Decision Tree: Customer Churn Analysis for a Loyalty Program using Data Mining Algorithm

Angela Siew-Hoong Lee [1[0000-0003-3388-2372]], Ng Claudia[2], Zuraini Zainol[3[0000-0002-6881-7039]] and Khin-Whai Chan[4]

[1,2,4] School of Science and Technology, Department of Computing and Information Systems, Sunway University, 47500 Selangor, Malaysia
[3] Department of Computer Science, Faculty of Defence Science and Technology, National Defence University of Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia

angelal@sunway.edu.my

Abstract. In the world of retailer, customers typically patronize multiple shops thus making loyalty programs a favorite among retailer to retain their customers. Loyalty programs are utilized across many different businesses as a marketing strategy to encourage customers to continuously shop or patronize the services provided by a certain organization. However, one of the biggest problem faced by these businesses is customer churn. The purpose of this research was to build a predictive model, which could predict customer churn, where visualization of data was generated to better understand the existing members and see the patterns and behavior demonstrated by members of the loyalty program. Through these, meaningful insights about the businesses' analysis on customers could be gathered and utilized for better actions which could be taken to address the issues which the company faces. At the end, based on the issues found, strategies were proposed to address the issues found. For this research, SAS Enterprise Miner was used to perform predictive analysis while Tableau was used to perform descriptive analysis.

Keywords: Decision Tree, Customer Analytics, Churn Analysis, Loyalty Program.

1 Introduction

Loyalty program is also known as a reward program; it is often offered by a company to their customers. Through a loyalty program, customer may have early access to latest products, special promotions, or discounted items, which will motivate the customers to join the loyalty program. For a customer to be part of the loyalty program, they would typically register with their personal information and will be given a unique identifier like a member ID or membership card, which they would use when making a purchase. A customer exhibits customer loyalty when they consistently use your services over an extended period. Customer loyalty is crucial to a company because it can positively impact long-term profitability.

Some customers who became a member might sometimes cease usage of the loyalty card after a period, which indicates that the customer may have stopped using the services provided by the loyalty program, these customers are known as churners. Losing customers or an increasing customer churn rate is one of the major problems companies may face because churn of good customers imposes irrecoverable damages for the company [1]. This is one of the reason customer loyalty is important, because when customer loyalty is present, customer churn rate will also reduce which in turn, retains customers. Many companies these days try to focus on customer acquisition to gain more new customers. However, it costs five times more to acquire new customers than to retain existing customers because companies often spend fortune on advertisements [2]. Therefore, the focus should shift from customer acquisition to customer retention and preventing churn.

1.1 Research Objectives and Problem Statement

There have been many researches done on identifying factors which can affect customer churn rate in loyalty cards on multiple different industries. However, there is less attention paid to improving loyalty programs and reducing customer churn rate, which in turn increase customer retention rate [3]. When customer churn rate is prevented, more customers will continue to patronize retailer services. Because retaining a customer imposes significant less cost than acquiring a new customer, it is better to focus on customer retention instead of customer acquisition [3].

Many malls have started providing loyalty programs with a main goal, to retain customers and keep them coming to the mall often. Retaining customers not as easy as it seems, due to all customers having different liking and preferences. When customers decide if they should patronize your services and use the loyalty card, they would always want rewards that can benefit them in return. Rewards can come in several different forms such as discounts, special promotions, vouchers, freebies, etc. Companies often face problems when they try to predict customer's churn behavior to understand factors which might retain or drive a customer away from using their services. This is a tedious task because it would require a lot of effort to predict customer's behavior based on the demographic data.

The objective of this research is to build a predictive model to predict customer's churn behavior and to propose potential strategies. This way, the company would be able to make better operational decisions and provide better rewards to the customer's liking hence potentially improve their overall experience on the loyalty program.

2 Literature Review

2.1 Loyalty Program

Loyalty programs are so common these days that they seem to be anywhere wherever we go. Whether be it in supermarkets, cinemas, or even restaurants all seem to provide membership benefits or rewards programs. A loyalty program is designed to

improve customer's satisfaction and commitment [4], it is defined as a system that offers rewards to customers, in aim and goals to increase customer loyalty [5]. Loyalty programs utilize psychological principles by rewarding customers to achieve desired behavior of the customer, such as increased transactions or constant repurchases of the company's product [5].

Way back in the late 18th century, customer loyalty program seemingly started as "premium marketing" [6]. Green Shield stamps was one of the very first retail loyalty programs, they awarded stamps when a purchase is made at selected retailers, the stamps could later be redeemed for products, the aim was to encourage customers to make purchases at participating stores [7]. The strategy of using stamps as customer loyalty programs continued to lead on until the early 20th century when retailers began to introduce new ways to attract customers [6]. In year 1929, Betty Crocker introduced the box top program where customers collect box top clippings that can be accumulated and exchanged for items [7]. In year 1981, the modern-day loyalty program was launched by American Airlines known as "Frequent Flier" which was regarded the first full scale customer loyalty program in the era [6]. After that, card-based loyalty programs began to gain popularity in the 1990s because of the convenience it provides rather than collecting stamps or currency and it is still popular up till today.

Loyalty programs today come in many different forms such as collecting points which can be redeemed by the company itself or third-party companies, cash back rewards, member exclusive sales, extra discounts, or tier levels of benefit depending on the different level of membership. Benefits provided by customer loyalty programs can be divided into two types; tangible benefits and intangible benefits. Tangible benefits are benefits that are monetary, it comes in form of rebates, exclusive discounts or coupons while intangible benefits are non-monetary benefits [8].

2.2 Customer Loyalty and Satisfaction

Customer loyalty is the relationship between a customer and a seller after the primary transaction, it is known as "a commitment to repeatedly buy or patronize a particular product or service consistently, which causes repetitive same-brand or same-family brand purchases despite any influences that might cause switching behavior. Customer satisfaction happens when a person feels like his goals were achieved or attained [9] it can be measured by the customer's attitude towards the company's products or services [10]. Customer loyalty can also be known as an outcome measured by the frequency of repurchases made over a period. Traditionally, by just doing your best to satisfy a customer's needs is a way to achieve customer loyalty, building customer loyalty means to develop a positive relationship which mutually rewards each both the organization and the customer, this can be done by ensuring that the customers feel certain about the company's commitment to them in a way that the customers themselves have to know that they are important to the company and that they are always placed first. Countless companies and organizations have been finding ways and solutions to retain their existing customers and at the same time, attain and attract new customers [9]. Therefore, customer loyalty is prioritized in many organizations as

it is a very important aspect as it would impact long-term profitability in a positive way, succeeding in achieving customer loyalty directly leads to successful customer retention which is very beneficial to the long-term growth of the company. Customer loyalty can only grow successfully when a company is aware of how their customers evolve and change each day, and to ensure that the customers knows exactly what benefits they are receiving from the company, positive connection between customers and the company makes the relationship between the two more stable and weighted, preventing other companies from winning the customers over.

3 Research Methodology

Figure 1 shows the data mining process that consists of seven main processes: (i) literature review, (ii) business understanding, (iii) data collection, (iv) data understanding, (v) data cleaning, (vi) data analysis, (vii) prediction model and, (viii) proposed strategies. The explanation of each stages will be explained in the next paragraph.

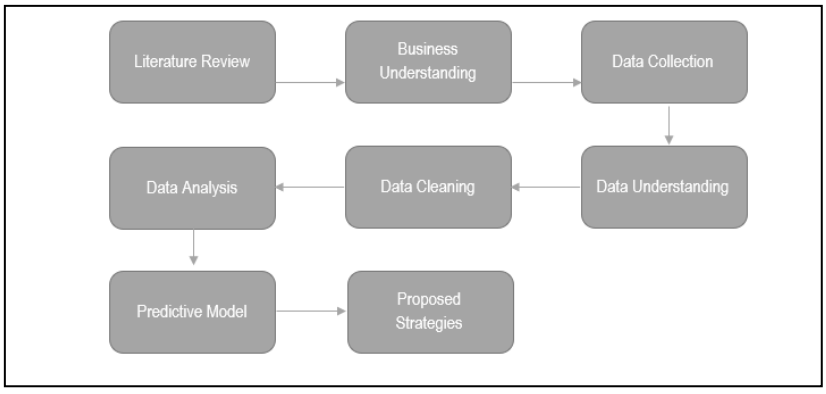

Fig. 1. Data Mining Process

A literature review was done to evaluate selected documents based on this topic to understand and identify the types of problems faced by other researchers while conducting the research, as well as to perceive the methods used by them to solve the problems. Based on the research done, it enabled us to broaden our understanding to perform this project. Business understanding allows us to identified and understand business problems. Data was collected from a loyalty card company. A total of four different datasets (CubeData, TransactionListing, MerchantListing & MemberListing) were given by the company. The duration of CubeData ranges from June 2016 – June 2017, TransactionListing has information for the entire year 2016, while MerchantListing and MemberListing contains the full data ranging from the beginning of the membership program until June 2017. Data understanding is to get familiar with the data to identify data quality problems and discover interesting insights about the

data, and to reveal relationships between subsets that can form hypothesis for hidden information. Data understanding is a crucial process to understand the requirements and the business. Because data provided is raw, it contained noisy data like irrelevant variables and outliers. Therefore, data cleaning is crucial and necessary. For this step, SAS Enterprise Guide was used to perform data cleaning. Data analysis was the next step which was performed using SAS Enterprise Miner. Data Analysis was divided into two parts; Descriptive Analysis and Predictive Analysis. Descriptive Analysis is where the data was understood and explain, graphical representation of data was generated and described. Predictive Analysis will be explained in the latter stages. In SAS Enterprise Miner, a predictive model was built to predict customer churn rate and their behavior. Decision tree was built based on the target variable set. Through the analysis and findings, proposed strategies were presented to the loyalvty company with valid explanation and reasons to justify each proposition, which could potentially impact the business.

4 Descriptive Analysis

The following models are based on the dataset CubeData which show the general overall statistics and information about the members. These graphical representations were generated using Tableau.

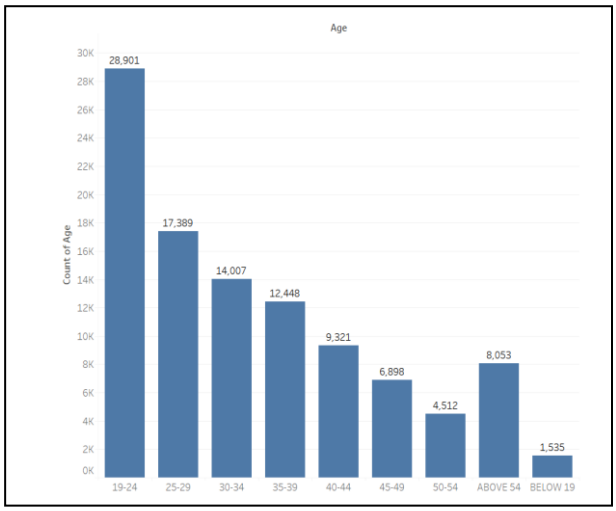

Fig. 2. Age Group of Customer

The above bar chart shows the overall age of customers (see Figure 2). It is observed that the number of customers is the highest with an amount of 28,901 at age ranged 19 – 24 which is the younger crowd, possibly students. From age range of 25 – 29 onwards the frequency of customers reduces steadily as the age increases until age

ABOVE 54 where a significantly higher frequency of customers was observed. Customers aged below 19 is the lowest at 1,535, this is because the company requires the customers to be at least 18 years old to register as a member. Hence, it explains the low number of customers in the age range as it only includes customers who are 18-19 years old.

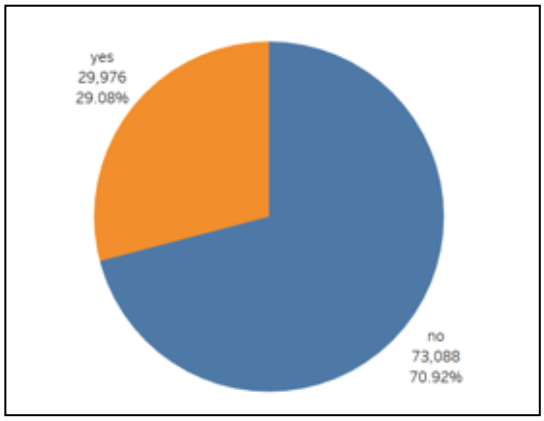

Fig. 3. Population of Churners and Non-Churners

Next, shows the population of churners and non-churners (see Figure 3). 'yes' indicates that they are churners while 'no' indicates that they are not churners. The members of this loyalty program are loyal members as there were only slightly more than a quarter of the population who have churned.

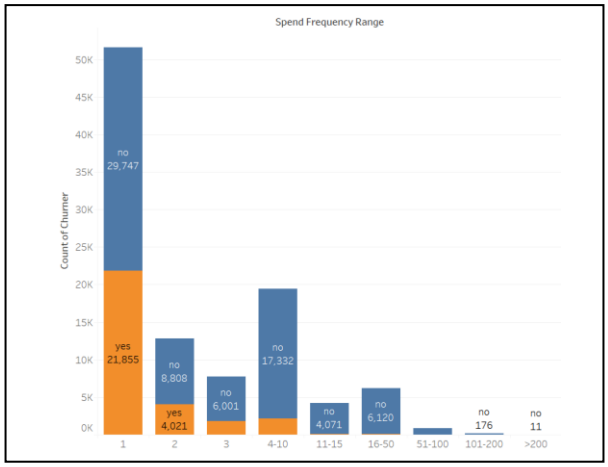

Fig. 4. Comparison of Churners vs Spend Frequency Range

When comparing Churner vs SpendFrequencyRange (see Figure 4), it is noticeable that most of the churners (yes) SpendFrequencyRange of 1, 2, 3, and 4-10 which means that they have only spent less than 10 times. Looking at SpendFrequencyRange of 1, it is almost a half-half distribution where there is a close number of churners and non-churners. This means that the members under the 'no' category could be new members while the members in the 'yes' category never continued using the card after once.

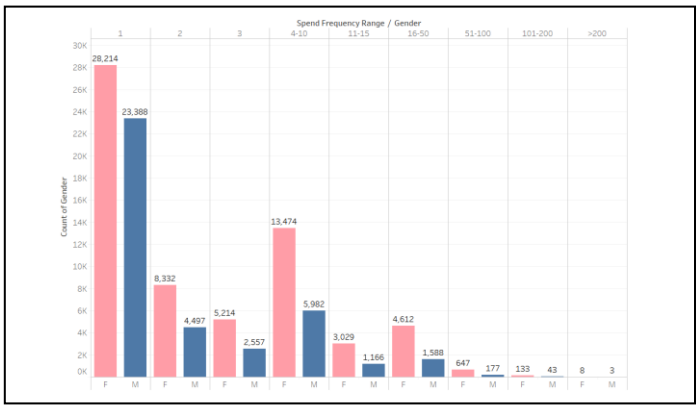

Fig. 5. Spend Frequency Range of Customer vs Gender

Above shows the spend frequency range of customers against gender (see Figure 5). No matter the age range, female always has higher spend frequency range compared to male.

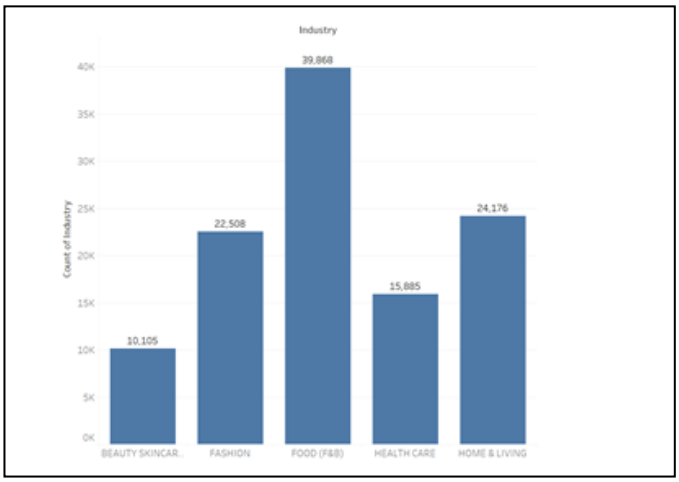

Fig. 6. Top 5 Merchant Industry

Lastly, this bar chart represents the Top 5 industries in this loyalty program (see Figure 6). According to the model above, it can be immediately concluded that the most popular industry is the **FOOD (F&B)** industry. Followed by **HOME & LIVING**, **FASHION, HEALTHCARE**, and lastly, **BEAUTY SKINCARE/PERFUMERY**.

5 Predictive Analysis

The predictive model which was constructed using SAS Enterprise Miner (see Figure 7). Starting from the left, the first node is CUBEDATA, which is the dataset 'CubeData' used for this model. The second node is Data Partition, this node was incorporated to separate the dataset into two distinct parts; 60% Training, 40% Validation. (Training is utilizing data containing target and input values to build and train predictive models, Validation is to validate the suitability of the data model created in Training.) After Data Partitioning, three different models were created - Default Tree, Depth & Branch Tree, and Maximum Tree. Decision trees were utilized because the target value 'CustomerActivity7to12Month' is binary, and three different decision trees were created to see if there are any notable differences between the three. The last node is Model Comparison, to compare and see which of the three models connected to it performs better.

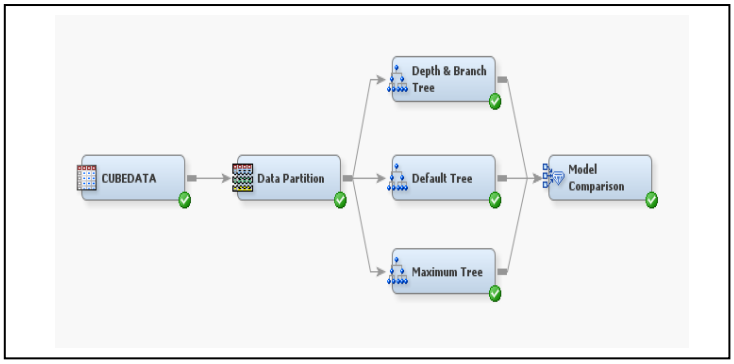

Fig. 7. Default Tree

5.1 Default Tree

There are 61839 observations used for Train, and 41225 for Validation (see Figure 8). The misclassification rate is 0.2625 for Train and 0.2615 for Validation. The average squared error for train and validation is 0.1740 and 0.1733 respectively which means that the both the misclassification rate and average squared error across the three partitions are not very significant.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churner		NOBS	Sum of Frequencies	61839	41225	
Churner		MISC	Misclassification ...	0.262488	0.261589	
Churner		MAX	Maximum Absolut...	0.97272	1	
Churner		SSE	Sum of Squared E...	21514.64	14286.02	
Churner		ASE	Average Squared ...	0.173957	0.173269	
Churner		RASE	Root Average Squ...	0.417081	0.416256	
Churner		DIV	Divisor for ASE	123678	82450	
Churner		DFT	Total Degrees of ...	61839		

Fig. 8. Default Tree

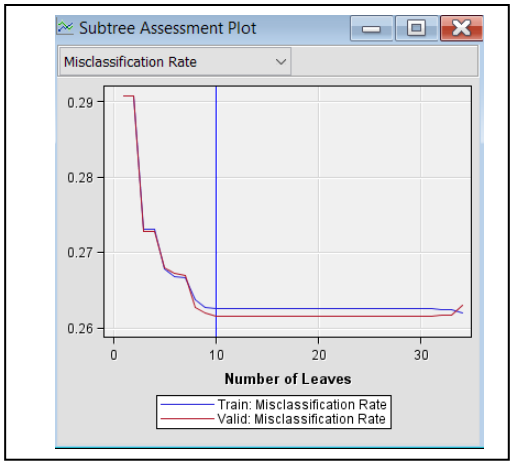

Fig. 9. Subtree Assessment Plot

The subtree assessment plot shows the Misclassification Rate against each sub-tree as the data splits (see Figure 9). Both Train and Valid started at 0.291 on Leaf 2 and dropped steeply until Leaf 3 where Valid has better performance than Train. Both Train and Valid meet again and dropped steeply from Leaf 4 to Leaf 5 where Train had better performance. Train and Valid met again at Leaf 7 and dropped gradually with Valid having the best performance.

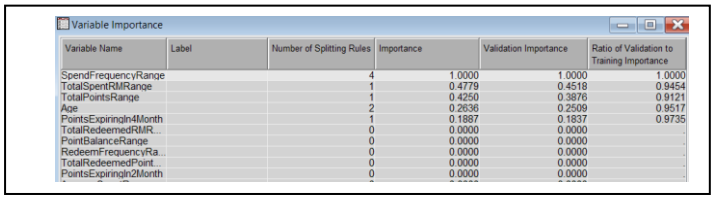

Fig. 10. Variable Importance

This shows us the importance of each variable towards the decision tree (see Figure 10). The most important variable starting from the top is SpendFrequencyRange, followed by TotalSpentRMRange, TotalPointsRange, and PointsExpir-ingIn4Month.

Which means that in Default Tree, these variables have highest influence towards the model.

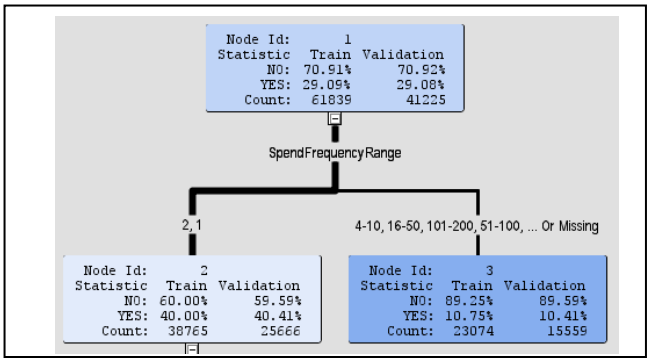

Fig. 11. First-Level Split

Firstly, for the chosen path the first-level split is at Node 1 to Node 2 and Node 3 (see Fig 11). The topmost node (Node 1) represents the whole sample of the loyalty program members. 'NO' represents non-churners while 'YES' represents churners, both Train and Validation data subsets show that there are 71% non-churners and 29% churners.

The thicker connecting lines represent the volume of record going to each node, this can be justified by the total Count for each node – Node 2 has total Count of 64,431 (38765 + 25666) while Node 3 has total Count of 38,633 (23074 + 15559).

In this case, there are more records going to Node 2 which means that more members have a SpendFrequencyRange of 1 or 2 compared to the other SpendFrequency-Range for Node 3. The lighter shading of the node indicates that the percentage of churners in Node 2 is higher compared to Node 3. In Node 2, 59.6% of members are non-churners, while 40.4% are churners.

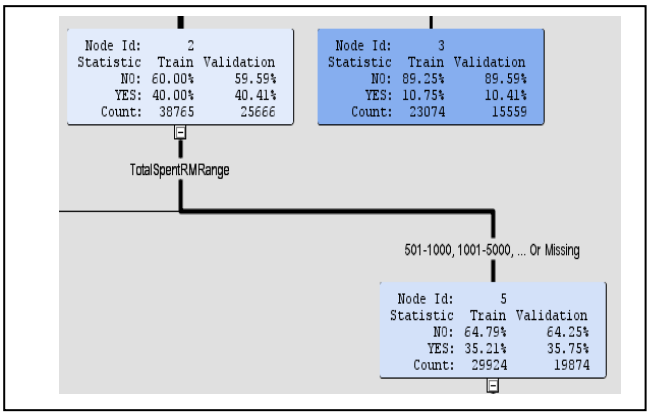

Fig. 12. Second Level Split

Figure 12 shows the second-level split. From Node 2, it splits to Node 5. Higher volume of members in Node 2 goes to Node 5, these members has TotalSpentRM-Range of 501-1000, 1001-5000, etc. In Node 5, there are 64.25% non-churners and 35.75% churners.

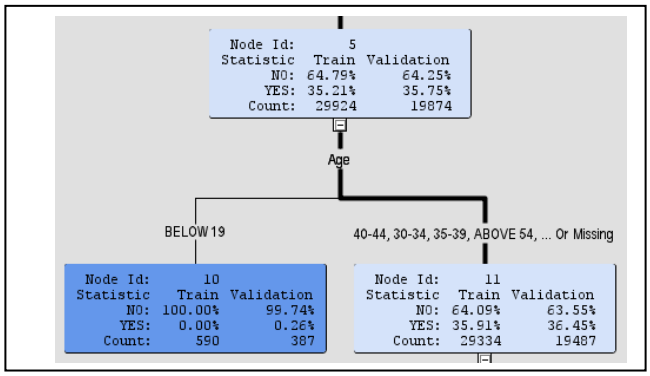

Fig. 13. Third Level Split

Next, the third-level split (see Figure 13). There are higher volume of members from Node 5 going to Node 11. This means that majority of members from Node 5 are Age of NOT BELOW 19. For Node 11, there are 63.55% non-churners and 36.45% non-churners.

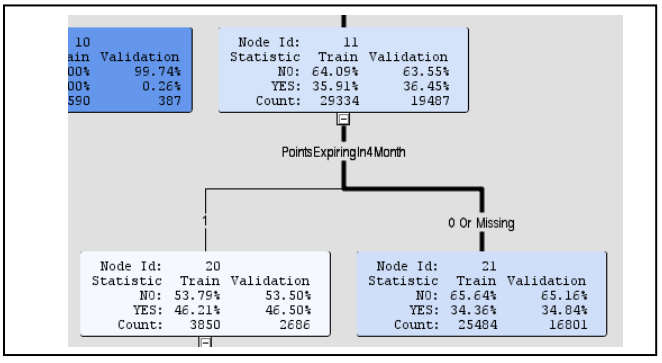

Fig. 14. Fourth Level Split

The fourth-level split from Node 11 to Node 21 and Node 20 (see Figure 14). The volume of members going to Node 21 is higher, this implies that majority of members

from Node 11 have PointsExpiringIn4Month of 0. For Node 21, there are 65.16% non-churners and 34.84% churners. The chosen path ends on Node 21.

6 Conclusion

The selected loyalty program for this research has been avidly promoting their loyalty program for the past few years, but in the beginning, there were not many participating merchants as compared to now where there are approximately 1,000 participating merchants in this program.

Based on the descriptive analysis, it allows understanding of the customer's background, characteristics, information and their spending pattern of participating merchants. These findings would be useful in identifying and targeting customers who are likely to churn and increase the possibility of customer retention. The descriptive analysis showed that most of the members of this loyalty program were loyal customers as there were only less than 30% who have churned. These churners consist mostly a of member who has spent less than 10 times. When further analyzed, it was found that the member does not feel that the loyalty program benefits them in any way after they have used it several times, and therefore ceased their usage.

It is understood that the younger the customers are, the more likely they would become a member of this program which the age group range from 19-34 represents almost 60% of the total members. The highest frequency of customers lies in the age group of 19 – 24 years old which are most probably students. The manager has also stated that generally there are higher number of female customers – as justified in the descriptive analytics previously, more than 60% of the members are females. As stated by him, females are more drawn to membership programs as compared to males. According to a survey done by CrowdTwist in 2016 [11] women are indeed more loyal customers and are stronger advocates of the loyalty program that they are using. In another research done by JakPat in 2016 [12], we have also observed that no matter which industry it is, females are always more inclined to participate in loyalty programs.

For the predictive analysis, Decision Trees were utilized for the model because the target 'Churner' is binary. Three decision trees with different settings were created to see the differences, but the Default Tree remains the best. Looking at the Variable Importance window for no matter which tree it is, the SpendFrequencyRange and TotalSpentRMRange which both represents member's spending behavior, are always the top two variables that shows most importance. This means that the number of times the member has spent and the total RM they have used while utilizing the card plays a massive impact on the prediction. It might be bias to just look at either one of these two variables because even though the member's SpendFrequencyRange is one of the highest, their TotalSpentRMRange might not be a lot. To add on, for some members who has very high TotalSpentRMRange could be property buyers who just signed up as a member for one-off benefit, it might not guarantee that they would continue patronizing or using the loyalty program. However, SpendFrequencyRange could tell more accurately whether a member is loyalty or not because if the member

continuously uses it often, it means that they are potentially loyal members. At the end of the day, the objective of the model was to identify customer churn, hence it justifies the emphasis on SpendFrequencyRange over TotalSpentRMRange.

Moving on to the next part of the analysis, clustering was performed to group members of similar characteristics or behavior together and to see which the most common behavior which member exhibit. According to the results gained from clustering, it seems that majority of female members, aged mostly between 19-39 spent the most in the FOOD (F&B) industry, this is considering both churners and non-churners. People love getting free stuff, and they would share it with their friends and families therefore by offering a sign-up incentive it would encourage customers to join.

Firstly, based on the finding from descriptive analysis, that members who spent less than 10 times is very possibly a churner, one of the possible strategy to prevent customers with this behavior from churning is to provide newly joined / signed up members with incentives dedicated specially to them. A few actions the company could take based on this strategy could be to provide the new members with rebates on the initial first few transactions. This could encourage the new members to get started on using the card more as they would want to use the rebate provided, while for the program, this can increase the likelihood of members to continuously utilize the membership program. Getting members engaged with the program is always the first step, once a member realizes the benefits, they could attain by using the card, they are more likely to continue patronizing.

Secondly, another strategy would be event-driven alerts. To send notifications or alerts to members of through the app to remind them of their point balance and the equivalent RM monetary value. The intention behind this idea was thought of because some members might have used the card previously but forgot about the points accumulated that can be used for redemption, while some members might not be aware of how much points they have collected thus far. This would benefit the members as they would be reminded to utilize the points they have collected, while for Sunway Pals it could encourage an increase of member's usage and increasing spend frequency.

Thirdly, this strategy is to encourage members engagement with the loyalty program. Increasing members spend frequency would increase members engagement. A way to increase the members spend frequency is to reward redeemable points to them after they have spent a certain amount. For example, to award RM10 worth of redeemable points to members for every RM200 they spent – in an accumulative manner. For members, they would feel that the loyalty program is worth their effort and participation. Through this research, we are able to identify that using these variables such as "spending frequency range, total spent range, total points collected and points expiring" are able to do an accurate and good prediction on customer churns especially in retail industry that deals with loyalty programs. We hope that with this analysis it will give practitioners and researcher in this field a better idea on targeting their customers and understand their spending behavior which may impact their loyalty program.

References

1. K. V. R. Reza Allahyari Soeini, "Applying Data Mining to Insurance Customer Churn Management," p. 11, 2012.
2. A. A. A. O. Oyeniyi, "Customer Churn Analysis In Banking Sector Using Data Mining," p. 10, 2015.
3. E.-P. L. D. L. F. Z. a. P. K. P. Richard J. Oentaryo, "Collective Churn Prediction in Social Network," p. 5, 2012.
4. B. A. R. A. K. O. N. A. M. Y. M. R. D. M. A. F. O. Ibhrahim Zakaria, "Loyalty Program," *The Relationship between Loyalty Program, Customer Satisfaction and Customer Loyalty in Retail Industry: A Case Study,* p. 8, 2013.
5. A. Toporek, "What Is a Loyalty Program? (And Will It Work For You)," CTS Service Solutions, 25 June 2012. [Online]. Available: http://customersthatstick.com/blog/what-is/what-is-a-loyalty-program/. [Accessed 31 May 2017].
6. A. McEachern, "A History of Loyalty Programs, and How They Have Changed," smile.io, [Online]. Available: https://blog.smile.io/a-history-of-loyalty-programs. [Accessed 31 5 2017].
7. L. Passport, "Evolution of Customer Loyalty Programs! – Loyalty Passport," Loyalty Passport, 27 March 2016. [Online]. Available: https://loyaltypassport.wordpress.com/2016/05/27/evolution-of-customer-loyalty-programs/. [Accessed 31 May 2017].
8. T. &. D. D. Mulhern, "Building Loyalty at Things Remembered," *The Journal of Consumer Marketing,* pp. 62-66, 2004.
9. A. Dehghan and T. B. Trafalis, "Examining Churn and Loyalty Using Support Vector Machine," *Business and Management Research,* vol. 1, no. 4, pp. 153-161, 2012.
10. S. E. Wyse, "Customer Satisfaction vs. Customer Loyalty," Snap Surveys, 26 June 2012. [Online]. Available: https://www.snapsurveys.com/blog/customer-satisfaction-customer-loyalty/. [Accessed 11 October 2017].
11. CrowdTwist, "Battle for the Sexes: Women Are More Brand Loyal Than Men, and Are Stronger Brand Advocates | Bulldog Reporter," 24 October 2016. [Online]. Available: https://www.bulldogreporter.com/battle-for-the-sexes-women-are-more-brand-loyal-than-men-and-are-stronger-brand-advocates/. [Accessed 25 October 2017].
12. JakPat, "Membership and Loyalty Programs - Survey Report - JAKPAT," 10 May 2016. [Online]. Available: https://blog.jakpat.net/membership-and-loyalty-programs-survey-report/. [Accessed 25 October 2017].