138955	$ (0.53)	0	$ -	$ -	$ -			**p(P|x, T)**	
50	2000	0	0.033321979	$ 0.15	1	$ 0.55	$ -	$ 0.55			0.4626	
49	1960	0	0.001228808	$ 0.54	1	$ 0.55	$ -	$ 0.55			**Base Revenue**	
56	2240	0	0.000938844	$ 0.54	1	$ 0.55	$ -	$ 0.55			$ 17,016.00	
43	1720	0	0.001258916	$ 0.54	1	$ 0.55	$ -	$ 0.55			**Economic Impact**	
44	1760	0	0.000539894	$ 0.54	1	$ 0.55	$ -	$ 0.55			$ 9,294.76	
52	2080	1	0.798572939	$ (9.03)	0	$ 12.00	$ 12.00	$ -			**Increase in Revenue**	
48	1920	0	6.68E-05	$ 0.55	1	$ 0.55	$ -	$ 0.55			55%	
58	2320	0	0.013755558	$ 0.39	1	$ 0.55	$ -	$ 0.55			**Total Targets**	
50	2000	0	0.037728959	$ 0.10	1	$ 0.55	$ -	$ 0.55			18454	
49	1960	0	0.001100885	$ 0.54	1	$ 0.55	$ -	$ 0.55			**Cost of Contacting**	
46	1840	0	0.015698985	$ 0.36	1	$ 0.55	$ -	$ 0.55			$ 92,270.00	

3. I also assumed that all customers would provide the same V(x). This would mean that the model underestimated the values for those who have more transactions in the 100 predictive locations. We can solve this by building a regression model that estimates the customer’s spending at the steakhouse. However, the data collection process might be costly and the variables used to predict their spending might overlap with what we currently use for this model, which may cause leakage.