

Data Mining Project

Business Problem:

A dataset from an airline company has been gathered in association with a direct mail offer to purchase cellular service. The data contains 15 variables identified to represent the airline customers' activity and spending amounts associated with the company. Can this data be used to determine the likelihood of acceptance of future purchase offers? If the customers can be classified as to being receptive to future purchasing offers using a data mining model, can it also be used for predicting potential offer acceptance from future customers.

Two models will be used for classification. The first being the Logic Regression model, the second being the Naïve Bayes method. These models are different in scope and the hope is that they will be able to identify, though the data, what are the strongest trends or traits for identifying potential customers.

First model used: Logistic Regression

The Logistic Regression model is used to classify the customers of the East/West Airlines dataset. The dependent variable that is being used is the field **Phone_sale** in the dataset. This is a dummy variable that indicated whether the customer purchased cellular service (binary; 1 being YES and 0 being NO).

First Run Criteria:

- Data is partitioned using standard partitioning

Standard Data Partition

Data source
Worksheet: data Workbook: EastWestAirlinesNN.xls
Data range: \$A\$1:\$P\$4986
Rows in data: 4985 # Columns in data: 16

Variables
☒ First row contains headers
Variables
Variables in the partitioned data
ID#
Topflight
Balance
Qual_miles
cc1_miles?
cc2_miles?
cc3_miles?
Bonus_miles
Bonus_trans

Partitioning options
☐ Use partition variable
☒ Pick up rows randomly Set seed ☒ 12345
Partitioning percentages when picking up rows randomly
☒ Automatic Training Set 60 %
☐ Specify percentages Validation Set 40 %
☐ Equal #records in training, validation & test set Test Set 0 %

Help OK Cancel

Click this to select / deselect the variable(s) from the variables list.

- Logical Regression is set up using all input variables except ID#
- Output variable is **Phone_sale** with the parameters of 2 classes and *Success* = 1
- *Cutoff Probability* is .5

Prior class probabilities

According to relative occurrences in training data

Class	Prob.	
1	0.136074891	<-- Success Class
0	0.863925109	

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.46679425	0.12598193	0	*
Topflight	0.11956786	0.17009272	0.48208261	1.12700975
Balance	-0.00000258	0.00000088	0.00346399	0.99999744
Qual_miles	0.0000081	0.00007019	0.90810871	1.00000811
cc1_miles?	-0.46405688	0.45709011	0.3099907	0.62872779
cc2_miles?	-0.1205029	0.30246475	0.69033301	0.88647449
cc3_miles?	-0.52707845	0.80508846	0.51267129	0.59032714
Bonus_miles	0.0000027	0.00000297	0.3636466	1.00000274
Bonus_trans	0.0226466	0.00855426	0.00811117	1.02290499
Flight_miles_12mo	0.00002691	0.00005653	0.63406718	1.00002694
Flight_trans_12	0.00016594	0.02493872	0.99469107	1.00016594
Online_12	0.13690205	0.07514681	0.06848618	1.14671576
Email	0.12936768	0.12047189	0.28289387	1.13810849
Club_member	0.17834058	0.1840755	0.33262265	1.19523239
Any_cc_miles_12mo	0.9987213	0.47296482	0.03471918	2.71480823

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)

0.5

(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	2	405
0	4	2580

Error Report			
Class	# Cases	# Errors	% Error
1	407	405	99.51
0	2584	4	0.15
Overall	2991	409	13.67

Validation Data scoring - Summary Report

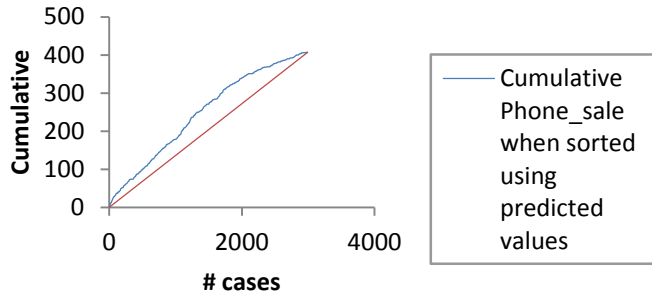
Cut off Prob.Val. for Success (Updatable)

0.5

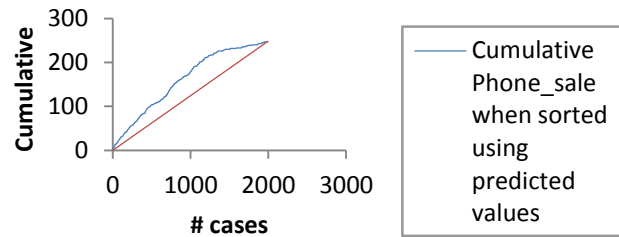
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	3	245
0	2	1744

Error Report			
Class	# Cases	# Errors	% Error
1	248	245	98.79
0	1746	2	0.11
Overall	1994	247	12.39

Lift chart (training dataset)



Lift chart (validation dataset)



First Run Summary:

The first run of the model didn't produce significant results. Since the cut-off value was .5, the number of records that were identified was quite low. From the training dataset lift chart you can see that there isn't a pronounced difference from the baseline sampling based on the cumulative average. However there is a more pronounced difference when the predicted values are applied to the validation dataset. Alternative measures need to be taken to see if there is a better way to fit the data to the model. A second and third pass is needed to see if oversampling and a better subset for classifying potential customers.

Second Run Criteria:

- Data is partitioned with oversampling

Partition with oversampling

Data source
 Worksheet: data Workbook: EastWestAirlinesNN.xls
 Data range: \$A\$1:\$P\$4986
 # Rows in data: 4985 # Columns in data: 16

Variables
☒ First row contains headers
 Variables:
 Variables in the partitioned data: ID#, Topflight, Balance, Qual_miles, cc1_miles?, cc2_miles?, cc3_miles?, Bonus_miles, Bonus_trans

Randomization options
 Set seed ☒ 12345
 Output variable: Choose one of the selected variables
 Phone_sale

Output options
 # Classes: 2 Specify Success class: 1
 % Success in data set: 13.13
 Specify % Success in training set: 50
 Specify % validation data to be taken away as test data:

Help OK Cancel

Click this to select / deselect the variable(s) from the selected variables list.

- No Test Dataset is created
- Logical Regression is set up using all input variables except ID#
- Output variable is **Phone_sale** with the parameters of 2 classes and *Success* = 1
- *Cutoff Probability* is .5

Prior class probabilities

According to relative occurrences in training data

Class	Prob.	
1	0.5	<-- Success Class
0	0.5	

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.67769891	0.1813543	0.00018632	*
Topflight	0.17474347	0.27489999	0.52499676	1.19094062
Balance	-0.0000038	0.00000143	0.00800034	0.99999619
Qual_miles	-0.00000935	0.00012163	0.93875253	0.99999064
cc1_miles?	-0.21904542	0.73659366	0.76617932	0.80328524
cc2_miles?	-0.26907218	0.4573082	0.55627555	0.76408809
cc3_miles?	0.04188573	1.48371804	0.9774785	1.04277527
Bonus_miles	0.00000074	0.00000529	0.88828313	1.00000072
Bonus_trans	0.03251926	0.01364959	0.01719858	1.03305376
Flight_miles_12mo	-0.00011767	0.00012088	0.33034292	0.99988234
Flight_trans_12	0.10380761	0.05616783	0.06457797	1.10938704
Online_12	0.27315056	0.21188705	0.19735189	1.31409812
Email	0.00110528	0.18165757	0.99514538	1.0011059
Club_member	0.15528461	0.29850167	0.60291475	1.16799033
Any_cc_miles_12mo	0.9335748	0.76881903	0.22463425	2.54358578

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)

0.5

(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	218	109
0	148	179

Error Report			
Class	# Cases	# Errors	% Error
1	327	109	33.33
0	327	148	45.26
Overall	654	257	39.30

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)

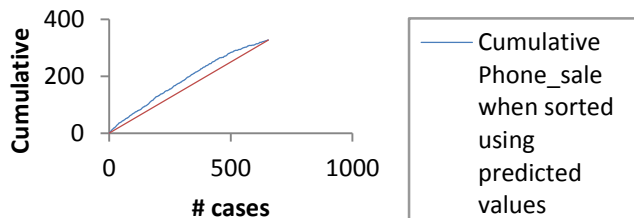
0.5

(Updating the value here will NOT update value in detailed report)

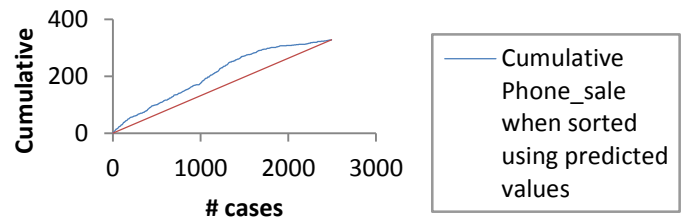
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	238	90
0	1047	1118

Error Report			
Class	# Cases	# Errors	% Error
1	328	90	27.44
0	2165	1047	48.36
Overall	2493	1137	45.61

Lift chart (training dataset)



Lift chart (validation dataset)



Second Run Summary:

Comparing the Lift Charts from the first and second run of the model, it doesn't appear that oversampling has made a significant difference. There are significant differences between the classification matrices of both the training and validation datasets, when oversampling is introduced. The error rate on the training dataset increased from 13.67% to 39.30% when going from Standard Partitioning to Oversampled Partitioning. The same is true for the validation training set, going from 12.39% to 45.61%. Based on the marked increases in the error rates, it appears that oversampling creates more misclassification than standard partitioning. A third pass of the model needs to be run using the *best subset* to see if variable reduction produces better results and a more refined classification model.

Third Run Criteria:

- Data is partitioned using standard partitioning

Standard Data Partition

Data source
Worksheet: data Workbook: EastWestAirlinesNN.xls

Data range: \$A\$1:\$P\$4986

Rows in data: 4985 # Columns in data: 16

Variables
☒ First row contains headers

Variables

Variables in the partitioned data
ID#
Topflight
Balance
Qual_miles
cc1_miles?
cc2_miles?
cc3_miles?
Bonus_miles
Bonus_trans

Partitioning options
☐ Use partition variable
☒ Pick up rows randomly Set seed ☒ 12345

Partitioning percentages when picking up rows randomly
☒ Automatic Training Set 60 %
☐ Specify percentages Validation Set 40 %
☐ Equal #records in training, validation & test set Test Set 0 %

Help OK Cancel

Click this to select / deselect the variable(s) from the variables list.

- Logical Regression is set up using all input variables except ID#
- Output variable is **Phone_sale** with the parameters of 2 classes and *Success* = 1
- *Cutoff Probability* is .5

Logistic Regression - Step 1 of 3

Data source
Worksheet: Data_Partition1 Workbook: EastWestAirlinesNN.xls

Data range: # Columns: 16

Rows
In training set: 2991 In validation set: 1994 In test set:

Variables
☒ First row contains headers

Variables in input data
ID#

Input variables
Topflight
Balance
Qual_miles
cc1_miles?
cc2_miles?
cc3_miles?
Bonus_miles

Weight variable:

Output variable:
Phone_sale

Classes in the output variable
Classes: 2 ☒ Specify "Success" class (necessary): 1

Specify initial cutoff probability value for success: 0.5

Help Cancel < Back Next > Finish

Specifies names of all the worksheets available in the selected workbook.

- *Exhaustive Search* is used to determine **Best SubSet**

Best Subset

☒ Perform best subset selection

Maximum size of best subset: 14 Number of best subsets: 1

Selection procedure
☐ Backward elimination ☐ Forward selection
☒ Exhaustive search ☐ Sequential replacement
☐ Stepwise selection

Stepwise selection options
FIN: FOUT:

Help OK Cancel

The procedure for the subset selection. Choose one among the five procedures available.

Best Subset:

Best subset selection

	#Coeffs	RSS	Cp	Probability	Model (Constant present in all models)														
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Best Subset	2	3000.518555	14.52719212	0.02007082	Constant	Bonus_trans													
Best Subset	3	2988.034424	4.03886509	0.36654642	Constant	Bonus_trans	cc_miles_12mo												
Best Subset	4	2984.598877	2.60216308	0.56737667	Constant	Bonus_trans	Online_12	cc_miles_12mo											
Best Subset	5	2981.705322	1.70763576	0.75381798	Constant	Balance	Bonus_trans	Online_12	cc_miles_12mo										
Best Subset	6	2979.762207	1.76386738	0.85502446	Constant	Topflight	Balance	Bonus_trans	Online_12	cc_miles_12mo									
Best Subset	7	2978.692383	2.69368362	0.88401949	Constant	Topflight	Balance	Bonus_trans	Online_12	Email	cc_miles_12mo								
Best Subset	8	2977.591797	3.59272766	0.91970873	Constant	Topflight	Balance	Bonus_trans	Online_12	Email	Club_member	cc_miles_12mo							
Best Subset	9	2976.885498	4.88619137	0.92923921	Constant	Balance	Bonus_miles	Bonus_trans	cc_miles_12mo	Online_12	Email	Club_member	cc_miles_12mo						
Best Subset	10	2976.065918	6.06633568	0.95527488	Constant	Balance	cc1_miles?	Bonus_miles	Bonus_trans	cc_miles_12mo	Online_12	Email	Club_member	cc_miles_12mo					
Best Subset	11	2975.553223	7.55346823	0.96800685	Constant	Topflight	Balance	cc1_miles?	Bonus_miles	Bonus_trans	cc_miles_12mo	Online_12	Email	Club_member	cc_miles_12mo				
Best Subset	12	2975.174072	9.17419052	0.98162221	Constant	Topflight	Balance	cc1_miles?	cc3_miles?	Bonus_miles	Bonus_trans	cc_miles_12mo	Online_12	Email	Club_member	cc_miles_12mo			
Best Subset	13	2975.013428	11.01349163	0.99327695	Constant	Topflight	Balance	cc1_miles?	cc2_miles?	cc3_miles?	Bonus_miles	Bonus_trans	cc_miles_12mo	Online_12	Email	Club_member	cc_miles_12mo		
Best Subset	14	2975	13.00005913	0.9938494	Constant	Topflight	Balance	Qual_miles	cc1_miles?	cc2_miles?	cc3_miles?	Bonus_miles	Bonus_trans	cc_miles_12mo	Online_12	Email	Club_member	cc_miles_12mo	
Best Subset	15	2975	15.00005913	1	Constant	Topflight	Balance	Qual_miles	cc1_miles?	cc2_miles?	cc3_miles?	Bonus_miles	Bonus_trans	cc_miles_12mo	Online_12	Email	Club_member	cc_miles_12mo	

Subset # 12 was chosen for further modeling.

Logistic Regression - Step 1 of 3

Data source
Worksheet: Data_Partition1 Workbook: EastWestAirlinesNN.xls

Data range: # Columns: 16

Rows
In training set: 2991 In validation set: 1994 In test set:

Variables
☒ First row contains headers

Variables in input data
ID#
Qual_miles
cc2_miles?
Flight_trans_12

Input variables
Topflight
Balance
cc1_miles?
cc3_miles?
Bonus_miles
Bonus_trans
Flight_miles_12mo

Weight variable:

Output variable:
Phone_sale

Classes in the output variable
Classes: 2 ☒ Specify "Success" class (necessary): 1
Specify initial cutoff probability value for success: 0.5

Help Cancel < Back Next > Finish

Specifies names of all the worksheets available in the selected workbook.

Prior class probabilities

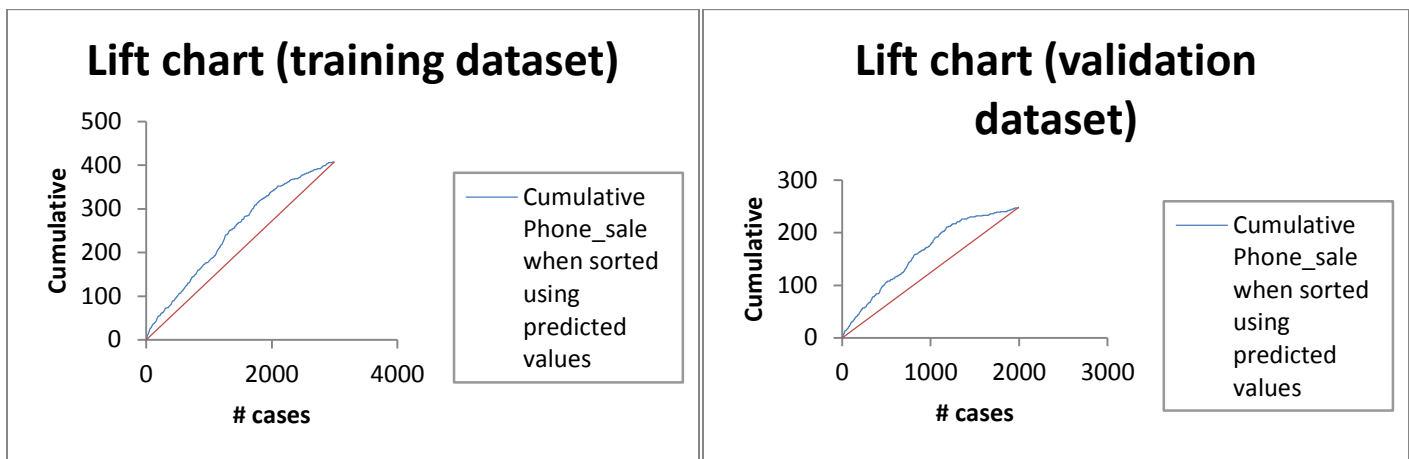
According to relative occurrences in training data

Class	Prob.	
1	0.136074891	<-- Success Class
0	0.863925109	

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.46693468	0.12569109	0	*
Topflight	0.12096554	0.16707435	0.46905211	1.12858605
Balance	-0.00000256	0.00000088	0.00351449	0.99999744
cc1_miles?	-0.35866421	0.37258124	0.33572468	0.69860888
cc3_miles?	-0.49333385	0.79923159	0.53706312	0.61058742
Bonus_miles	0.00000285	0.00000292	0.33033189	1.00000286
Bonus_trans	0.02251884	0.00804499	0.00512433	1.02277434
Flight_miles_12mo	0.00002727	0.00003759	0.46822545	1.0000273
Online_12	0.13761321	0.07501478	0.06658261	1.14753163
Email	0.13119143	0.12035125	0.27568173	1.14018607
Club_member	0.17767861	0.18382519	0.33376259	1.19444144
Any_cc_miles_12mo	0.88324195	0.37493223	0.01848597	2.41872835

Residual df	2979
Residual Dev.	2301.977539
% Success in training data	13.60748913
# Iterations used	9
Multiple R-squared	0.03257232



Third Run Summary:

When the Best Case subset was run, the statistical difference for classification was minimal as reflected in the lift charts. This was expected because the portioning of the data was the same for the first pass, as the third. What was significant was the identification of the relevant variables in the classification in the regression model summary.

Summary of the Logistic Regression Model when classifying the East/West Airlines Dataset

Using the 3 different Regression Models for classifying the data you can determine which variables can give you the best indication for customers that are receptive to a potential sales offer.

Regression Model # 1

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.46679425	0.12598193	0	*
Topflight	0.11956786	0.17009272	0.48208261	1.12700975
Balance	-0.00000258	0.00000088	0.00346399	0.99999744
Qual_miles	0.0000081	0.00007019	0.90810871	1.00000811
cc1_miles?	-0.46405688	0.45709011	0.3099907	0.62872779
cc2_miles?	-0.1205029	0.30246475	0.69033301	0.88647449
cc3_miles?	-0.52707845	0.80508846	0.51267129	0.59032714
Bonus_miles	0.0000027	0.00000297	0.3636466	1.00000274
Bonus_trans	0.0226466	0.00855426	0.00811117	1.02290499
Flight_miles_12mo	0.00002691	0.00005653	0.63406718	1.00002694
Flight_trans_12	0.00016594	0.02493872	0.99469107	1.00016594
Online_12	0.13690205	0.07514681	0.06848618	1.14671576
Email	0.12936768	0.12047189	0.28289387	1.13810849
Club_member	0.17834058	0.1840755	0.33262265	1.19523239
Any_cc_miles_12mo	0.9987213	0.47296482	0.03471918	2.71480823

Regression Model # 2

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.67769891	0.1813543	0.00018632	*
Topflight	0.17474347	0.27489999	0.52499676	1.19094062
Balance	-0.0000038	0.00000143	0.00800034	0.99999619
Qual_miles	-0.00000935	0.00012163	0.93875253	0.99999064
cc1_miles?	-0.21904542	0.73659366	0.76617932	0.80328524
cc2_miles?	-0.26907218	0.4573082	0.55627555	0.76408809
cc3_miles?	0.04188573	1.48371804	0.9774785	1.04277527
Bonus_miles	0.00000074	0.00000529	0.88828313	1.00000072
Bonus_trans	0.03251926	0.01364959	0.01719858	1.03305376
Flight_miles_12mo	-0.00011767	0.00012088	0.33034292	0.99988234
Flight_trans_12	0.10380761	0.05616783	0.06457797	1.10938704
Online_12	0.27315056	0.21188705	0.19735189	1.31409812
Email	0.00110528	0.18165757	0.99514538	1.0011059
Club_member	0.15528461	0.29850167	0.60291475	1.16799033
Any_cc_miles_12mo	0.9335748	0.76881903	0.22463425	2.54358578

Regression Model # 3

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.46693468	0.12569109	0	*
Topflight	0.12096554	0.16707435	0.46905211	1.12858605
Balance	-0.00000256	0.00000088	0.00351449	0.99999744
cc1_miles?	-0.35866421	0.37258124	0.33572468	0.69860888
cc3_miles?	-0.49333385	0.79923159	0.53706312	0.61058742
Bonus_miles	0.00000285	0.00000292	0.33033189	1.00000286
Bonus_trans	0.02251884	0.00804499	0.00512433	1.02277434
Flight_miles_12mo	0.00002727	0.00003759	0.46822545	1.0000273
Online_12	0.13761321	0.07501478	0.06658261	1.14753163
Email	0.13119143	0.12035125	0.27568173	1.14018607
Club_member	0.17767861	0.18382519	0.33376259	1.19444144
Any_cc_miles_12mo	0.88324195	0.37493223	0.01848597	2.41872835

The odds that are calculated for the identified variables are very small. These **are not** a strong indication that any of the variables will necessarily define a customer's potential. However they do give an indication of which variable are present or have some significance in our classification. The strongest indicator is the variable "**Any_cc_miles_12mo**", with the weakest being "**cc3_miles**".

Second model used: Naïve Bayes

The goal of using the Naïve Bayes model is the same as using the Logistic Regression Model; identify which are the strongest variables for identifying purchasing probability. The results of the two model will be compared to see which one may give the better indication.

First Run Criteria:

- Data is partitioned using standard partitioning (same as the when done in the Logistic Regression Model)
- Naïve Bayes was run with 12 of the input variables and **Phone_sale** as the output variable
- ID#, Balance, and Bonus_miles can't be run in this model because the dataset is too large and there are more than 1,000 different values.

Naive Bayes - Step 1 of 3

Data source
Worksheet: Data_Partition1 Workbook: EastWestAirlinesNN.xls

Data range: # Columns: 16

Rows
In training set: 2991 In validation set: 1994 In test set:

Variables
☒ First row contains headers

Variables in input data
ID#
Balance
Bonus_miles

Input variables
Topflight
Qual_miles
cc1_miles?
cc2_miles?
cc3_miles?
Bonus_trans
Flight_miles_12mo

Weight variable:

Output variable:
Phone_sale

Classes in the output variable
Classes: 2 ☒ Specify "Success" class (for Lift Chart): 1
Specify initial cutoff probability value for success: 0.5

Help Cancel < Back Next > Finish

Specifies names of all the worksheets available in the selected workbook.

- The default was taken for the prior class probabilities

Naive Bayes - Step 2 of 3

Prior class probabilities
☒ According to relative occurrences in training data
☐ Use equal prior probabilities
☐ User specified prior probabilities

Help Cancel < Back Next > Finish

This option will assign probabilities in proportion to occurrences of each class in the training data.

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)
---	-----	--

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	75	332
0	31	2553

Error Report			
Class	# Cases	# Errors	% Error
1	407	332	81.57
0	2584	31	1.20
Overall	2991	363	12.14

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)

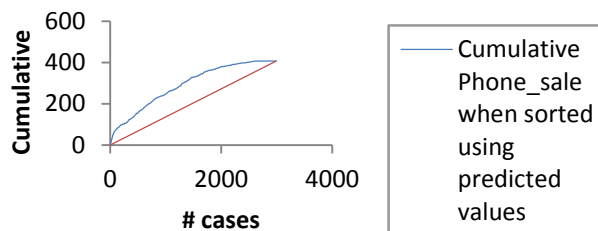
0.5

(Updating the value here will NOT update value in detailed report)

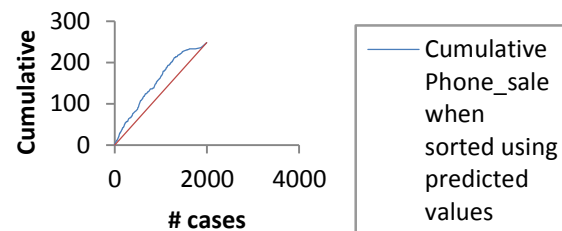
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	12	236
0	42	1704

Error Report			
Class	# Cases	# Errors	% Error
1	248	236	95.16
0	1746	42	2.41
Overall	1994	278	13.94

Lift chart (training dataset)



Lift chart (validation dataset)

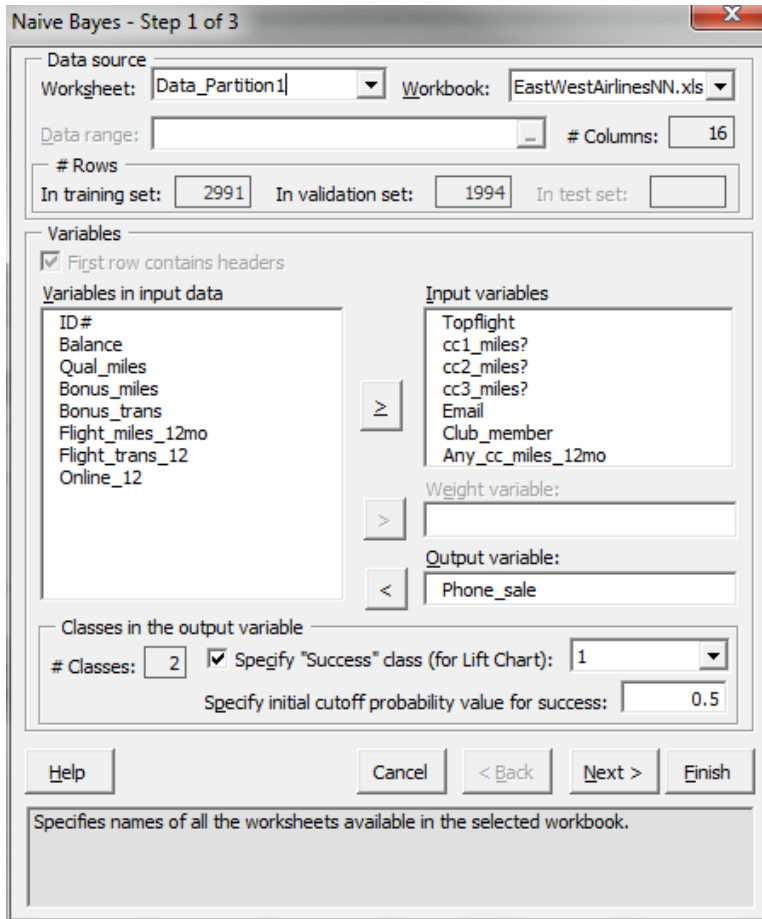


First Run Summary:

When looking at the Conditional Probabilities chart, too many variables were used to gather any meaningful significance. Although the Conditional Matrices reported a lower error rate for both the test and validation partition, in comparison to the Logistic Regression, that really didn't provide any significant information. The lift chart for the training data set indicated a better fit on the data than the validation dataset.

Second Run Criteria:

- Data is partitioned using standard partitioning
- Only the categorical variables are used to create the model



The dialog box is titled "Naive Bayes - Step 1 of 3". It contains the following fields and options:

- Data source:** Worksheet: Data_Partition1, Workbook: EastWestAirlinesNN.xls
- Data range:** # Columns: 16
- # Rows:** In training set: 2991, In validation set: 1994, In test set: (empty)
- Variables:**
 - ☒ First row contains headers
 - Variables in input data:** ID#, Balance, Qual_miles, Bonus_miles, Bonus_trans, Flight_miles_12mo, Flight_trans_12, Online_12
 - Input variables:** Topflight, cc1_miles?, cc2_miles?, cc3_miles?, Email, Club_member, Any_cc_miles_12mo
 - Weight variable:** (empty)
 - Output variable:** Phone_sale
- Classes in the output variable:**
 - # Classes: 2
 - ☒ Specify "Success" class (for Lift Chart): 1
 - Specify initial cutoff probability value for success: 0.5

Buttons: Help, Cancel, < Back, Next >, Finish

Specifies names of all the worksheets available in the selected workbook.

- All other defaults are taken as in the first run

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5
---	-----

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	0	407
0	0	2584

Error Report			
Class	# Cases	# Errors	% Error
1	407	407	100.00
0	2584	0	0.00
Overall	2991	407	13.61

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5
---	-----

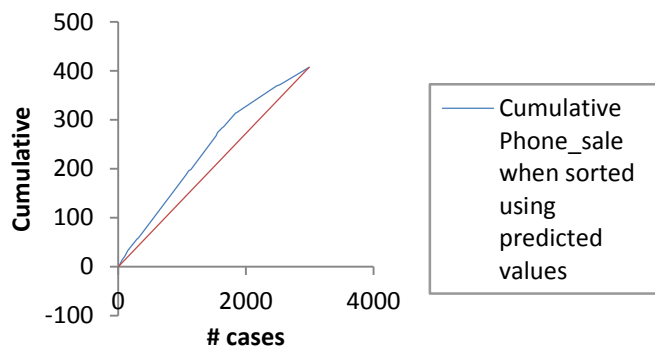
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	0	248
0	0	1746

Error Report			
Class	# Cases	# Errors	% Error
1	248	248	100.00
0	1746	0	0.00
Overall	1994	248	12.44

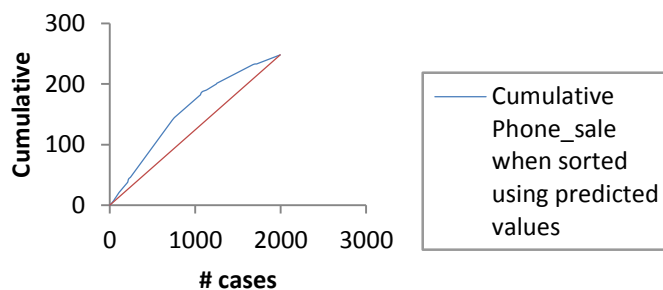
Conditional probabilities

Input Variables	Classes-->			
	1		0	
	Value	Prob	Value	Prob
Topflight	0	0.808353808	0	0.830882353
	1	0.191646192	1	0.169117647
cc1_miles?	0	0.343980344	0	0.510835913
	1	0.656019656	1	0.489164087
cc2_miles?	0	0.943488943	0	0.957043344
	1	0.056511057	1	0.042956656
cc3_miles?	0	0.995085995	0	0.995743034
	1	0.004914005	1	0.004256966
Email	0	0.312039312	0	0.361068111
	1	0.687960688	1	0.638931889
Club_member	0	0.891891892	0	0.910216718
	1	0.108108108	1	0.089783282
Any_cc_miles_12mo	0	0.319410319	0	0.495743034
	1	0.680589681	1	0.504256966

Lift chart (training dataset)



Lift chart (validation dataset)



Second Run Summary:

This model was generated to see if there was any way to more accurately classify the customers based solely on the conditional variables. The validation error rates for both the training and validation datasets were relatively the same as in the first run of the model. The lift charts show that the data fits a little bit better to the validation dataset than the training dataset, but there is no major difference. The Conditional Probabilities table does show some difference among the conditional variables however, it is slight. From this table, the variables with the strongest indicators are: ***cc1_miles***, ***Email***, and ***Any_cc_miles_12***. These three of variables showed the highest probability of predicting to be true in reference to the ***Phone_sale*** variable.

Summary of the Naïve Bayes model when classifying the East/West Airlines Dataset

Running the model two different ways did produce varied data. However, this was not an “apples to apples” comparison because the parameters were different between the two. When the model was generated using just the conditional variables, it gave a better indication of the stronger variables that could influence the outcome.

Project Summary

Using the Naïve Bayes and the Logistic regression models to classify the East/West Airlines dataset, produced different results. Several combinations of parameters and variables were run with each model. The goal was to find which variables could give the strongest indication of customers to be targeted for future sales offers and possibly indicate a hire response rate.

Neither model indicated any extreme indicators. Of all the models run, the most relevant was the Naïve Bayes model using just the constant variables and eliminating the numeric data. This was a better indication of which variables might generate a higher probability. In the Logistic regression model, it was hoped that the greater the value of the numeric variables, combined with more positive indication of the conditional variables would reveal a true indication of the customers' acceptance of a sales offer. However, the model did not indicate such results.