University of Southern California Marshall School of Business	Case 5 Ensemble Model
DSO 528 **Prof. Ansari**	**Team members: Andre Oliver, Guilherme Castro, Ying- Chu Lo, Penha Martins**

Executive Summary

1.1 Best model

The model proposed, named ensemble model, combined a neural network and a decision tree models. From a neural network perspective, we used only most statistical relevant variables, which are Recency, Income, Gender, New_State, Block_Party, and Art_Party. By using these variables and a two-layer model, we believe that we were able to build a model that takes into consideration if a customer is male or female, is located in the west coast, has an income higher than 150,000, bought the product in the last 12 months, so this customer would probably buy the product again, making business sense.

From a decision tree perspective, first, we split based on the state, because we could observe, analyzing the geographic map chart, distinct patterns of income among the states. Second, we used the specific split suggested by the software to identify the best type of party. Finally, we select the variable Recency, since that from a business point of view makes sense to assume that a customer who bought party products recently has a higher propensity to repurchase them in the future.

Furthermore, we also consider additional criteria to score and evaluate our best model. In the following sections, we will dive deep into those criteria. Primarily, our best model considers variables statistically significant; the expected profit is around $1,65 million -the highest total profit achieved and optimized the number of packages sent.

1.1.1 Relevance of the model

Combining decision tree and neural network models, we could design a new model that considers variables statistically significant (significance level of decision tree model is < .0001*), and that has a higher predictive power (neural network predictive power is 40.86%). Besides that, from a business perspective, the model also makes sense, since that we will target customers with higher income and in this sense with a higher propensity to repurchase the products.

The table below illustrated the significance level and R-Square (predictive power) for each model when applicable

Model	Ensemble	Neural Network	Decision Tree
Significance level	Not Applicable	Not Applicable	<.0001*

Model	Ensemble	Neural Network	Decision Tree
R-Square	Not Applicable	40.86%	12.49%

Also, one of the key KPIs to evaluate the model is the misclassification rate. In this sense, we could find the following results for training and testing data set.

Model	Ensemble	Neural Network	Decision Tree
Misclassification rate*	12.9%	7.7%	21.1%

*training set

Model	Ensemble	Neural Network	Decision Tree
Misclassification rate*	12.4%	7.2%	9.9%

*testing set

Clearly, we could see that the adoption of a neural network model contributed to reduce significantly the misclassification rate.

1.1.2 Expected profit

Comparing the models from a profit perspective, we could see that the ensemble model is expected to generate a profit of $ 1.65 million, which means 1.4x and 2.45x the expected profit associated with the neural network and the decision tree model respectively. Also, it is possible to achieve those results reducing the number of packages sent (decision tree vs. ensemble) which ultimately minimize the expected loss.

Variables	Ensemble (best model)	Neural Network	Decision Tree
Number of Packages sent	**89,500**	**18,500**	**100,000**
Revenue	1,842,750.00	682,500.00	1,437,732.34
Loss	(196,000.00)	(14,000.00)	(273,605.95)
Total Profit	$ 1,646,750.00	$ 668,500.00	$ 1,164,126.00

1.1.3 Conclusion

To conclude, we recommend the adoption of the ensemble model since that this model is expected to generate the highest total profit - $1.64 million, the misclassification rate is 12.9% and the number of package makes sense from business perspective.

JMP Model (ENSEMBLE)
Ensemble Model Average

i) Statistical KPIs of JMP Model – From Excel Printout

Other Metrics	Training	Testing
Accuracy %	87.10%	87.60%
True Positive Rate	72.32%	64.76%
False Positive Rate	11.04%	9.72%
Sensitivity (True Positive Rate)	72.32%	64.76%
Specificity (True Negative Rate)	88.96%	90.28%

KPI Chart	Training	Testing
R-Square	10.50%	7.90%
Accuary	87.10%	87.60%
Sensitivity	72.32%	64.76%
Specificity	88.96%	90.28%
Lift above 1	3.0403033	2.917050691

ii) a) Business KPIs of JMP Model – Training

Predicted number of Buyer	=	89500
Upper limit for packages sent	=	100000
Actual number of packages sent	=	89500

Propensity to buy the Package	=	45.251%
Propensity to not buy the Package	=	54.749%

Total Profit	=	$ 1,646,750

b) Business KPIs of JMP Model – Testing

Predicted number of Buyer	=	77500
Upper limit for packages sent	=	100000
Actual number of packages sent	=	77500
Propensity to buy the Package	=	43.871%
Propensity tonot buy the Package	=	56.129%
Total Profit	=	$ 1,373,000

iii) Confusion Matrix for Training

		Predicted		
		Not Buyer	Buyer	
Actual	Not Buyer	790	98	888
	Buyer	31	81	112
		821	179	1000

iv) Confusion Matrix for Testing

		Predicted		
		Not Buyer	Buyer	
Actual	Not Buyer	808	87	895
	Buyer	37	68	105
		845	155	1000

i) Profit and Loss Comparison between Ensemble, Neural Network and Logistic Regression

	Ensemble	Neural Network	Decision Tree
Number of Packages sent	**89,500**	**18,500**	**100,000**
Revenue	1,842,750.00	682,500.00	1,437,732.34
Loss	(196,000.00)	(14,000.00)	(273,605.95)
Total Profit	$ 1,646,750.00	$ 668,500.00	$ 1,164,126.00

i) Profit and Loss Comparison between Ensemble, Neural Network and Logistic Regression

	Ensemble	Neural Network	Decision Tree
Propensity to buy the Package	45.251%	81.081%	31.599%
Propensity to not buy the Package	54.749%	18.919%	68.401%

JMP Model (NEURAL NETWORK)
Training/Testing

i) Statistical KPIs of JMP Model – From JMP Printout

Training – **Success**

Measures	Value
Generalized RSquare	0.4939087
Entropy RSquare	0.4086615
RMSE	0.2466019
Mean Abs Dev	0.1247827
Misclassification Rate	0.0765766
-LogLikelihood	137.38105
Sum Freq	666

Validation – **Success**

Measures	Value
Generalized RSquare	0.4431569
Entropy RSquare	0.3596623
RMSE	0.2506347
Mean Abs Dev	0.1193159
Misclassification Rate	0.0718563
-LogLikelihood	75.781672
Sum Freq	334

Statistical KPIs of JMP Model – From Excel Printout

Other Metrics	Training	Testing
Accuracy %	92.34%	92.81%
True Positive Rate	40.54%	50.00%
False Positive Rate	1.18%	1.69%
Sensitivity (True Positive Rate)	40.54%	50.00%
Specificity (True Negative Rate)	98.82%	98.31%

ii) a) Business KPIs of JMP Model – Training

Predicted number of Buyer	=	18500
Upper limit for packages sent	=	100000
Actual number of packages sent	=	18500

Propensity to buy the Package	=	81.081%
Propensity to not buy the Package	=	18.919%

Total Profit	=	$ 668,500

b) Business KPIs of JMP Model – Testing

Predicted number of Buyer	=	12000
Upper limit for packages sent	=	100000
Actual number of packages sent	=	12000
Propensity to buy the Package	=	79.167%
Propensity tonot buy the Package	=	20.833%
Total Profit	=	$ 422,250

iii) Confusion Matrix for Training

Confusion Matrix

Actual Success	Predicted Count 0	1
0	585	7
1	44	30

iv) Confusion Matrix for Testing

Confusion Matrix

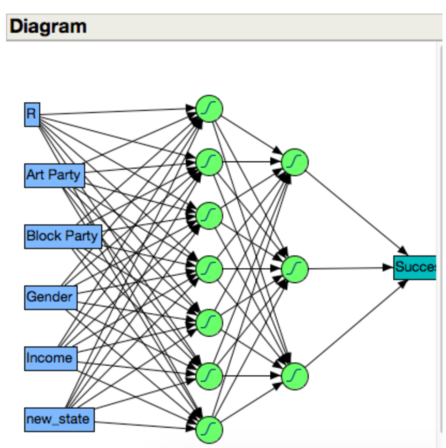
Actual Success	Predicted Count 0	1
0	291	5
1	19	19

v) Lift Table

Lift Table in Dollars	Training	Testing
Lift with respect to Baseline - JMP Model	4.456666667	2.815
Lift with respect to Baseline - My Best Model	6.063333333	5.773333333
Lift with respect to JMP Model - My Contribution	1.360508601	1.295437547
Overall Lift with respect to Baseline -My Best Model	6.063333333	5.773333333

Lift Table in Propensity	Training	Testing
Lift with respect to Baseline - JMP Model	7.297297297	7.125
Lift with respect to Baseline - My Best Model	2.743902439	2.683229814

vi) Neural Network Diagram



JMP Model (Decision Tree)
Training/Testing

Creating the enriched variables

1) State – by using the Graph Builder the team plotted the U.S. map against the Income variable to understand how the income is distributed across the states. After careful consideration, the team decided to divide the state variables into regions -> West, East, Central
2) HomeOwnership – changed from 5 different categories into only two -> Renters and Home Owners. The team didn't use this variable to split the decision tree because it was not relevant.
3) Urbanicity – changed from 6 different categories into only two -> Urban and Rural. The team didn't use this variable to split the decision tree because it was not relevant.

i) Statistical KPIs of JMP Model – From JMP Printout

Fit Details

Measure	Training	Definition
Entropy RSquare	0.1249	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1663	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3069	∑ -Log(ρ[j])/n
RMSE	0.2987	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1774	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1090	∑ (ρ[j]≠ρMax)/n
N	1000	n

Statistical KPIs of JMP Model – From Excel Printout

Other Metrics	Training	Testing
Accuracy %	78.90%	90.10%
True Positive Rate	75.89%	45.71%
False Positive Rate	20.72%	4.69%
Sensitivity (True Positive Rate)	75.89%	45.71%
Specificity (True Negative Rate)	79.28%	95.31%

ii) a) Business KPIs of JMP Model – Training

Predicted number of Buyer	=	134500
Upper limit for packages sent	=	100000
Actual number of packages sent	=	100000
Propensity to buy the Package	=	31.599%
Propensity to not buy the Package	=	68.401%

Total Profit	=	$ 1,164,126

b) Business KPIs of JMP Model – Testing

Predicted number of Buyer	=	45000
Upper limit for packages sent	=	100000
Actual number of packages sent	=	45000
Propensity to buy the Package	=	53.333%
Propensity tonot buy the Package	=	46.667%
Total Profit	=	$ 1,008,000

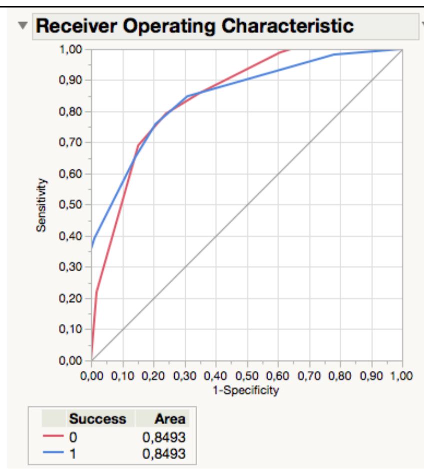
iii) Confusion Matrix for Training

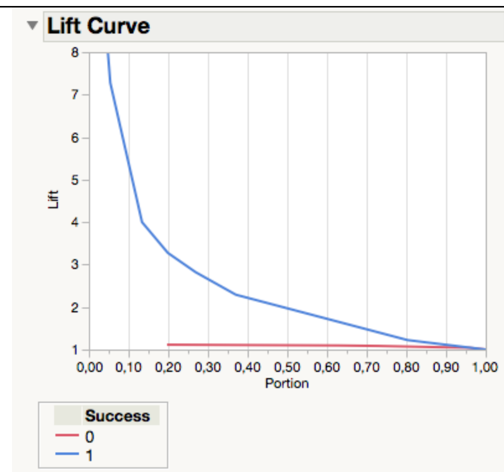
		Predicted		
		Not Buyer	Buyer	
Actual	Not Buyer	704	184	888
	Buyer	27	85	112
		731	269	1000

iv) Confusion Matrix for Testing

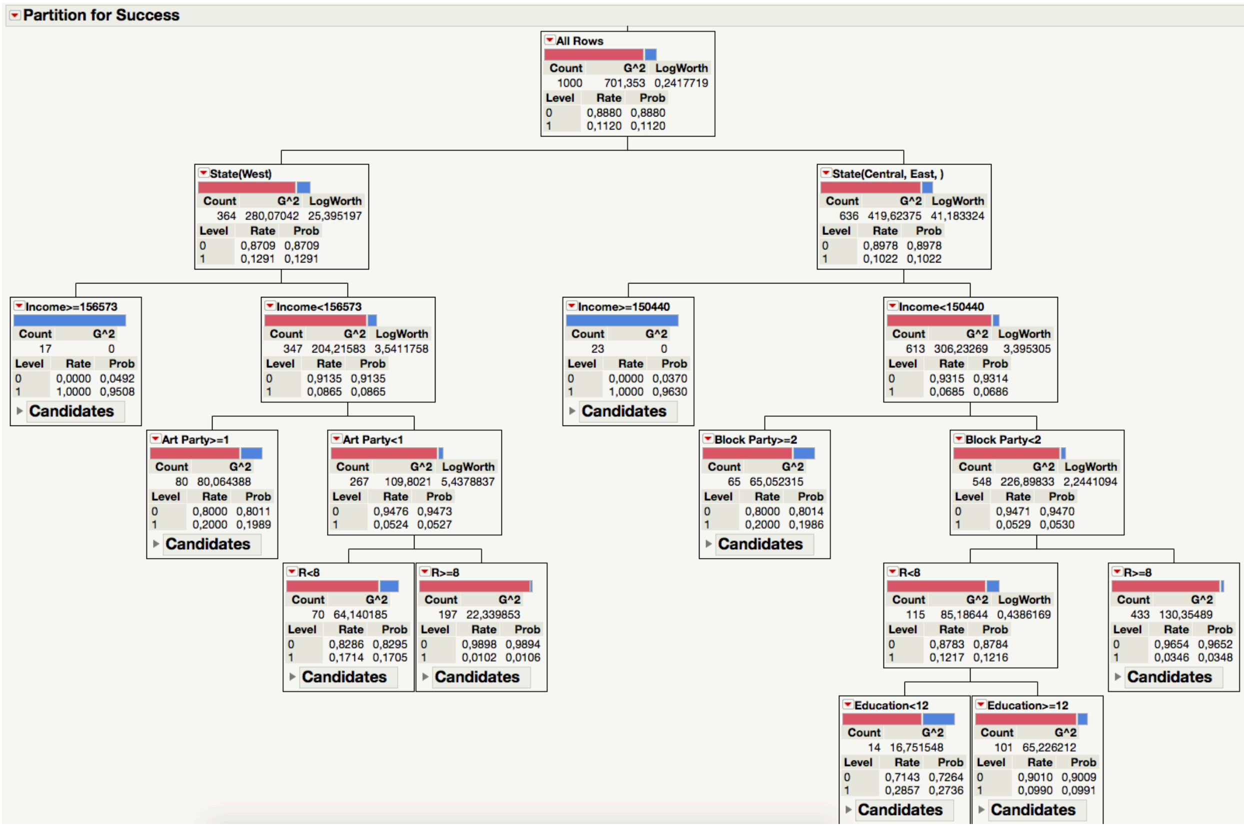
		Predicted		
		Not Buyer	Buyer	
Actual	Not Buyer	853	42	895
	Buyer	57	48	105
		910	90	1000

v) ROC and Lift Curve





i) Decision Tree


Partition for Success
All Rows
Count 1000 G^2 701,353 LogWorth 0,2417719
Level 0 Rate 0,8880 Prob 0,8880
Level 1 Rate 0,1120 Prob 0,1120
State(West)
Count 364 G^2 280,07042 LogWorth 25,395197
0 0,8709 0,8709
1 0,1291 0,1291
State(Central, East,)
Count 636 G^2 419,62375 LogWorth 41,183324
0 0,8978 0,8978
1 0,1022 0,1022
Income>=156573
Count 17 G^2 0
0 0,0000 0,0492
1 1,0000 0,9508
Candidates
Income<156573
Count 347 G^2 204,21583 LogWorth 3,5411758
0 0,9135 0,9135
1 0,0865 0,0865
Income>=150440
Count 23 G^2 0
0 0,0000 0,0370
1 1,0000 0,9630
Candidates
Income<150440
Count 613 G^2 306,23269 LogWorth 3,395305
0 0,9315 0,9314
1 0,0685 0,0686
Art Party>=1
Count 80 G^2 80,064388
0 0,8000 0,8011
1 0,2000 0,1989
Candidates
Art Party<1
Count 267 G^2 109,8021 LogWorth 5,4378837
0 0,9476 0,9473
1 0,0524 0,0527
Block Party>=2
Count 65 G^2 65,052315
0 0,8000 0,8014
1 0,2000 0,1986
Candidates
Block Party<2
Count 548 G^2 226,89833 LogWorth 2,2441094
0 0,9471 0,9470
1 0,0529 0,0530
R<8
Count 70 G^2 64,140185
0 0,8286 0,8295
1 0,1714 0,1705
Candidates
R>=8
Count 197 G^2 22,339853
0 0,9898 0,9894
1 0,0102 0,0106
Candidates
R<8
Count 115 G^2 85,18644 LogWorth 0,4386169
0 0,8783 0,8784
1 0,1217 0,1216
R>=8
Count 433 G^2 130,35489
0 0,9654 0,9652
1 0,0346 0,0348
Candidates
Education<12
Count 14 G^2 16,751548
0 0,7143 0,7264
1 0,2857 0,2736
Candidates
Education>=12
Count 101 G^2 65,226212
0 0,9010 0,9009
1 0,0990 0,0991
Candidates