Application of Logistic Regression to perform Churn Analysis-Propensity to attrite

Business Objective: John is the Customer Services and Relations head for a Multi brand retail store. He analyzed a couple of reports and got worried about losing his customers overtime. He thought over the different customer segments presented in the report and concluded that not all customers were worth retention. He identified the loyal and profitable customer segment and planned to develop a churn model to gauge the propensity of attrition of this customer segment.

He had plans to revise promotions and schemes for these customers based on the significant factors contributing to their attrition. John will use Churn model primarily to identify the customers next in line to attrite. He will then plan the promotions and strategies to retain them.

****

John model preparation will start with identification of responders and non-responders for his model. He will pull past customer's data from his CRM and set a timeline to classify customers in two groups.

Group1: Customers who have been in the system for more than 36 months but haven't done any transaction in past 12 months

Group2: Customers who have been in the system for more than 36 months but have done transaction in past 12 months

The assumption is that the loyal customers who have not done transaction in past 12 months are a case of churn and those who are still our customers are case of the retention. Since we are trying to identify the propensity of the customers who might leave, we will tag them as our responders (Case 1) and the retained customers as our non responders (Case 0).

The historical observation is that we lose close to 40% of the customers every year and we will use this information to keep the proportion of the responders and non-responders in our data sample.

The model will identify the important aspects of churn, known as the "Drivers of the churn" and give the propensity of churn for the customers.

Model building

Find as much attributes in CRM data as you can, and make a dataset of those attributes. The data should capture demographic details, transaction details, customer satisfaction, customer experience and other information if available.

List of attributes considered for this case:

The attributes have been denoted by a Label for ease of programming and reference in future

Attributes	Label
Region_cd	Attr1
Income	Attr2
Occupation	Attr3
Age	Attr4
Gender	Attr5
Marital_status	Attr6
Num_Cust_care_negative_exp	Attr7
Num_Cust_care_positive_exp	Attr8
Num_Promotions_Sent	Attr9
Num_Promotions_Used	Attr10
Cust_satisfaction_score	Attr11
Pymt_type	Attr12
Loyality_card	Attr13
Price_Sensitivitiy_Index	Attr14
Brand_Conscious_Index	Attr15
Weekend_shopper	Attr16
Frequent_hopper	Attr17
Average_value_per_Txn	Attr18

1. **Import data**

```
###Create working directory
setwd("C:/Users/babycorn/Documents/Churn Model-Retail ")

###Import the Raw data file into cust_data file
cust_data<-read.csv("churndata.csv")
### See the data summary (verify Data)
head(cust_data)
tail(cust_data)

summary(cust_data)
```

Cust_id	Responder	Attr1	Attr2	Attr3	Attr4	Attr5
Min. :1001	Min. :0.0000	B :1380	Min. : 101	A: 962	Min. : 1.00	Female:3941
1st Qu.:3001	1st Qu.:0.0000	H :1102	1st Qu.:2333	B:2345	1st Qu.: 20.00	Male :4059
Median :5000	Median :0.0000	A : 828	Median :4544	C:1494	Median : 40.00	NA
Mean :5000	Mean :0.2502	C : 828	Mean :4565	D: 960	Mean : 40.18	NA
3rd Qu.:7000	3rd Qu.:1.0000	E : 828	3rd Qu.:6834	E:1066	3rd Qu.: 60.00	NA
Max. :9000	Max. :1.0000	G : 828	Max. :8998	F:1173	Max. :176.00	NA
NA	NA	(Other):2206	NA's :26	NA	NA	NA
Attr6	**Attr7**	**Attr8**	**Attr9**	**Attr10**	**Attr11**	
Married :4450	Min. :1.000	Min. :1.000	Min. : 1.000	Min. : 1.000	Min. : 1.000	
Other : 581	1st Qu.:2.000	1st Qu.:2.000	1st Qu.: 3.000	1st Qu.: 3.000	1st Qu.: 3.000	
Unmarried:2965	Median :4.000	Median :4.000	Median : 5.000	Median : 6.000	Median : 5.000	
NA	Mean :4.001	Mean :4.001	Mean : 5.462	Mean : 5.524	Mean : 5.502	
NA	3rd Qu.:6.000	3rd Qu.:6.000	3rd Qu.: 8.000	3rd Qu.: 8.000	3rd Qu.: 8.000	
NA	Max. :7.000	Max. :7.000	Max. :10.000	Max. :10.000	Max. :10.000	
NA	NA	NA	NA	NA	NA	
Attr12	**Attr13**	**Attr14**	**Attr15**	**Attr16**	**Attr17**	**Attr18**
Card :1950	125:59:00	Min. :1.000	Min. :1.000	No :4080	No :3995	Min. : 5
Cash :1994	No :3952	1st Qu.:2.000	1st Qu.:1.000	Yes:3920	Yes:4005	1st Qu.: 55
Coupon:2014	Yes:3922	Median :3.000	Median :3.000	NA	NA	Median : 104
Other :2042	NA	Mean :2.976	Mean :2.571	NA	NA	Mean : 104
NA	NA	3rd Qu.:4.000	3rd Qu.:3.500	NA	NA	3rd Qu.: 152
NA	NA	Max. :5.000	Max. :5.000	NA	NA	Max. :1800
NA	NA	NA	NA's :7993	NA	NA	NA

The analysis from above report tells us that Var1 and Var27 will need missing value treatment. Further, the % of missing value is very high in Var27

% Missing = 4172/5000 = 83.44
As a rule, we generally drop variables with more than 50% missing values. Hence Var27 should be dropped out from the dataset.
For Var1 , % Missing = 25/5000 = 0.5% . We can impute this with either the mode value of the variable or any meaning value we can think of.

Export the above summary results to the work folder to prepare univariate report using the following code:

```
### Save the summary results in work folder
summ_cust_data.csv<-summary(cust_data)
write.csv(summ_cust_data.csv, "summ_cust_data.csv.csv")
```

Analyze the output for missing value, extreme values verification.
In the above case, we can see that % missing for Attr15 is 7993/8001 =99.9 % .We can safely drop this variable. As a rule, we generally drop variables with more than 50% missing values

and impute those with less than 50% missing values. In the above case. Attr2 can be imputed with the mean value of 4565.

Summary of the data preparation steps performed above:

1. Data Audit and EDA
2. Data Profiling

Data Audit and EDA: - The data audit report is the initial report that we prepared to understand the data well. This report will consist of descriptive statistics for all the variables in the dataset and also will helps in indentifying the missing value and levels of categorical predictive variables. This data Audit report serves as base for assessing the quality of the data we extracted and obtained from the client. Based on this report we can request for additional data which we seem to be important for our analysis and it will also help in dropping some insignificant variable. Refer the data audit report fields below for better understanding.

Variable
Data Type
Label
Total no. of records in dataset
No. of valid (non missing) cases
No. of missing values
Fill Rate
Count of Unique Values
No. of Levels
Mean
Std deviation
Median
Mode
Minimum
Q1(25th percentile)
Q3(75th Percentile)
99th Percentile
Maximum

Data Profiling: - Bivariate profiling assist in finding the frequency of each categorical variable with respect to the response variable. This would facilitate in binding/ grouping the categories which have same response rate so that the effect of that particular category can be captured in the model.

We prepare a report with following fields to further work on it.

Variable
Field categories
Frequency
Responses
Responses %

Data Treatment:

The first stage of any statistical modeling consists of data treatment activities. Approximately 80% of the entire modeling time is consumed by the data treatment techniques. Here we check the hygiene factor of our independent variables and try to make the data as exploitable as possible.

Before going to data treatment one has to find out the correlation between variables means finding the relationship of predictors with the response variable and also to find out the inter correlation among predictors. From this analysis we can exclude some of the predictors which are not important for the model building based on the significant correlation values. The first step of variable reduction happens in this stage and next is on basis of multicollinearity check. The variables selected from this step will undergo for further data treatment like missing value, extreme value treatment and multicollinearity check.

Missing value treatment: -

We should check that the independent variables have sufficient information to establish a significant relation with the dependent variables.

Some of the basic rules for Missing value treatment are as below:

1) If the independent variables have a large amount of missing value (More than 40%-50%), we drop that independent variable from our analysis, since no relation can be established between that independent variable and the dependent variable in question.
2) If the percentage of missing value lies between 10% -40%, we try to establish a separate relation between the dependent and independent variables to understand any hidden pattern.
3) If our predictors are categorical variable, then we can make that missing values as one category but we will miss the information since that category will not comes significant in the model so better treat the missing value with those category which has highest frequency among all the categories of a variable.

4) For quantitative independent variable, treat the missing values with central tendency like mean, median and mode value of that variable.
5) Various other methods like exploration, regression method etc. has also been used to treat the missing values.

Note: These missing values are represented by dots ('.') for numerical variable and blank for categorical variables in the SAS.

Extreme value treatment: - This step is done to understand the distribution of our independent variables. Presence of an abnormal value in one of the independent variables can affect the entire analysis (Leverage variable). The extreme values are treated by capping. Capping is required as sometimes, the variable may contain extreme values corresponding to some observations; whereas in reality, such values are unlikely to exist. Such values may be a result of wrong keying the data or may represent the default values. There may be cases of negative values which is logically incorrect for a given variable. In such cases, these values need to be capped because they may affect the mean of the variable drastically.
Some basic rules for capping are as below:

- Don't cap the value unless it is unrealistic
- Cap it to the next highest/lowest (realistic) value.
- Cap at a value so that the continuity is maintained.

Multicollinearity - When two or more independent variables are related between them, we tell that they have multicollinearity among each other. In technical terms, we say the one variable can be explained as a linear combination of other variables. Multicollinearity among independent variables does not allow the independent variables to explain their impact on the dependent variable optimally due to high internal impact. Keeping collinear variables in the model makes it unstable. In such a scenario we drop one of the variables from the model. Multicollinearity among the variables indicates that these explanatory variables are carrying a common set of information in explaining the dependent variable.

Detection of multicollinearity:

Variance Inflation Factor: We generally test the presence of multicollinearity using Variance Inflation Factor (VIF). Variance Inflation factor (VIF) is obtained by regressing each independent variable, say X on the remaining independent variables (say Y and Z) and checking how much of it (of X) is explained by these variables. Hence VIF = $1 / (1 - R^2)$.
If the VIF value be greater than 2 we drop the variable from the model.

Steps in multicollinearity check:

- Add all the independent variables in the model to explain the response
- Check for the variable which has highest VIF
- Keeping in that variable in mind, go to correlation table.
- Identify the variable (has highest VIF) has the highest correlation with some other variable.
- Drop the variable with higher VIF and repeat the procedure till you get the VIF < 2

Bi-variate analysis and variable reduction technique:

Perform the frequency analysis of each category in the data against the response variable (dependent variable).

The final analysis should prepare a report highlighting the Cumulative IV values as below:

variabl	Factor	Non-Re	Res	NER	ER	WOE	IV	CumIV	Flag
Attr10	10	623	1	0.103868	0.0005	-5.33727	0.551705	2.971843	1
Attr11	10	637	2	0.106202	0.000999	-4.66634	0.490914	2.092096	1
Attr6	Unmarrie	2100	869	0.350117	0.434066	0.21493	0.018043	0.562534	1
Attr7	7	838	318	0.139713	0.158841	0.128313	0.002454	1.797204	1
Attr9	10	583	2	0.097199	0.000999	-4.57776	0.440381	1.631267	1

Rebinding can be done based on the event rate by generating the heat map:

variable	Factor	Non-Res	Res	NER	ER	WOE	IV	CumIV	Flag	Rebins
Attr10	3	601	594	0.1002	0.296703	1.085564	0.213317	0.284603	0	1
Attr10	5	575	412	0.095865	0.205794	0.763933	0.083978	0.431405	0	2
Attr10	2	611	409	0.101867	0.204296	0.695898	0.07128	0.071286	0	2
Attr10	4	612	395	0.102034	0.197303	0.659433	0.062823	0.347427	0	2
Attr10	1	559	185	0.093198	0.092408	-0.00851	6.73E-06	6.73E-06	0	3
Attr10	7	579	2	0.096532	0.000999	-4.57088	0.43667	1.444133	0	4
Attr10	8	594	2	0.099033	0.000999	-4.59645	0.450609	1.894742	0	4
Attr10	6	646	1	0.107703	0.0005	-5.37352	0.576058	1.007463	0	4
Attr10	9	598	1	0.0997	0.0005	-5.29631	0.525396	2.420138	0	4
Attr10	10	623	1	0.103868	0.0005	-5.33727	0.551705	2.971843	1	4
Attr11	2	586	423	0.097699	0.211289	0.771332	0.087615	0.088778	0	1
Attr11	5	585	407	0.097533	0.203297	0.734481	0.077682	0.279626	0	1
Attr11	4	613	391	0.102201	0.195305	0.647622	0.060296	0.201944	0	1
Attr11	3	612	377	0.102034	0.188312	0.612792	0.05287	0.141648	0	1
Attr11	6	601	212	0.1002	0.105894	0.055271	0.000315	0.279941	0	2
Attr11	1	615	184	0.102534	0.091908	-0.10941	0.001163	0.001163	0	2
Attr11	7	584	2	0.097366	0.000999	-4.57947	0.441309	0.72125	0	3
Attr11	8	564	2	0.094031	0.000999	-4.54463	0.422797	1.144047	0	3
Attr11	9	601	2	0.1002	0.000999	-4.60817	0.457135	1.601183	0	3
Attr11	10	637	2	0.106202	0.000999	-4.66634	0.490914	2.092096	1	3
Attr6	Unmarrie	2100	869	0.350117	0.434066	0.21493	0.018043	0.562534	1	1
Attr6	Married	3742	708	0.623875	0.353646	-0.56765	0.153396	0.153396	0	2
Attr6	Other	156	425	0.026009	0.212288	2.099513	0.391095	0.544491	0	3
Attr7	5	836	580	0.13938	0.28971	0.731679	0.109994	1.79399	0	1
Attr7	4	894	563	0.14905	0.281219	0.634853	0.083908	1.683997	0	1
Attr7	7	838	318	0.139713	0.158841	0.128313	0.002454	1.797204	1	2
Attr7	3	877	276	0.146215	0.137862	-0.05883	0.000491	1.600089	0	2
Attr7	6	849	263	0.141547	0.131369	-0.07463	0.00076	1.79475	0	2
Attr7	1	842	1	0.14038	0.0005	-5.6385	0.788717	0.788717	0	3
Attr7	2	862	1	0.143715	0.0005	-5.66198	0.81088	1.599597	0	3
Attr9	3	599	413	0.099867	0.206294	0.725465	0.077209	0.080214	0	1
Attr9	5	592	403	0.0987	0.201299	0.712709	0.073123	0.216849	0	1
Attr9	4	618	399	0.103034	0.199301	0.659752	0.063512	0.143726	0	1
Attr9	2	619	228	0.103201	0.113886	0.09852	0.001053	0.003005	0	2
Attr9	7	594	191	0.099033	0.095405	-0.03733	0.000135	0.217296	0	2
Attr9	6	590	186	0.098366	0.092907	-0.0571	0.000312	0.217161	0	2
Attr9	1	612	177	0.102034	0.088412	-0.1433	0.001952	0.001952	0	2
Attr9	9	606	2	0.101034	0.000999	-4.61645	0.461805	1.190886	0	3
Attr9	10	583	2	0.097199	0.000999	-4.57776	0.440381	1.631267	1	3
Attr9	8	585	1	0.097533	0.0005	-5.27433	0.511784	0.729081	0	3

Data Modeling:

Once the data treatment is over we go ahead with the model building;

Development (Training) and Validation (Testing) Sample: -Before building the model on the data, divide entire data in the ratio of 70:30 as development sample and validation sample. Development sample is used to develop the model whereas validation sample will be used to

check validity of the model. Build the model on development sample data and use the estimates obtained from this model to score the validation sample. If the response captured from the validation sample is nearly equal to the response captured from the development sample then we can say that the model is robust in predicting the responses for future dataset.

Logistic Regression Model Building:-Logistic regression technique is used to assess the impact of independent variables and probability of event of interest. The approach is explained in the following steps

Ø Create the dummies and slope dummies for the categorical independent data if desired. Because most of the statistical analysis tools like E-Guide, E-Miner will take string variables directly. So in such cases no need to create dummies and here one of the categories will be converted into base category for comparison.

Ø Data transformation for continues independent variables if necessary.

Model Fit Criteria:

1. Use the Deviance or Hosmer & Lemeshow test statistics to check the validity of the model. Higher the "P" value better is the model. Proceed to next steps only if we have higher value of P.

2. Test the null hypothesis for the independent variables, i.e. all $\beta = 0$. P value should be significant (i.e. $p < 0.05$) to reject the null hypothesis and prove that β values are not equal to 0.

3. Check the concordance and Tie. The rule of thumb test is (Concordance+ ½ Tie) should be greater than 60%.

4. Check the significance of the estimates of each of the variable. If any of the estimates are not significant, variable with highest P value will be dropped and steps i to vi are repeated with the new set of variables. This process will continue until all the variables in the model have significant estimates.

5. Frame the equation with the significant variables. Odds ratio and probability value for each of the profile is estimated.

6. Specificity and Sensitivity of the model is assessed and ROC (Receiving Operating Characteristic) graph is plotted. Area under the ROC is an indication of how well the classification of good in to good and bad to bad is decided by the identified model.

7. Coefficient Stability: Coefficient stability is checked across development and validation sample. Once the model is performing satisfactorily on development sample, we use the same set of variables to model the validation sample. A robust model should perform equally well on validation sample too. Hence, the coefficients should be in a close range and should be of same sign.

8. Concordance: Consider a set of 100 individuals out of which 10 are the responders (denoted by 1) and 90 are non-responders (denoted by 0). Now we construct pairs for each responder with every non-responder. Hence, we get 900 such pairs (10*90 = 900). Using the model under development, we calculate the predicted response rate for each responder and non-responder in every pair. If responder's predicted probability is greater than non-responder's predicted probability, then the pair is concordant. If it is vice versa, then the pair is discordant and if both are equal, then the pair is tied. For a good model, the percent concordant pair lies above 65%.

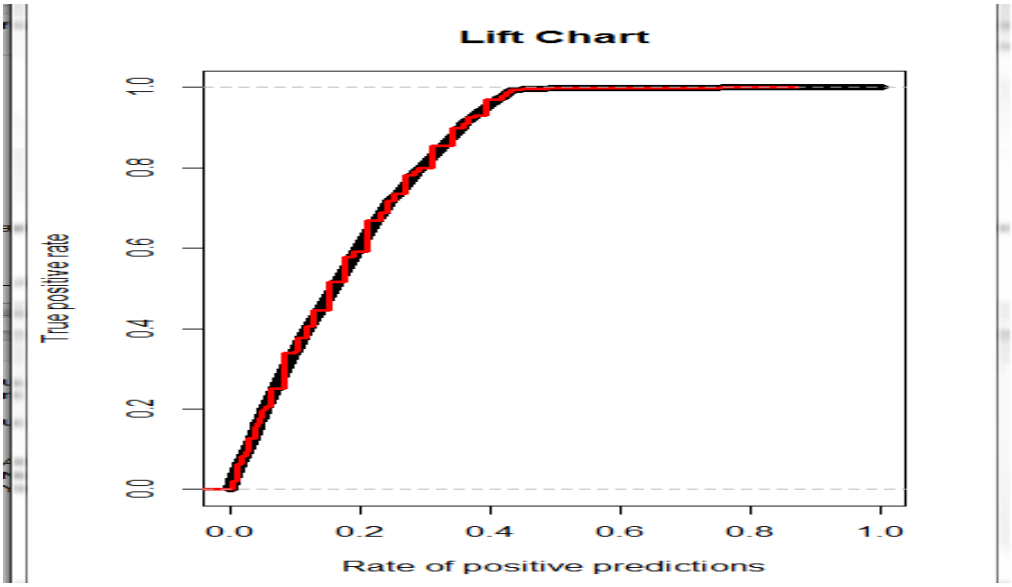
9. Gini Coefficient: The Gini coefficient is one which is used to test the model accuracy. It is calculated by using following formula. For good model the Gini coefficient should be in the range of 40-60%.

Gini=2C-1 Where C= Area under the curve (ie Concordance+1/2 of Tie)

10. Scoring: Satisfaction of the model comes when the model is doing well in terms of rank ordering, coefficient stability, Goodness of fit, Concordance and capturing both on development and validation samples.

Now, take the coefficients of variables obtained from a model run on development sample and use it to predict response rate of validation sample. This method is known as scoring of the model. Scoring provides a good idea about how the model will perform when applied to another data set. Here, we are concerned about the capturing of the responders, say in first 40 % of the population.
The model is used to predict the response rate for a set of new data is taken from a different time frame to test the validity of the rules suggested by the model. The model will be applicable to the profiles similar to the once already present in the sample data used for model development. Model validation is performed by taking the optimum threshold level of probability.

Lift/gains chart for churn model case:



The model is implemented and refreshed periodically to generate scores for the customers.