

# Delta Airlines Response Model Overview

**BUDT 733**

## **Group 5**

**Geoff Gray**

**Armando “Mannix” Litonjua**

**Matt McNamara**

**Tim Roach**

**Jason Thayer**

December 13, 2005



# Agenda

- Methodology
- Variables Examined
- Exploratory Analysis
- Logistic Regression Model
- Discriminant Analysis
- Recommendation

# Methodology

- **Predictive Goal:** Identify Delta Sky Miles members most likely to respond to a frequent flier mileage offer in exchange for the purchase of Sprint-Nextel wireless phones and service
  
- Created stratified sample from Delta Sky Miles members database
  - 1:6 ratio of buyers:non-buyers (includes all buyers from the population)
  - ~ 26K records total from a population of 4,000,000
  - 17 potential explanatory variables
  
- Divided stratified sample into Training / Validation / Test sets
  
- Performed exploratory analyses and built regression model using analysis sample
  
- Scored validation set using logistic regression and discriminant analysis models and compared lift curves
  
- Performed final check of chosen model using test data – scored test data using LR and DA model and compared lift to validation set

# Variables Examined

## 4 Classes of Explanatory Variables

### Account ID Variables

Account Number (unique ID; not a potential explanatory variable)

Zip Code

Age

### Member Status Variables

Elite Status Code (Silver; Gold; Platinum)

Crown Member Status

Million Mile Member Status

Mileage Range

### Partner Status Variables

Indicates membership with a partner company

- AMEX, Optima, MCI, Diners Club

### Transaction Recency Variables

Enrollment Recency

Domestic / International Flight Recency

Last Award Recency

Partner Transaction Recency

Hotel / Car Rental Recency

### Response Variable

“BUY\_IND” – Binary variable that indicates “1” if the member was a sprint product buyer.

Prior Probabilities: Actual buyers in the population =  $3,710 / 4,000,000 = 0.093\%$

Unequal misclassification costs:

- Cost of misclassifying a buyer as a non-buyer = \$900 (estimated lifetime value of subscriber)

- Cost of misclassifying a non-buyer as a buyer = \$2 (total cost to send mail to an individual Delta frequent flier)

# Exploratory Analyses

## ➤ Account ID Variables

- No meaningful predictive power for Zip Code or Age
- Age data missing for 80% of records

## ➤ Member Status Variables

ELITE_STATUS_CODE	% Buyers
None	13.6%
FO	17.0%
GM	21.7%
PM	24.7%
Overall Sample	14.3%

Meaningful  
Upward  
Trend

## Additional Observations

- Crown Members are 22% buyers – meaningful difference from the overall sample
- All Million Mile Members are also Elite Members
  - Million Mile Membership not used

MILEAGE_RANGE_CODE	% Buyers
0 - 13,999 miles	9.4%
14,000 - 19,999 miles	14.3%
20,000 - 23,999 miles	15.8%
+ 24,000 miles	16.8%
Overall Sample	14.3%

Overall  
Upward  
Trend

# Exploratory Analyses

## ➤ Partner Status Variables

- Each partner membership appears meaningful.
- AMEX members – only partner to indicate buyers are less likely.
- No meaningful correlation between the four.

Partner Status	% Buyers
OPTIMA_IND	22.2%
AMEX_IND	10.9%
DINERS_IND	23.1%
MCI_IND	27.6%
Overall Sample	14.3%

## ➤ Transaction Recency Variables

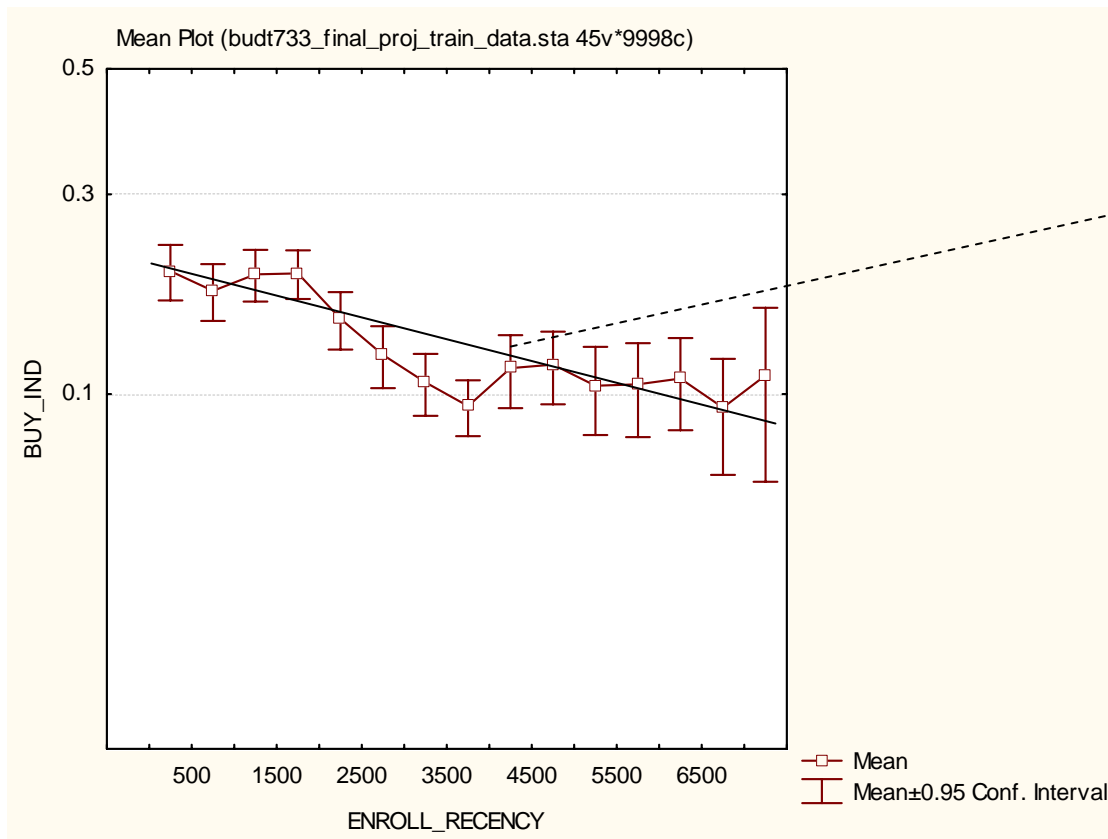
- All categorical values show the same trend – Customers who have purchased/flown/etc. more recently tend to have a higher percentage of buyers.
- 0-6 Month category appears to be the most meaningful.

Recency Category	% of Buyers					
	DOM_FLT	INTL_FLT	AWARD	PARTNER_TRANS	HOTEL_CAR_RENT	
0_6_MOS	16.9%	23.4%	22.4%	22.8%	20.2%	19.4%
7_12_MOS	12.8%	18.1%	18.5%	13.5%	19.9%	19.2%
13_24_MOS	14.1%	13.4%	15.4%	10.8%	15.4%	17.9%
24+_MOS	14.1%	15.8%	13.1%	9.4%	14.7%	12.8%
UNKNOWN	12.3%	13.5%	11.8%	5.9%	13.4%	13.4%
Grand Total	14.3%	14.3%	14.3%	14.3%	14.3%	14.3%

Hotel and Car Rentals count as Partner Transactions

# Exploratory Analyses

- Enrollment Recency is the only continuous variable – buyers tend to have a lower enrollment recency than non-buyers.



Fair/strong negative relationship



# Logistic Regression (cont'd)

## ➤ Coefficients for final LR model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.41753435	0.09168324	0	*
MCI_IND	0.64764637	0.06856698	0	1.91103768
ENROLL_RECENCY	-0.00012244	0.00001684	0	0.99987757
PARTNER_TRANS_IND	0.61538446	0.101902	0	1.8503679
PART_TRANS_REC_0_6	0.67535102	0.07945045	0	1.96472251
CELL_5_IND	-0.27596936	0.10984329	0.0119916	0.75883621

- CELL\_5\_IND used as a control variable to get proper estimates for the other coefficients

## ➤ Model Evaluation

- Lift curves showed the model prioritized buyers near the top (1.75x lift when mailing 40% of the available prospects)
- Confusion matrices not helpful
  - Model classified nearly all records as Non-buyers in Training, Validation, Test sets
  - Goal was prioritization rather than classification accuracy

# Discriminant Analysis

- Normalized Predictors to compare predictive utility
- Initial Model Ran with all 14 predictors
- Cutoff Value adjusted by prior probabilities and misclassification costs to minimize total overall cost.
  - Tends to place more importance on correctly identifying **non-buyers**.
- Overall % Error = 17.60%

## 14 Initial Predictor Variables

ELITE_STATUS_CODE_FO
ELITE_STATUS_CODE_GM
ELITE_STATUS_CODE_PM
CROWN_IND
MILEAGE_RANGE_0-13,999
OPTIMA_IND
AMEX_IND
DINERS_IND
MCI_IND
ENROLL_RECENCY
DOM_FLT_CAT_0_6_MOS
INTL_FLT_CAT_0_6_MOS
AWARD_CAT_0_6_MOS
PARTNER_TRANS_CAT_0_6_MOS

## Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	<b>0.76</b>
---	-------------

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	250	1162
0	529	7667

Error Report			
Class	# Cases	# Errors	% Error
1	1412	1162	82.29
0	8196	529	6.45
<b>Overall</b>	<b>9608</b>	<b>1691</b>	<b>17.60</b>

	prior P Misclass Cost	Total Cost
0.0009275	\$900	\$68.70
0.9990725	\$2	\$12.90
		\$81.59

# Discriminant Analysis

- Re-ran model with most discriminatory
  - Difference in buyer/non-buyer
- Re-Ran Model with just 6 predictors
- Results in actually lower total cost at an
- Overall % Error slightly larger = 17.70%

Predictive Power	14 Initial Predictor Variables
0.368	MCI_IND
0.290	PARTNER_TRANS_CAT_0_6_MOS
0.246	ENROLL_RECENCY
0.226	OPTIMA_IND
0.147	INTL_FLT_CAT_0_6_MOS
0.145	MILEAGE_RANGE_0-13,999
0.091	AWARD_CAT_0_6_MOS
0.059	ELITE_STATUS_CODE_GM
0.059	CROWN_IND
0.058	AMEX_IND
0.042	DINERS_IND
0.030	DOM_FLT_CAT_0_6_MOS
0.011	ELITE_STATUS_CODE_FO
0.007	ELITE_STATUS_CODE_PM

Cutoff at 1<sup>st</sup>  
6 variables

## Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable) **0.75**

Classification Confusion Matrix				
		Predicted Class		
Actual Class	1		0	
	1	276	1136	
0	565	7631		

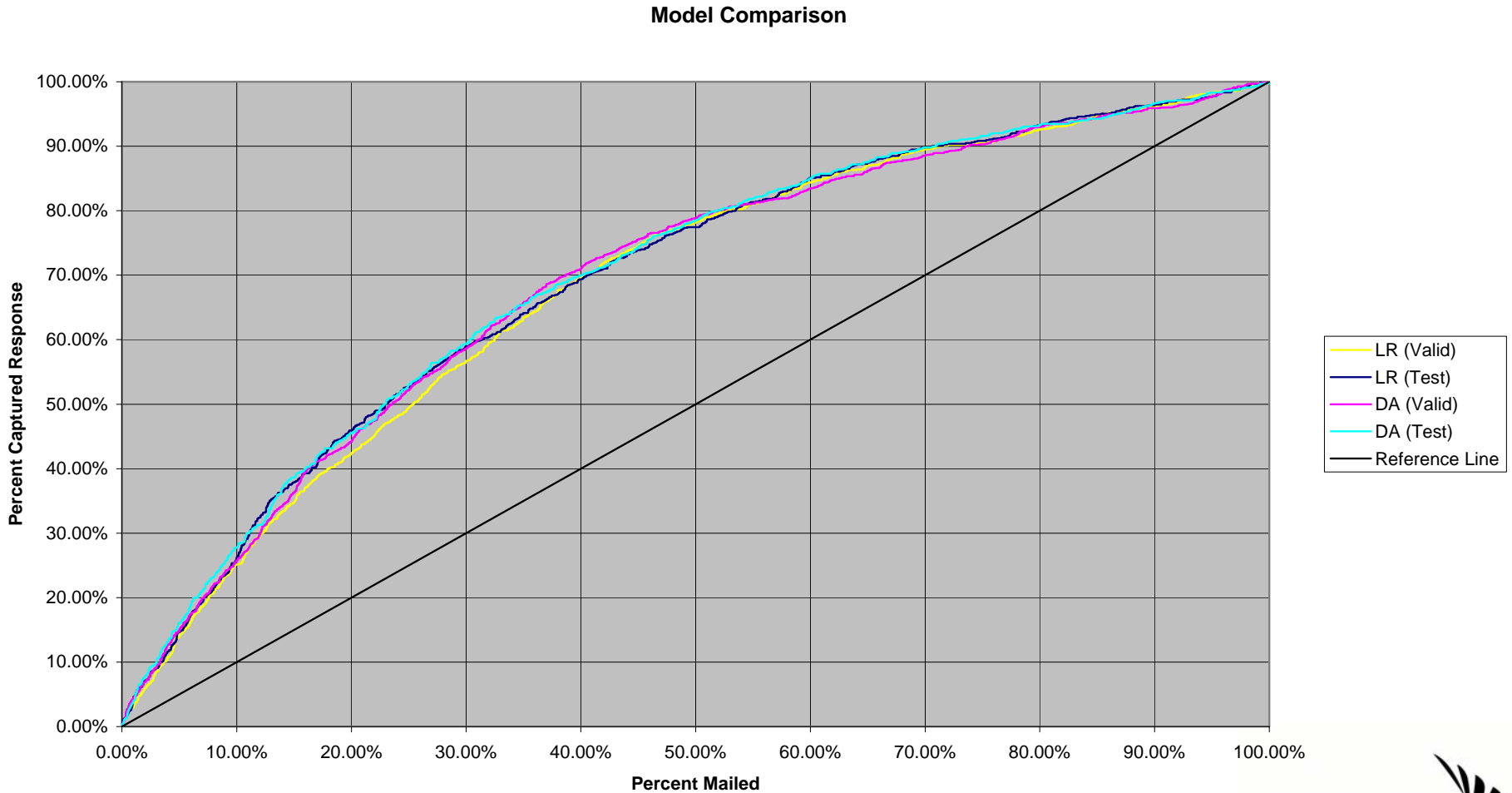
Error Report			
Class	# Cases	# Errors	% Error
1	1412	1136	80.45
0	8196	565	6.89
<b>Overall</b>	<b>9608</b>	<b>1701</b>	<b>17.70</b>

	prior P Misclass Cost	Total Cost
0.0009275	\$900	\$67.16
0.9990725	\$2	\$13.77
		\$80.93



# Model Comparison

Both LR and DA models do a good job of prioritizing buyers near the top of the mailing list



# Recommendations

- We recommend using the Logistic Regression model
  - Discriminant analysis normality assumptions violated due to large number of
- Three key predictors can be identified:
  - **MCI Membership**
  - **Partner Transaction Recency**
  - **Enrollment Recency**
- Additional Data collection that may be useful:
  - Complete Age Data
  - Income
  - Geographic Market