Business Analytics & Modeling

Final Project

Yiwen Chen

Expected Gross Profit from Customer Pool

For calculating the gross profit from 180,000 names, we first multiplied remaining 180,000 by 0.1 because our sample should be based on 10% of the pool. Therefore, we came up with a formula of 180,000 * 0.1 = 18,000. What we considered after this process is the fact that average spending per catalog mailed is $11.34 and it costs Tayko $2 to mail. To calculate gross profit, we subtracted out the cost of $2 from the amount of sales generated from a catalog, which is $11.34. Based on this net profit per a purchase, we multiplied it by the number of people in the pool, which is 18,000. Finally, we ended up having the expected gross profit of $168,120.

($11.34 - $2) * 18,000 = $168,120

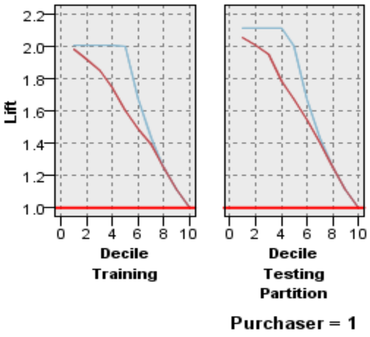
Classification Model for Distinguishing Purchaser/Non-purchaser

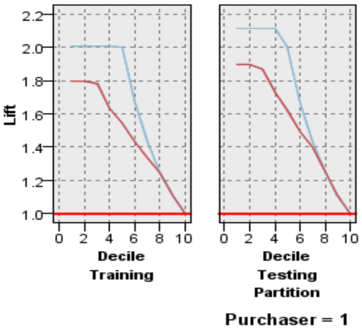
P("Purchaser") =

$$\frac{\mathbf{1}}{\mathbf{1+}\, e^{-(-2.014*source_{a(1)}+3.431*source_{h(1)}-1.675*source_{r(1)}-1.959*source_{u(1)}+2.538*Freq+0.915*Address_{is_{res(1)}}-1.057*WebOrderFlag(1))}}$$

(Refer to Appendix 1 for results from regression model)

In terms of the supervised techniques, Decision Tree and Logistic Regression models, we utilized regression model to analyze the predictors of being a purchaser. We compared the lift charts for the results from decision tree model and regression model: As graphs show, our predictive curve of training dataset from the regression model is higher than the one from the decision tree. The best lines for each model are the same, and the distance between our predictive curve and the best line from the regression model is closer than the one from decision tree model. Thus, we could conclude that utilizing the regression model is more accurate method, meaning we have higher probability to receive response using regression model than using random sample or decision tree model.





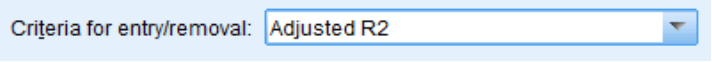
[Figure 1: Lift Chart from Regression Model] *[Figure 2: Lift Chart from Decision Trees]*

Our regression model results in better prediction than when we estimated the prediction based on the frequency of occurring class – We have 30.1% (80.2% - 50.1% = 30.1%) higher accuracy. *(Refer to Appendix 2)*

In utilizing the decision tree model, SPSS filtered out the insignificant variables and split the data into classes based on the least important indicators (In our case, DT splits our data based on the "Freq, WebOrderFlag and source_h, which means not all classes are equally important) (*Refer to Appendix 3*) Furthermore, our team suggest Tayko to have accurate classification of identify of the Purchaser. We interpret this result as meaning that manager could have specific probability of an individual as a purchaser based on the given variables, such as Gender, Address, Freq, etc. Managers could gain this specific answer by running the equation from the regression model.

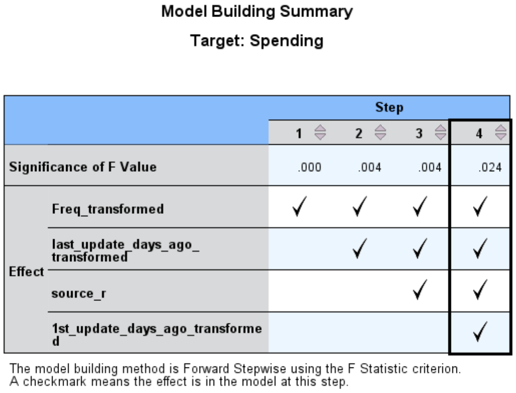
Variables for Predicting Spending among Purchasers

After running the linear regression model in SPSS, we came up with the top four influential variables for the "spending", such as Freq_transformed, last_update,source_r and 1st_update. After selecting "Significance of F value" and then "Adjusted R2" from the options for building model in SPSS, we reached to conclusion that all of these four variables are significant enough to be analyzed, which means its value is lower than 0.05. Additionally, the increased adjusted R-square shows that the combination of all these four variables could enhance the accuracy of our estimation.



Criteria for entry/removal: Adjusted R2

[Figure 3: Adjusted R2 from the model building options in SPSS]

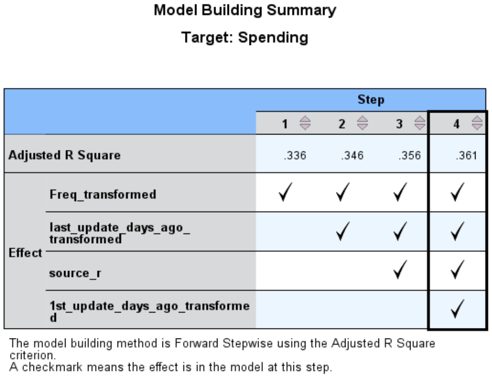


Model Building Summary

Target: Spending

		Step 1	Step 2	Step 3	Step 4
Significance of F Value		.000	.004	.004	.024
Effect	Freq_transformed	✓	✓	✓	✓
	last_update_days_ago_transformed		✓	✓	✓
	source_r			✓	✓
	1st_update_days_ago_transformed				✓

The model building method is Forward Stepwise using the F Statistic criterion.
A checkmark means the effect is in the model at this step.



Model Building Summary

Target: Spending

		Step 1	Step 2	Step 3	Step 4
Adjusted R Square		.336	.346	.356	.361
Effect	Freq_transformed	✓	✓	✓	✓
	last_update_days_ago_transformed		✓	✓	✓
	source_r			✓	✓
	1st_update_days_ago_transformed				✓

The model building method is Forward Stepwise using the Adjusted R Square criterion.
A checkmark means the effect is in the model at this step.

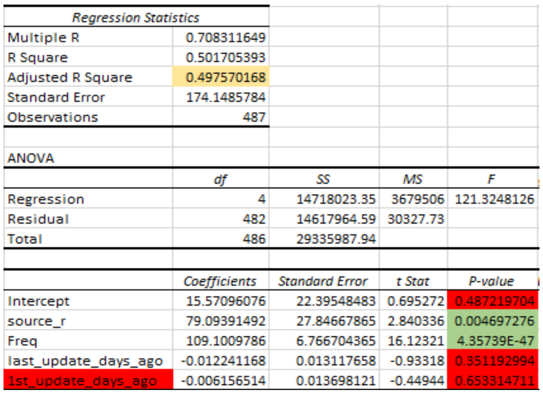
[Figure 4. "Variables are Significant in SPSS"]

[Figure 5. "R Square" from SPSS]

(Refer to Appendix 4 for regression formula)

Spending = 108.429*Freq – 0.053*last_update- 92.053*source_r(0)+0.035*1st_update+100.567

We put these top four variables into Excel Solver, the result turned out that the P-values for *last_update* and *1st_update* variables are greater than 0.05. So, we took out the less important one, *1st_update*, and re-run the model. It resulted in increase in the adjusted R square from 49.76% to 49.84% and all other variables are significant enough to be utilized since their p-value is less than 0.05). Along with the fact that the adjusted R square of 36.1% from SPSS model, we assume that the model made from Excel could be more accurate in making prediction.



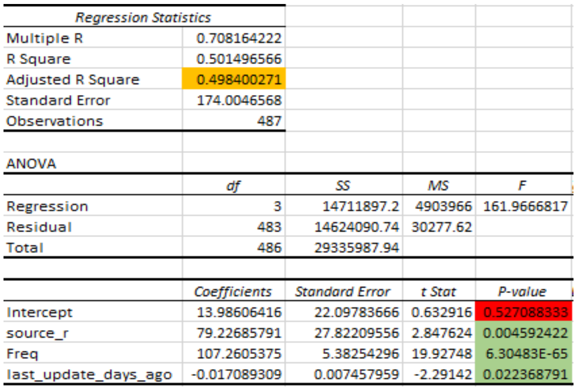
Regression Statistics	
Multiple R	0.708311649
R Square	0.501705393
Adjusted R Square	0.497570168
Standard Error	174.1485784
Observations	487

ANOVA

	df	*SS*	*MS*	*F*
Regression	4	14718023.35	3679506	121.3248126
Residual	482	14617964.59	30327.73	
Total	486	29335987.94		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	15.57096076	22.39548483	0.695272	0.487219704
source_r	79.09391492	27.84667865	2.840336	0.004697276
Freq	109.1009786	6.766704365	16.12321	4.35739E-47
last_update_days_ago	-0.012241168	0.013117658	-0.93318	0.351192994
1st_update_days_ago	-0.006156514	0.013698121	-0.44944	0.653314711

>>



Regression Statistics	
Multiple R	0.708164222
R Square	0.501496566
Adjusted R Square	0.498400271
Standard Error	174.0046568
Observations	487

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	14711897.2	4903966	161.9666817
Residual	483	14624090.74	30277.62	
Total	486	29335987.94		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	13.98606416	22.09783666	0.632916	0.527088333
source_r	79.22685791	27.82209556	2.847624	0.004592422
Freq	107.2605375	5.38254296	19.92748	6.30483E-65
last_update_days_ago	-0.017089309	0.007457959	-2.29142	0.022368791

[Figure 6: 1st Result after Excel Solver]

[Figure 7: 2nd (Final) Result From Excel Solver]

We put the formula of the result made from Excel Solver and calculated the estimated spending by using the variables from training dataset. It turned out that the difference between the formula estimation and the actual spending columns is "Error". Since we are looking for the total units of error instead of the situation of over-or under-spending, we calculated the absolute value for each of the error and summed them up based on the variables we found from the training dataset. Our estimated model showed 69785.63 units of total error, or could say 136.06 as average error estimation, in our testing dataset. *(Refer to Appendix 5)*

The estimation from the training dataset turned out to be useless. Based on the charts in SPSS Modeler, we found out that the curve for our spending estimation decreases in the testing dataset. *(Refer to Appendix 6)*

All being said, our team assumes that the number of data analyzed by the model is relatively limited. To build a better model for more accurate prediction, the dataset should include more information such as larger training pool.

Validation Data Analysis

To analyze the data from validation set, we imported the table from the first partition dataset in SPSS into Excel and take out the data showing "3_Validation" at "Partition" column.

To choose the first decile of the validation dataset, we sorted the order of "$LP-1" from largest to smallest and took out the top 53 (from: 525 *10%) rows of data. Then, we created a new column called Adjusted Probability of Purchase and calculated them by multiplying the "Purchasing Probability" from "$LP-1" with true response rate of 0.107. This adjusted probability returns ratios could be utilized at overall data consideration. Also, we created a new column in Excel called "Predicted Spending", gained by utilizing the spending formula we run out from the regression model in part 3 to estimate the predicted spending. The third step is to calculate the "Expected Spending", which could be gained by multiplying the new probability of purchasing by the predicted spending. We then summed this column up and came up with the result of "Total Expected Spending" -- $734.97 *(Refer to Appendix 7)*

By averaging adjusted purchasing probability from those 52 people, we gained the average adjusted purchase probability of 0.107 for the first decile in validation database. *(See Appendix7)* Compared to the first response rate of 0.053, the response rate from our model estimation is twice higher than the original one. Thus, we concluded that sending out the catalogs to targeted group of people based on our model would collect more response than sending to a pool randomly.

We also calculated all the adjusted purchase probability for all validation data and the average ratio of them is 0.053234*. (See Appendix7)* This response rate could be utilized in dealing with the expected spending from the 180,000 names pool. We realized that $734.97 would be spent from the top 10% of the validation database *(See Appendix7)*. In comparison with 180,000 names pool, we calculated the gross profit in the following logics.

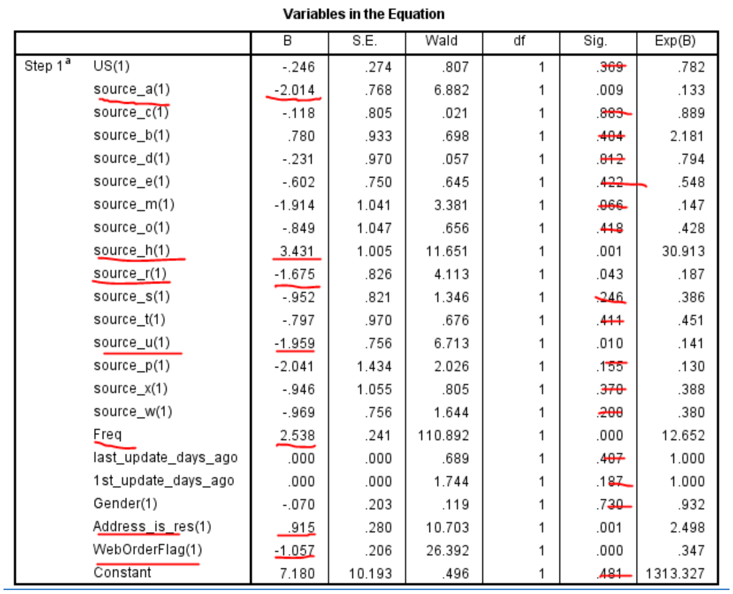
525/180,000 = 0.003

Total Expected Spending for 180,000: 743.97/10%/0.3% = $2,449,900

Gross Profit: $2,449,900 – $2*180,000 = $2,089,900

This new gross profit estimation is much higher than the result from the actual spend number given within the prompt ($168,120/0.1 = $1,681,200) because our model helps company aim at the targeted customer they should send out the catalogs. The average of spending we calculated based on the expected spend probability is $2,449,900/180,000 =$13.61. Compared to the average spending based on the actual number given within the prompt, which is $11.34, company using our model to send out the catalogs can collect higher revenue from customer spending because of the higher response rate. Also, targeted customers segments recognized by our model are useful at saving cost. Company could save cost at catalog printing and labor salary because company does not have to send out everyone but could be goal-oriented.

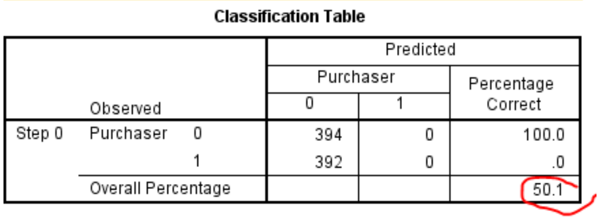
Appendix



Variables in the Equation

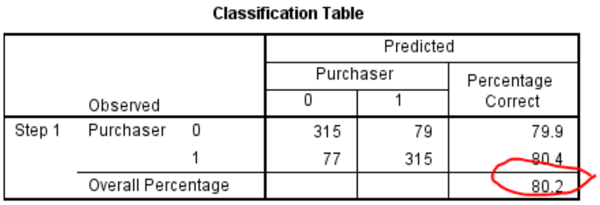
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	US(1)	-.246	.274	.807	1	~~.369~~	.782
	source_a(1)	-2.014	.768	6.882	1	.009	.133
	source_c(1)	-.118	.805	.021	1	~~.883~~	.889
	source_b(1)	.780	.933	.698	1	~~.404~~	2.181
	source_d(1)	-.231	.970	.057	1	~~.812~~	.794
	source_e(1)	-.602	.750	.645	1	~~.422~~	.548
	source_m(1)	-1.914	1.041	3.381	1	~~.066~~	.147
	source_o(1)	-.849	1.047	.656	1	~~.418~~	.428
	source_h(1)	3.431	1.005	11.651	1	.001	30.913
	source_r(1)	-1.675	.826	4.113	1	.043	.187
	source_s(1)	-.952	.821	1.346	1	~~.246~~	.386
	source_t(1)	-.797	.970	.676	1	~~.411~~	.451
	source_u(1)	-1.959	.756	6.713	1	.010	.141
	source_p(1)	-2.041	1.434	2.026	1	~~.155~~	.130
	source_x(1)	-.946	1.055	.805	1	~~.370~~	.388
	source_w(1)	-.969	.756	1.644	1	~~.200~~	.380
	Freq	2.538	.241	110.892	1	.000	12.652
	last_update_days_ago	.000	.000	.689	1	~~.407~~	1.000
	1st_update_days_ago	.000	.000	1.744	1	~~.187~~	1.000
	Gender(1)	-.070	.203	.119	1	~~.730~~	.932
	Address_is_res(1)	.915	.280	10.703	1	.001	2.498
	WebOrderFlag(1)	-1.057	.206	26.392	1	.000	.347
	Constant	7.180	10.193	.496	1	~~.481~~	1313.327

[Appendix 1: Model for classification of Purchaser]



Classification Table

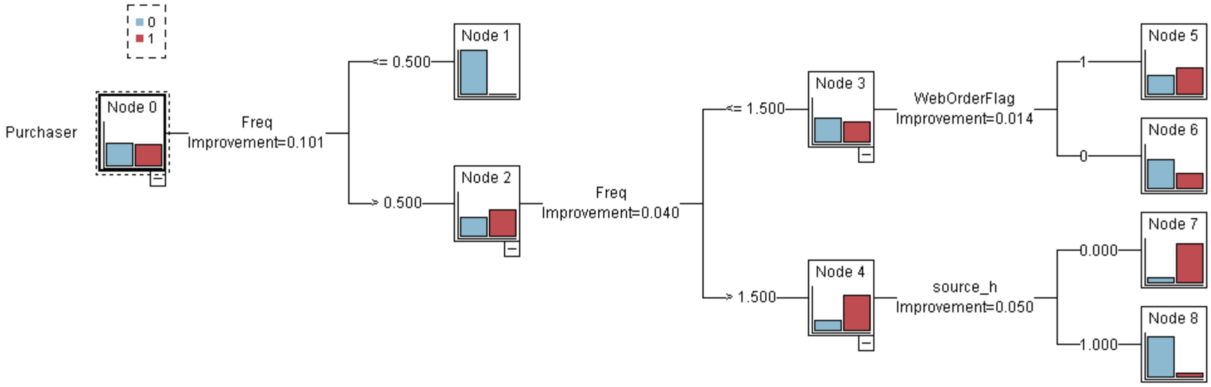
Observed			Predicted		
			Purchaser		Percentage Correct
			0	1	
Step 0	Purchaser	0	394	0	100.0
		1	392	0	.0
	Overall Percentage				50.1



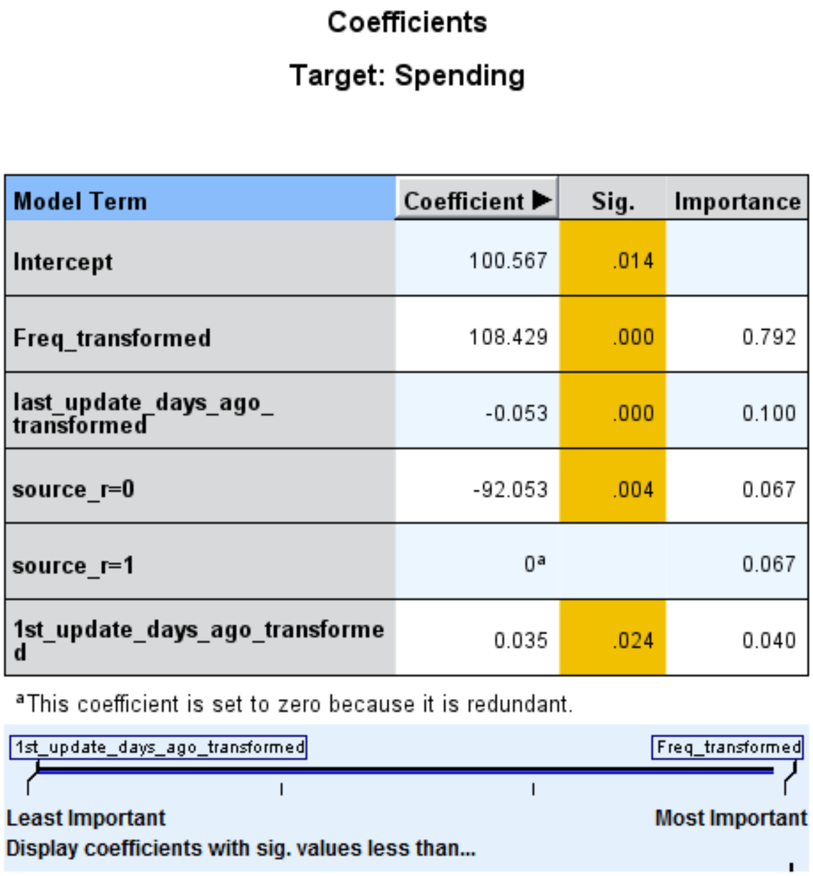
Classification Table

Observed			Predicted		
			Purchaser		Percentage Correct
			0	1	
Step 1	Purchaser	0	315	79	79.9
		1	77	315	80.4
	Overall Percentage				80.2

[Appendix 2 : Classification Table]



[Appendix 3: Decision Tree Classification]



Coefficients

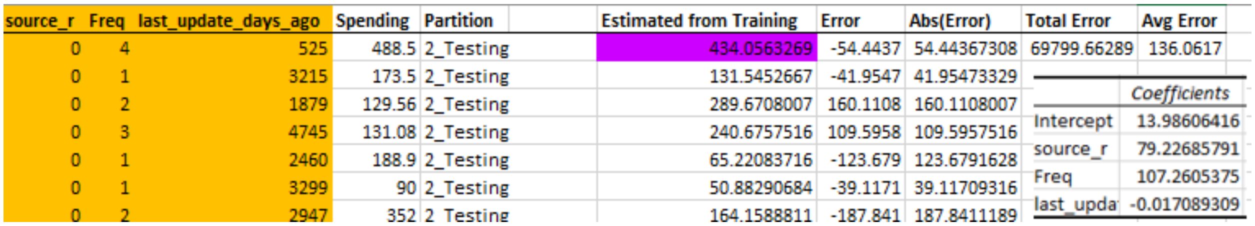
Target: Spending

Model Term	Coefficient ▶	Sig.	Importance
Intercept	100.567	.014	
Freq_transformed	108.429	.000	0.792
last_update_days_ago_transformed	-0.053	.000	0.100
source_r=0	-92.053	.004	0.067
source_r=1	0[a]		0.067
1st_update_days_ago_transformed	0.035	.024	0.040

[a] This coefficient is set to zero because it is redundant.

Display coefficients with sig. values less than...

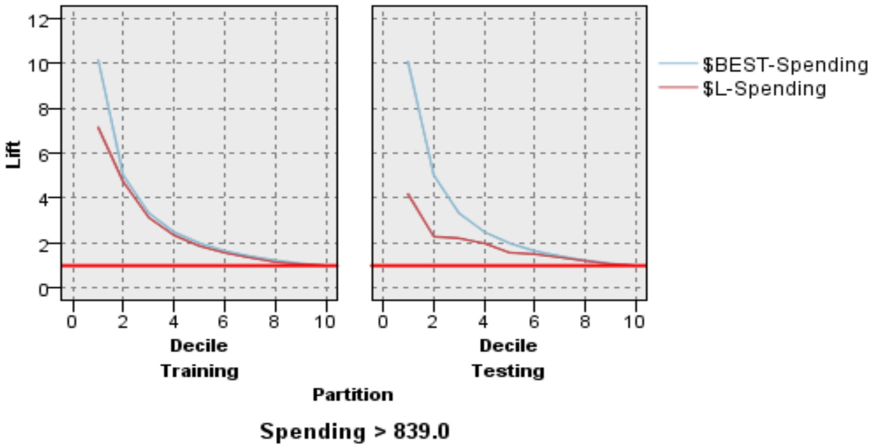
[Appendix 4: Coefficients Table for Spending Estimation from SPSS]



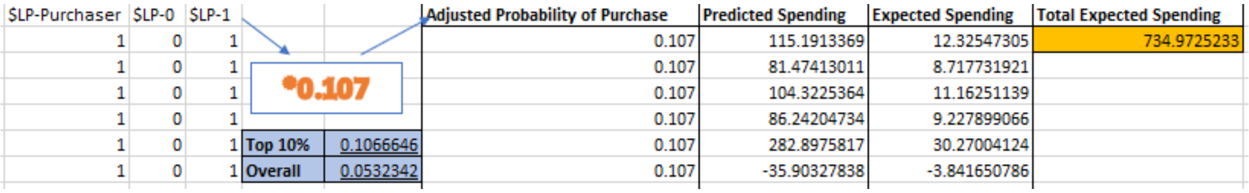
source_r	Freq	last_update_days_ago	Spending	Partition		Estimated from Training	Error	Abs(Error)	Total Error	Avg Error
0	4	525	488.5	2_Testing		434.0563269	-54.4437	54.44367308	69799.66289	136.0617
0	1	3215	173.5	2_Testing		131.5452667	-41.9547	41.95473329		
0	2	1879	129.56	2_Testing		289.6708007	160.1108	160.1108007		
0	3	4745	131.08	2_Testing		240.6757516	109.5958	109.5957516		
0	1	2460	188.9	2_Testing		65.22083716	-123.679	123.6791628		
0	1	3299	90	2_Testing		50.88290684	-39.1171	39.11709316		
0	2	2947	352	2_Testing		164.1588811	-187.841	187.8411189		

	Coefficients
Intercept	13.98606416
source_r	79.22685791
Freq	107.2605375
last_upda	-0.017089309

[Appendix 5: Error Estimation]



[Appendix 6: Lift Chart for “Spending” Model]



$LP-Purchaser	$LP-0	$LP-1			Adjusted Probability of Purchase	Predicted Spending	Expected Spending	Total Expected Spending
1	0	1			0.107	115.1913369	12.32547305	734.9725233
1	0	1	*0.107		0.107	81.47413011	8.717731921	
1	0	1			0.107	104.3225364	11.16251139	
1	0	1			0.107	86.24204734	9.227899066	
1	0	1	Top 10%	0.1066646	0.107	282.8975817	30.27004124	
1	0	1	Overall	0.0532342	0.107	-35.90327838	-3.841650786	

[Appendix 7: Validation Dataset Analysis Screenshot]

[Appendix8 “List of the people (sequence number) who are predicted to be in the top 10% of purchasing probability from the validation dataset”]

sequence_number
4
13

15
16
22
27
28
35
36
39
43
46
48
49
52
56
60
61
63
69
76
77
78
79
83
87
93
98
110
113
116
118
123
125
127
128
129
132
133
138
147
148
151
152
155

157
173
174
177
178
179
180