Classifying Biscuit Brand Switchers for Targeted marketing by a New Biscuit Manufacturer

BADM_B4_Minesweepers

Archana Rajan – 61410071

Kevin John – 61410072

Aditi Vaish – 61410848

Deepak Agnihotri – 61410408

Pranav Maranganty – 61410797

Contents

Executive Summary

- The stakeholder in this data mining project is Mine Sweeper Biscuits (MSB), a premium biscuit manufacturer based out of Denmark. While MSB has entered the Indian market through retail outlets, its sales have failed to take off due to the low product trial rate among Indian consumers.
- The business goal of this project is to be able to predict "brand loyalty" or absence of the same towards biscuit brands, for any new customer making purchases at a hypermart, supermart or other retail outlet. The premise behind such an exercise is that customers who are brand loyal will likely buy only one brand of biscuit, while those who are prone to "switch" brands will be more open to trying any new biscuit brands introduced.
- The need to develop a model for finding the right target segment is attributed to customer acquisition costs. Sending promotional offers, coupons and trial samples is costly for MSB, and hence targeted promotions ensure that the company does not waste its resources targeting "brand loyalists"
- The data mining goal of this project is to create a supervised learning algorithm wherein given certain data about new customers at a supermarket or hypermarket (demographic information and #SKUs, price and quantity of last two purchased baskets); we should be able to predict whether she is a brand "loyalist" or a brand "switcher".
- In order to accomplish this data mining goal, purchase data from the "Ready foods" department was aggregated at the basket level (1 Basket = 1 visit of a consumer). Data relating to #SKUs, average price and quantity purchased in the **last and second last purchase** was also derived for each customer. Note that in order to be included in this dataset the customer had to have made **at least 3 purchases** in the ready foods department. If more than 50% of the purchases were made of the same brand the customer was classified as a "brand loyalist", else a "brand switcher"
- The data was partitioned three ways: a training set for developing the model and validation and testing sets for determining accuracy and possibility of "over-fitting". Further a "holdout" set was created to test the final models developed. Standard data preparation methods such as missing data handling, transformation of categorical variables and creating binned variables (where necessary) where employed.
- Four models were tested: K-Nearest Neighbor, Naïve Bayes, Logistic regression and Classification trees. K-NN, Logistic regression and classification trees (CART) had low errors overall, however there was evidence of overfitting in K-NN. Thus both CART and logistic regression were deployed on the holdout set and based on our results and acquisition/promotion cost considerations we concluded that the misclassification costs were higher with CART. Thus, logistic regression was determined to be the model of choice (for predicting the brand loyalty/switching character of a new customer after her second purchase at the "Ready Foods" department).
- We recommend that this model be further updated by adding more customer demographic data (locality, area code etc) as and when available. The model can also be adapted to serve a similar business goal in other product categories of MSB.

Problem Description

Our stakeholder is Minesweeper Biscuits (MSB), a premium biscuit manufacturer based out of Denmark, looking to enter the Indian market. MSB has tie-ups with key distribution channels including supermarkets and hypermarkets, but has low sales in most of these outlets and is looking at ways to improve its top-line.

Business Goal

MSB is looking at Mumbai as its first location as a part of the expansion. Though it has a good brand name across the world, it is not so popular in India because of low trial rate. It is looking at improved and innovative marketing strategies to create a loyal customer base in Mumbai and has tied up with Hyper Market (HM) for expansion.

Sending promotional codes and other marketing activities are expensive and would not be effective unless there is a high probability that the customer switches. As a part of this project, MSB is trying to predict whether the biscuit customers in the hyper market are brand loyal or not. With this information, MSB will target the specific set of customers who are expected to switch in the near future, to buy MSB biscuits in the next purchase.

Data Mining Goal

A new column needs to be created in the existing dataset "Loyalist?" based on whether the customer has been a loyal customer of a particular brand. The data mining goal in this project will be to create a supervised model to predict the column "Loyalist" based on whether a new customer (i.e. a customer who has made 2 purchases) is a loyalist to any particular brand or not. The purchase patterns from the last 2 purchases and the demographics data of the existing customers will be used for this prediction. Since we are including the purchasing patterns of the customer, we will be able to predict the category of the customer only after 2 purchases in the Hyper Market.

Data Description

The data for this project was collected by Hansa Cequity Solutions. The initial data set contained details of the transactions by customers with loyalty cards over the last one year in the food department at a hyper market. The data included the transaction ID, customer ID, SKU ID, quantity, price, department, sub department, class etc. We were also provided with the customer details which include the Customer IDs, Date of Birth (DOB), Sex, Marital Status and location. These two data sets were linked based on the customer IDs.

Data Preparation

First, current customers were marked as "Loyalist" or not. For this analysis, if the customer has purchased biscuits at least 3 times from the HM and if more than 50% of these purchases were of the same brand, then we considered the customer to be a loyalist. It was assumed that the other times, the biscuit was out of stock or may be the customer purchased other brands for trial. For example, if a customer purchased biscuits 10 times and if he purchased the same brand more than 5 times, he is classified as a "Loyalist".

To predict whether the customer is a loyalist, we aggregated the purchases from the "Ready Food" department at the basket level. We also included the last purchase as well as the second last purchase of the customer to get a better picture of the purchase pattern. Note that data used here is of the last two purchases and not the first two purchases. As a result, we can only predict the behavior of a new customer (going forward) after he has made two purchases from the "Ready Food" department. Once we had the above data, we performed a transformation on the data to handle the missing values and then add the categorical components (married, sex). See Fig 1, for sample data set.

We performed some data visualizations for the loyalist's data across various input parameter, but we could not decipher any particular patterns across a single variable. See Fig 13, for sample data set.

Data Mining Solution

We ran different models and compared performance against benchmark (Naïve prediction). Following are the methods chosen:

- Logistic Regression
- K-Nearest Neighbor method
- Naïve Bayes
- Classification and Regression Trees (CART)

The data used contained 4843 unique customers. The Naïve prediction was brand loyalists: which consisted of 2886 customers (59.6%).1957 (40.4%) were determined to be brand switchers. The data was partitioned into training set (2172), Validation set (1303) and Test Set (868). 500 data points were set aside as "holdout" for model evaluation.

The key input variables used are shown below:

Customer Demographics	Historical Purchase Pattern (Ready Food)	Last Purchase (Ready Food)	Second Last Purchase (Ready Food)
Age Sex Marital Status Enrollment Store	Average Basket Price Average Basket Quantity Average Basket Unique Count Number of Baskets Standard Deviation of Basket Price Standard Deviation of Basket Quantity	Quantity Price Unique SKU Count	Quantity Price Unique SKU Count

Logistic Regression

A stepwise logistic regression was performed with 21 variables including variables from customer demographics, last purchase, total purchase and second last purchase. Based on the Cp values from the output (Fig 2) 20 coefficients model was chosen, as its Cp value was close to 20.

With 20 coefficients, the model performed better than the model with 21 coefficients and the output has been provided in the Appendix (Fig 3). The variables with high p values were still retained as the total % error increased as the variables were removed from the model.

Confusion matrices across the training, validation and test data looked very similar with the overall errors being 38.21%, 37.91% and 39.17% respectively (Fig 4).

K-Nearest Neighbor Method

A K-NN analysis was performed using 12 input variables: age, sex, marital status, enrollment date, average basket price, average basket quantity, number of baskets, last purchase (SKU unique count, price and quantity), and second last purchase (SKU unique count, price and quantity). The output variable was a classification into one of two classes: Loyalist (1) or Switcher (0).

Based on the model, the best K came out to be 5 as shown in (Fig 5) .The Confusion Matrix for the training data set had an overall error of 33.38%, while the validation and test data sets had higher error rates of 50.35% and 47% respectively (Fig 6).However, when this model was run on the holdout data set, there were some signs of over fitting (Fig 7).

Naïve Bayes Method

This method works only on categorical predictors. So in this case, initially the non-categorical variables such as Sex, Marital status were first converted to categorical variables. The predictors were subsequently binned. For this purpose, 99 bins were used for the sake of more granularity given the data and applied on the predictors to create the following predictors - Binned_Age, Binned_Sex, Binned_Average Basket Price, Binned_Average Basket Quantity, Binned_Std Dev of Basket Quantity, Binned_Std Dev of Basket Price, Binned_Last Purchase Quantity, Binned_Last Purchase Price, Binned_Second Last Purchase Price, Binned_Second Last Purchase Quantity. As seen in Fig 8, overall error was 33% in Training and 43% across Validation and Test indicating that there is no overfitting to the data.

Naive Bayes technique is known for its computational efficiency and can work well for large number of predictors. But for rate predictors (which due to partitioning might be missed out from training set) it assigns a zero probability and this can end up giving a biased probability of class membership in case of smaller data sets. Also it gives no insight into role of each predictor. Therefore, we decided **not to utilize this model** given the limited size of the data set and the fact that for multiple combinations of predictor variables, cut-off values the error rate did not fluctuate much.

CART- Classification tree

CART was carried out using 18 input variables. The best pruned tree had 3 decision nodes (Fig 9) with an error of 37.37%. Confusion matrices across the validation and test data looked very similar with the overall errors being 39.14% and 39.17% respectively (Fig 10)

However, we note that the tree-split changes every time we change the data composition (e.g. upon changing the seed for partitioning), and hence this may not be the best model to deploy.

Performance Evaluation

As stated before we chose to not use K-NN due to evidence of overfitting, and Naïve Bayes due to the limited size of the data set. The sensitivity and specificity across all models is seen in Fig 11.Based on the errors across validation and test data, CART and logistic regression looked to be the most promising

models. We evaluated the performance of both logistic regression and CART on the holdout model. The results are shown below:

Logistic Regression

Class	# Cases	# Errors	% Error
0	190	86	45.26%
1	310	102	32.90%
Overall	500	188	37.60%

CART

Class	# Cases	# Errors	% Error
0	190	144	75.79%
1	310	73	23.55%
Overall	500	217	43.40%

Key evaluation metric: We are more concerned with sensitivity (0-1) than specificity, as the misclassification cost is greater in the former case. Following are our misclassification cost assumptions:

- INR 120 for 0 → 1 (Loss in revenues because we fail to target a brand switcher due to misclassification of switcher as loyalist)
- INR 20 for 1 → 0 (Cost in providing coupons and samples to loyalists who will not switch their brand, because of misclassification of loyalists as switchers)

Hence due to the low overall error rate combined with higher sensitivity (misclassification of 0-1 is 45.26% in logistic regression as compared with 75.79% in CART); we recommend logistic regression as the optimal model to classify brand switchers.

We tried to create an **Ensemble** model using majority vote but this gave a greater percentage of error and hence was inferior to logistic regression. Also since the K-NN model had some amount of over fitting, we decided that the ensemble might not work well.

Conclusion and Recommendations

- We conclude that the model to be developed for classifying new customers should be Logistic Regression Model because of:
  - Low Misclassification Costs
  - Similar Accuracy across all Data
  - Better Overall Error and Sensitivity
- The model should be updated by updating the classifications and adding more data
- We recommend adding further demographics such as income level, address pin-code, locality, family size etc to improve the model as these could be good predictors of purchase behavior
- This model can subsequently be expanded to include other MSB products as well

Appendix 1

Fig 1: Sample Data Set

_No	Age	Sex	Enrollment Store	Marital Status	City	Average Basket Price	Average Basket Quantity	Average Basket Unique Count	Number of Baskets	StdDev of Basket Price	StdDev of Basket Quantity	Previous Purchase Quantity	Previous Purchase Price	Previous Purchase Unique SKU Count	Second Last Purchase Quantity	Second Last Purchase Price	Second Last Purchase Unique SKU Count	Loyalist
58	31.17	F	1001	N	Mumbai	518.250	4.667	15.000	3	318.941	3.300	3	90.00	1	30	855.00	9	1
11	48.25	M	1001	Y	Mumbai	1174.894	12.571	42.429	7	651.374	6.652	18	366.00	6	60	2002.26	20	1
56	38.00	M	1001	NA	Mumbai	1212.000	7.000	28.500	2	312.000	2.000	18	900.00	5	39	1524.00	9	1
29	36.00	F	1001	Y	Mumbai	328.500	4.000	16.500	2	106.500	1.000	9	222.00	3	24	435.00	5	0
17	42.00	M	1001	N	Mumbai	246.150	2.700	8.400	10	200.847	1.676	6	65.97	2	3	297.00	1	0
19	28.00	M	1005	N	Ahemdabad	163.515	2.500	7.500	2	105.015	1.500	12	268.53	4	3	58.50	1	0
35	45.00	M	1005	NA	Ahemdabad	1149.614	9.385	61.385	13	480.862	4.683	48	1295.85	16	39	1364.91	12	1
34	42.00	F	1005	NA	Ahemdabad	761.610	5.250	16.500	4	182.743	1.785	24	749.91	7	21	589.53	7	1
09	55.00	M	1005	Y	Ahemdabad	804.081	9.571	36.429	7	477.562	5.949	12	287.94	4	9	279.00	3	1

Fig 2: Cp values in Logistic Regression

#Coeffs	RSS	Cp	Probability	Model (Constant present in all models)												
				1	2	3	4	5	6	7	8	9	10	11	12	13
2	2175.311523	8.32338047	0.15128841	Constant	ber of Baskets	*	*	*	*	*	*	*	*	*	*	*
3	2169.262695	4.27173853	0.37602001	Constant	ber of Baskets	urchase Price	*	*	*	*	*	*	*	*	*	*
4	2164.897217	1.90422928	0.60290945	Constant	ber of Baskets	nsaction Date	urchase Price	*	*	*	*	*	*	*	*	*
5	2161.604492	0.60997301	0.77067214	Constant	ber of Baskets	nsaction Date	urchase Price	urchase Price	*	*	*	*	*	*	*	*
6	2240.932373	81.97475433	0	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	*	*	*	*	*	*	*
7	2240.272217	83.31429291	0	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	*	*	*	*	*	*
8	2240.266113	85.30818176	0	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	*	*	*	*	*
9	2240.260254	87.30232239	0	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	*	*	*	*
10	2239.917969	88.95987701	0	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	*	*	*
11	2239.44751	90.48919678	0	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	*	*
12	2167.473633	20.48184395	0.04206903	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	*
13	2167.340576	22.34872437	0.02704399	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
14	2166.164795	23.17239761	0.02391212	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
15	2163.45874	22.46508408	0.03637987	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
16	2156.688232	17.69142723	0.24410817	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
17	2155.811768	18.81455421	0.21520023	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
18	2155.493408	20.49604607	0.14036413	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
19	2153.184326	20.1858902	0.2043643	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
20	2150.459229	19.45952415	0.49816209	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price
21	2150	21.00008202	1	Constant	Age	Sex_F	Sex_M	nt Store_1001	nt Store_1002	arital Status_N	arital Status_Y	Email_Y	e Basket Price	asket Quantity	ber of Baskets	f Basket Price

Fig 3: Logistic regression with 20 predictors

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-47.1890297	22.87351227	0.03910859	*
Age	0.00544622	0.00423081	0.19799799	1.0054611
Sex_F	1.11986947	0.60941881	0.066121	3.06445408
Sex_M	1.03916395	0.60556686	0.08615863	2.8268528
Enrollment Store_1001	-0.67768991	0.83002526	0.41423193	0.50778866
Enrollment Store_1002	0.1495695	0.90910864	0.86931926	1.16133416
Marital Status_N	0.09804565	0.14699289	0.50476611	1.10301316
Marital Status_Y	0.07624547	0.14410833	0.59674686	1.07922745
Email_Y	-0.05565267	0.09803177	0.57023847	0.9458676
Average Basket Price	-0.00135054	0.00139278	0.33220983	0.99865037
Average Basket Quantity	-0.04568335	0.0698395	0.51303506	0.95534444
Number of Baskets	0.00852676	0.00164464	0.00000022	1.00856316
StdDev of Basket Price	0.00019505	0.00047311	0.68013567	1.00019503
StdDev of Basket Quantity	0.02833186	0.02601394	0.27610847	1.02873707
Last Transaction Date	0.00110265	0.00055753	0.04795689	1.00110328
Last Purchase Unique Count	0.01369065	0.01717211	0.42529947	1.01378477
Last Purchase Price	0.00012139	0.00010526	0.24879704	1.00012136
Last Purchase Quantity	-0.00228719	0.00381123	0.54842746	0.99771541
Second Last Purchase	-0.00255478	0.01313885	0.84582764	0.9974485
Second Last Purchase Price	0.00014995	0.00008951	0.09388413	1.00014997

Residual df	2152
Residual Dev.	2830.073486
% Success in training data	41.66666667
# Iterations used	15
Multiple R-squared	0.04078817

Fig 4: Training, Validation and Test data scoring reports for logistic regression

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.4

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	519	386
1	444	823

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
0	905	386	42.65
1	1267	444	35.04
Overall	2172	830	38.21

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.4

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	302	222
1	272	507

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
0	524	222	42.37
1	779	272	34.92
Overall	1303	494	37.91

Test Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.4

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	217	142
1	198	311

Error Report			
Class	# Cases	# Errors	% Error
0	359	142	39.55
1	509	198	38.90
Overall	868	340	39.17

Fig 5: Best K for K-NN

Value of k	% Error Training	% Error Validation	
1	0.00	44.67	
2	23.94	49.12	
3	23.16	45.13	
4	27.53	47.51	
5	28.68	43.13	<--- Best k
6	30.66	47.12	
7	30.34	43.90	
8	31.72	45.89	
9	32.27	43.75	
10	32.97	44.74	

Fig 6: Training, Validation and Test Data scoring reports for K-NN

Training Data scoring - Summary Report (for k=5)

Cut off Prob.Val. for Success (Updatable)	0.4

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	793	112
1	613	654

Error Report			
Class	# Cases	# Errors	% Error
0	905	112	12.38
1	1267	613	48.38
Overall	2172	725	33.38

Validation Data scoring - Summary Report (for k=5)

Cut off Prob.Val. for Success (Updatable)	0.4

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	376	148
1	508	271

Error Report			
Class	# Cases	# Errors	% Error
0	524	148	28.24
1	779	508	65.21
Overall	1303	656	50.35

Test Data scoring - Summary Report (for k=5)

Cut off Prob.Val. for Success (Updatable)	0.4

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	264	95
1	313	196

Error Report			
Class	# Cases	# Errors	% Error
0	359	95	26.46
1	509	313	61.49
Overall	868	408	47.00

Fig 7: Confusion Matrix for K-NN (Holdout data)

Classification Confusion Matrix – KNN Holdout		
	Predicted Class	
Actual Class	0	1
0	112	57
1	211	120

Error Report			
Class	# Cases	# Errors	% Error
0	169	57	33.72
1	331	211	63.75
Overall	500	268	57.80

Fig 8: Training, Validation and Test data scoring reports for Naïve Bayes

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	718	187
1	532	735

Error Report			
Class	# Cases	# Errors	% Error
0	905	187	20.66
1	1267	532	41.99
Overall	2172	719	33.10

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	342	182
1	387	392

Error Report			
Class	# Cases	# Errors	% Error
0	524	182	34.73
1	779	387	49.68
Overall	1303	569	43.67

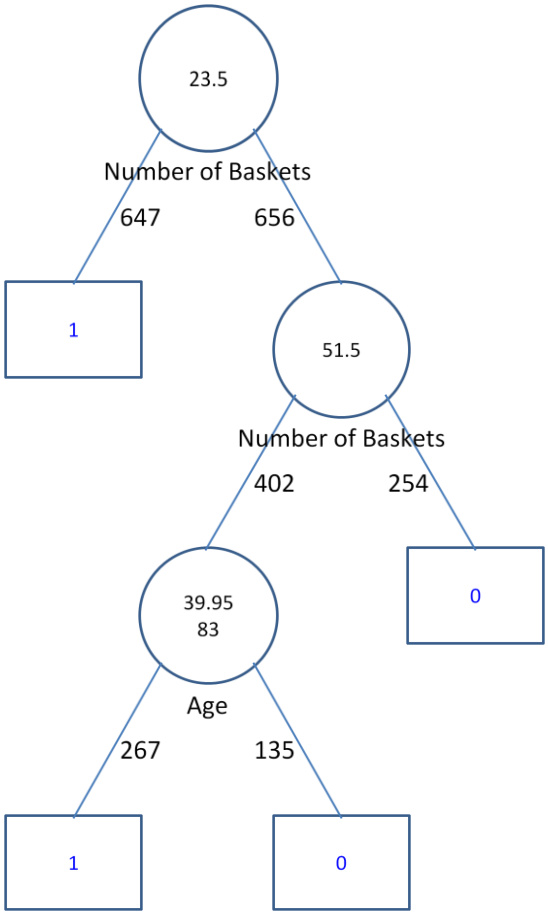
Test Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	241	118
1	261	248

Error Report			
Class	# Cases	# Errors	% Error
0	359	118	32.87
1	509	261	51.28
Overall	868	379	43.66

Fig 9: Best pruned classification tree



Best Pruned Tree Rules (Using Validation Data)

#Decision Nodes	3

#Terminal Nodes	4

Level	NodeID	ParentID	SplitVar	SplitValue	Cases	LeftChild	RightChild	Class	Node Type
0	0	N/A	ber of Baskets	23.5	1303	1	2	1	Decision
1	1	0	N/A	N/A	647	N/A	N/A	1	Terminal
1	2	0	ber of Baskets	51.5	656	3	4	0	Decision
2	3	2	Age	39.9583	402	5	6	0	Decision
2	4	2	N/A	N/A	254	N/A	N/A	0	Terminal
3	5	3	N/A	N/A	267	N/A	N/A	1	Terminal
3	6	3	N/A	N/A	135	N/A	N/A	0	Terminal

Fig 10: Training, Validation and Test data scoring reports for Naïve Bayes

Training Data scoring - Summary Report (Using Full Tree)

Cut off Prob.Val. for Success (Updatable)	0.4	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	905	0
1	0	1267

Error Report			
Class	# Cases	# Errors	% Error
0	905	0	0.00
1	1267	0	0.00
Overall	2172	0	0.00

Validation Data scoring - Summary Report (Using Best Pruned Tree)

Cut off Prob.Val. for Success (Updatable)	0.4	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	335	189
1	321	458

Error Report			
Class	# Cases	# Errors	% Error
0	524	189	36.07
1	779	321	41.21
Overall	1303	510	39.14

Test Data scoring - Summary Report (Using Best Pruned Tree)

Cut off Prob.Val. for Success (Updatable)	0.4	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	234	125
1	215	294

Error Report			
Class	# Cases	# Errors	% Error
0	359	125	34.82
1	509	215	42.24
Overall	868	340	39.17

Fig 11: Error rates across various models

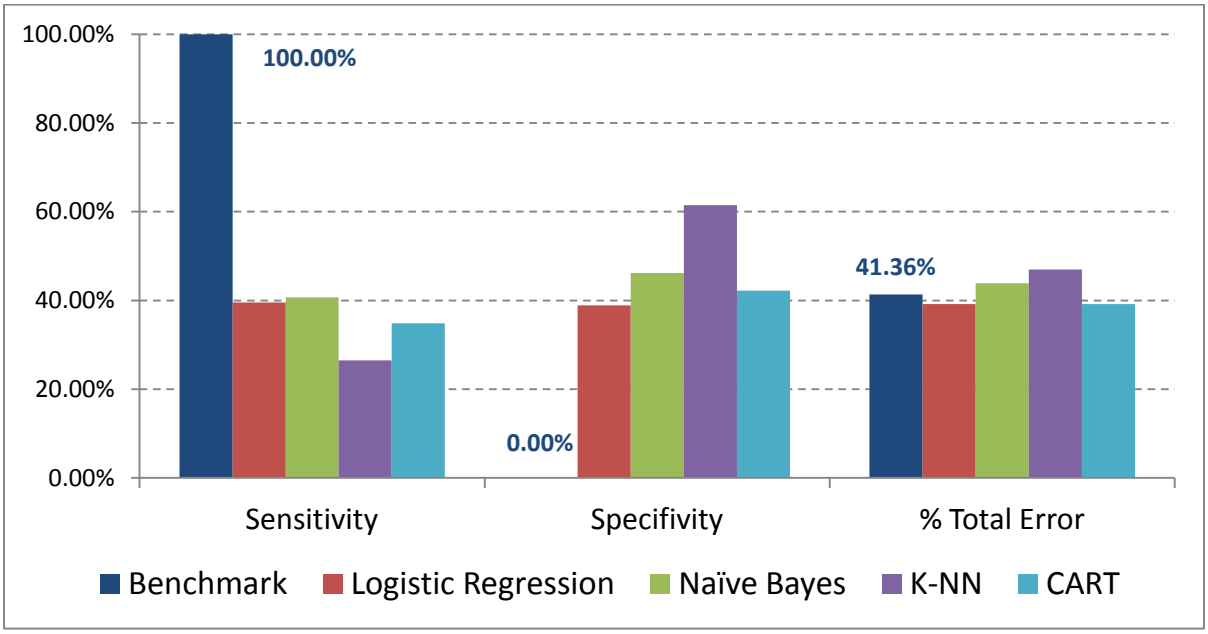


Fig 12: Misclassification costs of various models

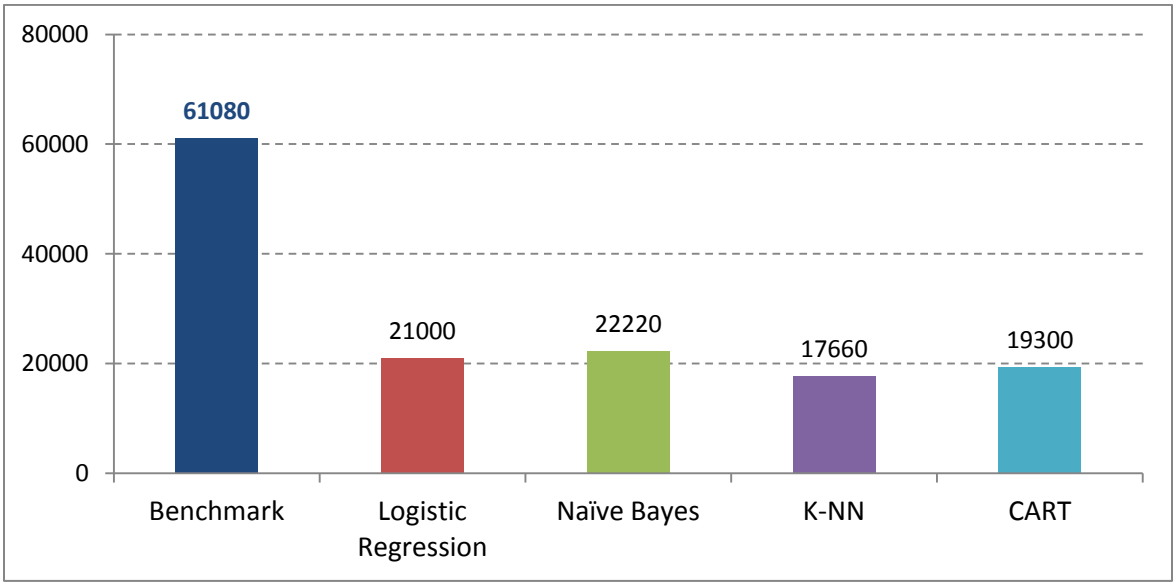


Fig 13: Data Visualizations