ASSIGN-4 CUSTOMER ANALYTICS

Vishanth Hari Raj 206310279

Q1:

$$logit(t) = -10.3659 - 0.0472 * Gender + 0.006 * FrequencyFoodWebsites + 0.0439 * FrequencyTravelWebsites + 0.0698 * RestaurantExp + 0.0341 * TravelExp + 0.0089 * EntertainmentExp - 0.0009 * Income - 0.0008 * Ethnicity$$

Q2: (refer the code)

	Gender	foodwebsites	travelwebsites	restaurantexp	travelexp	entertainmentexp	incm	ethnicdivneigh	y	predict	lift
0	0	158	57	70	63	0	13	0	1	0.525970	1.461026
1	1	187	57	70	63	0	0	29	0	0.554112	1.539201
2	1	313	57	45	63	0	0	0	0	0.320354	0.889874
3	1	310	57	70	63	0	0	14	1	0.723885	2.010792
4	0	37	48	70	63	0	0	0	1	0.268740	0.746499
...	...	...	...	...	...	...	...	...	...	...	...
295	1	141	57	55	63	0	13	0	0	0.251233	0.697870
296	0	232	48	50	26	0	26	14	0	0.073607	0.204465
297	0	302	57	70	63	0	0	14	1	0.723784	2.010512
298	0	84	42	70	63	0	0	0	0	0.272157	0.755993
299	1	385	48	60	37	0	0	14	0	0.361369	1.003803

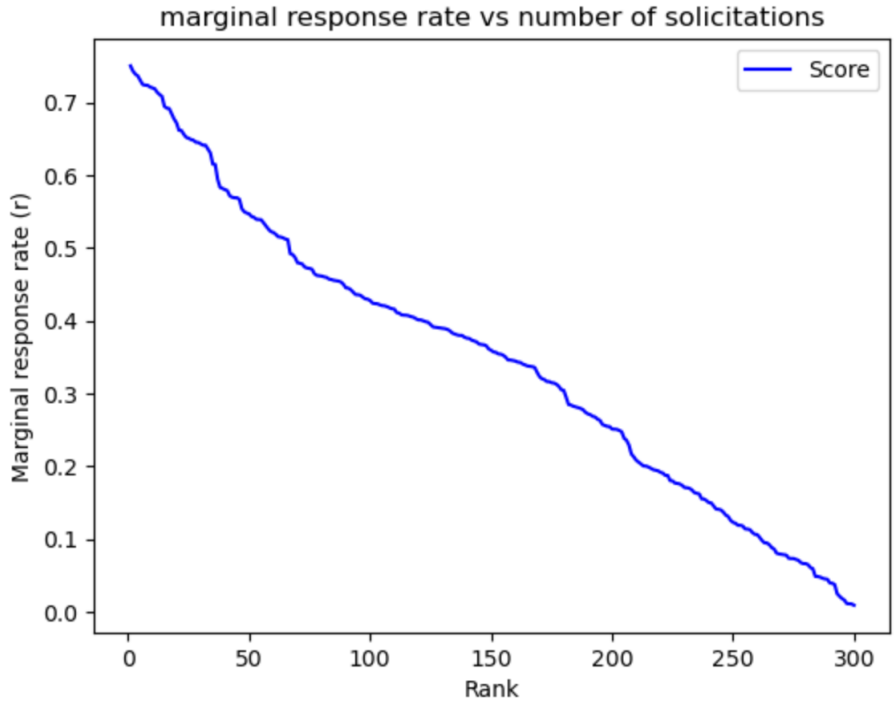
300 rows × 11 columns

Q3: (refer the code)

	Gender	foodwebsites	travelwebsites	restaurantexp	travelexp	entertainmentexp	incm	ethnicdivneigh	y	predict	lift
131	1	317	57	70	63	12	26	0	1	0.750384	2.084399
243	0	321	57	70	63	0	28	0	0	0.743278	2.064661
191	0	315	57	70	63	0	0	14	1	0.739021	2.052836
159	0	311	57	70	63	0	0	0	1	0.736668	2.046299
19	1	314	57	70	63	0	0	0	1	0.730938	2.030384
...	...	...	...	...	...	...	...	...	...	...	...
211	0	406	39	0	63	0	26	0	0	0.016224	0.045067
152	0	291	39	10	52	0	0	28	0	0.011335	0.031487
96	0	159	57	15	41	0	26	0	0	0.011060	0.030722
285	1	354	30	10	52	0	0	0	0	0.010868	0.030188
34	0	382	30	10	41	0	13	14	0	0.009051	0.025141

300 rows × 11 columns

Q4:



Q5:

Ratio above which it to make sense for the firm to contact customers = Solicitation Cost/ Customer Equity

= 12 / 30
= 0.4

121 people exist with probability>0.4

No of prospects holdout list that Melrose Chocolate House should contact **121. (Attached the code below)**

```
sol_cost=12
equity=30
m=sol_cost/equity
m
```

```
0.4
```

Let p is given by solicitation cost/equity. In this case solicitation cost is USD 12 and equity was USD 30. $p_n = \frac{solicitation\_cost}{equity}$

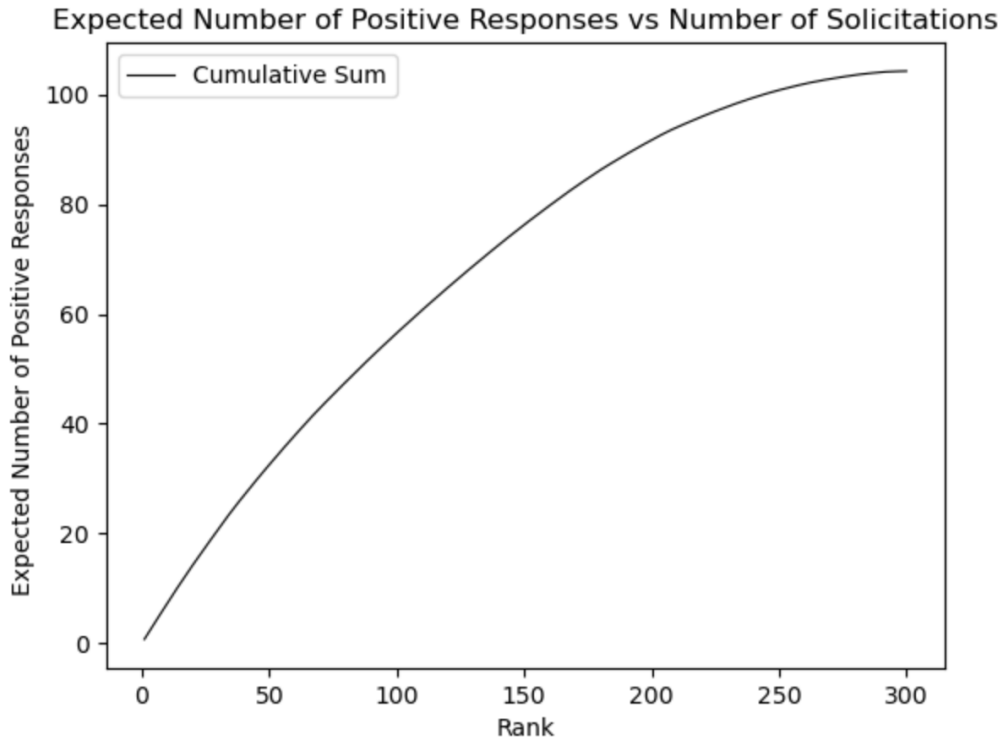
$p_n = \frac{12}{30}$

Now $p_n > 0.4$

```
num = len(plt_hld[plt_hld.predict>(m)]) # 121 people
print("No of prospects holdout list that Melrose Chocolate House should contact", num,)
```

```
No of prospects holdout list that Melrose Chocolate House should contact 121
```

Q6:



The curve depicting the expected number of positive responses vs. number of solicitations does not rise as fast as expected from the usual 80/20 rule. The curve does not seem to follow the general pareto rule we might expect to see. This might have to do with the effectiveness of the campaign and the general homogeneity of the lists that we have.

Effectiveness of the Predictive Model: The logistic regression model used to predict response probabilities may not capture all the relevant factors influencing customer behavior. Important variables might be missing, or the relationships between the variables and the outcome may be inaccurately modeled. This can result in a less steep curve.

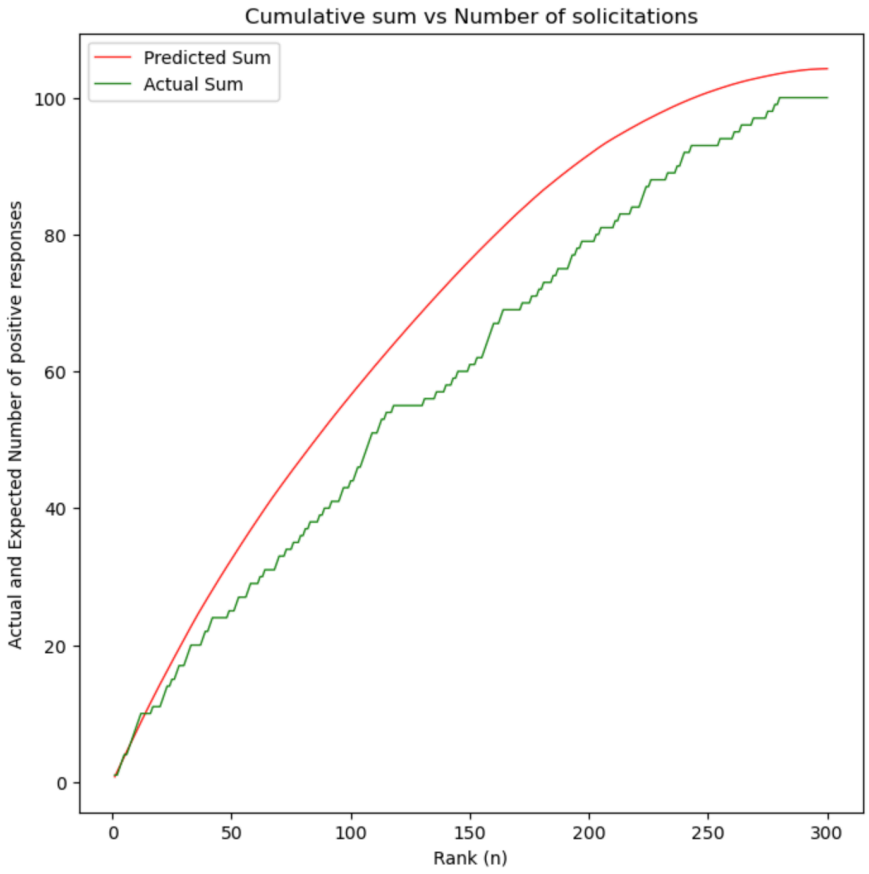
Homogeneity of the List: The holdout list used for testing may be more homogeneous than expected. If the list is composed of individuals with similar characteristics and behavior, the marginal differences in their response probabilities will be small. Consequently, the curve will not be steep at the beginning and will saturate later, as more solicitations are needed to achieve additional responses. This indicates that the population does not follow the typical Pareto distribution (80/20 rule) and instead exhibits less variance in responses.

Overall, the shape of the curve is influenced by the predictive model's accuracy and the characteristics of the dataset. In this case, the less steep curve suggests a more uniform response behavior among the individuals in the holdout list.

Q 7:

As there are only 40 boxes of pralines available for the introductory offer, we should target individuals with a cumulative score close to 40 (39.940648). If the firm contacts 64 customers with a cumulative sum of responses equal to or less than 40, we expect only 40 of them to convert and qualify for the offer. The actual number of conversions or positive responses, when the company contacts the top n clients with the highest predicted scores (r), is represented by the cumulative sum of y at rank n.

Q 8:



We observe that the actual curve has a steeper rise initially but flattens out more quickly compared to the expected curve. This indicates that the actual response rate is higher for the top-ranked prospects but declines more rapidly as we move down the list. In contrast, the expected curve rises more gradually and levels off more slowly, suggesting that the scoring model predicted a more uniform response rate across the entire list.

The differences in the two curves may have to do with the limitations of the logistic regression and to what data it was applied to. Elaborating on the second point, the loss function considered there is not the cumulative sum of probabilities but rather the point probabilities. The difference seems to be large as we are adding together errors from multiple points, and it gets stacked which is why we see the consistent gap. Elaborating on

the first point, the response function of logistic regression to the inputs has its limitations and cannot mimic exact behavior as it is quite biased, which is why we may be seeing the errors.

The discrepancies between the actual and expected responses have significant implications for the decision made in step 7, where we determined the number of prospects to target to maximize profits. Since the actual response rate is lower than expected after the initial top prospects, targeting more prospects based on the scoring model would result in lower profits than anticipated. Therefore, the decision made in Step 7 needs to be revised based on the actual response data to ensure more accurate targeting and profit maximization.

Q 9: (OPTIONAL)

TP = 0.2 * P (Number of top prospects)

TC = 0.8 * C (Number of conversions from top prospects)

Now, we want to calculate the average lift in the top 20% of prospects. Lift is defined as the ratio of the response rate for the top prospects to the overall response rate.

Given the Pareto rule, we know that 80% of conversions happen with just 20% of prospects.

20% P gives 80% C , 20% *P* gives 80% *C*

$$Average\ Lift\ = \left\{\left(\frac{1}{n}\right)\sum_{\{k=1\}}^{n} r_k \,/\left(\frac{C}{P}\right)\right\}$$

Using this rule, we can directly obtain the average probability of conversion for the top 20% of prospects or the numerator of equation 4 for the same:

0.8*C/0.2P* =Average Probability of Conversion

Substituting the values of TP and TC from the equations above, we get:

Lift = (0.8 * C) / (0.2 * P) / (C/P) = (0.8 * C * P) / (0.2 * C * P) =4

Average Lift is 4 for the top 20% of prospects.

Hence Proved that the Pareto 80-20 phenomenon holds then the average lift in the top 20% of the prospects will be 4.

Q 10: OPTIONAL

Using Generative AI Tools

A) The prompt or sequence of prompts that you entered into the generative tool:

Prompt 1: Logistic Regression on the Estimation List
"Write Python code to run a logistic regression predicting a binary outcome variable `y` based on the predictors `gender`, `visit_food`, `visit_travel`, `exp_restaurant`, `exp_travel`, `exp_entertainment`, `income`, and `ethnic_diversity`.

Prompt 2: Evaluating the Score Function
"write code to compute the predicted response probability `r` and the lift for each row?"

Prompt 3: Plotting Marginal Response Rate
"Rectify the error in my Python code to plot the curve for the marginal response rate vs. the number of solicitations made. (give my code)

Prompt 4: Cumulative Sum of Predicted Response Probability
"Rewrite my Python code to compute the cumulative sum of the predicted response probability `r` when the rows are sorted by lift. Plot this cumulative sum vs. the number of solicitations made."

Prompt 5: Comparing Predicted and Actual Responses
"Write Python code to plot the cumulative sum of actual responses and superimpose it on the cumulative sum of predicted response in same graph."

Prompt 6: Approach on 80-20
How to approach the problem: 80-20 phenomenon holds then the average lift in the top 20% of the prospects will be 4. A good starting point is to consider the situation where you have P prospects and C conversions. " proof on this with output I get (given the output)

B) An identification of which sequence of prompts was entered into which generative tool:

Prompts 1-5: ChatGPT (for generating code / rectify errors in my code)
Prompt 6: ChatGPT (On explanation on approach of the proof)

C) Generative tool's output from each prompt, identifying which output corresponds to which prompt and which generative tool:

Prompt 1:

model = smf.logit("y ~ "+'+'.join(est.columns[:-1]), data=est).fit() …. (code)

Prompt 2:

hld['lift']=hld.predict/est.y.mean()

Prompt 3:

plt.plot(plt_hld.index + 1, plt_hld.predict, color='blue', linewidth=1.5)
plt.xlabel('Rank')

```
plt.ylabel('Marginal response rate (r)')
plt.title('marginal response rate vs number of solicitations')
plt.legend(['Score'])
plt.show()
```

Prompt 4:

```
plt.plot(
   plt_hld.index + 1,
   plt_hld.pred_sum,
   color='black',
   linewidth=0.8
)
plt.xlabel('Rank')
plt.ylabel('Expected Number of Positive Responses ')
plt.title('Expected Number of Positive Responses vs Number of Solicitations')
plt.legend(['Cumulative Sum'])
plt.show()
```

Prompt 5:

```
plt.plot(plt_hld.index + 1, plt_hld.pred_sum, color='red', linewidth=0.9, label='Predicted
Sum')
plt.plot(plt_hld.index + 1, plt_hld.y_sum, color='green', linewidth=0.9, label='Actual Sum')

plt.xlabel('Rank (n)')
plt.ylabel('Actual and Expected Number of positive responses')
plt.title('Cumulative sum vs Number of solicitations')
plt.legend()
plt.show()
```

Prompt 6:

Assume PP as the total number of prospects and CC as the total number of conversions.
Lift$i$=riAverage Response RateLift$i$=Average Response Rateri
Average Response Rate=CPAverage Response Rate=PC
Lift$i$=ri(CP)Lift$i$=(PC)ri
Average Lift=(1n∑k=1nrk)(CP)Average Lift=(PC)(n1∑k=1nrk)
where nn is the period for which we want to define our average lift.