Indian School of Business

Model to predict MVAS Likelihood

Team Famous Five

- Raghuraman Chandrasekhar
- Sagar Gupta
- Saurabh Choudhary
- Sudhanshu Dharmadhikari
- Yash Chandwani

Contents

Executive Summary 3

BUSINESS OBJECTIVES 4

DATA PREPARATION 5

DATA MINING APPROACH & METHODS 6

Logistic Regression 6

K Nearest Neighbor (KNN) 6

BENCHMARKING + CRITIQUE 8

CONCLUSIONS AND RECOMMENDATIONS 8

APPENDIX - TECHNICAL ANALYSIS 9

Executive Summary

The Indian mobile telephony market is a classic example of a volume based strategy. Average revenue per user (ARPU) is one of the lowest, while the subscriber figures are second only to China. The market itself is highly fragmented with more than 5 players holding less than 90% of the market. In this scenario, identifying new sources of revenues is critical for survival.

This study focuses on identifying potential customers for the mobile value added services (MVAS) – an increasingly significant revenue driver for Indian operators. As most customers have already adopted MVAS in some form, identifying pockets of opportunity while maintaining accuracy of prediction is essential to reduce promotion costs.

Post data preparation and standardization, two data mining approaches were adopted, Logistic Regression and K Nearest Neighbor (KNN). Lift and Decile charts were used to compare the performance of the models. Benchmarking of the model was done based on the Naïve Majority class for comparison with the output of our selected model

The study shows that smartphone ownership and monthly mobile expenditure are major influencers for MVAS adoption and the fact should be central to the promotion strategy.

BUSINESS OBJECTIVES

The global telecom industry has gone through a revolution – from selling simple voice services, the mobile phone is changing the way people share and consume information. India is no exception to this norm. However, in this highly competitive market, identifying the right customer mix and the right promotion strategy is critical to keep acquisition costs down. Therefore, companies must be able to identify customers who are engaged with existing products or applications and target them for new and innovative product offerings.

The idea of this exercise is to identify the likelihood of a customer towards buying MVAS offerings based on past data of similar customers. The report is ideally targeted at large Indian mobile operators who want to identify growth opportunities and identify the most likely response set for the MVAS offering.

As per the available data, a major challenge faced by the businesses is that the proportion of non MVAS customers to the total number is very small. Hence, accuracy of prediction is essential to reduce promotion costs. Additionally, for existing customers, non-availability of historical data made it difficult to establish a financial model.

The model explains the factors that drive the MVAS adoption amongst the customers. The model also explains the strengths of the above mentioned variables in driving adoption of MVAS. Hence, the client can understand the importance of each factor and try to influence the most important ones for maximizing the chances of MVAS adoption by the customers.

The model ranks the survey respondent’s likelihood to respond to MVAS. The ranking obtained through probability would help the company in identifying the customers with high probability of adoption. This is information will be very helpful to the client, as in real life marketing budgets are limited and the company would like to focus marketing on customers where the probability of success is high.

DATA PREPARATION

The following steps were followed for data preparation

Initial Analysis: All the variables were studied and their values observed for relevance to the prediction. Knowledge about telecom usage patterns was valuable in understanding the importance of the independent variables. The analysis also considered the availability of the predictors on a regular basis (e.g. age, usage pattern, etc.)

Identification: Values key to the model were identified and put through further analysis and the following variables were shortlisted:

- Num.Mobiles_CatNo.
- Mobile.Type..Primary._CatNo.
- Network.duration_CatNo.
- last.handset.purhcased_CatNo.
- pref.Apple_CatNo.
- Current.handset.brand_CatNo.
- Freq.of.changing.handsets_CatNo.
- Monthly.expenditure.on.mobile.service_CatNo.
- Usual.top.up.size.Pre.Paid.User._CatNo.
- Average.SMSes.per.day_CatNo.
- Age_CatNo.
- Yearly.household.income_CatNo.

Missing Value checks: In the identified variables, missing value checks were performed on the data

Reduce Categories: Once the variables were identified, the values in these variables were categorized using XLMiners's reduce Categories function

Data Partitions: After the categorized were reduced for all the required predictors, data was partitioned into Training, Validation and Test datasets (50%, 30%, 20% respectively). A test partition was created as we were using K-NN which uses the validation data to fit the model and an independent hold-out set is required to test the goodness of fit.

DATA MINING APPROACH & METHODS

Logistic Regression

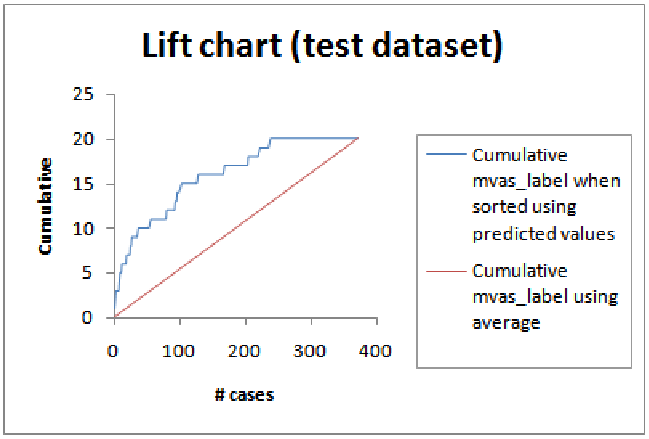
PROS

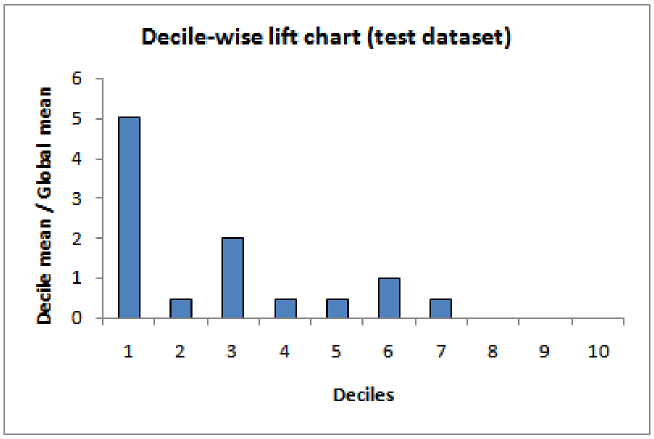
- The output of interest is binary and hence Logistic Regression is a good fit
- Its easy to understand and quantify impact of drivers well using Lift and Decile Charts
- Sensitivity Analysis with different cut-off based on business situations helps

CONS

- The business process of the model is lengthy and it is very important to have the right data preparation done for the model
- The Non-linearities and interactions within the data needs to be handled manually

Charts





K Nearest Neighbor (KNN)

PROS

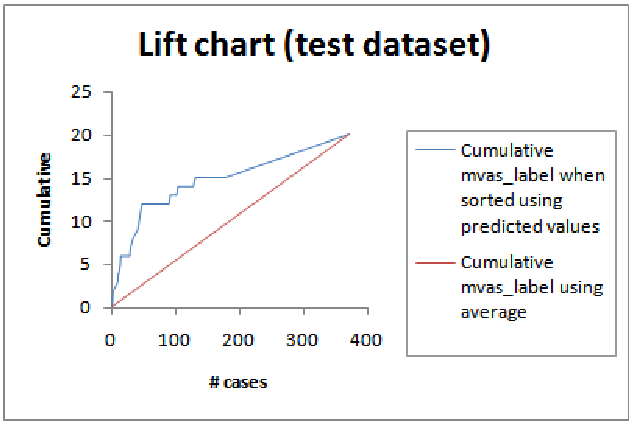
- Since the response rates in our data is low, KNN could be a good fit here

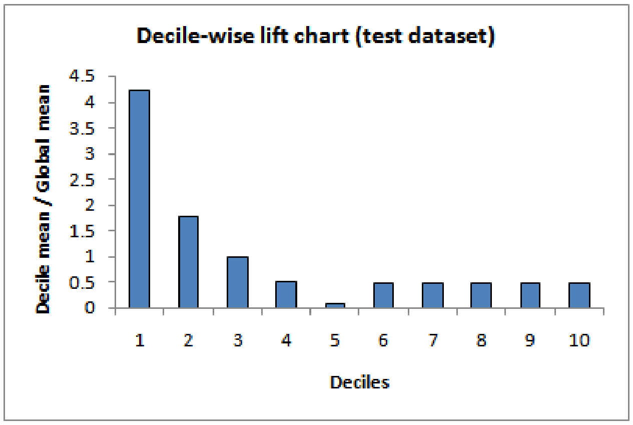
- Its easy to understand and quantify impact of drivers well using Lift and Decile Charts
- Can adapt cut-off cater to various business scenarios

CONS

- Existence of irrelevant attributes in the data could make the model vulnerable to noise

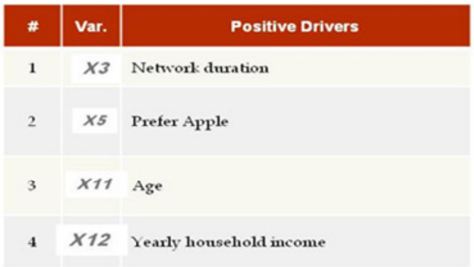
Charts



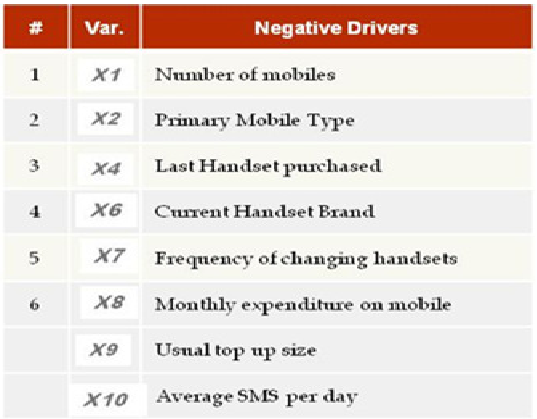


As we can see from the Decile charts above, the top 3 deciles capture more Non MVAS subscribers much better in logistic regression as compared to KNN. Therefore, it is recommended to adapt the Logistic Regression model for the study.

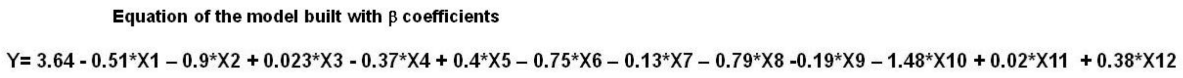
Based on the outcome of the models, we recommend the following model:



#	Var.	Positive Drivers
1	X3	Network duration
2	X5	Prefer Apple
3	X11	Age
4	X12	Yearly household income



#	Var.	Negative Drivers
1	X1	Number of mobiles
2	X2	Primary Mobile Type
3	X4	Last Handset purchased
4	X6	Current Handset Brand
5	X7	Frequency of changing handsets
6	X8	Monthly expenditure on mobile
	X9	Usual top up size
	X10	Average SMS per day



Equation of the model built with β coefficients

$$Y= 3.64 - 0.51*X1 – 0.9*X2 + 0.023*X3 - 0.37*X4 + 0.4*X5 – 0.75*X6 – 0.13*X7 – 0.79*X8 -0.19*X9 – 1.48*X10 + 0.02*X11 + 0.38*X12$$

BENCHMARKING + CRITIQUE

Ideally the model should outperform a random guess. We observe both the models capture over 70% of the Non-MVAS users within the top three deciles indicating a fair degree of accuracy.

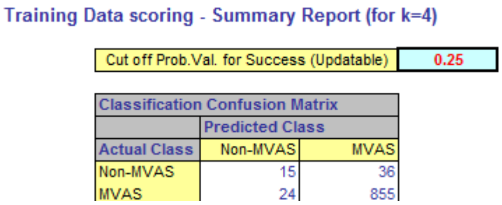
CONCLUSIONS AND RECOMMENDATIONS

1. The sophistication of the model depends on the data quality. Therefore the managers must focus on getting the twelve selected variables. This can be done during the on-boarding process (e.g. Age, handset type, etc.) or by monitoring usage (network duration, monthly expenditure, etc.)
2. The probability scores presented by the model can then be used to plan a tailor made campaign to target customers. Based on the sensitivity of the model, an ideal segment would be smartphone users who use SMS heavily and spend more than INR 750 a month.
3. The model can be further improved by:
   a. Adding time series data along with greater financial information to add potential customer life time values (CLV) in the model
   b. Do a study specific to the most profitable MVAS types (such as caller tunes, astrology, etc.) to further customize the offer.

APPENDIX - TECHNICAL ANALYSIS

K-NN Model

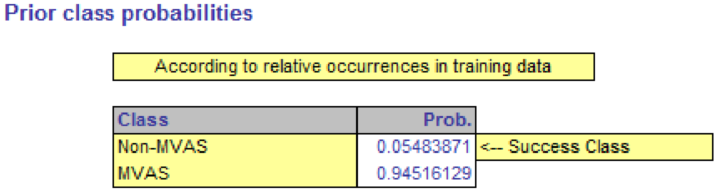
This is essentially a classification problem. To estimate the probability of adoption of MVAS, the first step in our approach was to use K-NN model. The output of K-NN model is as follows:



Training Data scoring - Summary Report (for k=4)

Cut off Prob.Val. for Success (Updatable)	0.25

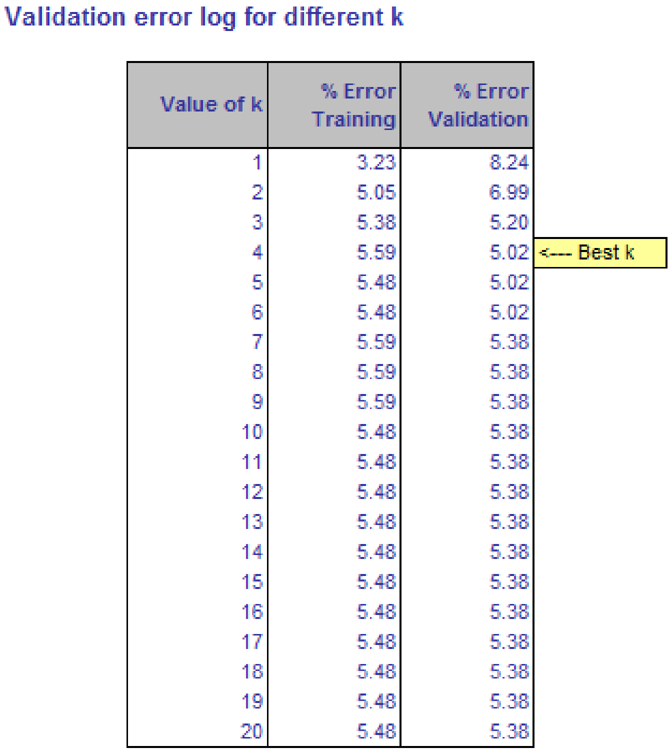
Classification Confusion Matrix		
	Predicted Class	
Actual Class	Non-MVAS	MVAS
Non-MVAS	15	36
MVAS	24	855



Prior class probabilities

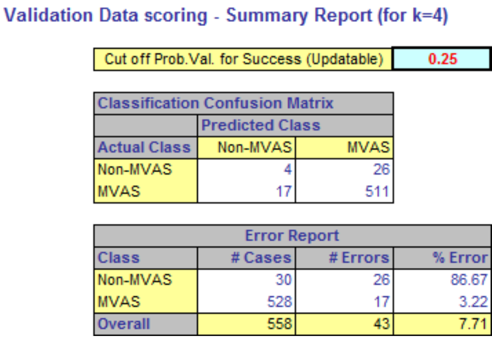
According to relative occurrences in training data

Class	Prob.	
Non-MVAS	0.05483871	<-- Success Class
MVAS	0.94516129	



Validation error log for different k

Value of k	% Error Training	% Error Validation	
1	3.23	8.24	
2	5.05	6.99	
3	5.38	5.20	
4	5.59	5.02	<--- Best k
5	5.48	5.02	
6	5.48	5.02	
7	5.59	5.38	
8	5.59	5.38	
9	5.59	5.38	
10	5.48	5.38	
11	5.48	5.38	
12	5.48	5.38	
13	5.48	5.38	
14	5.48	5.38	
15	5.48	5.38	
16	5.48	5.38	
17	5.48	5.38	
18	5.48	5.38	
19	5.48	5.38	
20	5.48	5.38	

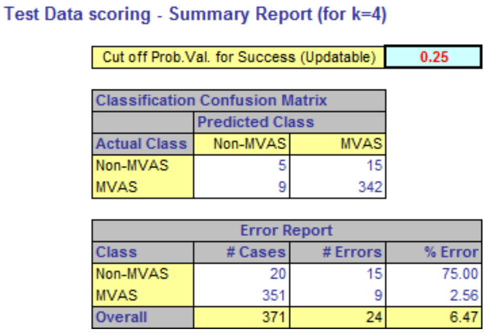


Validation Data scoring - Summary Report (for k=4)

Cut off Prob.Val. for Success (Updatable)	0.25

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Non-MVAS	MVAS
Non-MVAS	4	26
MVAS	17	511

Error Report			
Class	# Cases	# Errors	% Error
Non-MVAS	30	26	86.67
MVAS	528	17	3.22
Overall	558	43	7.71

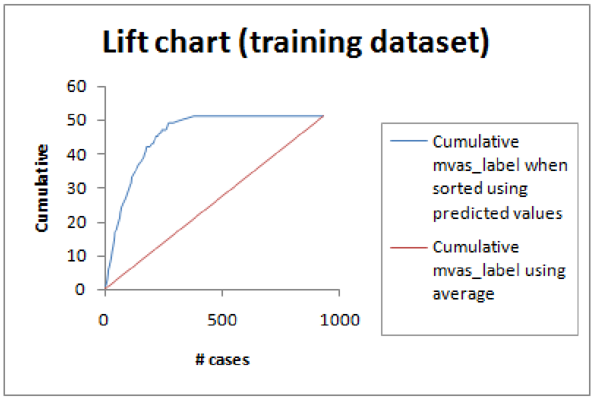


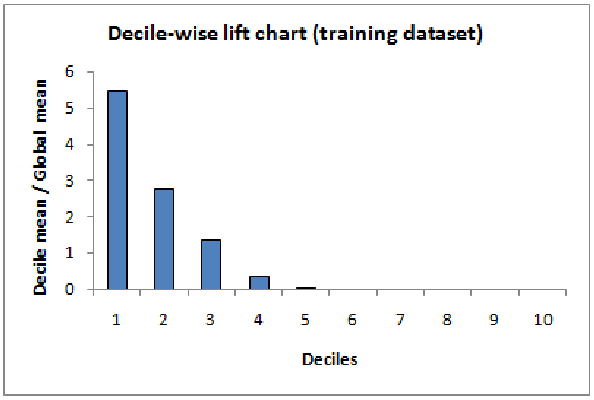
Test Data scoring - Summary Report (for k=4)

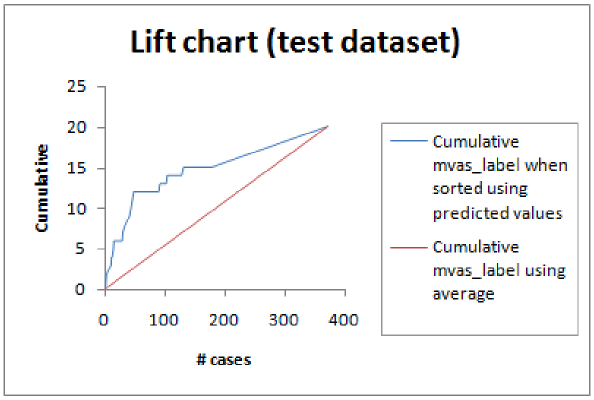
Cut off Prob.Val. for Success (Updatable)	0.25

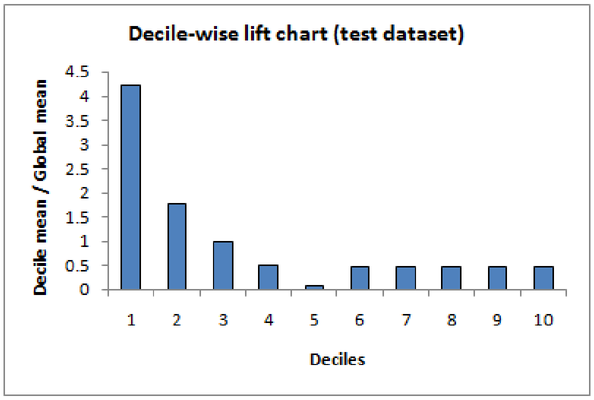
Classification Confusion Matrix		
	Predicted Class	
Actual Class	Non-MVAS	MVAS
Non-MVAS	5	15
MVAS	9	342

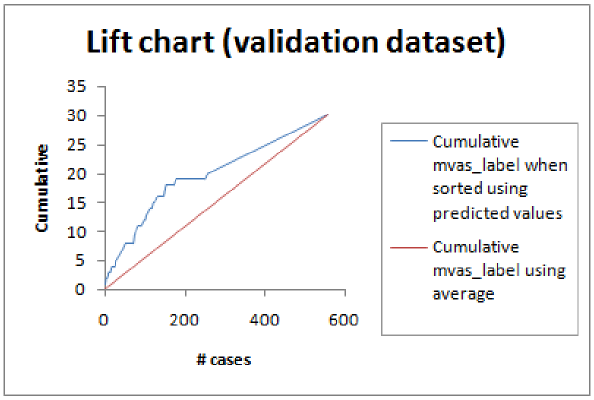
Error Report			
Class	# Cases	# Errors	% Error
Non-MVAS	20	15	75.00
MVAS	351	9	2.56
Overall	371	24	6.47

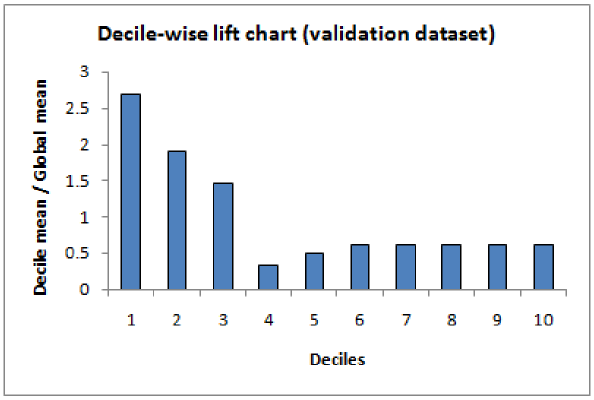








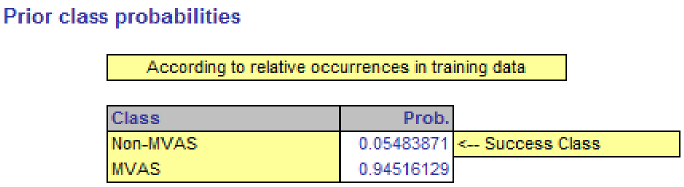




As seen from the table comparing % error in validation for different values of k, k=4 was chosen as the optimum criteria for K-NN. From the lift charts, we can see that although the model performs quite well on training data, its performance on validation and training data is also reasonably good.

Logistics Regression

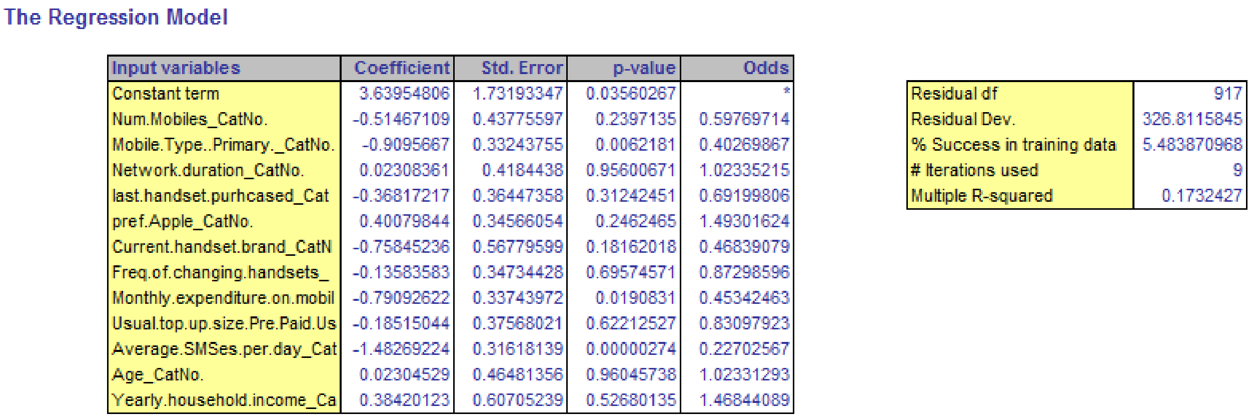
Although the K-NN model provides satisfactory results, we have explored the possibility of using logistics regression for more accurate prediction. The results of logistics regression are as follows:



Prior class probabilities

According to relative occurrences in training data

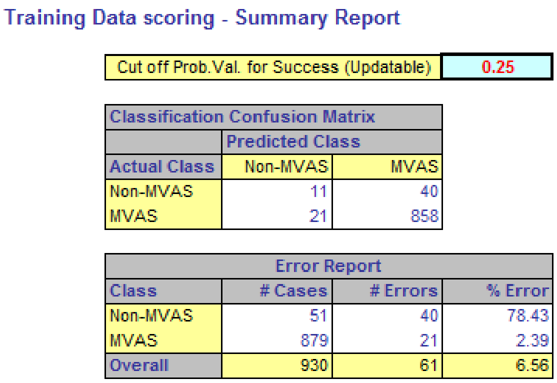
Class	Prob.	
Non-MVAS	0.05483871	<-- Success Class
MVAS	0.94516129	



The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	3.63954806	1.73193347	0.03560267	*
Num.Mobiles_CatNo.	-0.51467109	0.43775597	0.2397135	0.59769714
Mobile.Type..Primary._CatNo.	-0.9095667	0.33243755	0.0062181	0.40269867
Network.duration_CatNo.	0.02308361	0.4184438	0.95600671	1.02335215
last.handset.purhcased_Cat	-0.36817217	0.36447358	0.31242451	0.69199806
pref.Apple_CatNo.	0.40079844	0.34566054	0.2462465	1.49301624
Current.handset.brand_CatN	-0.75845236	0.56779599	0.18162018	0.46839079
Freq.of.changing.handsets_	-0.13583583	0.34734428	0.69574571	0.87298596
Monthly.expenditure.on.mobil	-0.79092622	0.33743972	0.0190831	0.45342463
Usual.top.up.size.Pre.Paid.Us	-0.18515044	0.37568021	0.62212527	0.83097923
Average.SMSes.per.day_Cat	-1.48269224	0.31618139	0.00000274	0.22702567
Age_CatNo.	0.02304529	0.46481356	0.96045738	1.02331293
Yearly.household.income_Ca	0.38420123	0.60705239	0.52680135	1.46844089

Residual df	917
Residual Dev.	326.8115845
% Success in training data	5.483870968
# Iterations used	9
Multiple R-squared	0.1732427

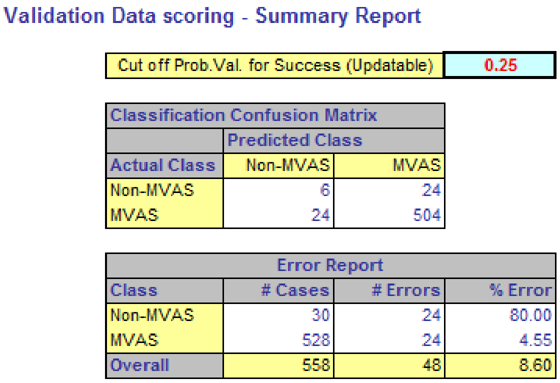


Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.25

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Non-MVAS	MVAS
Non-MVAS	11	40
MVAS	21	858

Error Report			
Class	# Cases	# Errors	% Error
Non-MVAS	51	40	78.43
MVAS	879	21	2.39
Overall	930	61	6.56

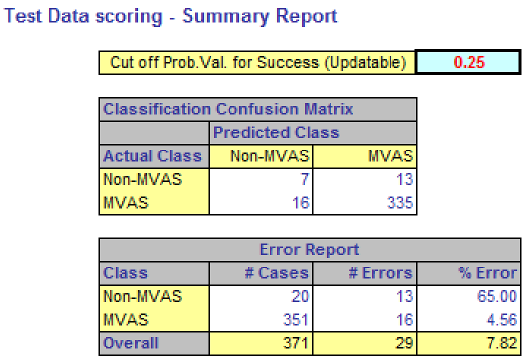


Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.25

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Non-MVAS	MVAS
Non-MVAS	6	24
MVAS	24	504

Error Report			
Class	# Cases	# Errors	% Error
Non-MVAS	30	24	80.00
MVAS	528	24	4.55
Overall	558	48	8.60

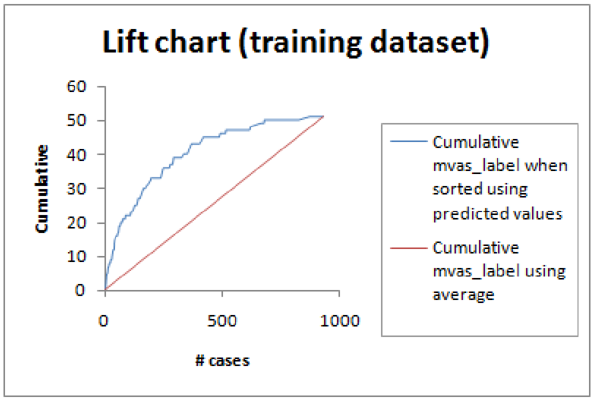


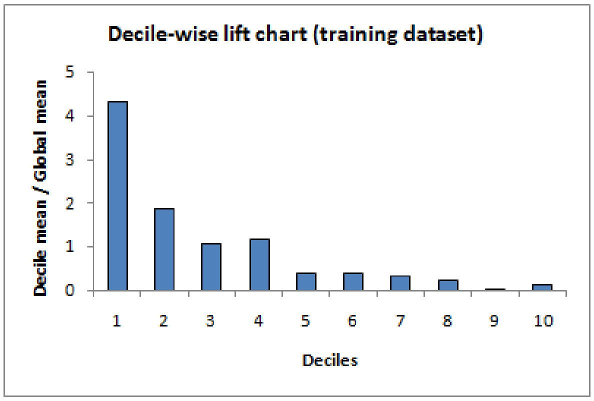
Test Data scoring - Summary Report

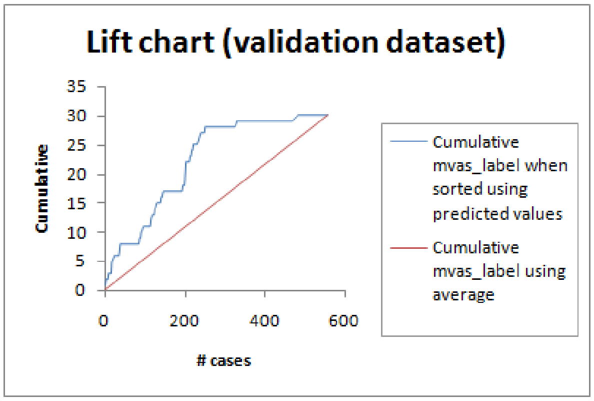
Cut off Prob.Val. for Success (Updatable)	0.25

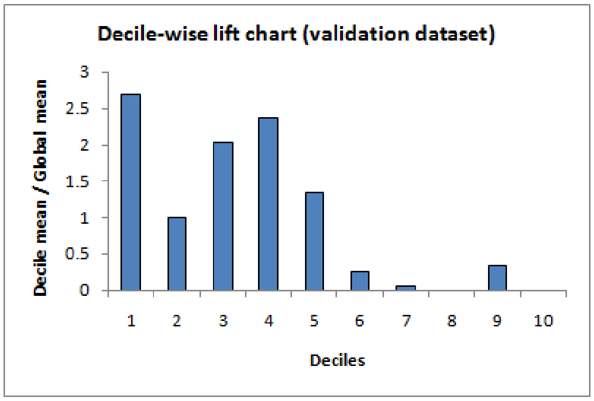
Classification Confusion Matrix		
	Predicted Class	
Actual Class	Non-MVAS	MVAS
Non-MVAS	7	13
MVAS	16	335

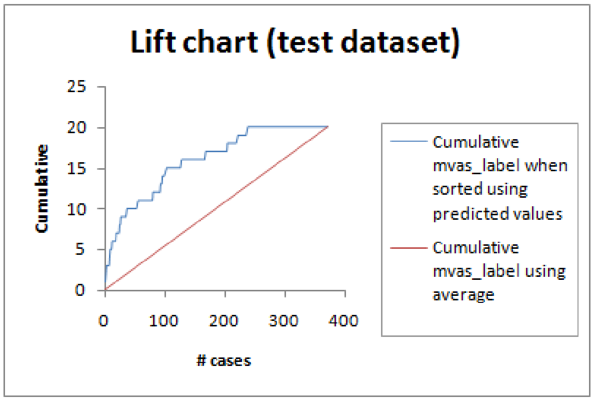
Error Report			
Class	# Cases	# Errors	% Error
Non-MVAS	20	13	65.00
MVAS	351	16	4.56
Overall	371	29	7.82

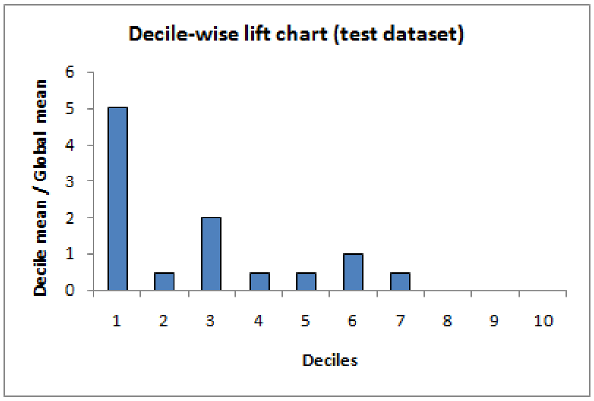












As can be seen from the lift charts, the LR model fits the test data quite well.

On comparing the Decile-wise lift charts for K-NN model and LR model, we find that the top three deciles capture more non VAS subscribers in logistics as compared to K-NN. Hence, we recommend the model based on Logistic Regression.