MIS784 – Marketing Analytics – Uses of and the Difference between Logistic Regression and RFM Models in Marketing Analytics

- Karan Thakker

Introduction:

In today's world, almost everything runs on data. How stores are laid out, what customers see as soon as they enter the store, every text message, e-mail, advertisement is sent to one based on "data"

In the case of Runte PLC, their "loyalty program" is not only beneficial to the customer by giving them certain discounts, but beneficial to the organization by providing them with several datapoint, which would help them to run customized offers – i.e if a certain customer has a higher average spending in the bakery section, the algorithm will send that customer more tempting offers in the bakery section.

Many key decisions can be made based on such data models, as addressed throughout the literature review and the report.

The report analyses the data by logistic regression and RFM models and points out the differences (if any) between the accuracy of the two models.

Literature Review:

While carrying out marketing analytics, a lot of times analysts come across "data". This data might contain several independent variables – and certain models must be used in order to make sense out of a vast amount of data. This literature review talks about 2 techniques that can be used to understand the probability of "churners" while considering certain datapoints that have been provided. These two data analytics techniques are – "linear regression" and "Recency, Frequency and Monetary (RFM) modelling"

Companies greatly benefit from using certain data analysis techniques to gain certain knowledge about their customers. This data can help the companies a great amount to increase their customer retention, and understand exactly where they face shortcomings, in order to fix them.

(Liang, et al., 2017)

This journal entry's purpose checks if the high health benefits outweigh the disadvantages of the high pricing of organic food. It is an attempt to understand if organic food customers are more responsive to a specific type of promotional campaign. The studies concluded that the use of effective sales promotional had a large impact on customers who were in a decision-making process. Effective campaigns essentially converted churners into non-churners, and a linear regression model was of great help to predict non churners and churners in this case.

(Kiguchi, et al., 2022)

Education Technology and Digital game-based learning (DGBL) industries face a large problem – Extremely high churn rates. In order to address the high churn rate, churn prediction is an important aspect for companies in this industry. The dataset used in the journal is from a popular Japanese DGBL company.

Churn prediction is extremely important in order manage retention of customers in a highly competitive industry, and the researchers used techniques such as Linear Regression (and 2 more techniques) in order to predict customer churn. This allows companies to better target their marketing campaigns towards customers who are churners. This makes learners understand the importance of models that predict customer churns in businesses and marketing analytics.

(Rahim, et al., 2021)

In this journal, the researchers applied RFM techniques and modelling techniques in order to better understand customer behaviours. The research uses various machine learning techniques, along with RFM in order to segment customer in the studies, and it showed that the RFM attributes were of a highly persuasive nature.
This study showed that RFM attributes play an extremely important part in predicting if a customer will churn or not churn.
(Asllani & Halstead, 2015)

The following journal investigates if a vast amount of data can be used to target their advertisements at customers using a data-analytical approach (the RFM model)
This research was carried out to find inappropriate and appropriate RFM segments based only on profitability, marketing objectives and budget constraints.
The research can be used as a model to enable the transformation of purchasing history data (RFM) into a useful decision model that can be applied to any market situation and even a scenario when a tight budget has been imposed.
(McCarty & Hastak, 2007)

These linear regression techniques have been compared with one another in the past.
The study does an investigation of RFM, Linear regression and CHAID methods on 2 different datasets, one from a multi-division catalogue marketer and the other of non-profit organization that has made a recent solicitation for donations from its members. Both studies were carried out on data that was randomly taken from the entire sample.
The findings from this journal state that: In the second study, RFM was effective across all tests, and performed as well as the other two techniques, however in the first study, when the response rates were low, and the sample size was relatively smaller (30%), RFM was a bit more inaccurate compared to the other methods.
However, the journal also suggests that RFM is relatively easier, cost and time efficient compared to the other two methods.
(Coussement, et al., 2012)

This journal investigated how problems with data accuracy affected RFM analysis, Logistic regression and decision trees methods of data modelling.
How this journal became relevant to the learner's literature review is that it also compared the RFM and Logistic regression models under several different scenarios.
According to the analysis, the sensitivity of these two models directly depended on the response rates in the datasets.
Logistic regression seemed less or equally sensitive to data inaccuracies in low response rate scenarios, but highly sensitive in high response rate scenarios, but still quite comparable to one another.

The information provided can be seen in various parts of the report, making readers understand how models like Logistic regression and RFM models can be used in real life situations.
The information provided in these literature reviews are quite comparable to the results found in the data analysis, showing the accuracies of both models to be quite close to one another, only different from one another by razor-thin margins.

The Dataset and Variables:

The dataset given to the learner is of a multinational supermarket chain called Runte PLC, which is headquartered in Hertfordshire, England.
Runte launched its loyalty scheme, the Runte Clubcard in 2004, which allows them to collect a vast amount of data about its customers transactions while shopping at their stores.
The rewards program allows the company to collect various datapoints such as : customer ID (unique identification), Number of Purchases, t.last (time between first and last purchase during observation period), t.active (time between first purchase and last day of observation period), customer spendings across 9 different categories, Loyalty (Customer loyalty level) , service failures, their socio-economic status (spending capacity of each customer) and churn (1 – they will defer from Runte PLC, 0 – They will continue to be Runte PLC's customers).
"Churn" is the only dependant variable in the data, while all other 16 variables are independent. The dataset contains 30,000 rows of customer data that needs to be analysed.

Types of Analysis and Methodology:

Logistic Regression:

Logistic Regression is a type of analysis when the predicted outcome of the dependant variable is either a 1 and a 0, and it has to be compared with one or more independent, nominal variables.
Whenever a researcher considers a "multivariable problem", a logistic regression is the approach that one can use to describe the relationship of several independent variables to a dependant variable – "churn" in the case of our question. The reason for popularity of the logistic model is that they are designed to give an estimate of probability between 1 and 0, and that it can give out a graph describing the output while considering the effects of several factors. (Klein & Kleinbaum, 2010)

Methodology:

The learner first had to convert the categorical variable "Loyalty" into a numeric variable. Dummy coding was used to do this, where the base variable was platinum – this means whenever "Platinum" shows under loyalty, the dummy variable columns read out 0,0.
The Data was then divided into two parts, the first 20,000 rows were then copied onto a new sheet as values only, so that a logistic regression function could be run onto it. Before running the function, the learner deleted the "Candidate ID" column as it was only used to identify customers (irrelevant to the outcome) and "Loyalty" column as it was already dummy coded.
The learner had to download an add-in into Microsoft excel called "Real Statistics" before being able to run the function. The learner then had to use "Logistic and Probit" regression across all 20,000 rows and 17 columns while using the "Solver" analysis type.

Testing Methodology:

After the logistic and probit analysis, the learner receives and output of 17 coefficients – 1 a constant, and 16 others for each independent variable present in the dataset.
The learner multiplies uses b0 + sumproduct((b1:b16),(data in each row)) to find the logit value for each of the 10,000 rows, which will help one to find the estimated probability.
The formula to find the estimated probability for each of the rows is =exp(logit)/1+exp(logit).
If the estimated probability value for a row is greater than 0.5, the churn value is 1 and if the estimated probability value for a row is lesser than 0.5, the churn value is 0.
The actual meaning of this data can be analysed by using the confusion matrix. It is a simple way to test the accuracy of the logistic regression model. It uses "countifs" to make a matrix to count how many of the predicted churn values match with the actual churned values.

The model's performance is assessed by calculating the following values:

- Rate of Accuracy: n(Actual churn = Predicted churn)/Total number of columns
- Rate of Misclassification: n(Actual churn is not equal to predicted churn)/Total number of columns
- Rate of Sensitivity = n(Number of correctly predicted churners)/n(total churners)
- Rate of specificity = n(Number of correctly predicted non churners)/n(total non-churners)

Lift Analysis:

The lift analysis is used to compare the logistic regression model to a random baseline model. Lift analysis models are used to prove the fact that the proposed model i.e. the logistic regression model is more efficient than the random baseline model.
The actual churn values for each row are compared (in increasing percentages) with the random baseline model.

The Recency Frequency Monetary (RFM) Model:

An RFM model is a model that helps analysts measure customer lifetime value – which analysts derive from the purchase data of the customers. The 3 factors that are required to create an RFM model are:

- Recency: The time between their last purchase and one prior to it
- Frequency: The number of total purchases made by a customer
- Monetary: The money spent by the customer

(Chou & -Chen Chang, 2022)

Methodology:

First, the data for RFM had to be allotted.
Recency: t.active – t.last = time elapsed until the most recent purchase
Frequency: the number of purchases made by the customer is the frequency
Monetary: Average of spending across all the categories
These values were individually sorted in ascending order by recency, purchases and average spending and allotted numbers from 1-5. (2000 per number, as there are 10,000 rows in total)
Finally, these values were combined using the "concatenate" command which gave the learner the final RFM value.

Testing:

The accuracy of this method can be tested by comparing the churn rate in the RFM model to a random baseline model. Visualising this data next to each other will show the which model detects all churners more efficiently.

Results and Interpretation:

Logistic Regression Model:

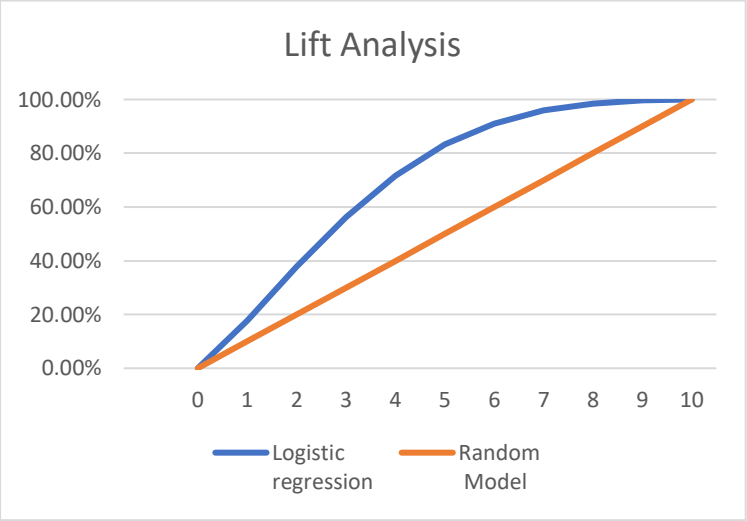
CONFUSION MATRIX			
	Predicted churn		
Actual churn	0	1	Total
0	4737	1012	5749
1	1117	3134	4251
Total	5854	4146	10000
MODEL PERFORMANCE			
Rate of Accuracy		78.71%	
Rate of Misclassification		21.29%	
Rate of Sensitivity		73.72%	
Rate of Specificity		82.40%	

The rate of accuracy of the logistic regression model is 78.71%, Which means it predicted 78.71% of the churners and non-churners correctly.
21.29% of the churners and non-churners were predicted incorrectly.
The rate of sensitivity is 73.72%, which means that many % of the churners were predicted accurately.
The rate of specificity is 82.40%, which means that many % of non-churners were predicted correctly.

Logistic Regression Model Vs Random Baseline Model

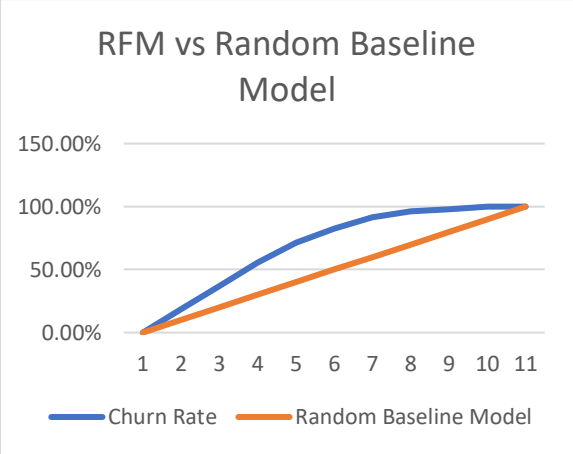
Number of decile	Logistic regression	Random Baseline Model
0	0.00%	0.00%
1	17.85%	10.00%
2	37.80%	20.00%
3	56.25%	30.00%
4	71.54%	40.00%
5	83.30%	50.00%
6	91.08%	60.00%
7	95.91%	70.00%
8	98.42%	80.00%
9	99.58%	90.00%
10	100.00%	100.00%
Total	4251	

The logistic regression clearly performs a lot better than a random model, predicting all of the churners a lot faster based on the lift analysis graph below:



The Recency, Frequency, Monetary (RFM) Model vs Random Baseline Model:

Number of deciles	Actual Churns	Churn
0	0	0.00%
1	792	18.64%
2	773	36.82%
3	801	55.67%
4	671	71.46%
5	480	82.75%
6	371	91.48%
7	201	96.21%
8	73	97.93%
9	78	99.76%
10	10	100.00%
Total	4250	



RFM vs Random Baseline Model

The churn rate of the RFM model is compared against the random baseline model in order to compare and visualize the efficiencies of one model against the other.
The performances of both, the RFM and Logistic regression model are almost similar, shown by them equally outperforming the random baseline model.

Limitations:

The RFM model is limited by the fact that the learner does not have a model that gives the learner values of recency, frequency, monetary organize them based on their importance.

Conclusion:

The assignment analyses the predicted customer churn using the Logistic Regression and RFM models. The literature review takes the reader through existing research about the topics of analytical models.
The methodology takes the reader through the exact steps used in the analysis and interprets and reflects on the results found in the analysis.
The findings in the analysis conclude that neither one of the analytical models are better at predicting the customers in this case of data.

Future research must be carried out across various levels of responses for customers in order to include more scenarios that the research has been carried out.

- Bibliography

- Akinci, S., Kaynak, E., Atilgan, E. & Aksoy, . S., 2005. Where does the logistic regression analysis stand in marketing literature?. *Marketing Literature*, 537(5/6), pp. 537-567.

- Asllani, A. & Halstead, D., 2015. Effect of sales promotions. *Academy of marketing study journal*, 19(3), pp. 49-62.

- Chou, T.-H. & -Chen Chang, S., 2022. The RFM Model Analysis for VIP Customer: A Case Study of Golf Clothing Brand. *International Journal of Knowledge Management*, 18(1), pp. 2-3.

- Coussement, K., Van den Bossche, F. V. d. B. A. & De Bock , K. W., 2012. Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, 67(67), pp. 2752-2757.

- Kiguchi, M., Saeed, W. & Medi, I., 2022. Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. *Applied Soft Computing*, 118(108491), pp. 1-19.

- Klein & Kleinbaum, D., 2010. Introduction to Logistic Regression. *Statistics for Biology and Health*, pp. 5-6.

- Liang, A. R.-D., Yang, W., Ji Chen, D. & Chung, Y. F., 2017. The effect of sales promotions on consumers' organic food response. *Effect of sales promotions*, 119(6), pp. 1247-1262.

- McCarty, J. A. & Hastak, M., 2007. Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research* , 60(656-662), pp. 659-661.

- Rahim, M. A., Mushafiq, M., Khan, S. & Arain, Z. A., 2021. RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Customer Servies*, 61(102566), pp. 1-7.

- Safari, F., Safari, N. & Montazer, G. A., 2016. Customer lifetime value determination based on RFM model. *Marketing Intelligence & Planning*, 34(4), pp. 446-461