Name: Alisha Shahane
StudentID: 200311941
UnityID: asshahan

CSC 591 ADBI

Homework Solution

Data Pre-processing

The first step I did was splitting the data into train and test sets. All the pre-processing was performed only on training set.

Pivot table for 'Category' gave the following results,

Category	Competitive?
Antique/Art/Craft	0.50
Automotive	0.37
Books	0.55
Business/Industrial	0.64
Clothing/Accessories	0.51
Coins/Stamps	0.41
Collectibles	0.55
Computer	0.65
Electronics	0.74
EverythingElse	0.25
Health/Beauty	0.16
Home/Garden	0.68
Jewelry	0.37
Music/Movie/Game	0.57
Photography	0.88
Pottery/Glass	0.31
SportingGoods	0.70
Toys/Hobbies	0.53

I grouped the categories as follows,

New Category	Mean Range	Old Categories
Cat_1	[0.15 to 0.20)	Health/Beauty
Cat_2	[0.25 to 0.30)	EverythingElse
Cat_3	[0.30 to 0.35)	Pottery/Glass
Cat_4	[0.35 to 0.40)	Automotive, Jewelry
Cat_5	[0.40 to 0.45)	Coins/Stamps
Cat_6	[0.50 to 0.55)	Antique/Art/Craft, Clothing/Accessories, Toys/Hobbies, Books
Cat_7	[0.55 to 0.60)	Collectibles, Music/Movie/Game
Cat_8	[0.60 to 0.65)	Business/Industrial
Cat_9	[0.65 to 0.70)	Home/Garden, Computer, SportingGoods
Cat_10	[0.70 to 0.75)	Electronics
Cat_11	[0.85 to 0.90)	Photography

Pivot table for 'currency' gave following results,

currency	Competitive?
EUR	0.53
GBP	0.76
US	0.50

I grouped currency as follow,

New Currency	Mean Range	Old Currency
Cur_1	[0.50 to 0.60)	EUR, US
Cur_2	[0.6 to 0.70)	EGBP

Pivot table for 'duration' gave following results,

duration	Competitive?
1	0.58
3	0.47
5	0.68
7	0.46
10	0.54

I grouped them as follows,

New Duration	Mean Range	Old Durations
Dur_1	[0.40 to 0.50)	3, 7
Dur_2	[0.50 to 0.60)	1, 10
Dur_3	[0.60 to 0.70)	5

Pivot table for 'endDay' gave following results,

endDay	Competitive?
Fri	0.49
Mon	0.66
Sat	0.41
Sun	0.45
Thu	0.66
Tue	0.5
Wed	0.40

I grouped them as follow,

New endDay	Mean Range	Old endDay
Day_1	[0.40 to 0.45)	Wed, Sun, Sat
Day_2	[0.45 to 0.50)	Tue, Fri
Day_3	[0.65 to 0.70)	Mon, Thu

Problem 1

I build the model after grouping the data as mentioned above. I obtained the following results,

Predictor	Estimate	StdError	zValue	p>|z|
(Intercept)	-0.3124	0.7559	-0.4133	0.6794
sellerRating	0.0000	0.0000	-1.6278	0.1036
ClosePrice	0.1053	0.0112	9.3972	0.0000
OpenPrice	-0.1155	0.0119	-9.7012	0.0000
Category_cat_9	0.2109	0.6973	0.3025	0.7623
Category_cat_7	0.4287	0.6765	0.6337	0.5263
Category_cat_6	0.4020	0.6749	0.5957	0.5514
Category_cat_8	1.0162	0.9789	1.0381	0.2992
Category_cat_4	-0.1727	0.6923	-0.2494	0.8031
Category_cat_5	-0.6400	0.8733	-0.7329	0.4636
Category_cat_11	0.9766	1.4314	0.6823	0.4951
Category_cat_1	-1.0521	0.8037	-1.3091	0.1905
Category_cat_2	-0.7178	1.0976	-0.6540	0.5131
Category_cat_10	0.7806	0.8091	0.9647	0.3347
currency_cur_1	-1.0310	0.3165	-3.2578	0.0011
endDay_day_3	0.6155	0.1904	3.2331	0.0012
endDay_day_2	0.4002	0.1845	2.1695	0.0300
Duration_dur_3	0.8605	0.2492	3.4524	0.0006
Duration_dur_1	0.1960	0.2036	0.9628	0.3356

Based on the results, I chose 'Category_cat_8' (highlighted yellow) as my predictor for fit.single model.

And I chose ClosePrice, OpenPrice, currency_cur_1, endDay_day_3, endDay_day_2 and Duration_dur_3 as my predictors for reduced model.

Q1) fit.single

$\beta_o = 0.1$ and $\beta_1 = 0.45$

X_h *is* Business/Industrial

a) Probabilities

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} = \frac{1}{1 + e^{-0.45x - 0.1}}$$

b) Odds

$$\text{odds} = \frac{\text{Prob}(Y = \text{yes})}{1 - \text{Prob}(Y = \text{Yes})} = e^{-0.45x - 0.1}$$

c) Logit

$$\text{logit} = \log\left(\frac{\text{Prob}(Y = \text{yes})}{1 - \text{Prob}(Y = \text{Yes})}\right) = \log(e^{-0.45x - 0.1}) = -0.45x - 0.1$$

Q2) fit.all

Choosing four predictors with highest estimates.

$\beta_o = -0.31$

$\beta_1 = 1.01$ X_1 : Category_cat_8 $\beta_2 = 0.97$ X_2 : Category_cat_11

$\beta_3 = 0.86$ X_3 : Duration_dur_3 $\beta_4 = 0.78$ X_4 : Category_cat_10

$$\text{let } e^z = e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)} = e^{1.01x_1 - 0.97x_2 + 0.86x_3 + 0.78x_4 - 0.31}$$

a) Probabilities

$$p = \frac{1}{1 + e^z} = \frac{1}{1 + e^{1.01x_1 - 0.97x_2 + 0.86x_3 + 0.78x_4 - 0.31}}$$

b) Odds

$$\text{odds} = \frac{\text{Prob}(Y = \text{yes})}{1 - \text{Prob}(Y = \text{Yes})} = e^{1.01x_1 - 0.97x_2 + 0.86x_3 + 0.78x_4 - 0.31}$$

c) Logit

$$\text{logit} = \log\left(\frac{\text{Prob}(Y = \text{yes})}{1 - \text{Prob}(Y = \text{Yes})}\right) = \log\left(e^{1.01x_1 - 0.97x_2 + 0.86x_3 + 0.78x_4 - 0.31}\right)$$

$$= 1.01x_1 - 0.97x_2 + 0.86x_3 + 0.78x_4 - 0.31$$

Q3) Odds Ratio

The coefficient for highest predictor i.e. Business/Industrial in fit.all is 1.01
Let S be the sum of all terms of other predictors.

$$\text{odds ratio} = \frac{\text{odds}(X_h + 1, X_2, X_3, \ldots X_q)}{\text{odds}(X_h, X_2, X_3, \ldots X_q)} = \frac{e^{S + 1.01(X_h + 1)}}{e^{S + 1.01X_h}}$$

$$\text{odds ratio} = e^{1.01} = e^{\beta_1} = e^{\text{coefficient of predictor with highest estimate}} = 2.745$$

Thus, one unit increase in a predictor variable causes log-odds of response variable to change by a ratio of the coefficient of that predictor.

We can say that the logistic regression coefficients give the change in log odds of response variable for unit increase in a predictor variable, holding other predictors constant. If it was linear regression, it would have led to change equal to coefficient of that predictor in overall prediction of the model.

Q4) fit.reduced

Reduced model was built on ClosePrice, OpenPrice, currency_cur_1, endDay_day_3, endDay_day_2 and Duration_dur_3.

The anova test for comparing reduced and fit.all model gave a p value of 0.0053. This indicates that the models are significantly different.

Q4) Dispersion

$$\frac{\text{observed variance}}{\text{expected variance}} = 0.473 < 1$$

Since the ratio is less than 1, the model is not over dispersed.