Predicting Customer Churn at Myer

Introduction

Myer (stylized as MYER) is a prominent Australian department store chain with a presence in all Australian states and territories. The company operates 56 department stores throughout Australia and offers a wide range of products, including clothing, footwear, cosmetics and fragrance, homewares, electrical items, furniture, toys, books and stationery, and travel goods. Myer also runs a two-level loyalty scheme, with silver and gold membership tiers, allowing them to collect valuable data on customer purchasing behaviour. This data enables Myer to create personalized offers and execute targeted customer retention campaigns.

In this case study, I have conducted various analyses and modelling tasks to predict customer churn and evaluate the effectiveness of the models and methods employed to retain Myer's valuable customers. I have analysed this customer data to predict customer churn using logistic regression and evaluate the performance of the predictive model with RFM and random selection methods. Customer churn refers to customers who stop purchasing from a business. Understanding and predicting churn is crucial for businesses to implement targeted retention strategies and maintain customer loyalty.

Literature Review

Binomial (binary) logistic regression is a regression method applied when the dependent variable is not continuous, but instead represents a state that can either occur or not, or falls into specific categories. Logistic regression is utilized to forecast discrete outcomes by considering both continuous and categorical variables. In cases where there are more than two categories for the dependent variable, multinomial logistic regression is employed to address such scenarios. (Mutanen, 2006) Logistic regression is inherently designed to be an interpretable model, as it allows for a straightforward examination of its coefficients to determine how much and in which direction each feature impacts the ultimate prediction. (Kamil Matuszelański, 2022)

The logistic regression model has demonstrated improved predictive performance in churn prediction. Enhancing the results can be achieved by raising threshold values and carefully selecting relevant features using different combinations. (Nikita Bagul, 2021)

The logistic regression model has become the preferred method for predicting binary outcomes. (Tian-Shyug Lee a, 2006) Logistic regression modeling is a widely recognized method that has significant appeal for several reasons: It offers a straightforward solution for calculating posterior probabilities, which is not the case with the probit model. It is user-friendly and delivers fast and reliable outcomes. (Wouter Buckinx, 2005)

The RFM (Recency, Frequency, and Monetary) model is a behavior-oriented approach employed to analyze customer behavior and subsequently make predictions based on the data within a database. These three variables are considered behavioral indicators and can be employed as segmentation criteria by evaluating customers' attitudes towards the product, brand, benefits, or even their loyalty using the database. (Jo-Ting Wei, 2010)

"Recency" is a time-based measure that reflects how recently a customer has completed a purchase. It is commonly regarded as the most influential predictor of future customer behavior among the RFM variables. (Miglautsch, 2000)

"Frequency" serves as an indicator of the depth of a customer's relationship with the company. It naturally measures behavioral loyalty, as a customer's loyalty is higher when they have made repeated purchases from the company. In other words, the more frequent the customer's purchases, the greater their loyalty. (V.L. Miguéis a, 2012)

"Monetary" represents the measure of the expenditure made by customers at a specific company. Generally, within the RFM framework, this dimension is considered to have relatively less predictive power, although it remains valuable when combined with the other two variables for prediction purposes. (Rud, 2000)

Despite the existence of more statistically advanced methods, RFM ranks as the second most widely utilized approach by direct marketers, trailing only cross tabulations in popularity. (John A. McCarty, 2007) When compared to statistical modeling, RFM may be considered a cost-effective and generally dependable method because it doesn't demand highly skilled personnel, which can add to the expenses. (Ronald G Drozdenko, 2002)

Methodology

Data Description

For the purpose of this case study, I have access to a dataset comprising 30,000 randomly selected Myer customers, split into a training set of 21,000 customers and a test set of 9,000 customers. This dataset covers customer information and behaviours between January 1st, 2018, and December 31st, 2018, and includes 17 variables.

Adopted Analytical Techniques

To construct a model for predicting customer churn in the Myer case study, I have adopted the logistic regression and RFM (Recency, frequency and monetary) models.

Model building process for logistic regression:

Logistic regression is a statistical method commonly used for binary classification tasks, such as predicting whether a customer will churn (1 for churner or 0 for non-churner). Here's an overview of the adopted analytical technique:

- Data Splitting

  The training set has been used to train the logistic regression model, while the test set has been used to evaluate its performance.

- Feature selection

  The logistic regression model incorporates all available features. It assigns interpretable coefficients to each feature, signifying their influence on predicting churn. Features with small coefficients have minimal impact on the model when applied to the test data.

- Model Building

  The dependent variable (Churn) is modelled as a function of independent variables (features) using the logistic function. The output is a probability score between 0 and 1, which represents the likelihood of a customer churning.

- Model Training

  During the training phase, the model deduces coefficients for each feature that maximize the likelihood of the observed churn patterns. This logistic regression model is applied to the training dataset to compute the coefficient values for all variables, which are subsequently utilized on the test data to calculate estimated probabilities. Customers are categorized as either 1 (churners) or 0 (non-churners) based on whether their estimated probability exceeds 0.5 or not. Additionally, the model provides a covariance matrix, convergence information, a classification table illustrating successful and unsuccessful predictions and observations, as well as an ROC table and curve.

- Model Evaluation

  Evaluation of performance of the logistic regression model on the test set is done using metrics related to the confusion matrix. These metrics include accuracy, misclassification, sensitivity and specificity rate

Model building process for RFM (Recency, Frequency, Monetary) Analysis

The RFM method is a traditional approach to segment and target customers based on their recent purchase behaviour, purchase frequency, and monetary value. While it's not a predictive model like logistic regression, it can complement the churn prediction efforts. Here's an overview of the adopted analytical technique:

- Recency (R)
  Calculation of the "Recency" score for each customer, representing how recently they made a purchase. The time gap is calculated between the last purchase date and the end of the observation period, December 31, 2018 using the variables T.active and T.last. A score is given based on this recency value, where a lower score indicates a more recent purchase (e.g., 1 for the most recent, 5 for the least recent). The customers are divided into 5 recency segments each containing 20% customers.

- Frequency (F)
  Calculation of the "Frequency" score for each customer, representing how often they make purchases. The total number of purchases made by each customer during the observation period is used as the frequency metric. Assign a score based on this purchase frequency, where a lower score indicates more frequent purchases (e.g., 1 for the highest frequency, 5 for the lowest frequency). The customers are divided into 5 frequency segments each containing 20% customers.

- Monetary (M)
  Calculation of the "Monetary" score for each customer, representing how much they spend using sum of the total spending of each customer across all product categories during the observation period. Assign a score based on this total spending, where a lower score indicates higher monetary value (e.g., 1 for the highest spending, 5 for the lowest spending). Again, segment customers into 5 monetary segments each containing 20% customers.

RFM Segmentation

The RFM (Recency, Frequency, Monetary) scores were combined to create an RFM segment for each customer, resulting in a three-digit code, such as "231," where 2 represented recency, 3 represented frequency, and 1 represented monetary value. Further evaluation was conducted on the spending patterns of the top 10%, 20%, and 30% of customers. The test data was then sorted based on the RFM code, ranging from 555 to 111.

Lift Chart

A lift chart, also known as concentration, was used to compare the effectiveness of a predictive model (logistic regression in this case) against RFM and the random model. It showed the concentration of churners in each segment.

- Customers were divided into 10 deciles (e.g., top 10%, 20%, 30%, etc.), based on their rankings.
- Churn rates were calculated for each percentile group by determining the proportion of churners within that group. The logistic regression calculates the cumulative percentage on the basis of sorting done by the estimated probability in descending order. The RFM Model shows the cumulative percentage on the basis of the RFM code generated in descending order from 555 to 111.
- The lift chart was created with the x-axis representing the percentile groups and the y-axis representing the churn rate.

The lift chart shows how much better the model (logistic regression or RFM) or the random model is at identifying churners compared to random chance. It also indicates which approach performed better across different percentiles.

Results

Logistic regression model

The coefficient values obtained from the logistic regression model represent the strength and direction of the relationship between each predictor variable and the likelihood of customer churn. Here's a concise interpretation of the results affecting churn:

- The most influential predictor is "Loyalty" with a negative coefficient of -0.5301. This suggests that gold loyalty members are significantly less likely to churn compared to silver loyalty members, indicating the importance of loyalty programs in retaining customers.
- The second significant variable is "Total Profit" with a positive coefficient of 0.0734. Higher total profits generated by customers during the observation period are associated with a decreased likelihood of churn, emphasizing the value of high-profit customers.
- "Purchase" has a negative coefficient of -0.0444, indicating that as the number of purchases increases, the probability of churn decreases slightly. This suggests that active customers with frequent purchases are less likely to churn.

- The variable "CF.spent" (spending on Cosmetics and Fragrance) also plays a role, with a negative coefficient of -0.0259. Higher spending in this category is associated with a reduced likelihood of churn.
- Lastly, "T.last" (the time gap between the first and last purchase) has a negative coefficient of -0.1113. Customers who made their last purchase more recently are less likely to churn, underlining the importance of recent engagement.

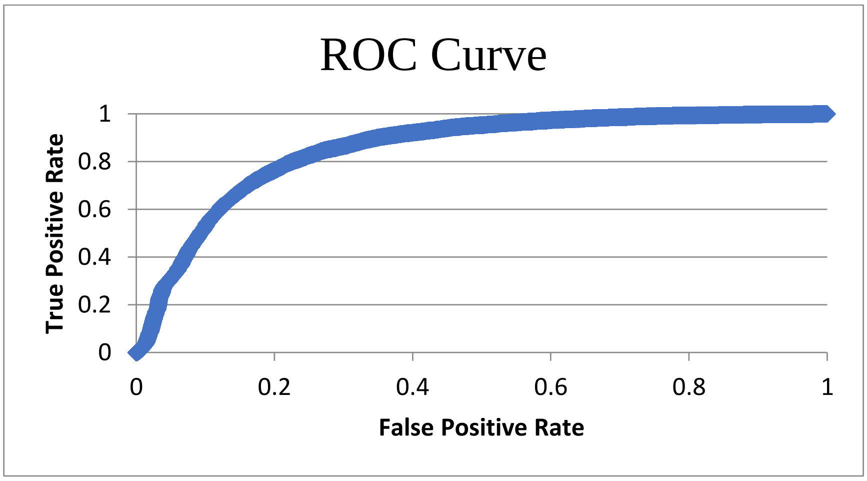
Performance Metrics

Accuracy Rate: The accuracy rate of 80.12% suggests that the logistic regression model performs well in overall classification, correctly predicting a majority of customer outcomes.

Misclassification Rate: The misclassification rate of 19.88% indicates that there is room for improvement in the model's predictive accuracy. Reducing misclassifications would be beneficial to Myer's retention efforts.

Sensitivity Rate (True Positive Rate or Recall): The sensitivity rate of 74.49% means that the model is reasonably effective at identifying customers who are likely to churn. This is crucial for targeted retention strategies, as it identifies a significant portion of actual churners.

Specificity Rate (True Negative Rate): The specificity rate of 84.43% signifies that the model is good at correctly identifying customers who are not likely to churn. This minimizes the risk of expending resources on customers who are unlikely to leave.

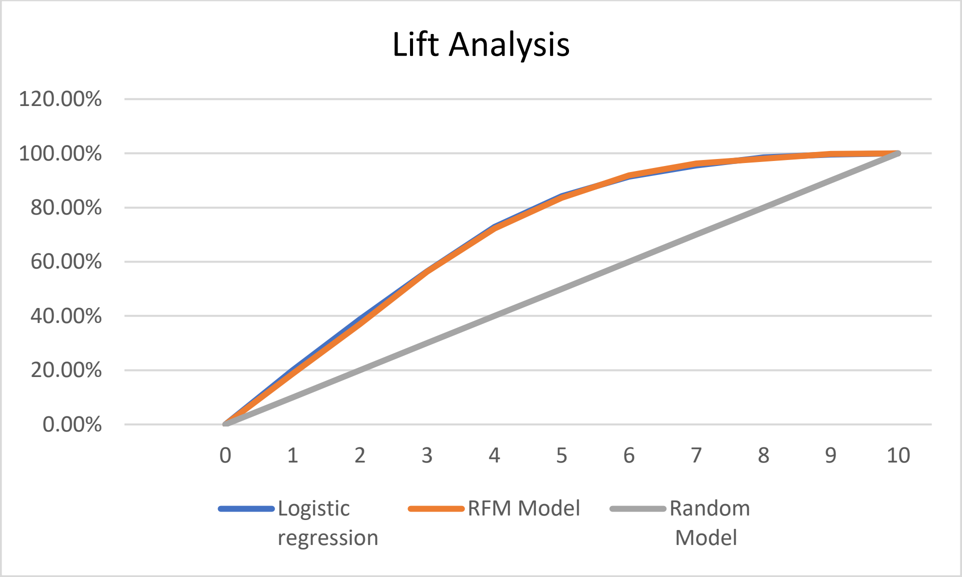


The ROC Curve assesses a classification model's ability to differentiate between positive and negative cases. An ideal ROC curve achieves a true positive rate of 1 and a false positive rate of 0. The provided graph suggests that the classifier is performing effectively, accurately distinguishing between positives and negatives. The greater the deviation of the curve from the diagonal line, the stronger the model's discriminatory power.

RFM model

The total spending generated from 9000 customers combined is $ 1411330.89. The total spending of the top 10% of the customers is $442218.22. The customers in this segment are those with the highest RFM scores being the most loyal customers. The top 20% customers contribute $6,17,415.78 and the top 30% contributes $9,20,722.

Lift Analysis



- Logistic Regression
  At the 1st decile, this model concentrates approximately 19.97% of the outcomes. By the 9th decile, it captures 99.62% of the outcomes.

- RFM Model
  The RFM model follows a similar trend to logistic regression, with 18.71% of outcomes concentrated in the 1st decile and 99.74% in the 9th decile.

- Random Model:
  The random model, as expected, distributes outcomes more evenly across deciles. It allocates 10.00% of outcomes to the 1st decile and 100.00% by the 10th decile.

The Logistic Regression and RFM Model exhibit similar performance, with the Logistic Regression model performing slightly better at some deciles. The random model, on the other hand, does not provide meaningful predictions and merely captures churners randomly, which is not useful for practical purposes.

Conclusion

Based on the analysis conducted several key conclusions can be drawn:

- Model Performance: Both the logistic regression model and the RFM model show promise in predicting customer churn. They outperform a random model and are effective at identifying high-risk customers. The cumulative concentration analysis, especially the lift chart, demonstrates the superiority of the logistic regression and RFM models in concentrating churners in the top deciles. This means that these models are effective in identifying customers who are more likely to churn.
- Key Predictors: In the logistic regression model, loyalty level, total profit, purchase frequency, spending on cosmetics and fragrance, and recency of purchases were identified as key predictors of churn. These factors can help Myer in targeting and retaining customers effectively.
- Segmentation with RFM: The RFM model, while not a predictive model, provides valuable insights into customer segments based on recency, frequency, and monetary factors. This can aid in tailoring retention strategies for different customer groups.

Recommendations

Based on the analysis the following recommendations can be made:

- Targeted Marketing Campaigns:
  Instead of running broad marketing campaigns, Myer should focus on targeted marketing efforts. Utilize the insights gained from the logistic regression model to identify high-risk customers who are more likely to churn.
- Loyalty Program Enhancement:
  Given the significant influence of the "Loyalty" variable in the logistic regression model, Myer should invest in enhancing its loyalty program. Consider providing exclusive benefits, discounts, or rewards to silver and gold members, with a stronger focus on gold members who have shown to be less likely to churn.
- Retention Messaging:
  Develop targeted retention campaigns that address the specific needs and concerns of customers who are at risk of churning. Send personalized emails or notifications with messages emphasizing the value of their continued relationship with Myer and offering incentives to stay.
- Customer Segmentation:
  Continue using the RFM model to segment customers based on recency, frequency, and monetary value. Identify the segments that contribute the most to revenue and profit (e.g., the top 10% or 20%). Focus marketing efforts on retaining customers within these high-value segments.
- Monitoring Churn Risk:
  Implement a system for real-time or periodic monitoring of customer churn risk. This proactive approach allows Myer to identify potential churners early and take immediate action to retain them, reducing the need for extensive marketing campaigns.
- Feedback Mechanisms:

Encourage feedback from customers who have churned to understand the reasons behind their decision. Use this feedback to make necessary improvements in products, services, or the overall customer experience, addressing the root causes of churn.

- Data Analytics and AI:
  Explore advanced data analytics and AI techniques to gain deeper insights into customer behaviour and churn prediction. Machine learning models can be refined over time to improve accuracy and identify subtle patterns.

References

John A. McCarty, M. H. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0148296306002323?via%3Dihub

Jo-Ting Wei, S.-Y. L.-H. (2010). A review of the application of RFM model. *African Journal of Business Management*. Retrieved from https://academicjournals.org/article/article1380555001_Wei%20et%20al.pdf

Kamil Matuszelański, K. K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. Retrieved from https://www.mdpi.com/0718-1876/17/1/9

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management* . Retrieved from https://link.springer.com/article/10.1057/palgrave.jdm.3240019

Mutanen, T. (2006). *Customer churn analysis – a case study.* Technical Research Centre of Finland. Retrieved from https://publications.vtt.fi/julkaisut/muut/2006/customer_churn_case_study.pdf

Nikita Bagul, P. B. (2021). *Retail Customer Churn Analysis using RFM.* International Journal of Engineering Research & Technology (IJERT). Retrieved from https://www.researchgate.net/profile/Priya-Surana/publication/366248713_Retail_Customer_Churn_Analysis_using_RFM_Model_and_K-Means_Clustering/links/63997a27484e65005b0a46d2/Retail-Customer-Churn-Analysis-using-RFM-Model-and-K-Means-Clustering.pdf

Ronald G Drozdenko, P. D. (2002). *Optimal Database Marketing: Strategy, Development, and Data Mining.* Sage Publications. Retrieved from https://books.google.com.au/books?hl=en&lr=&id=hEvghMsfl8AC&oi=fnd&pg=PR15&dq=Drozden#v=onepage&q&f=false

Rud, O. (2000). Data mining cookbook: Modeling data for marketing, risk and customer relationship management. *Wiley*. Retrieved from https://books.google.com.au/books?hl=en&lr=&id=HYveCsbG3ZoC&oi=fnd&pg=PT8&ots=lGWNdrm7R4&sig=PnkCDL_vnEuTuPDGC5eed67TaTA&redir_esc=y#v=onepage&q&f=false

Tian-Shyug Lee a, C.-C. C.-C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*. Retrieved from https://www.sciencedirect.com/science/article/pii/S016794730400355X

V.L. Miguéis a, D. V. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417412005969#b0125

Wouter Buckinx, D. V. (2005). *Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting.* European Journal of Operational Research. Retrieved from https://www.sciencedirect.com/science/article/pii/S0377221703009184