Relay Foods Management

HW 2 Presented by: Donovan Sullivan, Tim Hulak, Eduardo Robles, Cyrus Garrett, and Anthony Ferraiolo

The Pre-Regression

	custid <fct>	retained <int>	created <fct>	firstorder <fct>	lastorder <fct>	esent <int>	eopenrate <dbl>	eclickrate <dbl>	avgorder <dbl>
1	6H6T6N	0	9/28/12	8/11/13	8/11/13	29	100.00000	3.448276	14.52
2	APCENR	1	12/19/10	4/1/11	1/19/14	95	92.63158	10.526316	83.69
3	7UP6MS	0	10/3/10	12/1/10	7/6/11	0	0.00000	0.000000	33.58
4	7ZEW8G	0	10/22/10	3/28/11	3/28/11	0	0.00000	0.000000	54.96
5	8V726M	1	11/27/10	11/29/10	1/28/13	30	90.00000	13.333333	111.91
6	2B6B83	1	11/17/08	10/12/10	1/14/14	46	80.43478	15.217391	175.10

The above data represents a portion of the data prior to running it through our regression analysis

Key Finding: The Hit Ratio

Model 1

- Based on the model, each of the variables were statistically significant and yielded a hit rate of 0.94

Model 2

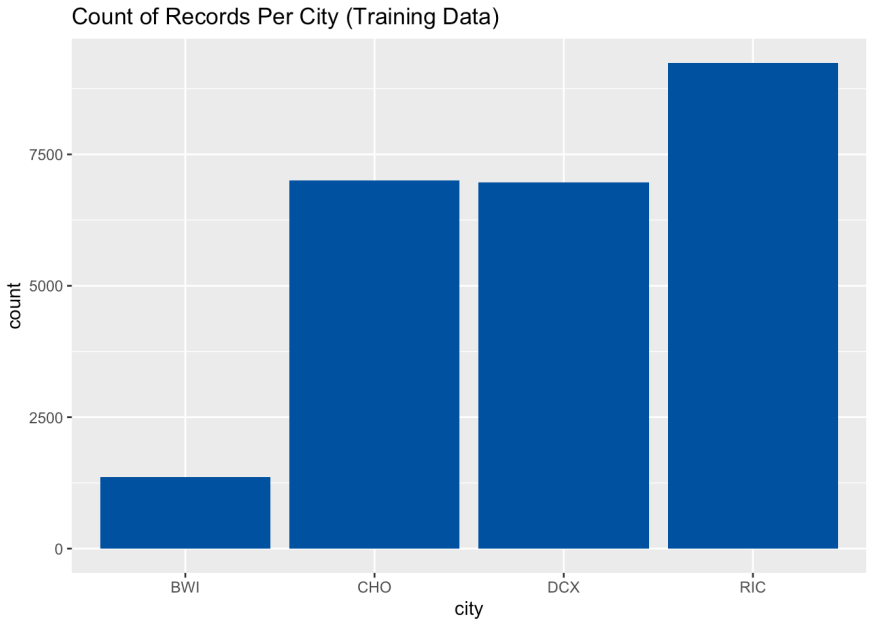
- All variables were statistically significant, with the exception of orderfreq (which had a p-value of 0.64). This model yielded a hit rate of 0.

Model 3

- esent was statistically significant and yielded a hit rate of 0.93. However, this is rather deceptive and will be addressed in a later slide.

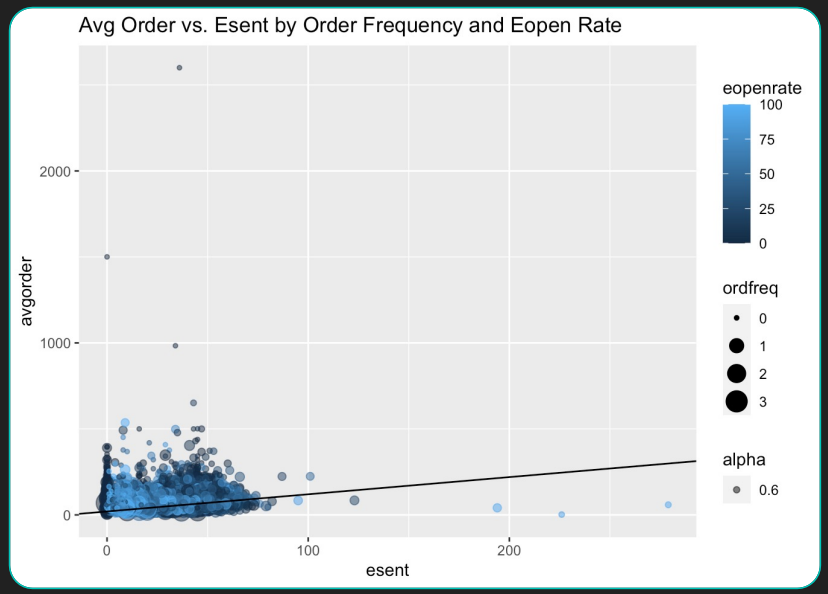
Geographic Impact

- We found that the most overrepresented geography in the Train data was Richmond while the most underrepresented was Baltimore



Bringing Order(s) to it…

- The chart to the right shows the correlation to the orders and the average number of emails received
- We found that there are outliers showing that opened emails may not have had an impact on orders
- As such the number of emails may not be related to the average order as strongly as the slope indicates



Question 5: (part 1)

Why is esent a strong predictor of retention?

- ✓ esent is a strong predictor of retention because it is correlated directly to the other measured variables.
- ✓ Without esent, interactive metrics would lack a baseline by which to be measured.
- ✓ Esent also provides a proxy for a customer's dormancy duration.
- ✓ Esent is also a controlled variable dictated by Relay that can be adjusted and therefore is a true measure of their intentional engagement.
- ✓ It can also provide an understanding of when productive engagement is hurt by the volume of emails sent.

Question 5: (part 2)

Do you see any issues with using esent as a predictor for retention?

- The issue with using esent alone is that it does not measure the extent of the engagement nor the customer's interaction with marketing.

Question 5: (part 3)

Recommended transformations of esent that can overcome the issues of using esent as a predictor?

By using esent as the baseline in comparisons with eopenrate, eclickrate, firstorder, lastorder, and avgorder the probabilities can be discovered for the following questions, improving target marketing:

- How many emails sent on average equals retained?
- How many emails sent, opened, and orders equals retained?
- What is the average length of dormancy of customers between first and second-order?

Question 6: (Reg1)

Does the sign of the coefficient for avgorder, ordfreq, and weekend make sense? What consumer behavior explanation can you provide for the sign of these coefficients?

Reg1 results:

Avgorder = -0.0035576

Ordfreq= -0.6074376

Weekend= N/A

- The negative coefficient values suggest that there might be a negative correlation.
- It also shows that of the two, ordfreq has a more negative correlation than avgorder.
- In the context of our example, this means the average order size for the customer is less than 1 as well as the number of orders by customer tenure.

Question 6: (Reg2)

Does the sign of the coefficient for avgorder, ordfreq, and weekend make sense? What consumer behavior explanation can you provide for the sign of these coefficients?

Reg2 results:

Avgorder = 0.002025

Ordfreq= 0.073195

Weekend= N/A

- The positive coefficient values suggest that there is be a positive correlation.
- It also shows that of the two, ordfreq has a more positive correlation than avgorder.
- In the context of our example, this means the average order size for the customer is greater than 1 as well as the number of orders by customer tenure.

Question 6: (Reg3)

Does the sign of the coefficient for avgorder, ordfreq, and weekend make sense? What consumer behavior explanation can you provide for the sign of these coefficients?

Reg3 results:

Avgorder = -0.0034973

Ordfreq= -0.5967557

Weekend= 0.2640833

- In the context of our example, this mean the weekend has a positive effect on the orders.

Question 7: Recommendations?

- Start by doing a more thorough analysis of the customers and look for the strongest correlated variables and try to maximize those in the greater customer population.
- These variables have the strongest positive correlation, and intuitively, anyone with a subscription is at least marginally satisfied with their service or they would cancel the subscription. Within the 'refill' and 'doorstep' subsets
- Additionally, it may be worth examining what products or product categories are most popular on a given day, and plan sales around those days.
- For example: Sprouts Grocery, sees a significant bump in business on Wednesdays when fresh produce is delivered, and so plans their new weekly deals starting on that date.

Appendix

```
##
## Call:
## glm(formula = retained ~ ., family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1905   0.0257   0.0615   0.1650   2.8987
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept) -2.5348036  0.0777869 -32.587 < 0.0000000000000002 ***
## esent        0.2110782  0.0032626  64.696 < 0.0000000000000002 ***
## eclickrate   0.0192004  0.0017564  10.932 < 0.0000000000000002 ***
## avgorder    -0.0035576  0.0007572  -4.698 0.000002624313868324 ***
## ordfreq     -0.6074376  0.2313531  -2.626              0.00865 **
## paperless    0.5252447  0.0641304   8.190 0.000000000000000261 ***
## refill       0.7693665  0.1040461   7.394 0.000000000000141969 ***
## doorstep     0.8659042  0.1602369   5.404 0.000000065207401008 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24944.7  on 24578  degrees of freedom
## Residual deviance:  8472.7  on 24571  degrees of freedom
## AIC: 8488.7
##
## Number of Fisher Scoring iterations: 7
```

Logistic Regression

- Use esent, eclickrate, avgorder, ordfreq, paperless, refill, doorstep as independent variables to estimate the model using train data. Report the model coefficients. Predict retention, and calculate hit rate in the test data.

```
##
## Call:
## glm(formula = retained ~ ., family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9348   0.4069   0.6109   0.6243   0.9097
##
## Coefficients:
##             Estimate Std. Error z value             Pr(>|z|)
## (Intercept) 0.668536   0.039065  17.113 < 0.0000000000000002 ***
## avgorder    0.002025   0.000453   4.469       0.000007846323 ***
## ordfreq     0.073195   0.156046   0.469                0.639
## paperless   0.809074   0.033121  24.428 < 0.0000000000000002 ***
## refill      0.867731   0.079470  10.919 < 0.0000000000000002 ***
## doorstep    0.851283   0.131893   6.454       0.000000000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24945  on 24578  degrees of freedom
## Residual deviance: 23946  on 24573  degrees of freedom
## AIC: 23958
##
## Number of Fisher Scoring iterations: 5
```

Logistic Regression (con't)

- Use avgorder, ordfreq, paperless, refill, doorstep as independent variables to estimate the model using train data. Report the model coefficients. Predict retention, and calculate hit rate in the test data.

```
##
## Call:
## glm(formula = retained ~ ., family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1122   0.0255   0.0594   0.1703   1.9304
##
## Coefficients:
##              Estimate Std. Error z value            Pr(>|z|)
## (Intercept) -2.117615   0.042094  -50.31 <0.0000000000000002 ***
## esent        0.211453   0.003121   67.76 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24944.7  on 24578  degrees of freedom
## Residual deviance:  8972.3  on 24577  degrees of freedom
## AIC: 8976.3
##
## Number of Fisher Scoring iterations: 7
```

Logistic Regression (con't)

- Use esent alone as independent variables to estimate the model using train data. Report the model coefficients. Predict retention, and calculate hit rate in the test data.

```
##
## Call:
## glm(formula = retained ~ ., family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1794   0.0253   0.0617   0.1647   2.7711
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept) -2.6036352  0.0797664 -32.641 < 0.0000000000000002 ***
## esent        0.2115356  0.0032734  64.622 < 0.0000000000000002 ***
## eclickrate   0.0192718  0.0017602  10.949 < 0.0000000000000002 ***
## avgorder    -0.0034973  0.0007535  -4.641 0.000003463420704471 ***
## ordfreq     -0.5967557  0.2313996  -2.579              0.00991 **
## paperless    0.5192569  0.0642167   8.086 0.000000000000000617 ***
## refill       0.7607544  0.1040587   7.311 0.000000000000265522 ***
## doorstep     0.8471091  0.1606152   5.274 0.000001333694804433 ***
## weekend      0.2640833  0.0654575   4.034 0.000054736449631580 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24944.7  on 24578  degrees of freedom
## Residual deviance:  8456.3  on 24570  degrees of freedom
## AIC: 8474.3
##
## Number of Fisher Scoring iterations: 7
```

Logistic Regression (con't)

- Create a dummy variable called weekend which is 1 if favday is Friday, Saturday or Sunday, and 0 otherwise. Use esent, eclickrate, avgorder, ordfreq, paperless, refill, doorstep, and weekend as independent variables to estimate the model using train data. Report the model coefficients, and predict retention, and calculate hit rate in the test data.