Lab #3

By

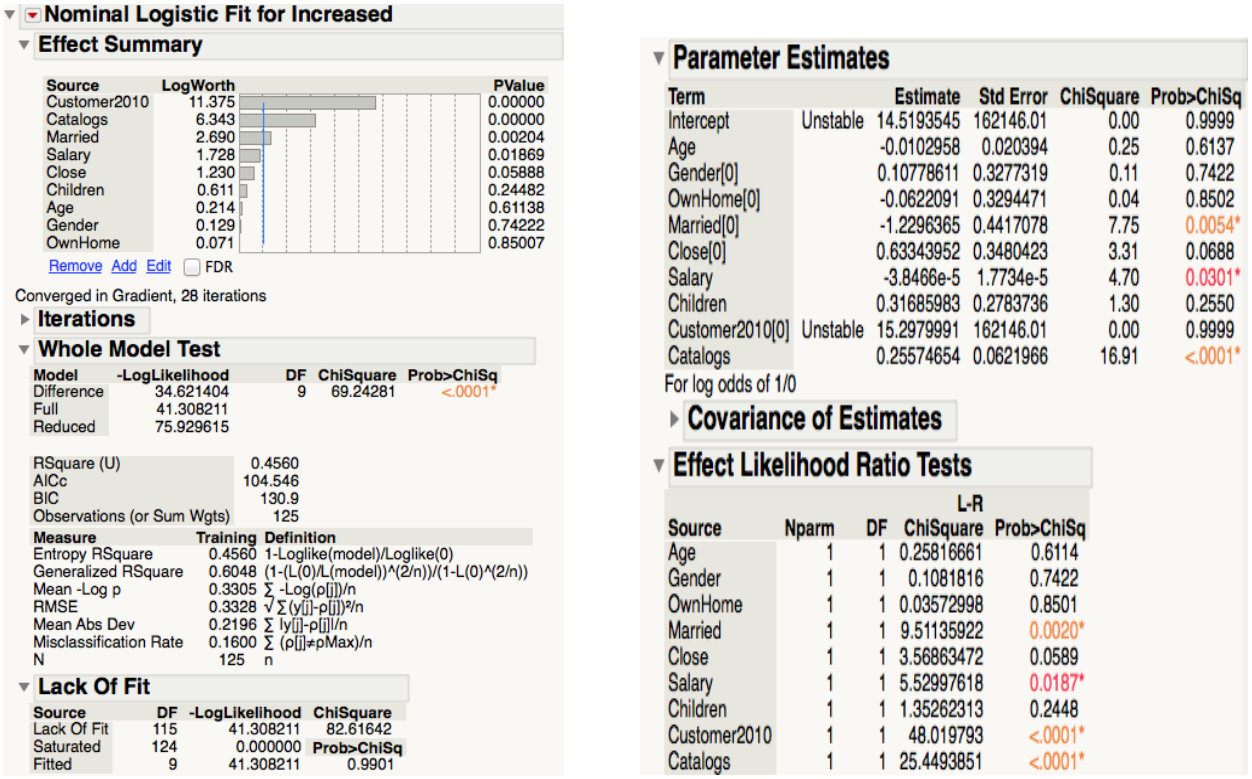
Olivia Rancour (orr2)
Rishi Reddy (rc81)
Nikhil Preeth Birra (nb35)
Vir Benipal (bs31)

Professor Prince

Lab 3- Multiple Logistic Regression

Task 1: Build a Logistic Regression Model

- JMP, Analyze → Fit Model → Increased (Y) and Age, Gender, OwnHome, Married, Close, Salary, Children, Customer2010, Catalogs (Construct Model Effects) → Run
- To get 50% (125 cases) of the data for training and 50% for testing (125 cases): Rows → Row Selection → Select Randomly (.5) → Right Click Highlighted Row (Exclude/Unexclude)



Nominal Logistic Fit for Increased

Effect Summary

Source	LogWorth	PValue
Customer2010	11.375	0.00000
Catalogs	6.343	0.00000
Married	2.690	0.00204
Salary	1.728	0.01869
Close	1.230	0.05888
Children	0.611	0.24482
Age	0.214	0.61138
Gender	0.129	0.74222
OwnHome	0.071	0.85007

Remove Add Edit FDR

Converged in Gradient, 28 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	34.621404	9	69.24281	<.0001*
Full	41.308211			
Reduced	75.929615			

RSquare (U)	0.4560
AICc	104.546
BIC	130.9
Observations (or Sum Wgts)	125

Measure	Training	Definition
Entropy RSquare	0.4560	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6048	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3305	∑ -Log(ρ[j])/n
RMSE	0.3328	√∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2196	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1600	∑ (ρ[j]≠ρMax)/n
N	125	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	115	41.308211	82.61642
Saturated	124	0.000000	Prob>ChiSq
Fitted	9	41.308211	0.9901

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	14.5193545	162146.01	0.00	0.9999
Age		-0.0102958	0.020394	0.25	0.6137
Gender[0]		0.10778611	0.3277319	0.11	0.7422
OwnHome[0]		-0.0622091	0.3294471	0.04	0.8502
Married[0]		-1.2296365	0.4417078	7.75	0.0054*
Close[0]		0.63343952	0.3480423	3.31	0.0688
Salary		-3.8466e-5	1.7734e-5	4.70	0.0301*
Children		0.31685983	0.2783736	1.30	0.2550
Customer2010[0]	Unstable	15.2979991	162146.01	0.00	0.9999
Catalogs		0.25574654	0.0621966	16.91	<.0001*

For log odds of 1/0

Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Age	1	1	0.25816661	0.6114
Gender	1	1	0.1081816	0.7422
OwnHome	1	1	0.03572998	0.8501
Married	1	1	9.51135922	0.0020*
Close	1	1	3.56863472	0.0589
Salary	1	1	5.52997618	0.0187*
Children	1	1	1.35262313	0.2448
Customer2010	1	1	48.019793	<.0001*
Catalogs	1	1	25.4493851	<.0001*

Figure 1

Task 2: Interpret the results

(1) Use the whole model and lack of fit tests. Identify significant parameter estimates and interpret their odds ratio.

In JMP, go to Analyze and Fit Model then we have taken outcome variable as "Increased" (Y) and other above considered independent variables to X "Construct Model Effects".

The below are the whole model test, lack of fit test and parameter estimates of the logistic regression model.

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	34.621404	9	69.24281	<.0001*
Full	41.308211			
Reduced	75.929615			

RSquare (U)	0.4560
AICc	104.546
BIC	130.9
Observations (or Sum Wgts)	125

Measure	Training	Definition
Entropy RSquare	0.4560	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6048	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3305	$\sum$ -Log(ρ[j])/n
RMSE	0.3328	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2196	$\sum$ ly[j]-ρ[j]l/n
Misclassification Rate	0.1600	$\sum$ (ρ[j]≠ρMax)/n
N	125	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	115	41.308211	82.61642
Saturated	124	0.000000	Prob>ChiSq
Fitted	9	41.308211	0.9901

Figure 2

From whole mode test, we can predict if the model fits better than constant response probabilities. This view can be viewed as analogous to the Analysis of Variance report for a continuous response model. It is a specific likelihood-ratio Chi-square test that evaluates how well the categorical model fits the data. The negative sum of natural logs of the observed probabilities is called the negative log-likelihood (–LogLikelihood). The negative log-likelihood for categorical data plays the same role as sums of squares in continuous data. Twice the difference in the negative log-likelihood from the model fitted by the data and the model with equal probabilities is a Chi-square statistic. This test statistic examines the hypothesis that the x variable has no effect on the responses.

Values of the Rsquare (U) (sometimes denoted as R2) range from 0 to 1. If R2 is high that means the model is a good fit however this is very rare in categorical models. The model is a good predictor of buy. Based off the Lack of Fit test, if Prob>ChiSq is closer to 1.0, the parameters are a good fit.

Hypothesis testing:
H0 = Not a good model(X has no effect in Y). None of the parameters have a significant effect on the probability of success.
Ha = It is a good model (X has an effect on Y)
From the parameter estimates, the significant variables are Married, Salary, Customer2010, Catalogs

Hypothesis Test for Married:
Ho: $\beta 1 = 0$
Ha: $\beta 1 \neq 0$
P-value is <0.0001 which is less than 0.05. We reject Ho.

Hypothesis Test for Salary:
Ho: $\beta 2 = 0$
Ha: $\beta 2 \neq 0$
P-value is <0.0001 which is less than 0.05. We reject Ho.

Hypothesis Test for Customer2010:
Ho: $\beta 3 = 0$
Ha: $\beta 3 \neq 0$
P-value is <0.0001 which is less than 0.05. We reject Ho.

Hypothesis Test for Catalog:
Ho: $\beta 4 = 0$
Ha: $\beta 4 \neq 0$
P-value is <0.0001 which is less than 0.05. We reject Ho.

From the above hypothesis test the significant parameters can be confirmed as married,salary,customer2010 and catalog.
The odds ratio is a common effect indicator. To get the odds ratio: JMP, Analyze → Fit Model → Increased (Y) and Age, Gender, OwnHome, Married, Close, Salary, Children, Customer2010, Catalogs (Construct Model Effects) → Run → Red Triangle → Odds Ratio
For task success odds of 1 versus 0, the odds increase the probability that spending is greater this year than last year. As seen below.

Nominal Logistic Fit for Increased

Odds Ratios

For Increased odds of 1 versus 0
Tests and confidence intervals on odds ratios are likelihood ratio based.

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Age	0.989757	0.949297	1.029556	1.0103489
Salary	0.999962	0.999924	0.999994	1.0000385
Children	1.37281	0.809162	2.440358	0.7284329
Catalogs	1.291425	1.156198	1.480557	0.7743382

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Age	0.567641	0.057164	4.963068	1.7616779
Salary	0.009893	0.000103	0.483624	101.08004
Children	2.587208	0.529793	14.53318	0.386517
Catalogs	99.82691	13.63255	1168.565	0.0100173

Odds Ratios for Gender

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
1	0	0.80608	0.7422	0.2181179	2.9560522
0	1	1.2405716	0.7422	0.338289	4.5846754

Odds Ratios for OwnHome

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
1	0	1.1324894	0.8501	0.3113407	4.2480769
0	1	0.8830105	0.8501	0.2354006	3.2119159

Odds Ratios for Married

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
1	0	11.696305	0.0020*	2.3331887	77.480391
0	1	0.0854971	0.0020*	0.0129065	0.428598

Odds Ratios for Close

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
1	0	0.2817095	0.0589	0.0658585	1.0474865
0	1	3.5497566	0.0589	0.9546663	15.184063

Odds Ratios for Customer2010

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
1	0	5.156e-14	<.0001*	0	0.0158199
0	1	1.939e+13	<.0001*	63.211696	.

Figure 3

2) The most important predictors are married,customer2010(says if purchase is made last year or not) and catalog. From the above predictors the important characteristics for predicting the increased variable are Demographic characteristics and Psychographic characteristics.The customer can be categorized into demographic category such as married which is an important predictor .The psychographic characteristics can be deducted from the catalogs sent to the customers, the items bought which gives the interests ,activities and opinion .

Task 3: Evaluate the model--If above 45 Degree, pass

(1) ROC, Lift, and Confusion Matrix

- From JMP Output *LRT* → Turn on: Odds Ratio, ROC Curve (level 1 is positive), Lift Curve, Confusion Matrix

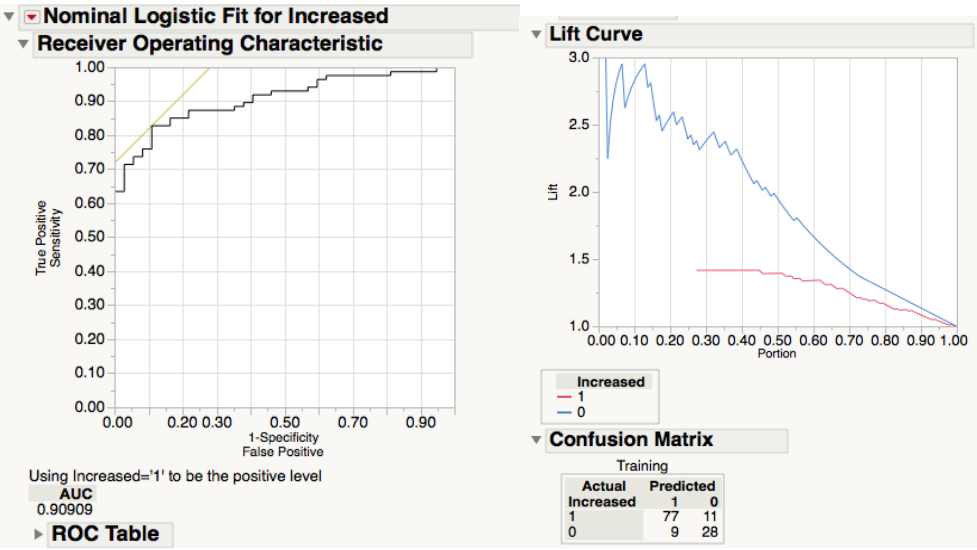


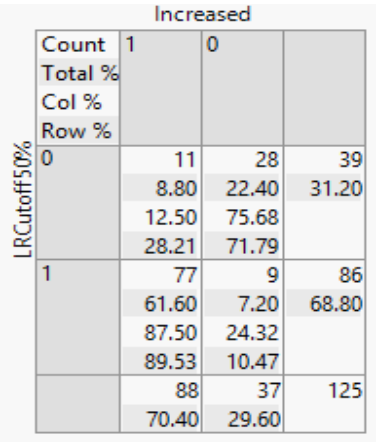
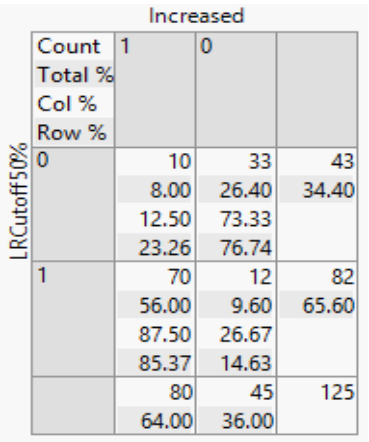
Figure 4

The ROC curve is above the base line and the Area under Curve (AUC) = 0.90909 which is greater than 0.5, Hence we can say that the model has good accuracy

If we consider the Lift Curve, the lift curve of Increased = 1 and Increased = 0 is above the base line. So based on the LIFT, we can say that the model has good accuracy.

(2) Save Probability Formula

- From JMP Output → *LRT* → Save Probability Formula → This gave us four new columns: Lin[1], Prob[1], Prob[0], and Most Likely Increased
- Confusion matrix for training data and a confusion matrix for testing data: JMP → Fit Y by X → Increased (Y) and Most Likely Increased (X)

Confusion Matrix for data set (training)	Confusion Matrix for data set (testing)
	

(3) Trial and error to find a better cutoff probability that increase accuracy in confusion matrix. In JMP, create a new columns like "Cutoff30%" with formula: If (Prob[1]>0.3] 1, else 0, etc.

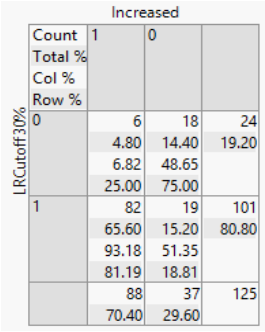
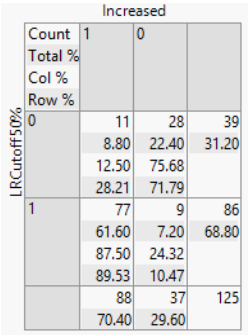
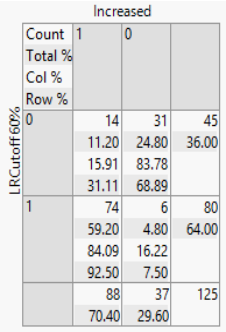
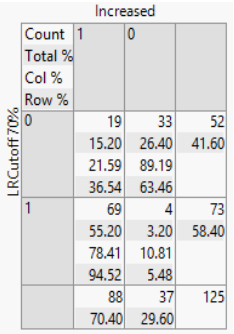
	Cutoff30%	**Cutoff50%**	**Cutoff60%**	**Cutoff70%**
Confusion Matrix (for training data)				
Total Accuracy	(82+18)/125= 80%	(77+28)/125= 84%	(74+31)/125= 84%	(69+33)/125= 81.6%
Accuracy Increased = 1	82/88=93.1%	77/88=87.5%	74/88=84%	69/88=78.4%
Accuracy Increased = 0	18/37=48.6%	28/37=75.6%	31/37=83.7%	33/37=89.1%

Table 1: Confusion Matrix for Training data

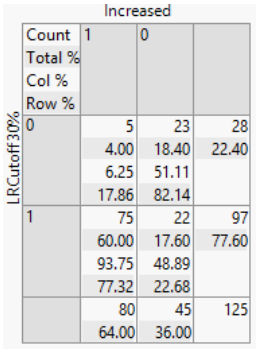
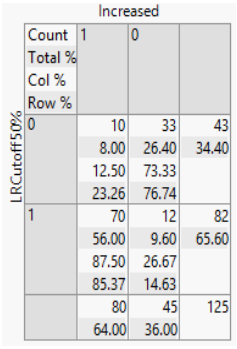
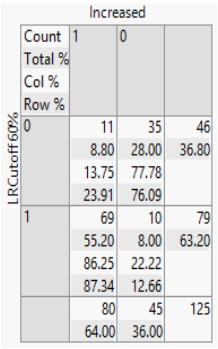
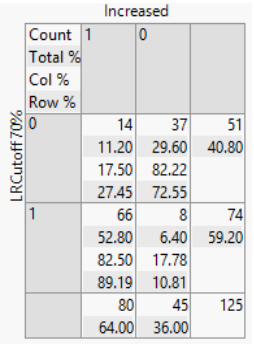
	Cutoff30%	Cutoff50%	Cutoff60%	Cutoff70%
Confusion Matrix (for testing data)				
Total Accuracy	(75+23)/125= 78.4%	(70+33)/125= 82.4%	(69+35)/125= 83.2%	(66+37)/125= 82.4%
Accuracy Increased = 1	75/80=93.7%	70/80=87.5%	69/80=86.2%	66/80=82.5%
Accuracy Increased = 0	23/45=51.1%	33/45=73.3%	35/45=77.7%	37/45=82.2%

Table 2: Confusion Matrix for Training data

Task 4: Explain how Terry can use these results to make a new advertising plan.

From the above parameter estimates, effective likelihood ratio tests and hypothesis testing we have the significant attributes married,salary,catalogs and customer2010. So to increase the sale Terry's village should advertise gifts ,outdoor knick-knack, home decor and holiday items to attract the married with salary 20,000$ to 60,000$.They can also target this section by sending catalogs with items which they can buy. By targeting married couples they can sell home decor, holiday packages and, children's items etc.