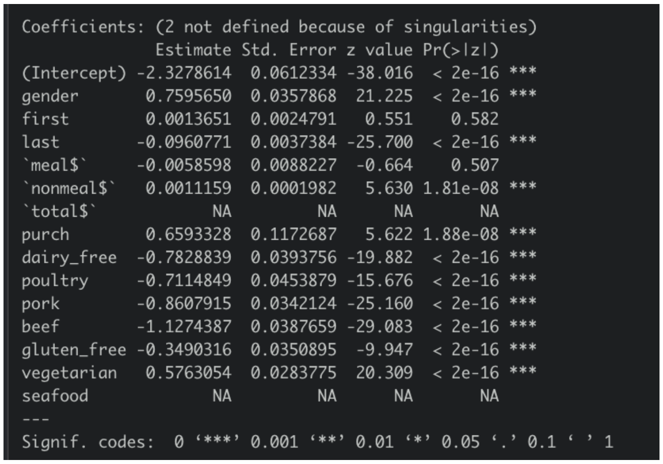
Customer Analytics Using R
Jinyu Li
Meals A La Minute

Part I: Logistic Regression

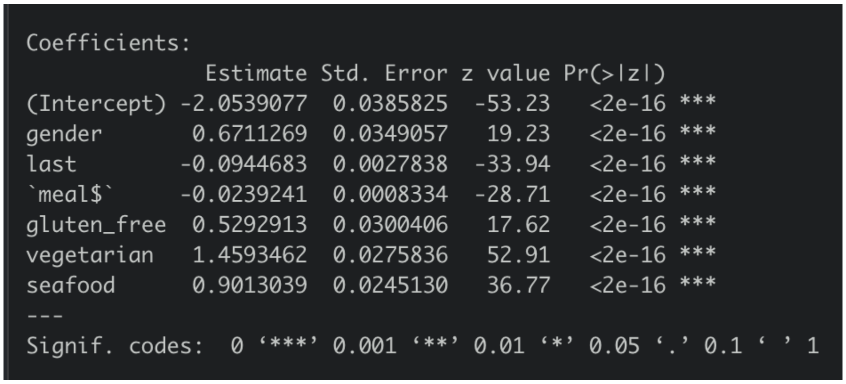
(1) We set 'buyer' as the dependent variable as we want to predict whether or not the consumers will ultimately decide to purchase the product. Also since we are working with logistic regression which only supports binary dependent variables, the buyer variable is the best choice that we have.

(2)



```
Coefficients: (2 not defined because of singularities)
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3278614  0.0612334 -38.016  < 2e-16 ***
gender       0.7595650  0.0357868  21.225  < 2e-16 ***
first        0.0013651  0.0024791   0.551    0.582
last        -0.0960771  0.0037384 -25.700  < 2e-16 ***
`meal$`     -0.0058598  0.0088227  -0.664    0.507
`nonmeal$`   0.0011159  0.0001982   5.630 1.81e-08 ***
`total$`            NA         NA      NA       NA
purch        0.6593328  0.1172687   5.622 1.88e-08 ***
dairy_free  -0.7828839  0.0393756 -19.882  < 2e-16 ***
poultry     -0.7114849  0.0453879 -15.676  < 2e-16 ***
pork        -0.8607915  0.0342124 -25.160  < 2e-16 ***
beef        -1.1274387  0.0387659 -29.083  < 2e-16 ***
gluten_free -0.3490316  0.0350895  -9.947  < 2e-16 ***
vegetarian   0.5763054  0.0283775  20.309  < 2e-16 ***
seafood             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As seen above, we noticed that some variables were significant while others were not. Thus we decided to pick variables that were deemed statistically significant (based on the Signif. codes), as well as ones that made sense since we are trying to predict the purchase of a specific product type.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.0539077  0.0385825  -53.23   <2e-16 ***
gender       0.6711269  0.0349057   19.23   <2e-16 ***
last        -0.0944683  0.0027838  -33.94   <2e-16 ***
`meal$`     -0.0239241  0.0008334  -28.71   <2e-16 ***
gluten_free  0.5292913  0.0300406   17.62   <2e-16 ***
vegetarian   1.4593462  0.0275836   52.91   <2e-16 ***
seafood      0.9013039  0.0245130   36.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(male = 1, female = 0)

As we can see from the two sets of data, the second analysis shows that all independent variables are highly significant (p-value = 0), as compared to simply using all given variables as independent (as some are not as significant as others, like poultry, pork, beef, etc.). We also decided against using zip codes and areas since they seem to complicate our model beyond the scope of our analysis. Hence we decided that our independent variables are gender, last (recency variable), meal$ (since white risotto is a meal and hence we are not concerned about $ earned from non-meals), and # of purchases of gluten-free, vegetarian, and seafood versions. We decided to focus on the latter three variables so that our model can be based on the eating habits and patterns of MALM's consumers so that they can use the results of the regression to target consumers according to what trend in consumption they seem to follow.

(3) All chosen variables have a p-value of 0, which is < 0.05 and hence significantly impacts buyer behavior.

The variable for gender has a positive b value, which shows us that the gender variable has a direct relationship with the number of buyers i.e. male customers are more likely to respond to emails than females. The odds of a male customer responding are exp(0.67) = 1.95 times more than a female customer. Hence MALM can differentiate their email campaigns based on gender, and how to appeal more to female customers than males.

The variable for recency, defined by the 'last' variable has a negative b value. Hence we infer that the variables have an inverse relationship, i.e., the higher the weeks of the last purchase, the lower the probability of the buyer purchasing the White Truffle Risotto meal. Exp(-0.095) = 0.90, which implies that the odds of more recent customers responding are 0.9 times that of less recent customers, holding all other variables fixed.

The variable for total dollar spent on meal kits, defined by 'meal$' has a negative value and hence an inverse relationship i.e., more money spent on meal kits, lower the probability of customers purchasing the white risotto meal. Exp (-0.024) = 0.97, which means that the odds of consumers responding to emails are 0.97 times that of spending money on meal kits. While there is a negative relationship for this variable, 0.97 is very close to 1, and hence can be interpreted as not affecting purchasing behavior as much.

The variable for the number of gluten-free meals ordered has a positive value and hence a direct relationship with consumer response. Exp (0.53) = 1.69, which implies that consumers who buy gluten-free meals from MALM have 1.69 times higher chances of responding to emails about white risotto than those who don't. Hence the gluten-free market segment can be a good area of target for MALM's email campaign to increase profits.

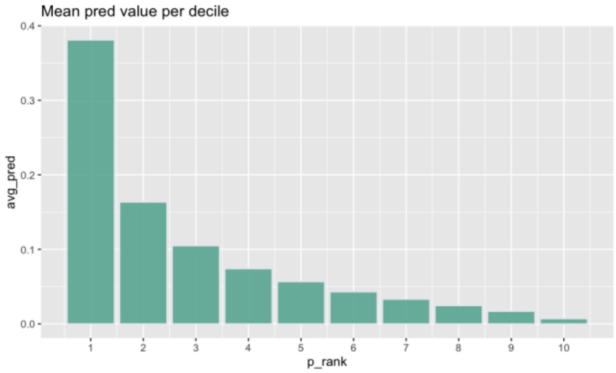
The variable for the number of vegetarian meals ordered has a positive value and hence a direct relationship with consumer response. Exp (1.46) = 4.3, which implies that consumers who buy vegetarian meals from MALM have 4.3 times higher chances of responding to emails than those who don't. Hence MALM's vegetarian consumer base is an extremely opportunistic target for MALM's email campaign to increase return on marketing expenditure and increase profit.

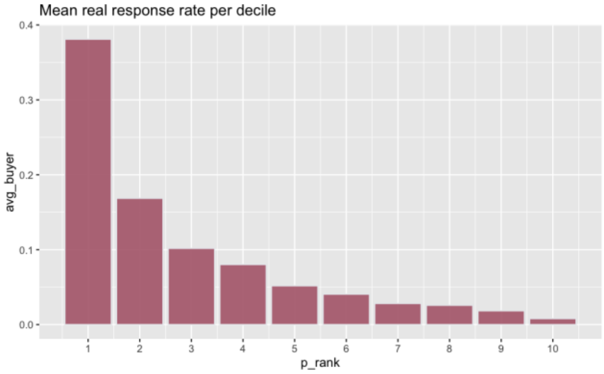
The variable for the number of seafood meals ordered has a positive value and hence a direct relationship with consumer response. Exp (0.90) = 2.45, which implies that consumers who buy seafood meals from MALM have 2.45 times higher chances of responding to emails than those who don't. Hence those with a preference for seafood and following the pescetarian diet within MALM's consumer base are another great target for MALM's email campaign.

Part II: Decile Analysis of Logistic Regression Results

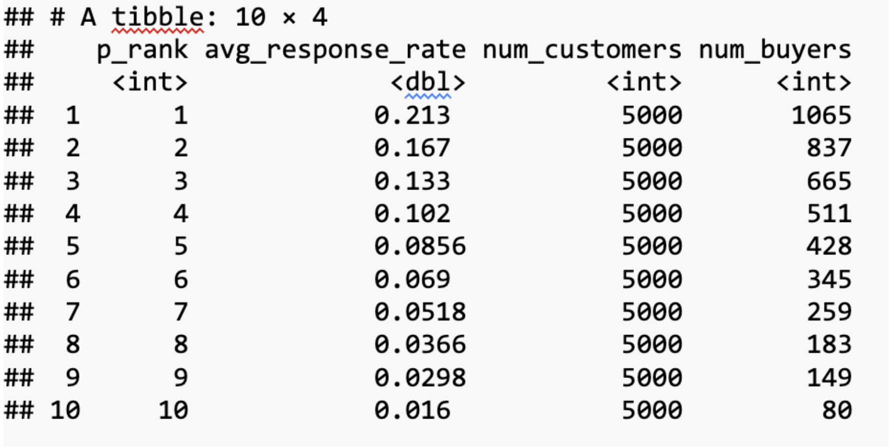
(4)

p_rank <int>	avg_pred <dbl>	avg_real <dbl>
1	0.381007052	0.3810
2	0.163423489	0.1686
3	0.104853413	0.1018
4	0.074066127	0.0802
5	0.056681734	0.0518
6	0.043004603	0.0406
7	0.033081104	0.0282
8	0.024547514	0.0258
9	0.016825874	0.0184
10	0.006909089	0.0080





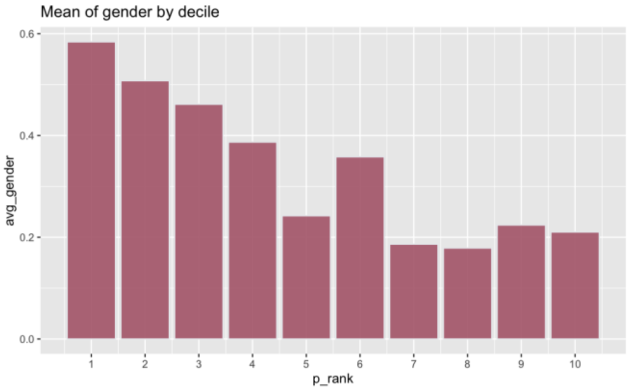
(5)

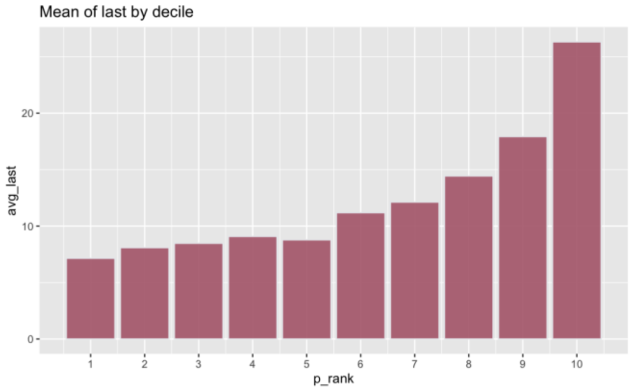


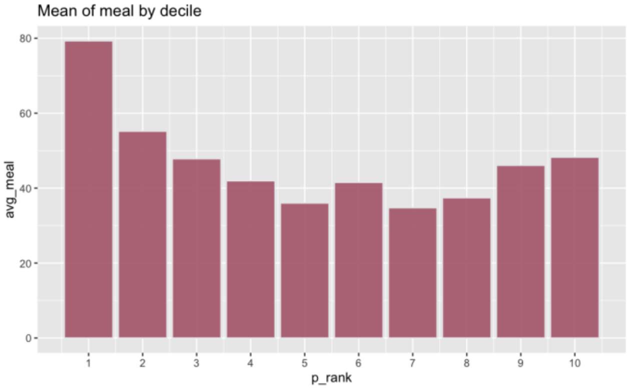
```
## # A tibble: 10 × 4
##     p_rank avg_response_rate num_customers num_buyers
##      <int>             <dbl>         <int>      <int>
##   1      1            0.213           5000       1065
##   2      2            0.167           5000        837
##   3      3            0.133           5000        665
##   4      4            0.102           5000        511
##   5      5            0.0856          5000        428
##   6      6            0.069           5000        345
##   7      7            0.0518          5000        259
##   8      8            0.0366          5000        183
##   9      9            0.0298          5000        149
## 10     10            0.016           5000         80
```

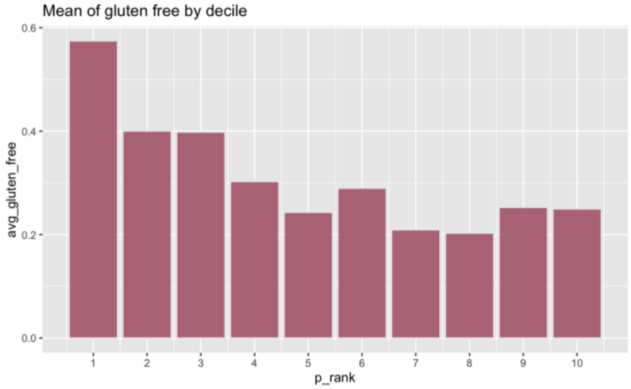
Unlike RFM analysis, the customer base is divided into an equal number of segments for decile analysis. Out of these, it is clear that higher-ranking customers have a much higher response rate and are much more likely to become buyers than those in the lower ranks.

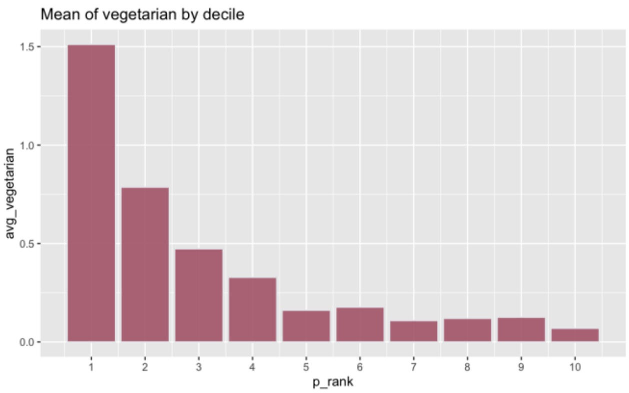
(6)

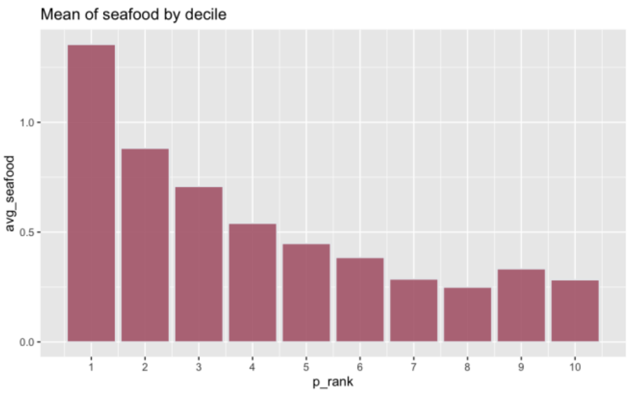

Mean of gender by decile


Mean of last by decile


Mean of meal by decile

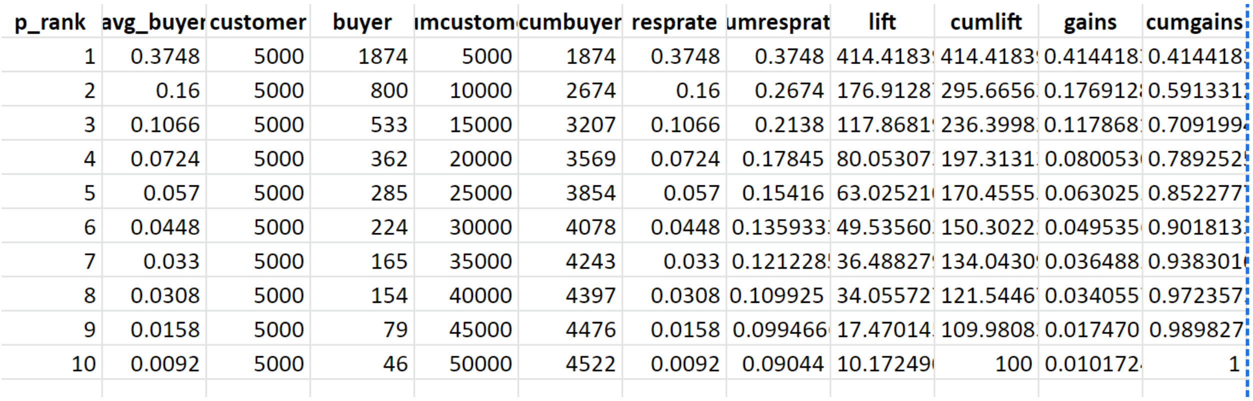

Mean of gluten free by decile





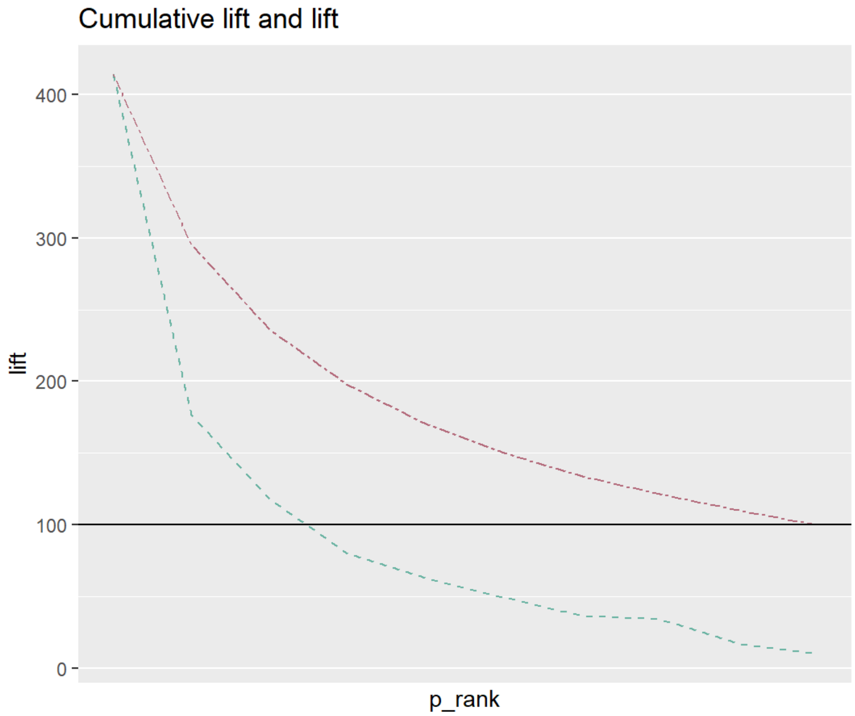
As seen from the graphs, it is clear that the results from our logistic regression analysis match the decile analysis for each independent variable. The graphs show the highest purchase probability for ranks 1,2 and 3, indicating that high-ranking customers buy various types of meals (seafood, gluten-free and vegetarian), spend more money on meals, are more recent in these purchases, and have a higher probability of being male than female. Since we are working with a large amount of real-life data, slight variations can be seen e.g. for rank 6. However, overall our decile analysis backs the results of the logistic regression.

Part III: Lifts and Gains

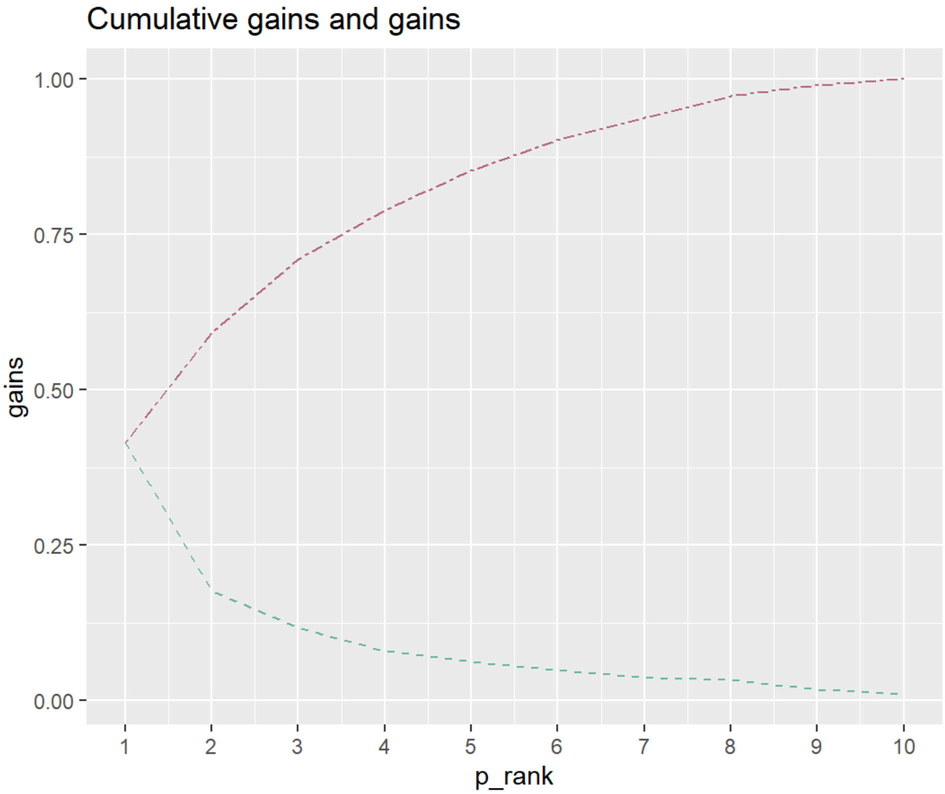


p_rank	avg_buyer	customer	buyer	umcustom	cumbuyer	resprate	umresprat	lift	cumlift	gains	cumgains
1	0.3748	5000	1874	5000	1874	0.3748	0.3748	414.4183	414.4183	0.414418	0.414418
2	0.16	5000	800	10000	2674	0.16	0.2674	176.9128	295.6656	0.176912	0.591331
3	0.1066	5000	533	15000	3207	0.1066	0.2138	117.8681	236.3998	0.117868	0.709199
4	0.0724	5000	362	20000	3569	0.0724	0.17845	80.05307	197.3131	0.080053	0.789252
5	0.057	5000	285	25000	3854	0.057	0.15416	63.02521	170.4555	0.063025	0.852277
6	0.0448	5000	224	30000	4078	0.0448	0.135933	49.53560	150.3022	0.049535	0.901813
7	0.033	5000	165	35000	4243	0.033	0.121228	36.48827	134.0430	0.036488	0.938301
8	0.0308	5000	154	40000	4397	0.0308	0.109925	34.05572	121.5446	0.034055	0.972357
9	0.0158	5000	79	45000	4476	0.0158	0.099466	17.47014	109.9808	0.017470	0.989827
10	0.0092	5000	46	50000	4522	0.0092	0.09044	10.17249	100	0.010172	1

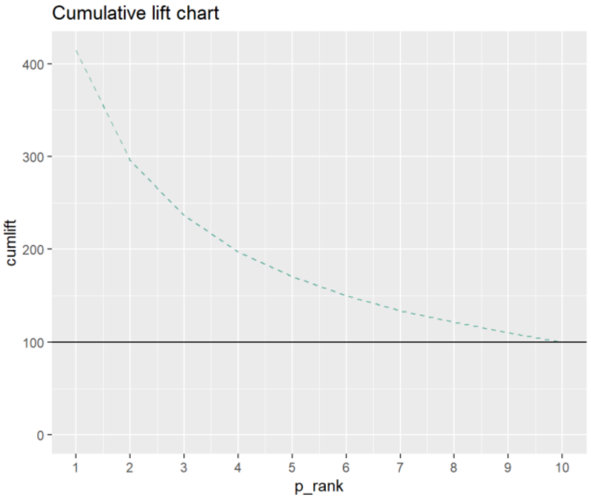
(7)

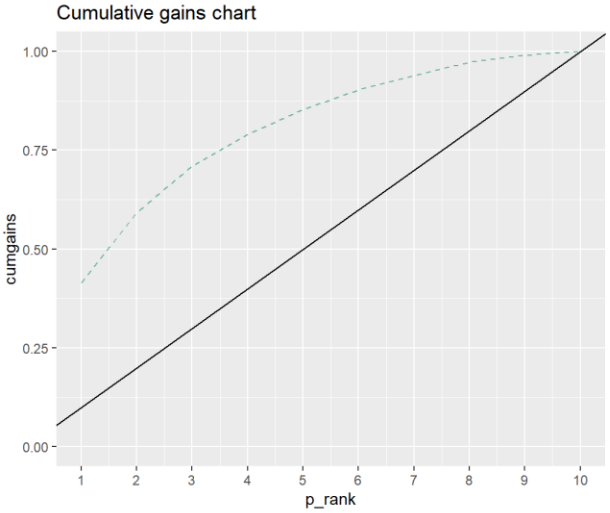

Cumulative lift and lift
400
300
200
100
0
lift
p_rank

(8)


Cumulative gains and gains
1.00
0.75
0.50
0.25
0.00
gains
1
2
3
4
5
6
7
8
9
10
p_rank

(9)


Cumulative lift chart


Cumulative gains chart

The lifts chart shows that our predicted model performs better than randomly targeting customers. Since values for all of the groups in each decile are greater than 100 (or equal to, for rank 10), they can be considered for focused marketing strategies by the company.
The lift for the rank 1 decile is ~410, which indicates that targeting only these customers would result in 4.10 times the number of buyers we would gain by randomly selecting customers for email marketing. Similarly, the relative probable increase in buyers for each decile can be extrapolated from the cumulative lift chart. The lift for the rank 10 decile is ~100, which means targeting this customer segment would result in about the same number of buyers that we would gain from random targeting.

The gains chart helps in identifying the total percentage of buyers we can expect if we target the customers of each of the deciles. As shown in the graph, we can observe that using a modeled strategy would yield 40% of buyers by targeting ~10% of customers. The graph peaks towards the end, where the model yields close to 100% of buyers by targeting 90% of customers. In terms of profitability, it might be better to target 60% of customers, which would result in gains in the form of ~88% of buyers.

Both graphs also show successful models due to the high area between the lift/gains curves and the no model baselines.

Part IV: Profitability Analysis

The following analysis is based on:
Cost of emailing offer to customer (includes CRM platform maintenance and content creation costs): $0.50
Selling price (shipping included): $18.00
Ingredient and preparation costs paid by Meals à la Minute: $9.00
Shipping costs: $3.00

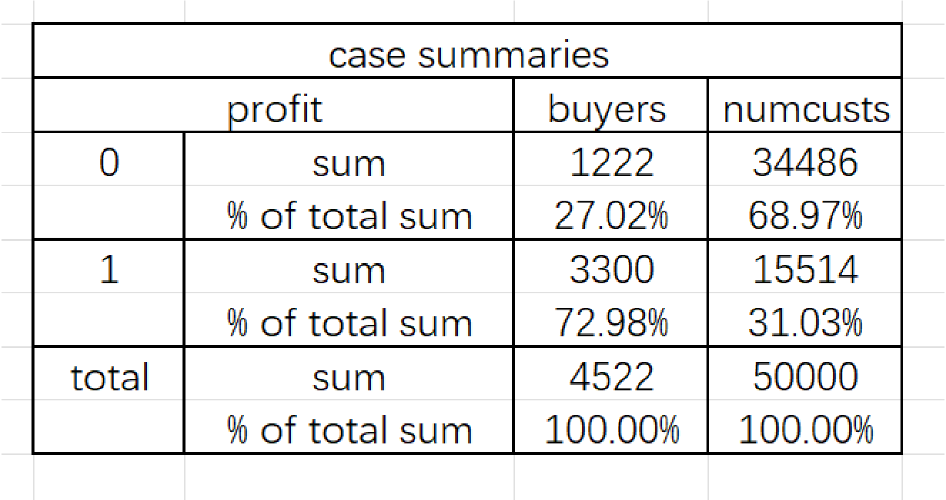
(10)

Breakeven response rate = cost per email / (selling price - production cost - shipping) = 0.5/(18-9-3) = 8.3%

(11)

R script: malm_d$target <- ifelse(malm_d$pred >= 0.083333, 1, 0)

(12)



case summaries			
profit		buyers	numcusts
0	sum	1222	34486
	% of total sum	27.02%	68.97%
1	sum	3300	15514
	% of total sum	72.98%	31.03%
total	sum	4522	50000
	% of total sum	100.00%	100.00%

(13)

Gross profit = profit per unit * quantity sold - cost per email * number of emails sent = 6*3300 - 0.5*11514 = $12043

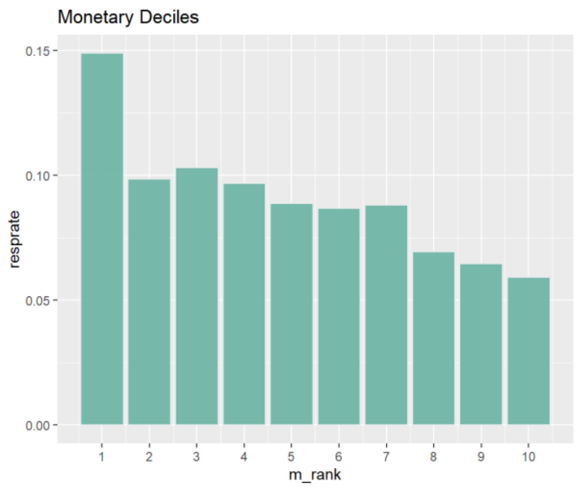
Gross sales = selling price * quantity sold 18*3300 = 59400

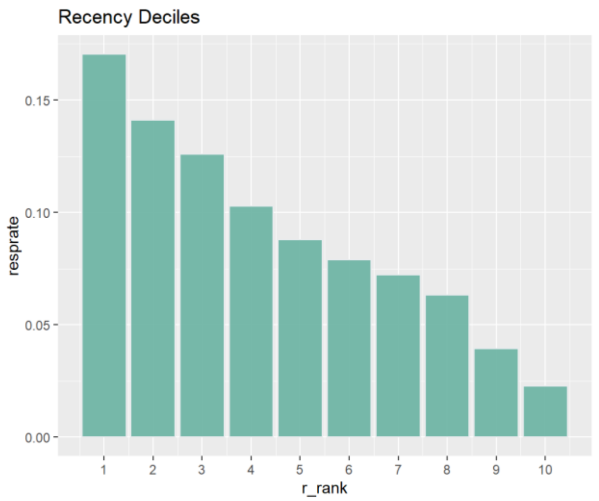
gross profit / gross sales = 20.27%

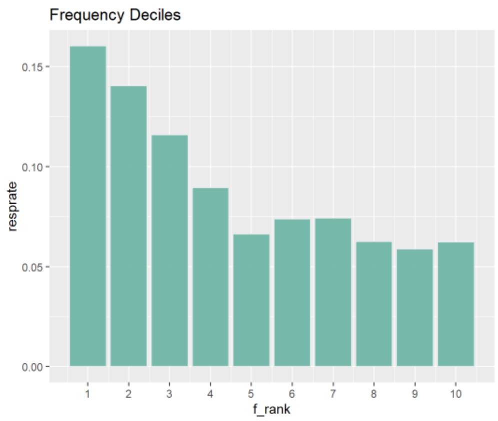
Return on marketing expenditure = gross profit / (cost per email * number of emails sent) = 12043 / (0.5*11514) = 155.25%

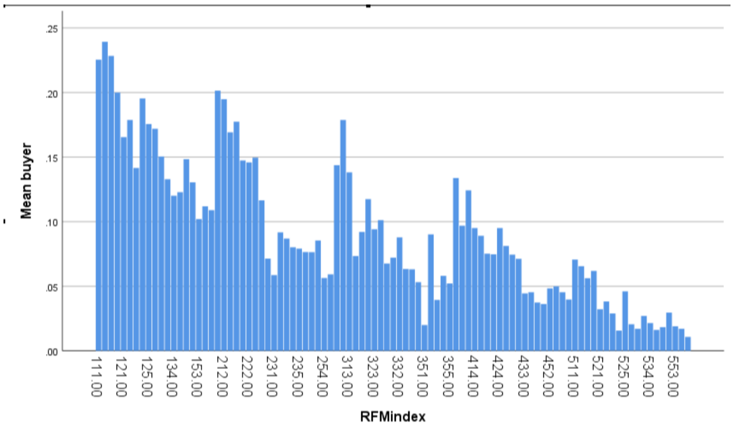
Part V: RFM Analysis vs. Logistic Regression

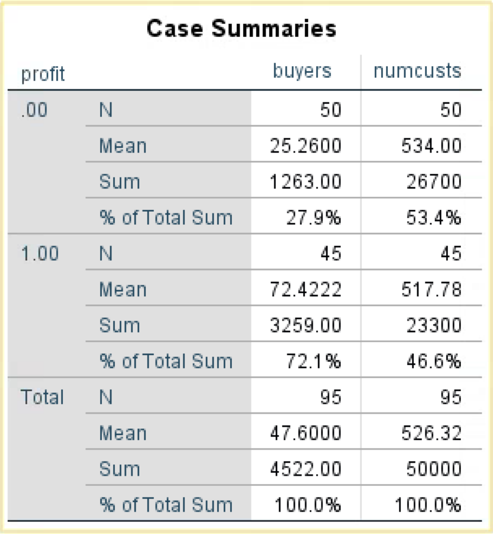
(14)











Case Summaries

profit		buyers	numcusts
.00	N	50	50
	Mean	25.2600	534.00
	Sum	1263.00	26700
	% of Total Sum	27.9%	53.4%
1.00	N	45	45
	Mean	72.4222	517.78
	Sum	3259.00	23300
	% of Total Sum	72.1%	46.6%
Total	N	95	95
	Mean	47.6000	526.32
	Sum	4522.00	50000
	% of Total Sum	100.0%	100.0%

Profitability analysis:
Gross profit per sale = 18-3-9 = 6
Gross profit = $6.00×3,259 – $0.50×23,300 = $7,904
Gross sales = $18.00×3,259 = $58,662
gross profit / gross sales = $7,904 / $58,662 = 13.47%
Return on marketing = $7,904 / ($0.50×23,300) = 67.85%

We can conclude that logistic analysis is a more precise way of analyzing KPIs than RFM analysis, which reports much less profitability. MALM should consider doing logistic regression analysis in addition to RFM for a better understanding of their data. Logistic regression also takes more variables into consideration and allows for clear market segmentation.

(15) Based on the data generated above, it is clear that MALM should consider logistic regression analysis in addition to their current RFM analysis.

Logistic regression allows us to test the significance of each variable, which helps in accurate decision-making regarding which independent variables to choose for our model. After generating the values, we concluded that our model would consider gender, money spent on meals, recency value, and the number of purchases for vegetarian, gluten-free, and seafood meals since the latter are three distinct types of eating patterns observed within consumers.

Additionally, conducting a profitability analysis of our model showed that targeting 15514 customers resulted in x buyers, with a gross profit of $12043 and a return on marketing expenditure of $155.25, while earning a net revenue of $3.64 per order.

Our marketing strategy recommendations would revolve around targeting each of the consumer segments with an aim to expand profits on consumers that have the most likely chance of becoming buyers. The email campaign can also work on expanding its consumer base by targeting customers outside of the defined segments. For e.g., since the non-vegetarian segment of their total market can be emailed with more frequency, offered higher discounts and rewards on returns.