Marketing Data Analytics

Master's in business Analytics and Project Management, University of Connecticut

OPIM-5604-Predictive Modeling

Prof. Sulin Ba

December 6th, 2021

This report developed by:

Amit Anand Kumar, Nitish Ledala, Sathwik Pendyala, Shylesh Pala, Yogesh Thupda

TABLE OF CONTENTS

EXECUTIVE SUMMARY:

The primary objective of the modeling in the Marketing dataset considered is to "Predict who will respond to an offer/service which is being advertised in the campaign".

By predicting the customer's response and categorizing them into customer groups before the marketing campaign, we will be able to significantly boost the marketing campaign's efficiency by increasing the conversion rate. This way, we would be able to target the potential customers within the campaign budget.

Based on the results of our modeling efforts to predict whether a customer will accept or reject an offer, we have concluded that the Group B significant variables models, specifically the nominal logistic model, are some of the most powerful models for determining whether a customer will accept or reject the next campaign offer.

The ensemble model came in second, and the neural network model came in third. These were the most popular models.

Accuracy is maximized with the least amount of overfitting. After evaluating the models' results, we interpret the models' findings.

We recommend the following to our business associates:

- Consumers who accepted more campaign offers in past campaigns are more inclined to accept the next campaign's offer.
- Customers who have recently purchased something are more prone to accept the following advertising offer.
- In addition, clients who invest more in wines and meat products are more likely to accept the next campaign offer
- Customers with an annual income of more than $75,000 are more likely to accept the offer, with the single status being the most crucial factor.

PROBLEM STATEMENT:

Several marketing campaigns are being conducted regularly to attract customers to buy a product. Some turn out to be successful, and some will not be as successful as the company expected them to be. A good marketing campaign can be the game-changer in making a product successful. In Many cases, the campaigns done are not reaching the consumers who would buy the potential Product. The products used in the marketing campaign are Wines, Fruits, Meat, Fish, Sweets, and Gold Products. The campaign manager for the company wants to know where most of the funds are diverted and doesn't want any of the resources to be squandered when targeting the wrong people. Hence, we propose to analyze the marketing data to gain valuable insights for finding the target customers with more potential to buy.

METHODOLOGY

We have undertaken the five-step SEMMA process (Sample, Explore, Modify, Model, and Assess) as the methodology for this project.

SAMPLE

We have chosen the "Marketing Analytics" dataset from Kaggle, with 28 columns and 2240 rows. The data dictionary can be found in Appendix A.

EXPLORE

With the help of data visualization, we tried to understand the data and identified if there were any correlations between the different predictor variables. We also tried to see if there were any abnormalities within our data.

We started with exploring the data by building several graphs using graph builder to gain better insights into our data. Some of the insights of the visualizations are as follows:

Income vs Type of Purchases:

People are split into groups based on their earnings. They are classified into five different groups: low income, lower middle income, middle income, middle income, upper middle income, and high income. The income ranges for the categories are as follows: '$1,730 - $32,011', '$32,011 - $44,529', '$44,529 - $58,482', '$58,482 - $71,819' & '$71,819 - $666,666. From the graph, we can infer that the lower and lower-middle-income people visit the websites the most but have the Least Amount of web purchases. The high and upper middle income less frequently visit the website's but do the highest number of purchases. The people in the middle-income purchase the highest with deals and offers, and the high-income purchase the least. The high and upper-middle-income groups make the most catalog and in-store purchases. This information provides relevant details about the expenditure the customers within different salary brackets make on different platforms.

Income & Marital Status vs Response

We also derived the Purchase Behavior of People based on their Marital Status and Salary brackets. From this visualization, we can conclude that people who are single or Divorced from the High-income bracket are more likely to accept the marketing campaign offer.

Total Amount spent Vs. Customers duration associated in months:

From this visualization, we have classified our customers into four segments.

1. Premium Loyal Customers
2. Inherently Loyal Customers
3. High-Valued Customers
4. Low-Valued Customers

Premium Customers are associated with the company for more than 100 months and spend the most among our customers.

Inherently Loyal Customers are those customers associated with the company for more than 100 months; however, they do not spend much on the customers.

High-valued Customers are not associated with the company for a longer duration, but the amount spent is high.

Low-Valued Customers are the ones who are not associated with the company for a long time, and they have not spent money.

Campaign Acceptance vs Response:

From the above graph, we can infer that if a customer accepts an offer in Campaign 1, they are subsequently not accepting the offer in Campaign 2 & 3. However, they accept the offer in Campaign 4 & Campaign 5.

A similar pattern can be observed even when a customer rejects an offer in campaign one they reject in the subsequent campaigns, i.e., in Campaign 2&3. Still, the conversion rate is high in Campaign 5, irrespective of the acceptance of the offer in campaign 4.

We can conclude that Campaign 5 is outperforming all the other campaigns.

Geographical Segmentation (Locations Vs. Mean Acceptance of all the 5 Campaigns):

In campaign 1, "Spain," "Canada" performed well.

In campaign 3, "Germany," "India," "Spain," "United States" performed exceptionally well.

In campaign 4, "Spain," "India," "Germany," "Canada" performed quite well.

In campaign 5, "Australia," "Canada," "Spain" performed quite well.

Campaign 2 was not doing very well in any of the countries.

Canada, Germany, and Spain responded well to the Campaigns regarding the overall average performance.

We will be targeting the customers in the Premium - High Income Level.

MODIFY:

We found 24 missing values in the 'income' column. Since the size of the Dataset is limited, we performed auto imputation for these values. These were replaced using Automated data imputation. Column mean or Mode imputation was found to be the best fit.

DATATYPE

To Understand the behavior of the customer more accurately, the following column datatypes are changed accordingly

- Changed Education from Nominal to Ordinal variables. This is done to categorize and predict the values based on the education levels of the customer.
- Changed Income to Ordinal, we have divided all the people into five income groups.

MISSING VALUES

We found 24 missing values in the 'income' column. Since the size of the dataset is limited, we performed auto imputation for these values. These were replaced using Automated data imputation. Column mean or Mode imputation was found to be best fit.

OUTLIERS

Outliers were observed in the following columns

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds
- NumDealsPurchases
- NumWebPurchases
- NumCatalogPurchases
- NumWebVisitsMonth

Again, since the Dataset is limited, We decided to impute them by using the continuous fit feature offered by JMP. The continuous fit feature is very robust. We could have applied the conventional method of imputing these outliers using the median values, but the continuous fit feature offered by JMP is more robust. Hence, we decided to go ahead with this

The columns MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds were handled using 'Fitted SHASH Distribution' and the columns NumWebPurchases, NumWebVisitsMonth were handled using Fitted Normal 3 Mixture Distribution.

MODEL

We used a systematic approach to investigate different modeling strategies. Essentially, the modeling strategy was to throw everything at the wall and see what sticks. We ran all the other models that were explained to us in the class, but at the same time, we also ran SVM to check if it makes any difference in the accuracy. The different models we created are Logistic Regression, Decision Tree, Bootstrap Forest, Neural Network, SVM, and K Nearest Neighbors.

Before we created models, we split 60% of our data into the Training set, 20% into the Validation set, and 20% into the Test set. Our Target variable is 'Response' which is binary. The target variable 'Response' tells us if the customer accepts the marketing campaign offer.

We utilized each of these modeling strategies with each predictor variable in the Dataset to anticipate the customers' responses. Some of these predicted models were excellent, while others fell short. Many of the models produced comparable outcomes. We also encountered models where there was an overfit in the training Data. The model exploration data are in Appendix C.

Logistic Regression

We created our first model using Logistic Regression. We discovered that the initial run of the logistic regression model revealed that roughly half of the variables were inconsequential in predicting the target variable. We kept the cutoff of our P Values as 0.15 for the variables which are contributing for our model. The effect summary showed that only 12 columns are contributing to the target variable.

We can infer from the fit details table, The misclassification rate over here is 10.79%. We can tell from the misclassification rate that the accuracy of our model is almost 90%.

REDUCED LOGISTIC REGRESSION MODEL

Afterwards, we ran the model with the 12 variables which contributed maximum towards the target variable.

The revised Logistic Regression model produced significantly better results. So based on this derivation, we decide to run an optimized model for all other techniques considering only significant variables by logistic regression method.

The variables that were the most significant in predicting the target variable are:

CompleteAcceptedCmp, Recency Mnt Wines, MntMeatProducts, Customer Relationship with company, Marital Status, Education status, Number of Catalog Purchases were the variables in order of relevance.

Based on this we decided to run all the models with only these 12 variables. The ranking of the best models is given in Appendix C.

ASSESS

We divided the models into two groups to rank their performance. Models built including all variables were placed in Group A, while models with just significant variables were placed in Group B. The correctness of the performance was then ranked in ascending order by partition. Overfitting between the training and validation partitions was also investigated. Overfit models were given the lowest overall rating. As a result, we were able to properly examine the performance of our models and interpret the data to make conclusions.

RESULTS

Based on the results of our modeling efforts to predict whether a customer will accept or reject an offer, we have concluded that the Group B significant variables models, specifically the nominal logistic model, are some of the most powerful models for determining whether a customer will accept or reject the next campaign offer.

As stated earlier, we built the models using two sets of data, one with all the predictor variables and the other with only the significant values, as described in the paper's modeling section. This successfully divided the models into two groups for analysis.

All predictor variables in Group A were included, regardless of their significant value.

Our Training, validation, and Test groups had 1,344,448 and 448 participants, respectively.

The difference in misclassification rates between the Training and Validation groups was used to account for overfitting. Regarding deriving the model outcome from Test data considering that training data is prone to overfitting, we found that the Logistic Regression and Neural Model NtanH(3)Linear2 has the same approximate accuracy with a misclassification rate of 9.60 %. The next best models are Partition, SVM, Bootstrap Forest, K Nearest Neighbors in this order. The Bootstrap Forest has the maximum accuracy in the training data. Still, we are not going with this model because it is highly overfitting in the training data, with the accuracy dropping from 95% to 87% and 88% invalidation and test data.

Group B consists of models with only significant values. Here the Logistic regression has the lowest misclassification rate of 9.60 %. The next best models here are SVM, Neural, Bootstrap Forest, Partition, K Nearest Neighbors in this order.

CONCLUSIONS

Finally, based on the Model results, we are going with the Logistic regression after considering factors including optimum misclassification rate, Least overfitting, and better interpretation and understanding of the model. Even though Neural Network had similar accuracy to Logistic regression in Group A, we are not going with it because Neural network is a kind of backbox that is difficult to interpret.

Based on the results of our modeling efforts to predict whether a customer will accept or reject an offer, we have concluded that the Group B significant variables models, specifically the nominal logistic model, are some of the most powerful models for determining whether a customer will accept or reject the next campaign offer. Important Takeaways from the Model are:

- Consumers who accepted more campaign offers in past campaigns are more inclined to accept the next campaign's offer.
- Customers who have recently purchased a product are more inclined to accept the following advertising offer.
- In addition, clients who invest more in wines and meat products are more likely to accept the next campaign offer.
- Customers with an annual income of more than $75,000 are more likely to accept the offer, with the single status being the most important factor.

RECOMMENDATIONS:

To make the next campaign a success, these are the following recommendations we would provide to our business associates:

- The company should provide additional promotions and deals on wine and meat goods. Given that high-income individuals are more likely to buy a product in a shop and middle-income people are more likely to buy a product online with a discount.
- It is suggested that middle and above-middle-income people get digital adverts in the next marketing campaign to enhance sales.

- Customers in the above medium income category who are unmarried are more likely to take up the offer than others, so firms should target them to enhance sales and reduce marketing campaign costs.
- Offer Lucrative offers to customers who have not accepted offers in previous campaigns. By doing this, the customer of this category will be more likely than ever to accept the upcoming campaigns.

REFERENCES

https://www.kaggle.com/jackdaoud/marketing-data

Shmueli, Galit, et al. Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro. John Wiley & Sons, Inc., 2017.

APPENDIX

APPENDIX A : DATA DICTIONARY

Description of all the labels in the dataset.

ID – Unique ID of all the customers in the database

Year_Birth - Birthdate of each customer

Education - Qualification of each customer

Marital_Status - explains the marital status of each customer

Income - Yearly Household Income of customer

Kidhome - Number of kids in the Customer’s Household

Teenhome - Number of teenagers in the Customer’s Household

Dt_Customer - date of customer’s enrolment with the company

Recency - Number of days since the last purchase is made.

MntWines - Amount Spend on Wine Products since last two years

MntFruits - Amount Spent on Fruits since last two years

MntMeatProducts - Amount Spend on Meat Products since last two years

MntFishProducts - Amount Spend on Fish Products since last two years

MntSweetProducts - Amount Spent on Sweet Products since last two years

MntGoldProds - Amount Spend on Gold since last two years

NumDealsPurchases - Number of purchases made with discounts

NumWebPurchases - Purchases happened over company website

NumCatalogPurchases - Number of purchases made using catalogs

NumStorePurchases - Number of purchases made directly in store

NumWebVisitsMonth - Number of web visits to the company's website in the last month

AcceptedCmp1 - If Customer accepted offer in the 1st campaign

AcceptedCmp2 - If Customer accepted offer in the 2nd campaign

AcceptedCmp3 - If Customer accepted offer in the 3rd campaign

AcceptedCmp4 - If Customer accepted offer in the 4th campaign

AcceptedCmp5 - If Customer accepted offer in the 5th campaign

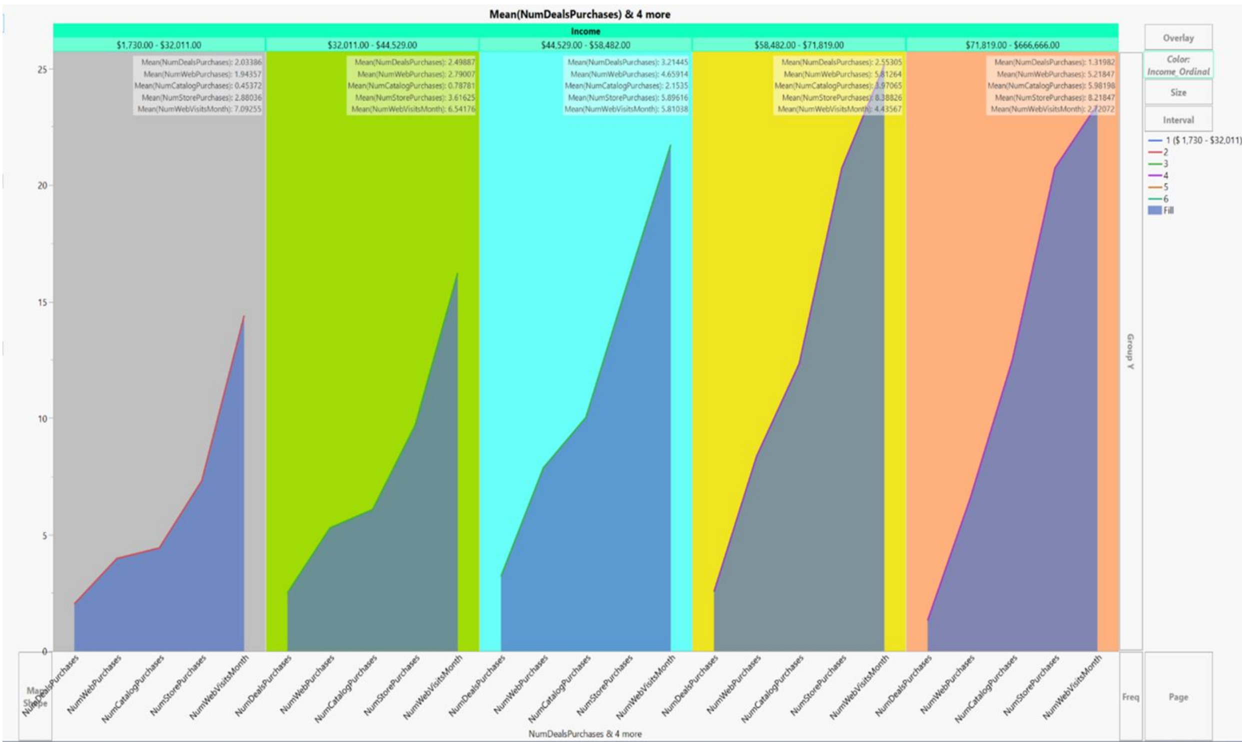
Response - **Response is the Target Variable which explains if the Customer accepted the offer in the Last Campaign**

Complain - If the Customer raised any complaint in the last two years.

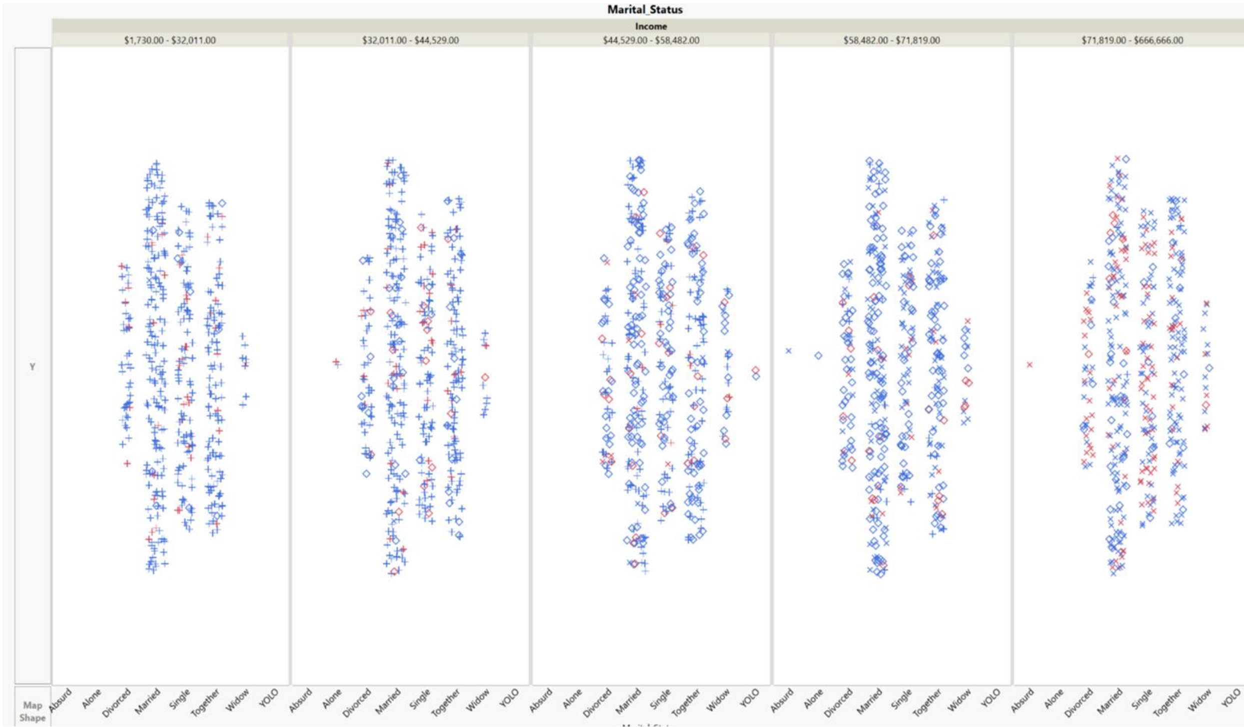
Country - Country from which the Customer belongs.

APPENDIX B: EXPLORE

Income vs Type Of Purchases:

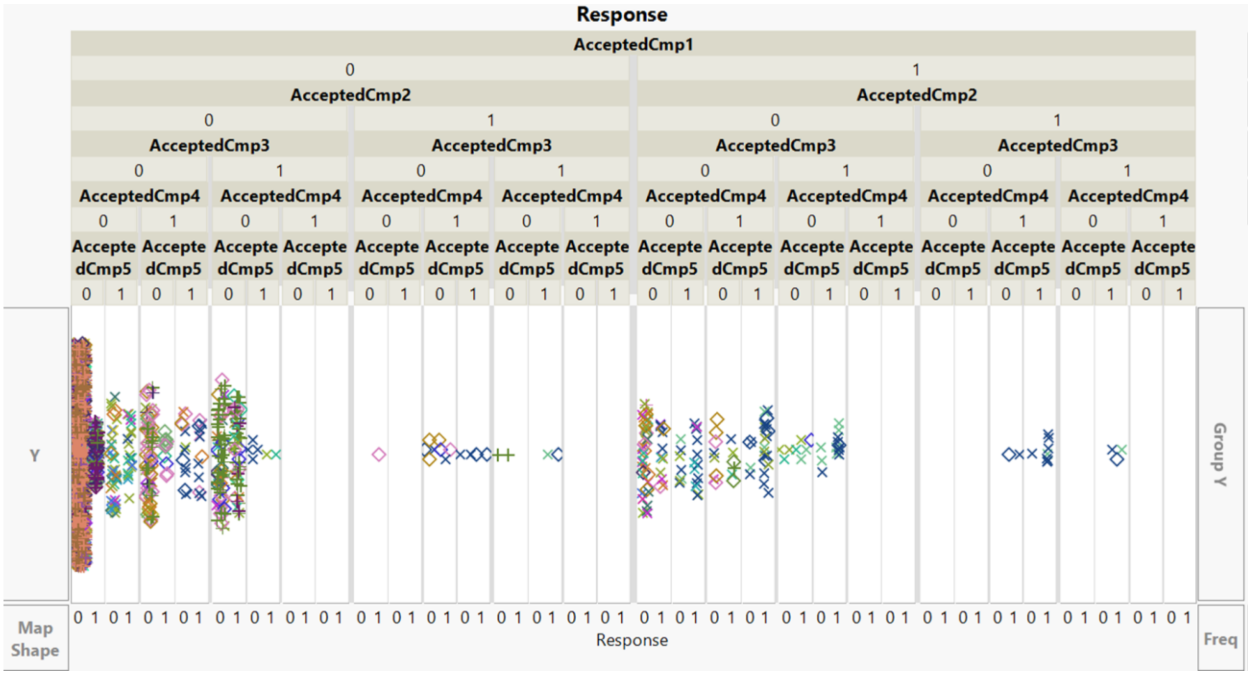


Income & Marital Status vs Response



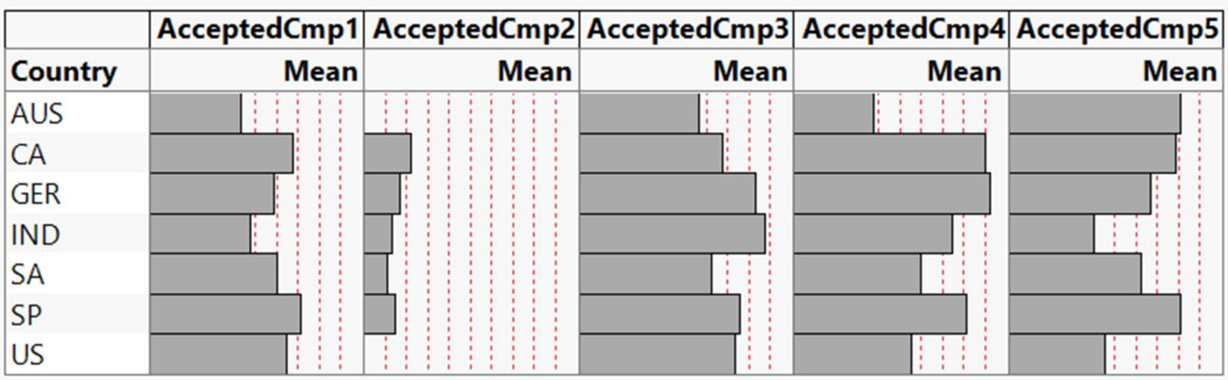
Total Amount spent Vs. Customers duration associated in months:

Campaign Acceptance vs Response:



<u>Geographical Segmentation (Locations Vs. Mean Acceptance of all the 5 Campaigns):</u>

	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5
Country	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**
AUS	0.04375	0	0.05625	0.0375	0.08125
CA	0.0671641791	0.0223880597	0.0671641791	0.0895522388	0.078358209
GER	0.0583333333	0.0166666667	0.0833333333	0.0916666667	0.0666666667
IND	0.0472972973	0.0135135135	0.0878378378	0.0743243243	0.0405405405
SA	0.059347181	0.0118694362	0.0623145401	0.059347181	0.0623145401
SP	0.0712328767	0.0146118721	0.0757990868	0.0812785388	0.0812785388
US	0.0642201835	0	0.0733944954	0.0550458716	0.0458715596



APPENDIX C: MODELING

LOGISTIC REGRESSION

Nominal Logistic Fit for Response

Effect Summary

Source	LogWorth	PValue
CompleteAcceptedCmp	35.189	0.00000
Recency	20.178	0.00000
MntWines(N)	13.290	0.00000
Customers relationship with the company in Months	7.873	0.00000
Marital_Status	6.255	0.00000
Education_ordinal	5.727	0.00000
MntMeatProducts(N)	5.639	0.00000
NumCatalogPurchases(N)	4.954	0.00001
NumWebVisitsMonth(N)	3.383	0.00041
MntGoldProds(N)	1.517	0.03039
NumWebPurchases(N)	1.134	0.07344
Imputed_Income	0.886	0.13011
Dependents	0.509	0.31002
Year_Birth	0.388	0.40936
NumDealsPurchases(N)	0.341	0.45562
MntFishProducts(N)	0.300	0.50104
MntFruits(N)	0.188	0.64882
Country	0.174	0.66997
Complain	0.076	0.84038
MntSweetProducts(N)	0.018	0.95946

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4228	0.0795	0.3951	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5268	0.1156	0.4906	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2443	0.4058	0.2382	$\sum$ -Log(ρ[j])/n
RASE	0.2737	0.3011	0.2708	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1481	0.1658	0.1459	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.1079	0.1183	0.0960	$\sum$ (ρ[j]≠ρMax)/n
N	1344	448	448	n

REDUCED LOGISTIC REGRESSION

Nominal Logistic Fit for Response

Effect Summary

Source	LogWorth	PValue
CompleteAcceptedCmp	39.211	0.00000
Recency	20.855	0.00000
MntWines(N)	14.754	0.00000
MntMeatProducts(N)	8.745	0.00000
Customers relationship with the company in Months	8.268	0.00000
Marital_Status	6.084	0.00000
Education_ordinal	5.688	0.00000
NumCatalogPurchases(N)	4.504	0.00003
NumWebVisitsMonth(N)	2.705	0.00197
MntGoldProds(N)	1.430	0.03713
NumWebPurchases(N)	1.098	0.07983
Imputed_Income	0.847	0.14234

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4124	0.0536	0.3482	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5160	0.0788	0.4400	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2487	0.4172	0.2567	$\sum -\mathrm{Log}(\rho[j])/n$
RASE	0.2763	0.3036	0.2743	$\sqrt{\ } \sum(y[j]-\rho[j])^2/n$
Mean Abs Dev	0.1508	0.1687	0.1481	$\sum \lvert y[j]-\rho[j] \rvert/n$
Misclassification Rate	0.1086	0.1228	0.0960	$\sum (\rho[j]\neq\rho\mathrm{Max})/n$
N	1344	448	448	n

PARTITION

Leaf Report

Response Prob

Leaf Label	0	1
CompleteAcceptedCmp>=1&Recency<20	0.2875	0.7125
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp>=3	0.2075	0.7925
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months>=104	0.5616	0.4384
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months<104	0.8863	0.1137
CompleteAcceptedCmp<1&Recency<19	0.8013	0.1987
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months>=111	0.7835	0.2165
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months<111	0.9612	0.0388

Response Counts

Leaf Label	0	1
CompleteAcceptedCmp>=1&Recency<20	17	44
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp>=3	5	22
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months>=104	43	34
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months<104	102	13
CompleteAcceptedCmp<1&Recency<19	157	39
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months>=111	72	20
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months<111	746	30

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2696	0.2411	0.2535	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3573	0.3268	0.3320	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3091	0.3346	0.2940	∑ -Log(ρ[j])/n
RASE	0.3020	0.3153	0.2940	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.1829	0.1933	0.1780	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1176	0.1339	0.1094	∑ (ρ[j]≠ρMax)/n
N	1344	448	448	n

Confusion Matrix

Training

Actual Response	Predicted Count 0	Predicted Count 1
0	1120	22
1	136	66

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.981	0.019
1	0.673	0.327

Validation

Actual Response	Predicted Count 0	Predicted Count 1
0	369	7
1	53	19

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.981	0.019
1	0.736	0.264

Test

Actual Response	Predicted Count 0	Predicted Count 1
0	381	7
1	42	18

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.982	0.018
1	0.700	0.300

REDUCED PARTITION

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2696	0.2411	0.2535	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3573	0.3268	0.3320	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3091	0.3346	0.2940	∑ -Log(ρ[j])/n
RASE	0.3020	0.3153	0.2940	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1829	0.1933	0.1780	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1176	0.1339	0.1094	∑ (ρ[j]≠ρMax)/n
N	1344	448	448	n

Column Contributions

Term	Number of Splits	G^2	Portion
CompleteAcceptedCmp	2	181.2916	0.5910
Recency	2	66.4803104	0.2167
Customers relationship with the company in Months	2	59.0075757	0.1923
Education_ordinal	0	0	0.0000
Marital_Status	0	0	0.0000
Imputed_Income	0	0	0.0000
MntWines(N)	0	0	0.0000
MntMeatProducts(N)	0	0	0.0000
MntGoldProds(N)	0	0	0.0000
NumWebPurchases(N)	0	0	0.0000
NumCatalogPurchases(N)	0	0	0.0000
NumWebVisitsMonth(N)	0	0	0.0000

Leaf Report

Response Prob

Leaf Label	0	1
CompleteAcceptedCmp>=1&Recency<20	0.2875	0.7125
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp>=3	0.2075	0.7925
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months>=104	0.5616	0.4384
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months<104	0.8863	0.1137
CompleteAcceptedCmp<1&Recency<19	0.8013	0.1987
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months>=111	0.7835	0.2165
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months<111	0.9612	0.0388

Response Counts

Leaf Label	0	1
CompleteAcceptedCmp>=1&Recency<20	17	44
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp>=3	5	22
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months>=104	43	34
CompleteAcceptedCmp>=1&Recency>=20&CompleteAcceptedCmp<3&Customers relationship with the company in Months<104	102	13
CompleteAcceptedCmp<1&Recency<19	157	39
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months>=111	72	20
CompleteAcceptedCmp<1&Recency>=19&Customers relationship with the company in Months<111	746	30

Confusion Matrix

Training

Actual Response	Predicted Count 0	Predicted Count 1
0	1120	22
1	136	66

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.981	0.019
1	0.673	0.327

Validation

Actual Response	Predicted Count 0	Predicted Count 1
0	369	7
1	53	19

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.981	0.019
1	0.736	0.264

Test

Actual Response	Predicted Count 0	Predicted Count 1
0	381	7
1	42	18

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.982	0.018
1	0.700	0.300

NEURAL NETWORK

Model NTanH(3)NLinear2(2)

Training

Response

Measures	Value
Generalized RSquare	0.4901265
Entropy RSquare	0.3879225
RASE	0.2806873
Mean Abs Dev	0.1622937
Misclassification Rate	0.108631
-LogLikelihood	348.15847
Sum Freq	1344

Validation

Response

Measures	Value
Generalized RSquare	0.4505108
Entropy RSquare	0.3475965
RASE	0.2932212
Mean Abs Dev	0.1754325
Misclassification Rate	0.1138393
-LogLikelihood	128.85092
Sum Freq	448

Test

Response

Measures	Value
Generalized RSquare	0.471411
Entropy RSquare	0.3770868
RASE	0.2755276
Mean Abs Dev	0.1563268
Misclassification Rate	0.0959821
-LogLikelihood	109.89225
Sum Freq	448

Confusion Matrix

Actual Response	Predicted Count 0	1
0	1100	42
1	104	98

Confusion Rates

Actual Response	Predicted Rate 0	1
0	0.963	0.037
1	0.515	0.485

Confusion Matrix

Actual Response	Predicted Count 0	1
0	364	12
1	39	33

Confusion Rates

Actual Response	Predicted Rate 0	1
0	0.968	0.032
1	0.542	0.458

Confusion Matrix

Actual Response	Predicted Count 0	1
0	378	10
1	33	27

Confusion Rates

Actual Response	Predicted Rate 0	1
0	0.974	0.026
1	0.550	0.450

REDUCED NEURAL

Model NTanH(3)NLinear2(2)

Training

Response

Measures	Value
Generalized RSquare	0.4567975
Entropy RSquare	0.3571025
RASE	0.2869804
Mean Abs Dev	0.1702274
Misclassification Rate	0.1130952
-LogLikelihood	365.68935
Sum Freq	1344

Validation

Response

Measures	Value
Generalized RSquare	0.469576
Entropy RSquare	0.3649416
RASE	0.2936831
Mean Abs Dev	0.178734
Misclassification Rate	0.1183036
-LogLikelihood	125.42523
Sum Freq	448

Test

Response

Measures	Value
Generalized RSquare	0.468021
Entropy RSquare	0.3739333
RASE	0.2774609
Mean Abs Dev	0.1606598
Misclassification Rate	0.1026786
-LogLikelihood	110.44858
Sum Freq	448

Confusion Matrix

Actual Response	Predicted Count 0	Predicted Count 1
0	1096	46
1	106	96

Confusion Rates

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.960	0.040
1	0.525	0.475

Confusion Matrix

Actual Response	Predicted Count 0	Predicted Count 1
0	363	13
1	40	32

Confusion Rates

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.965	0.035
1	0.556	0.444

Confusion Matrix

Actual Response	Predicted Count 0	Predicted Count 1
0	379	9
1	37	23

Confusion Rates

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.977	0.023
1	0.617	0.383

K-NN

Training

K	Count	RSquare	Misclassification Rate	Misclassifications
1	1344	-0.0165	0.15402	207
2	1344	0.13465	0.14881	200
3	1344	0.20379	0.12574	169
4	1344	0.21815	0.13021	175
5	1344	0.224	0.12202	164
6	1344	0.23215	0.12426	167
7	1344	0.2406	0.12500	168
8	1344	0.23941	0.11830	159 *
9	1344	0.25226	0.11979	161
10	1344	0.25233	0.11830	159

Validation

K	Count	RSquare	Misclassification Rate	Misclassifications
1	448	0.00786	0.15848	71
2	448	0.08543	0.16295	73
3	448	0.11885	0.15402	69
4	448	0.13591	0.14509	65
5	448	0.15912	0.14063	63
6	448	0.19398	0.14286	64
7	448	0.20591	0.13170	59 *
8	448	0.2202	0.13616	61
9	448	0.22779	0.14063	63
10	448	0.24717	0.13393	60

Test

K	Count	RSquare	Misclassification Rate	Misclassifications
1	448	-0.0104	0.13393	60
2	448	0.15136	0.15179	68
3	448	0.17072	0.11384	51
4	448	0.19729	0.11607	52
5	448	0.23412	0.10714	48 *
6	448	0.22632	0.11607	52
7	448	0.23062	0.12054	54
8	448	0.23762	0.12500	56
9	448	0.22644	0.12946	58
10	448	0.22063	0.12277	55

Confusion Matrix for Best K=7

Training

Actual Response	Predicted Count 0	Predicted Count 1
0	1114	28
1	140	62

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.975	0.025
1	0.693	0.307

Validation

Actual Response	Predicted Count 0	Predicted Count 1
0	369	7
1	52	20

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.981	0.019
1	0.722	0.278

Test

Actual Response	Predicted Count 0	Predicted Count 1
0	375	13
1	41	19

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.966	0.034
1	0.683	0.317

REDUCED KNN

K Nearest Neighbors

Response

Model Selection

Training

K	Count	RSquare	Misclassification Rate	Misclassifications	
1	1344	-0.042	0.16071	216	
2	1344	0.07307	0.16443	221	
3	1344	0.1371	0.13616	183	
4	1344	0.17329	0.13542	182	
5	1344	0.19743	0.13021	175	
6	1344	0.21309	0.13318	179	
7	1344	0.2323	0.12872	173	
8	1344	0.24637	0.12649	170	
9	1344	0.25102	0.12277	165	
10	1344	0.25955	0.12128	163	*

Validation

K	Count	RSquare	Misclassification Rate	Misclassifications	
1	448	0.01601	0.15625	70	
2	448	0.1188	0.16741	75	
3	448	0.11963	0.14732	66	
4	448	0.1404	0.15179	68	
5	448	0.1894	0.13616	61	*
6	448	0.2194	0.14063	63	
7	448	0.24364	0.13839	62	
8	448	0.24006	0.14063	63	
9	448	0.25126	0.13839	62	
10	448	0.26509	0.14286	64	

Test

K	Count	RSquare	Misclassification Rate	Misclassifications	
1	448	0.03524	0.12277	55	
2	448	0.17135	0.12946	58	
3	448	0.18485	0.11607	52	
4	448	0.1901	0.12946	58	
5	448	0.23008	0.12277	55	
6	448	0.2756	0.11830	53	
7	448	0.26762	0.11607	52	
8	448	0.28547	0.11384	51	
9	448	0.28581	0.10938	49	*
10	448	0.2988	0.11384	51	

Confusion Matrix for Best K=5

Training

Actual Response	Predicted Count 0	Predicted Count 1
0	1103	39
1	136	66

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.966	0.034
1	0.673	0.327

Validation

Actual Response	Predicted Count 0	Predicted Count 1
0	365	11
1	50	22

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.971	0.029
1	0.694	0.306

Test

Actual Response	Predicted Count 0	Predicted Count 1
0	376	12
1	43	17

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.969	0.031
1	0.717	0.283

BOOTSTRAP

Minimum Size Split: 5

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.6664	0.3408	0.3359	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7549	0.4429	0.4265	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1412	0.2906	0.2615	$\sum$ -Log(ρ[j])/n
RASE	0.1938	0.3029	0.2857	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1122	0.1813	0.1626	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0499	0.1272	0.1116	$\sum$ (ρ[j]≠ρMax)/n
N	1344	448	448	n

Confusion Matrix

Training

Actual Response	Predicted Count 0	Predicted Count 1
0	1142	0
1	67	135

Actual Response	Predicted Rate 0	Predicted Rate 1
0	1.000	0.000
1	0.332	0.668

Validation

Actual Response	Predicted Count 0	Predicted Count 1
0	373	3
1	54	18

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.992	0.008
1	0.750	0.250

Test

Actual Response	Predicted Count 0	Predicted Count 1
0	381	7
1	43	17

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.982	0.018
1	0.717	0.283

Term	Number of Splits	G^2	Portion
CompleteAcceptedCmp	214	95.6754936	0.1694
Recency	456	78.7361595	0.1394
Customers relationship with the company in Months	381	52.6044901	0.0932
MntMeatProducts(N)	269	51.3370662	0.0909
Marital_Status	333	32.5499676	0.0576
MntGoldProds(N)	228	31.337092	0.0555
MntWines(N)	208	24.6555722	0.0437
NumWebVisitsMonth(N)	251	24.6421392	0.0436
NumDealsPurchases(N)	228	20.7962281	0.0368
Country	189	18.1847325	0.0322
Year_Birth	205	18.0217048	0.0319
MntFishProducts(N)	183	17.3348752	0.0307
NumCatalogPurchases(N)	202	17.208862	0.0305
MntSweetProducts(N)	181	16.8973093	0.0299
NumWebPurchases(N)	194	16.1532118	0.0286
Education_ordinal	185	14.9256803	0.0264
Dependents	212	14.2024014	0.0252
MntFruits(N)	167	14.1604962	0.0251
Income_Ordinal	92	5.11898594	0.0091
Complain	2	0.094705	0.0002

REDUCED BOOTSTRAP

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.6581	0.3299	0.3768	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7479	0.4307	0.4711	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1447	0.2954	0.2454	$\sum$ -Log(ρ[j])/n
RASE	0.1980	0.3029	0.2781	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1140	0.1815	0.1574	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0506	0.1317	0.1094	$\sum$ (ρ[j]≠ρMax)/n
N	1344	448	448	n

Confusion Matrix

Training

Actual Response	Predicted Count 0	Predicted Count 1
0	1138	4
1	64	138

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.996	0.004
1	0.317	0.683

Validation

Actual Response	Predicted Count 0	Predicted Count 1
0	372	4
1	55	17

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.989	0.011
1	0.764	0.236

Test

Actual Response	Predicted Count 0	Predicted Count 1
0	379	9
1	40	20

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.977	0.023
1	0.667	0.333

SVM

Support Vector Coefficients

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5694	0.3570	0.2743	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6697	0.4609	0.3565	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1823	0.2835	0.2858	∑ -Log(ρ[j])/n
RASE	0.2327	0.2864	0.2848	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1060	0.1422	0.1358	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0804	0.1228	0.1116	∑ (ρ[j]≠ρMax)/n
N	1344	448	448	n

Confusion Matrix

Set Probability Threshold

Training

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.993	0.007
1	0.564	0.436

Misclassification Rate 0.0908

Actual Response	Predicted Count 0	Predicted Count 1
0	1134	8
1	114	88

Validation

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.992	0.008
1	0.667	0.333

Misclassification Rate 0.1138

Actual Response	Predicted Count 0	Predicted Count 1
0	373	3
1	48	24

Test

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.979	0.021
1	0.683	0.317

Misclassification Rate 0.1094

Actual Response	Predicted Count 0	Predicted Count 1
0	380	8
1	41	19

Model Launch

Support Vector Machine Model 1

Model Summary

Response	Response
Validation Method	Validation Column
Kernel Function	Radial Basis Function

Measure	Training	Validation	Test
Number of rows	1344	448	448
Sum of Frequencies	1344	448	448
Misclassification Rate	0.0907738	0.1138393	0.109375
Number of Support Vectors	441	441	441

Estimation Details

Cost	1
Gamma	0.05556

REDUCED SVM

Model Summary

Response	Response
Validation Method	Validation Column
Kernel Function	Radial Basis Function

Measure	Training	Validation	Test
Number of rows	1344	448	448
Sum of Frequencies	1344	448	448
Misclassification Rate	0.0900298	0.1116071	0.1049107
Number of Support Vectors	421	421	421

Estimation Details

Cost	1
Gamma	0.08333

Training

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.993	0.007
1	0.559	0.441

Misclassification Rate 0.0900

Actual Response	Predicted Count 0	Predicted Count 1
0	1134	8
1	113	89

Validation

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.989	0.011
1	0.639	0.361

Misclassification Rate 0.1116

Actual Response	Predicted Count 0	Predicted Count 1
0	372	4
1	46	26

Test

Actual Response	Predicted Rate 0	Predicted Rate 1
0	0.990	0.010
1	0.717	0.283

Misclassification Rate 0.1049

Actual Response	Predicted Count 0	Predicted Count 1
0	384	4
1	43	17

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5082	0.3585	0.3152	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6122	0.4626	0.4034	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2081	0.2828	0.2697	∑ -Log(ρ[j])/n
RASE	0.2449	0.2899	0.2815	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1187	0.1472	0.1398	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0863	0.1116	0.1004	∑ (ρ[j]≠ρMax)/n
N	1344	448	448	n

MODEL COMPARISON:

Group A– All Variables

Training (N= 1344)				
Rank	Model Name	Misclassification Rate	Total Accuracy	RASE
1	Bootstrap Forest	4.99	95.01	0.1938
2	SVM	8.63	91.37	0.2449
3	Logistic	10.79	89.21	0.2731
4	Neural	10.86	89.14	0.2806
5	Partition	11.76	88.24	0.3020
6	K Nearest Neighbors	12.50	87.50	

Validation (N= 448)				
Rank	Model Name	Misclassification Rate	Total Accuracy	RASE
1	Neural	11.38	88.62	0.2932
2	Logistic	11.83	88.17	0.3011
3	SVM	12.28	87.72	0.2864
4	Bootstrap Forest	12.72	87.28	0.3029
5	K Nearest Neighbors	13.17	86.83	
6	Partition	13.39	86.61	0.3153

Test (N= 448)				
Rank	Model Name	Misclassification Rate	Total Accuracy	RASE
1	Logistic	9.60	90.40	0.2708
2	Neural	9.60	90.40	0.2755
3	Partition	10.94	89.06	0.2940
4	SVM	11.16	88.84	0.2848
5	Bootstrap Forest	11.16	88.84	0.2857
6	K Nearest Neighbors	12.05	87.95	

Group B – Reduced Variables

Training (N = 1344)				
Rank	Model Name	Misclassification Rate	Total Accuracy	RASE
1	Bootstrap Forest	5.06	94.94	0.1980
2	SVM	8.63	91.37	0.2449
3	Logistic	10.86	89.14	0.2763
4	Neural	11.30	88.70	0.2869
5	K Nearest Neighbors	13.02	86.98	
6	Partition	11.76	88.24	0.3020

Validation (N = 448)				
Rank	Model Name	Misclassification Rate	Total Accuracy	RASE
1	SVM	11.16	88.84	0.2899
2	Neural	11.83	88.17	0.2936
3	Logistic	12.28	87.72	0.3036
4	Bootstrap Forest	13.17	86.83	0.3029
5	Partition	13.39	86.61	0.3153
6	K Nearest Neighbors	13.61	86.39	

Test (N= 448)				
Rank	Model Name	Misclassification Rate	Total Accuracy	RASE
1	Logistic	9.60	90.40	0.2743
2	SVM	10.04	89.96	0.2815
3	Neural	10.26	89.74	0.2774
4	Bootstrap Forest	10.94	89.06	0.2781
5	Partition	10.94	89.06	0.2940
6	K Nearest Neighbors	12.27	87.73	