International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Consumer Adoption and Retention for New Generation Tobacco Products - A Machine Learning Approach

Dr. Vinay M. R.[1], Naveen Kumar T[2]

[1]Ph. D. (Econometrics), M. B. A (Finance), Senior Data Scientist, Data & Analytics Unit (DNA), Infosys Technologies Ltd., Bengaluru, Karnataka, India http: //www.linkedin. com/pub/dr - vinay - m - r/28/653/aab

[2]B. E. (Electronics and Communication), Data Analyst, Data & Analytics Unit (DNA), Infosys Technologies Ltd., Bengaluru, Karnataka, India https: //www.linkedin. com/in/naveen - kumar - t - 438b37182/

Abstract: ***There is a significant transformation in the tobacco industry recently towards reducing the health and hazardous specific risk through an array of new innovations and inventions. This is also in compliance with many regulations and restrictions introduced by governing authorities towards safeguarding tobacco consumers' interest. In this regard, the tobacco industry has witnessed substantial portfolio diversification including introduction of new generation product categories called New Categories. The tobacco companies are making huge investment to popularize these products attracting new as well as already existing FMC consumers in order to cope up with global health interest and to promote less risky products. This is also, to cultivate the habit of consuming nicotine free products among consumers. Though, they are witnessing partial success in this upliftment process, there is a high rate of attrition in the New Category segment in very short period of time. Given that continuously changing nature of consumer behavior and their preference, it is imperative to the business to retain the on boarded consumers and convert them as long - term purchasers; else forego huge investment as well as consumer upliftment. In this regard, many tobacco businesses are using conventional analytical options for identifying such vulnerable consumers; however not able to succeed as they are reactive and ineffective in nature. Hence, there is a strong need for proactive Machine Learning models which can predict the probable churning consumers well in advance. A comprehensive Machine Learning models which capture multidimensional aspects like trend, pattern, demographic, behavioral, product attributes, loyalty, usage, and ownership specific factors makes the solution more robust in predicting the probable churn well in advance. As such, we have focused our study on retaining new consumers proactively in the New Category segment. This paper with an empirical analysis discusses the problem statement, business case, solution approach, modelling, integration, implementation, and recommendations of the Machine Learning solution.***

Note: We extend our sincere thanks to DNA Analytics leadership team for their vision, guidance, and continuous support throughout this research and special thanks to ***Nayan Kumar Mohanbhai Rathod****, Lead Analyst, Infosys Technologies Ltd. for his contribution and support throughout this study.*

Keywords: Tobacco Industry, Data Science, Machine Learning, Consumer Attrition, Supervised Learning, Predictive Modeling, Consumer Segmentation, Probability of Churn, Segmentation Profiling, Retention Measures.

1. Introduction

Tobacco industry is witnessing substantial transformation in the market, specifically in the product portfolio segment as there is a lot of concern by governing authorities as well as due to increased awareness among the consumers. New players are entering the market with unique products and existing players are also introducing innovative products frequently to bring positive transformation and upliftment from hazardous risk and health issues in consumers. The traditional cigarette which are combustible in nature has been joined by products which are consumables in nature spanning new and unique categories like first - generation e - cigarettes (closed system cig - a - likes, which replicate the look and feel of combustibles); vapours, tanks, and mods (VTMs) with open refillable systems, tobacco - heated products, and licensed medicinal products etc. There is a huge marketing investment from tobacco companies to introduce these new category products and attract new consumers. Butit is observed that there is a frequent change in consumer preference and purchase behavior driven by many factors. Though, tobacco companies are able to convert significant number of consumers, theyare not continuing with the business for longer period of time because of many reasons. This is indeed a matter of concern as there would be a threat to the faster transformation and upliftment. This would ultimately result in reducing the interest of tobacco companies in new innovations and in adoption of new consumers for New Category products.

Hence, there is a need for proactive prediction of such vulnerable consumers and preventing them from churning to reduce the negative impact associated with it. In addition, there is a need for early detection of such consumers in advance well before they exit the segment. Many companies are following conventional and descriptive approaches which are partially helping the business because of their limitations. Moreover, they are reactive in nature and result in complete failure to avoid such attrition. It is imperative that any effective solution should be proactive in nature and should consider multidimensional aspects and factors in identifying such vulnerable consumers. The solution should help the business effectively and constructively to mitigate the churn and to make use of the prospective consumers for the sales maximization. In this regard, we propose an advanced solution based on Machine Learning approach which would proactively predict such vulnerable consumers

we will in advance. This also takes into account multi – dimensional aspects and their possible influence on the consumer churn.

Here, in this case, we have considered new consumers who are purchasing New Category consumables compatible to their device in one of the emerging markets. It is observed that their attrition rate is very high in a particular group. Business is currently using conventional analytics but not able reduce the attrition rate. To avoid such ineffective and delayed decision making, we have proposed Machine Learning predictive modeling approach. This approach would help business to proactive identify vulnerable consumers and take the retention measures as per the model recommendations. This solution will clearly articulate the root cause and strategies based on the significant drivers of churn specific to that market. It will also be an effective solution in terms of on time and most frequently updated solution. The solution can also be customized for any market with required changes. Thus, proposed solution helps tobacco businesses to reduce the New Category consumer attrition from multidimensional aspects perspective and thus help the transformation and upliftment to the maximum extent.

Problem Statement

This paper focuses on New Category consumers retention and thus enhancing sales of lesser risk tobacco consumables. Our study specifically focuses on the new consumers who purchases the New Category consumables for the first time. The study has been done for one of the largest and developed countries. The firm is facing approximately41% attrition in one of the consumers groups leading to massive return of the consumers to conventional FMC segment. This is also causing lot of speculation in the New Category market.

The business wanted to reduce the attrition rate to the minimal extent through identifying the vulnerable consumers well in advance. They were also looking for business strategies based on the predictive and logical solution which can be proactive in nature.

Business Case

We have selected one of the largest tobacco manufacturers globally for our study. Also, we have created a hypothetical business case, which would be truly representative of actual scenario. We have **used synthetic data to perform the analysis and build a predictive model. We are declaring here that we have not used any actual client specific data, business scenario or market information. This is completely a hypothetical but ideal market scenario.**

We would like name this **hypothetical Tobacco company as ABC Tobacco** and henceforth this hypothetical name will be used throughout this paper. ABC Tobacco, one of the largest manufacturers and sellers of Tobacco products based out ofUSA. It has operations in 150 countries and having highest number of Brands and varieties of products. They manufacture and sell both conventional FMC and New Category products in the market. The approximate distribution of the market share is 70% and 30% by FMC and NC products sequentially.

Recently, ABC Tobacco has come with innovative and unique New Category products which include vapour, oral and heating products. The New Category products are being sold across the globe in various markets. ABC Tobacco is investing huge amount in marketing of these New Category products. It is also trying to woe its FMC Category consumers with various offers to purchase the New Category products.

However, the New Category segment is having a significant attrition rate leading to huge revenue loss and lower return on marketing investments. This has negatively impacted ABC Tobacco company and reduced its interest towards new innovations pertaining to less risk and hazardous products. In this regard, business is looking for an advanced, proactive solution which will help them to identify the probable churning consumers and to build strategies to reduce the churn in the New Category segment. Obviously, our key focus of the study is to reduce attrition in the New Category segment.

Business has defined attrition as If the purchasers purchasing competitors' products on month 6 from their first month purchase of consumables of ABC Tobacco company, then it is called purchasers attrition. The objective here is to reduce the attrition rate of such consumers and increase the retention. It is a market level analysis using monthly aggregate data. The scope of the analysis is limited to new consumers of New Category products. Also, we have targeted only those consumers who are purchasing the New Category products in convenient stores only.

We have proposed a multilayered advanced analytical approach comprises of three major steps. Below are the key analytics objectives and steps:

- **Understand** attrition distribution across various levels and identify the factors leading to high attrition.
- **Predict** user retention/attrition based on classification using various demographic, usage, pricing, spent etc.
- **Categorize** purchasers based on the likelihood of attrition and provide the value at risk for next best action along mitigation measures.

Finally, the analysis results will be used to develop **strategies** to mitigate the higher churn and increase the retention.

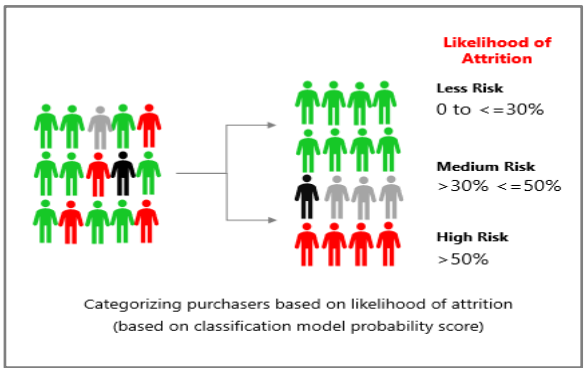

Figure 1: Analytics Objectives

2. Literature Review

Attrition reduction and loyalty in tobacco industry has been a point of lot of research interest exploring various data, segments, markets, techniques and approaches. We have reviewed few of the studies related to this space and given our findings below.

John Davesin his paper "Cigarette brand loyalty and purchase patterns: An examination using US consumer panel data", analyzes panel dataset on cigarette purchasing. This paper concludes that cigarette purchasing exhibits Negative Binomial Distribution with many infrequent buyers and fewer frequent buyers. Cigarette brands do exhibit high loyalty compared to other consumer categories. The study sheds more light on consumer behavior towards a product with addictive qualities and known harmful effects. James and Erika, et al. [8] in their paper 'Cigarette brands with flavour capsules in the filter: trends in use and brand perceptions among smokers in the USA, Mexico and Australia, 2012–2014'describes trends, correlates use, and consumer perceptions related to the product design and innovation of flavour capsules in cigarette filters. The results indicate that use of cigarette with flavour capsules is growing, is associated with misperceptions of relative harm, and differentiates brands in ways that justify regulatory action. Philip and Donald, et al. [2] in their paper 'Quantifying Brand Loyalty: Evidence from the Cigarette Market' quantify consumer brand loyalty, using quasi - experimental variation in the prices of premium brand cigarettes relative to Native brands. Using data from the NYS - ATS and the NHP, this estimates that although there is substantial brand switching, the results show that cigarette consumers display a high degree of brand loyalty. Many premium - brand consumers were willing to pay at least an extra $4.35 per pack rather than switch to an untaxed Indian manufactured cigarette brand. Dr. Sajjan in his paper, 'Consumer Behaviour towards Cigarette Smoking in Kolkata Region Knowing the fact its Injurious for Health: A Case Study' explains increasing consumption of cigarette is a concern for the health status of Kolkata people. In this context, the study evaluates the predicting factors that are influencing people to be accustomed a definite purchasing behavior. Different aspects of brand loyalty such as taste, flavor, design and filters or strength creates brand loyal customers that ultimately superimpose the awareness of people about negative impacts of cigarette smoking.

Scope of Analysis and Data Availability

As explained in the Business Case, we have considered one of the largest markets for the study and analysis is done at the national level. Below is the scope of our study:

- **Consumers**: We have considered those consumers who are buying New Category devices for the first time on a particular month which is called M0. In addition to this, a filter has been applied to get those consumers who were using competitors consumables in the previous month which is called M - 01.
- **Distribution Channel**: The study is focused only on 1 of the 3 major distribution channels which is Convenient Store purchases. Also, those consumers who are purchasing offline are only considered.
- **Category**: The data is collected at store level when new purchasers purchase New Category products.
- **Data Available**: There are backward and forward tracking for these new purchasers. There is one - month backward tracking which is called M - 01 which includes their previous month purchases, purchased volume and category of the purchase. Also, there is 6 months of forward tracking which includes their respective month purchases, purchased volume and category of purchases. Essentially, this is month level of aggregated data provided by the business. Thus, to summarize purchasers purchase new device on month 0 and their consumption pattern (consumables) is being tracked for the next 6 months.
- **Period of Data**: The period of data available to us is for 21 months from Oct 2020 to Jun 2022.
- **Products Attributes Data**: Along with the above information, we have a one - time Product Attribute table provided by business which includes product attributes such as Taste, Barrel, Tar, Tar group, Price group, Capsule, Length and Pack price etc. The Product Attribute table further merged with the raw data in the monthly consolidated master file.

Proposed Approach

We had set the hypothesis after verifying the data that there are multi - dimensional factors which are driving the attrition. However, these factors and their role can be measured and predicted using the machine learning models. Thus, we have created a comprehensive and a multi - layer approach to safeguard the revenue of ABC Tobacco company through a sophisticated retention modeling solution.

We propose three - layer approach to the above business case which includes understanding data, predicting risk score and segmentation of consumers. Understanding would include exploration and initial analysis of provided data. After that, predicting consumer purchase using Machine Learning model and create individual risk score. Finally, create segments of consumers based on the risk score.

The above said approach can be developed using 7standard Data Science ML steps. This has been explained in detail below:

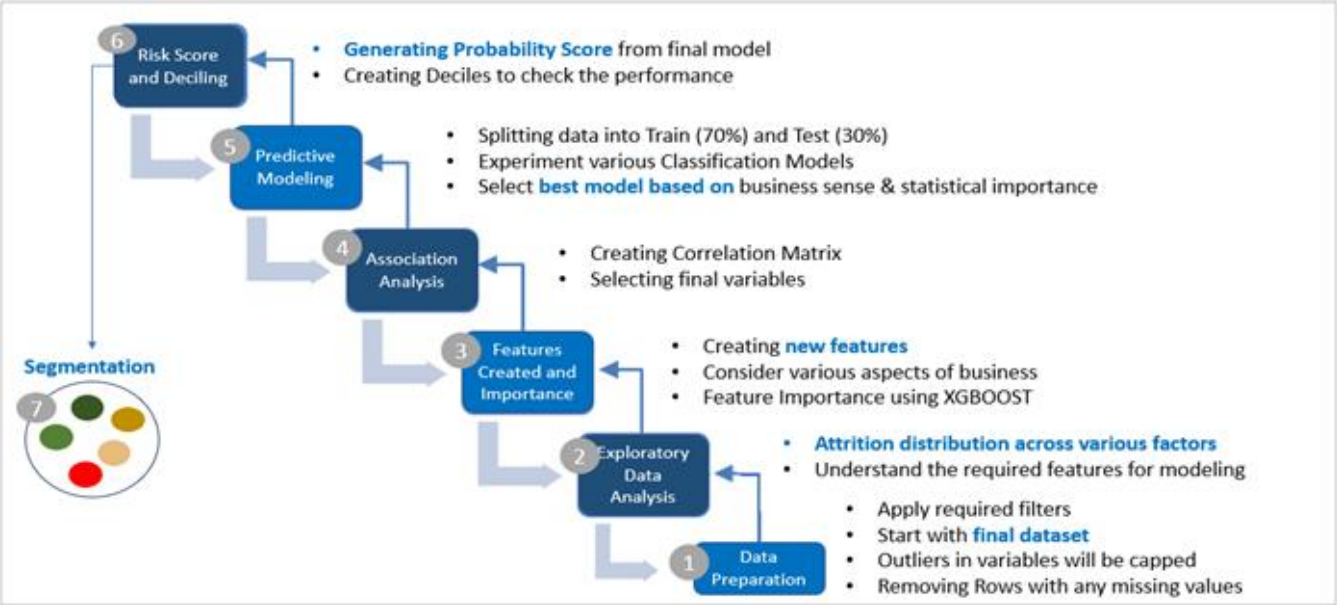

Figure 2: Data Science Steps

We have followed 7 Data Science standard approach starting with data preparation and applying required filters. In the second step, exploratory data analysis will be done to understand the data and required features. Feature engineering for feature creation and assessing the features importance to select the important variables will be done in the 3rd step. Correlation analysis to identify the highly correlated variables and exclude one of them to reduce the multicollinearity problem will be done in the step 4.

In the next step, the data is split into two samples – training and testing samples. The appropriate Machine Learning model is selected based on the different criteria like model performance, consistency across samples and accuracy. As the model output, the risk score will be generated at individual consumer level which would reflect consumer probability to churn. The risk score ranges between 0 to 1, closure to one the score, there is more risk of that individual to churn.

In the final step, we will create segments of the consumers based on their risk score – Low Risk, Medium Risk and High – Risk segments. This is to proactively identify the probable churn consumers well in advance before they churn. This helps the business to be most updated in terms of the vulnerable consumer groups on time. At the end, we would recommend the business team of the ABC Tobacco company with the strategies for each group.

Solution Development

Data Preparation

The data preparation has been done based on standard Data Science approach. The provided files were first verified for the quality of the data, then all the monthly files were appended to create one master file. Initial filters were applied to the master file to create required data sample and the final data used for further analysis. Below is the detailed data preparation procedure followed:

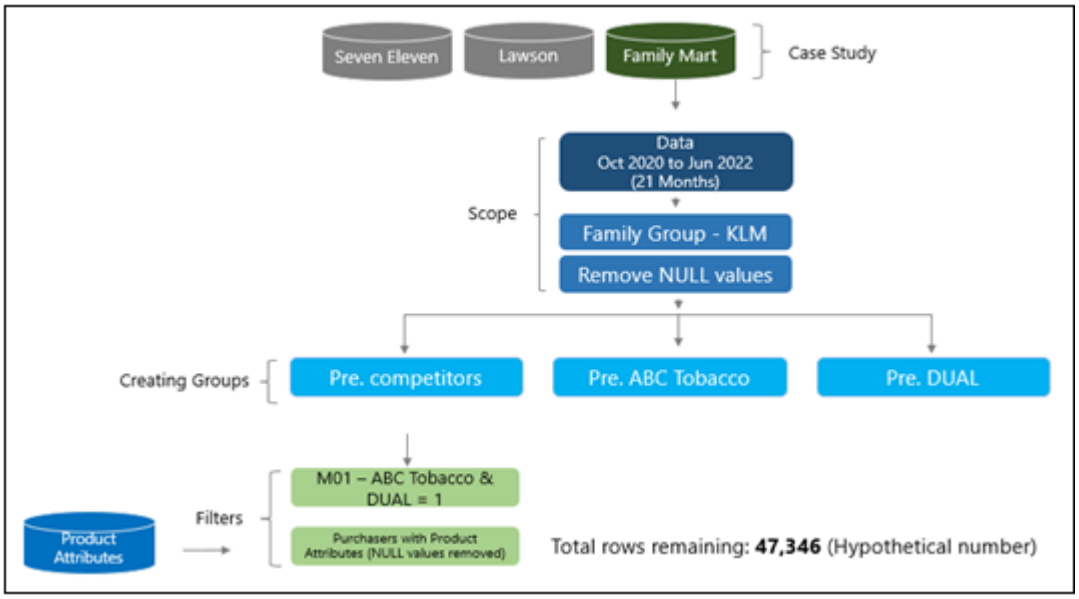

Figure 3: Data Preparation

We have considered Family Mart segment for this case study which is one of the three major retail stores chain in the market. The purchasers who are visiting and purchasing the New Category devices in the Family Mart stores have been considered for the analysis. The data has been collected for the period of 21 months i. e., Oct 2020 to Jun 2022. The required filters suggested by the business team like required family, consumer group have been applied. In the next stage, if any cells of the row are having empty values, then the entire row has been removed. We did not find any extreme values in the data; hence outlier treatment was not required. The duplicate values are removed after reconciliation. After this, the purchasers who were having only previous month (M - 01) status as competitors' consumption have been selected as this group is having higher attrition rate. In the next stage, purchasers who are consuming either ABC Tobacco or competitors consumables are selected. Finally, product attribute data has been merged to the master file and final data is prepared for the modeling purpose.

Attrition Definition and Dependent Variable

Attrition is defined as if the purchasers have not purchased the ABC Tobacco consumables on M06 and purchased only competitors consumables then they are called churned consumers, and this has been referred as attrition. A new binary variable is created based on this status – if the purchasers are purchasing consumables, then 1 is assigned else 0 is assigned. This column has been considered as dependent variable in the pre modeling and modeling stage. The assumption here is that if the purchaser is not purchasing on M06, then he will not come back and purchase again. This assumption is made because of that the purchasers purchase behavior has been tracked for only 6 months.

Exploratory Data Analysis

Exploratory Data Analysis has been done for both independent and dependent variables. Firstly, dependent variable attrition has been plotted to see the pattern by quarter and at aggregate level. Below graph shows the distribution of the Attrition across the study period at aggregate level as well as quarterly.

It is an important indication that business is witnessing higher attrition on the early months of first purchase. This pattern shows that purchasers mainly purchase because of the attractive offers and discounts and then starts moving out to

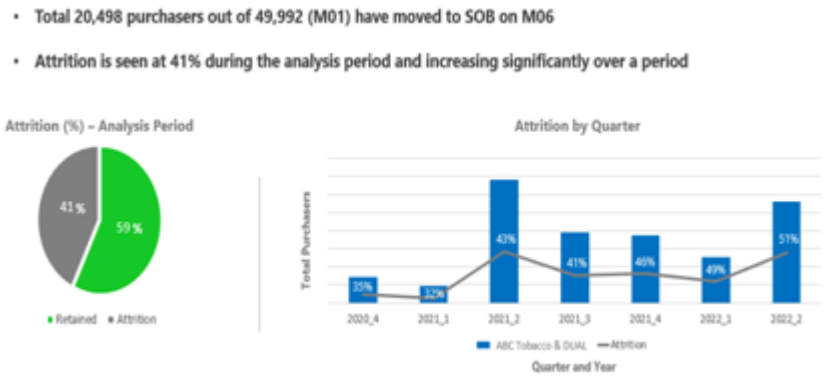

Figure 4: Distribution of Attrition (%)

As we can see in Figure 4, the attrition is randomly distributed over a period. The highest attrition rate has been recorded in the quarter 2 of 2022. As we are focusing on the short window of analysis period, we are not exploring more on the seasonality factors. The attrition rate for the entire analysis period Oct 2020 to Jun 2022 is 41%. During this analysis period, total 20,498 purchasers out of 49,992 purchasers have moved out to competitors on the month 6 of their first purchase.

Figure 4.1: Attrition Waterfall (%)

their previous status. Now the question is, how can business retain them in the same stream for longer duration. It is very important to explore, identify and take the retention measures at the early stage only before major chunk of the consumers leave. Further, the Exploratory Data Analysis is done to understand the attrition distribution across various features like demographic features (age and gender), product attributes (brand and taste), consumption specific features (loyalty, category consumption status), ownership and usage specific, purchaser behavior specific (recency, frequency, consistency, monetary value), Price Groups etc.

The results of Exploratory Data Analysis which is done for individual features at various level are summarized below:

- **Consumption features:** Purchasers of ABC Tobacco on M01 have lesser tendency for attrition and they move partially to DUAL status on M06. Purchasers of DUAL have more tendency towards moving to competitors is very high. Purchasers with FMC Category status on previous month are having higher attrition compared with of New Category. NC purchasers on M01 are having higher tendency of continuing in NC Category on M06.

- **Demographic features:** It is observed that the attrition is almost distributed similarly across age and gender groups. The age groups 20 and 60 are having relatively higher tendency to churn.
- **Product attributes:** Brand XYZ is having relatively higher attrition compared with the other Brands of ABC Tobacco. Attrition rate seen higher in Regular Taste and relatively lower in Menthol and Flavor Tastes.
- **Purchaser behavior features:** The key indicators of the attritions are low frequency, low spent, inconsistent, and not a recent status of the purchasers. Lower these indicators, higher the probability of churn.
- **Pricing specific:** Attrition rate differs mainly between higher price (ASP and PRE) and lower price (VFM and LOW) groups. More attrition is seen in the higher price groups ASP and PRE.
- **Usage and Ownership specific:** Attrition is seen relatively higher in the purchasers who are using FMC followed by Multiplatform currently. Attrition is not significantly differing when purchasers owning single device across the manufacturers. More the number of devices owning along with ABC Tobacco device, higher the attrition is seen.

Feature Engineering and Variable Selection

We have created 35 features and then selected the final list based on both domain knowledge, business input and the statistical importance. There are two types of features –1. Casual Features and 2. Contributing Features. The casual features are the one which would drive purchasers' attrition directly, while contributing features are indicative in nature and represents the trend and pattern of attrition.

The different categories of features are created using multidimensional aspects for the modeling purpose – 1. Consumption specific, 2. Ownership specific, 3. Product characteristics specific, 4. Pricing specific, 5. Purchase behavior specific, 6. Demographic specific. The created features are screened for the complete understanding of distribution, frequency, and association among themselves. Based on the understanding, the feature engineering process was initiated. We decided to create binary attributes for these categorical features based on criteria explained by assigning 1 and 0. Table 1 shows the list of features that we have created, and their importance score calculated using XGBOOST COVER metric.

Table 1: Feature Engineering – Importance Score

Features	Features Category	Description	Score
ABC_TOBACCO_45	Device Ownership	Purchasers owning ABC TOBACCO + 4 to 5 competitors devices	686
PRICE_GROUP_PRE_M01	Price Group	Premium Consumable purchasers	478
CURUSE_NC_THP_ONLYSPLATFORM	Current Use	NC THP ONLYSPLATFORM users	457
BRAND_XYZ_M01	Product Characteristics - Brand	XYZ purchasers	432
ABC TOBACCO_SPENT_15000_M01_M04	Purchasers Behaviour - Spent	ABC TOBACCO M01 to M04 Spent - High Spenders	387
BRAND_ABC_M01	Product Characteristics - Brand	ABC purchasers	366
RECENT	Purchasers Behaviour - Recency	Purchased in M03 to M04	325
Taste_REGULAR_M01	Product Characteristics - Taste	REGULAR purchasers	300
FREQUENT	Purchasers Behaviour - Frequency	> 3 times purchasers purchased in M01 to M04	286
Taste_FLAVOR_M01	Product Characteristics - Taste	FLAVOUR purchasers	267
CONSISTENT	Purchasers Behaviour - Consistency	Purchasers purchased in all months M01 to M04	245
BRAND_OTH_M01	Product Characteristics - Brand	OTH purchasers	234
PRICE_GROUP_VFM_M01	Price Group	VFM Consumables purchasers	223
ABC TOBACCO_SPENT_1_3000_M01_M04	Purchasers Behaviour - Spent	ABC TOBACCO M01 to M04 Spent - Low Spenders	210
Taste_MENTHOL_M01	Product Characteristics - Taste	MENTHOL purchasers	198
M01_STATUS_ABC TOBACCO	Purchasers Group	M01 - ABC TOABCCO Status purchasers	189
M_01_M01_FMC_Flag	Loyalty	M-01 and M01 FMC purchasers	171
PRICE_GROUP_ASP_M01	Price Group	ASP Consumables purchasers	170
CURUSE_NC_THP_MULTIPLATFORM	Current Use	NC THP MULTIPLATFORM users	166
M_01_M01_NC_Flag	Loyalty	M-01 and M01 NC purchasers	151
CURUSE_NC_THP_FMC_ONLY	Current Use	NC THP FMC users	143
PRICE_GROUP_LOW_M01	Price Group	LOW Consumables purchasers	138
OWN_ABC_Tobacco	Device Ownership	Purchasers owning only ABC Tobacco device	119
ABC TOBACCO_SPENT_3001_15000_M01_M04	Purchasers Behaviour - Spent	ABC TOBACCO M01 to M04 Spent - Medium Spenders	108
ABC_TOBACCO_123	Device Ownership	Purchasers owning ABC TOBACCO + up to 3 competitors devices	98

The above features are having binary format in nature and assigned with 1 and 0 for yes or no consequently. We have created 35 variables in the beginning, however, XGBOOT COVER metric dropped 10 variables because of multicollinearity. Finally, we have got 25 variables for the model development purpose.

Association Analysis

We have performed Tetrachoric correlation to identify the high correlation among the independent features. We have selected Tetrachoric correlation technique because all our independent variables are binary in nature. The general correlation techniques are not suitable for this kind of data. Below is the result of Tetrachoric correlation analysis:

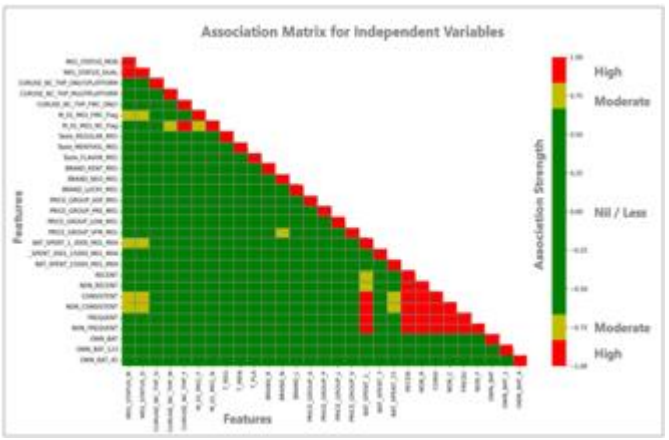

Figure 5: Association Analysis

The correlation coefficient ranges from 0 to 1, closure to 1 indicates high correlation while closure to 0 indicates low or no correlation. The correlation metric is plotted in Figure 5. As we can see most of the independent variable are either having low or moderate correlating stating that there will not be any multicollinearity issue in the if we include these features directly. At the same, time, as we can see in the figure, few features are having high correlation among themselves (consistent and non-consistent, frequent, and non-frequent, recent, and non-recent), hence one of them have been dropped from further analysis. Finally, the remaining features are included in the model development process.

Model Development

Model development consists of various steps - stating with splitting the data for train and testing purpose then fitting appropriate model, comparing the performance metrics among the various models after developing suitable models, selecting the best fit model, and finally generating risk score.

Sampling

This step focuses on deciding the sampling structure and procedure. We have splitted the data in two – training data and testing data. The entire data set for the period of Oct 2020 to Jun 2022 is splitted in to 70% and 30% respectively. The sampling is done using random sampling method. The distribution of attrition in both training and testing data equivalent to 41% of churn and 59% of non - churn purchasers. As the distribution is meeting the standard requirements, we have not boosted the attrition rate using any techniques and used as is.

Modeling

In this section, we will describe the Machine Learning models that we have explored and developed in order identify the factors influencing attrition and generate the risk score at individual consumer level. The key focus here is to capture the attrition probability at maximum and to reduce the false positives and false negatives.

We have experimented 4 supervised models as we are having the labeled data. The labels which assigned are explained in the feature engineering section. Below are the supervised models we have explored – 1. Logistic Regression, 2. Random Forest, 3. Decision Tree, and 4. XGBOOST. All these algorithms gave very similar results and carrying the similar performance metrics. However, we have finalized Logistic Regression model as it was having relatively highest Recall values. Below is the comparison of performance metrics of all the four models:

Table 2: Model Performance Metrics

Models	Accuracy		Precision		Recall		f1 Score	
Train/Test Results	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	**0.83**	**0.83**	**0.80**	**0.79**	**0.87**	**0.87**	**0.83**	**0.82**
Random Forest	0.84	0.83	0.81	0.80	0.85	0.85	0.83	0.82
Decision Tree	0.84	0.83	0.81	0.80	0.85	0.85	0.83	0.82
XGBOOST	0.84	0.83	0.81	0.80	0.85	0.85	0.83	0.82

We are not explaining the algorithm part here in detail as our focus is more on identifying the factors which driving attrition and predict the attrition well in advance. As explained above we have selected, Logistic Regression model for the further processing. Below is the confusion metrics for the selected Logistic Regression model.

Figure 6: Confusion Metrics

As shown, model is performing consistently in both Train and Test data and minimized the misclassification successfully. Hence, this model is finalized for the further processing.

Significant Variables

After selecting Logistic Regression model, we have tried to improve the model performance through number of iterations with various set of variables. Also, based on the business input, we have included new variables and removed the few of the selected variables. Below is the list of variables that we have included in the various iterations and their respective results. The Accuracy and Recall remains same at 0.83 and 0.87 respectively across all the iterations shown below. However, any one model from various iteration of different set of variables can be selected based on the business requirement. Currently, Model number 4 has been selected for further analysis.

Table 3: Comparison of Different Iterations of Model

Comparison of Different Iterations of Model (co-efficient and p values)					
Significant Features	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Loyalty – FMC	0.26 (0.01)	0.13 (0.04)	0.17 (0.03)	0.26 (0.04)	0.13 (0.03)
Current Use – NC THP Multiplatform	1.25 (0.00)	0.52 (0.00)	0.55 (0.00)	0.46 (0.00)	0.37 (0.00)
Recency	-1.25 (0.00)				
Current Use – NC THP FMC only	1.12 (0.00)	0.69 (0.00)	0.39 (0.00)	0.54 (0.00)	0.57 (0.00)
Consistency	-3.05 (0.00)	-3.49 (0.00)	-2.56 (0.00)	-2.23 (0.00)	-2.68 (0.00)
Purchaser Behavior – Spend < 3000	0.43 (0.00)	0.82 (0.00)	0.89 (0.00)	0.89 (0.00)	0.33 (0.00)
Price Group – ASP	0.07 (0.02)	0.15 (0.01)	0.20 (0.03)	0.13 (0.02)	0.56 (0.03)
Product Characteristics – Brand XYZ	0.44 (0.05)	0.42 (0.06)	0.44 (0.05)	0.37 (0.05)	0.25 (0.04)
Device Ownership – ABC Tobacco with/without Competitors	-0.17 (0.04)	-012 (0.00)			
Device Ownership – Comp1 with/without other Competitors			0.12 (0.02)		
Device Ownership – Only Comp1 and Other NC				0.30 (0.01)	0.39 (0.01)
Age 20					0.51 (0.00)

Note: The Accuracy and Recall remains same at 0.83 and 0.87 respectively across all the experimentations show above. However, the models with different set of combinations can be selected based on the business requirement. Currently, Model 4 has been selected for further analysis.

Removed features *Added features*

Deciling

The deciling has been done using the model number 4 from the above list. The deciles were created through equally dividing the population in to 10 parts after arranging the probability of churn score in descending order. As shown below, first five deciles would cover 85% of cumulative attrition in both Train and Test data.

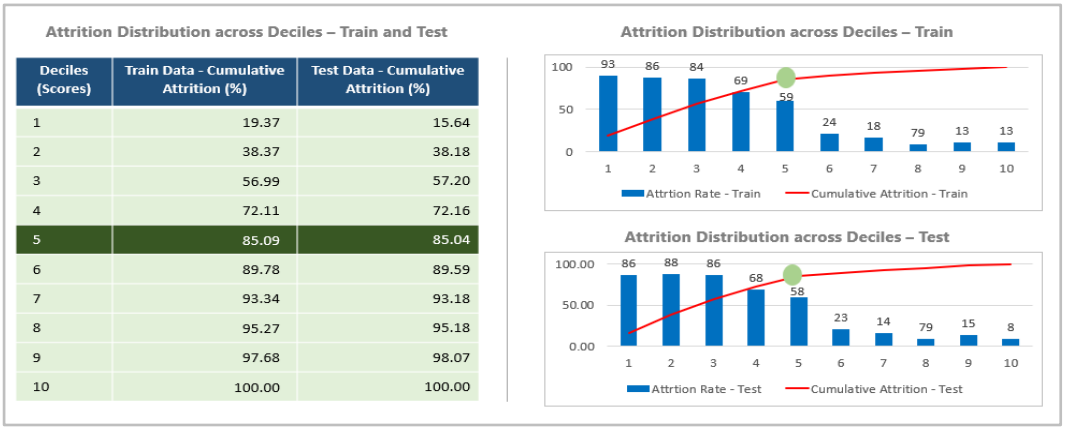

Figure 7: Deciling

The deciling is also used to compare the model performance on Training and Testing data. As shown above in the figure, the model is performing similarly on both the samples. However, for the effective decision - making purpose, we have created consumers segments based on the risk score. This is explained in the next section.

3. Segmentation and Profiling

Segmentation

The individual purchasers have been grouped in to three segments based on therange of probability risk score generated by the model –1. High Risk segment, 2. Medium Risk segment, 3. Low Risk segment. The entire population is sorted in descending order or risk and then equally divided in to three Segments. Below is the brief explanation of the 3 segments:

- **High Risk segment** – This segment includes the purchasers with high probability risk score and have the high tendency towards churn. In this case, 83% of the total purchasers have churned.
- **Medium Risk segment**: This segment carries the moderate risk of churn, with 44% of the purchasers have churned during the analysis period.
- **Low Risk segment**: This is most valuable consumers segment as consumer are very loyal and prosperous to the business. This segment can be targeted for the sales expansion and revenue maximization.

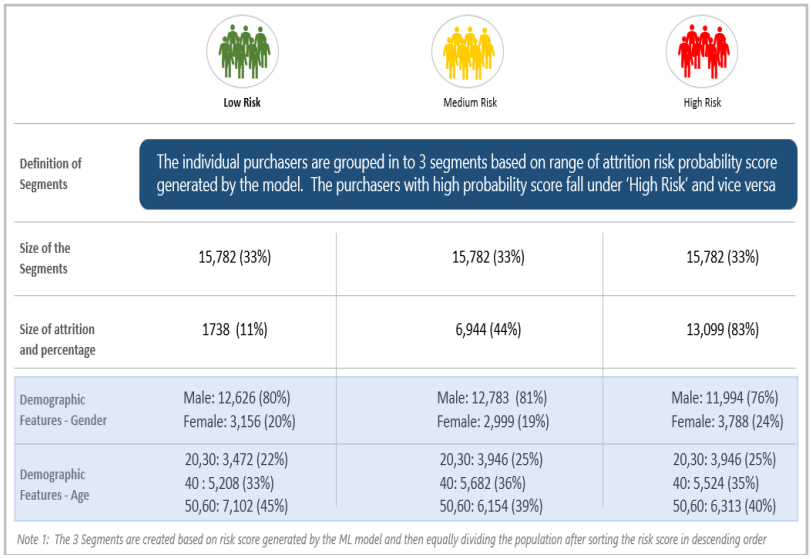

Figure 8: Consumer Segmentation

As shown in the figure no 8, all the three segments are having equal number of consumers but the attrition rate among the segments differ significantly based on their probability score and type of consumers falling the respective segment.

Profiling

We have done profiling for High - Risk segment in order to give a clear explanation to the business on how exactly the risk population is distributed considering the associated factors.

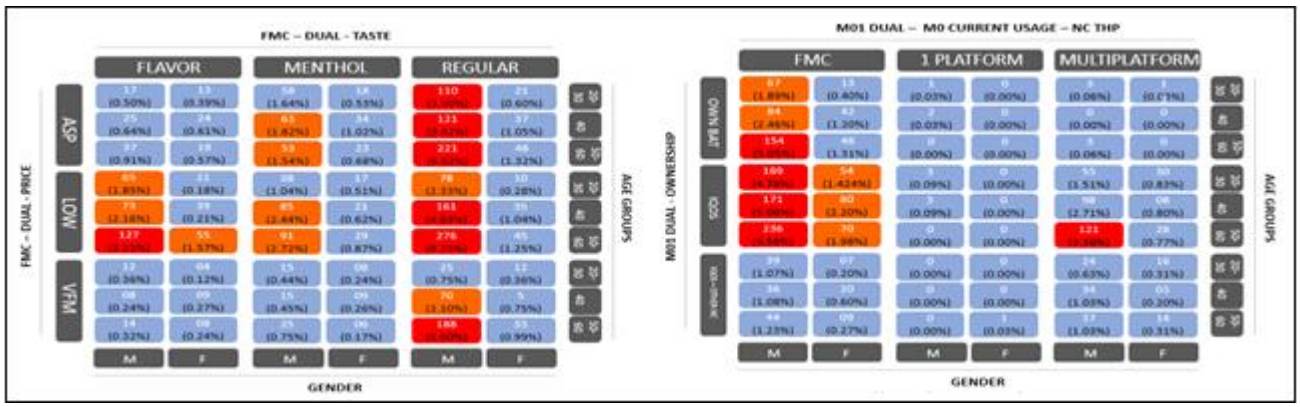

Figure9. Segmentation Profiling

Segmentation

The detailed profiling of the High - Risk segment has been created using number of parameters like Category, Taste, M01 Status, M01 Category, gender, age groups, ownership, current usage etc. Below are the few observations from High - Risk segment profiling:

It is observed that, FMC DUAL compared to NC Dual is having higher purchasers' distribution. Hence, we given more information below on FMC DUAL segment to help business to target their retention measures on specific groups:

- **M01 Status (DUAL) –M01 Category (FMC) M01 Taste–Price Group - Gender - Age:** REGULAR Taste, 40 and 50 & 60 Age Groups, Male, LOW & ASP Price Groups contributing 28.93%. FLAVOR & Menthol Tastes, 50 & 60 Age, LOW Price Group is contributing 13.95%. REGULAR Taste, 50 & 60 Age Groups, VFM Price Group, Male is contributing 5.24%
- **M01 Status (DUAL) M0 Current Use - M0Ownership - Gender - Age:** Device Ownership COMP1& OWN ABC Tobacco and Current Use of NC THP FMC only are having higher risk population distribution (22%). COMP1 MULTIPLATFORM users, 50 & 60 Age Group, Male is contributing 3.36%. Purchasers of Current Use 'ONLY 1 PLATFORM' Group is relatively very less population distribution.

4. Key Findings and Recommendations

Below are key findings from the study and the recommendations based on the business input and domain knowledge:

- **Attrition distribution**: Attrition is seen high in the consumer group who are consuming competitors' consumable in the previous month of device purchase. This explains that purchasers who purchasing either ABC Tobacco or DUAL on M01, but their status is competitors' purchaser in pervious month, are having higher attrition rate than other groups.
- **Factors associated with attrition:** There are two categories of factors which are influencing attrition –

Casual factors and contributing factors. **Casual factors are** Demographic, Consumption specific features, Current Use, Device Ownership, Price Groups, Brands, Category Loyalty. **Contributing factors** are recency, frequency, consistency and spent.

- **Significant Variables**: Predictive Model has identified the significant factors which are driving the attrition rate –FMC Loyalty, Current Use NC THP Platform, Current Use NC THP FMC, consistency, lower spent, ASP Price Group, Brand XYZ, Device Ownership and Age Group 20.
- **Machine Learning Model**: Many models are explored, and Logistics Regression is finalized based on model parameters. The model successfully able to classify, predict and generate the risk score.
- **Consumer Segments**: Segmentation of the consumers has proven to be a significant solution as the consumers were classified based on the probability of risk. The high - risk segment should be the main target to the business to introduce retention measures and target most valuable consumer segment for the business expansion.
- **Retention Measures**: Target FMC loyal purchasers base, current users with NC THP Multiplatform and NC THP FMC only, Non recent, Non consistent, low spenders and multiple device owners, ASP Price Group, XYZ Brand and Regular Taste purchasers for intensive promotional activities

Majority of the High - Risk population belong to Age Group 20 & 60 followed by 40. With respect to product characteristics, Price Group ASP followed by PRE and with respect to Taste REGULAR followed by MENTHOL and with respect to BRAND XYZ followed by ABCare having higher risk population distribution.

5. Conclusion

The empirical analysis from this study proves that Tobacco industries can make use of advanced Machine Learning predictive modeling approach for proactive decision making. The consumer segments which are created based on the risk score generated by the Predictive model, help business to identify the vulnerable and most prospective consumers well in advance. It is observed that many factors like demographic, behavioral, product attributes, loyalty, usage, and ownership, pricing specific are driving the attrition. Hence, the solution can be even more robust as the multidimensional factors are considered while developing the model.

References

[1] Anderson, S. J., Glantz, S. A., & Ling, P. M. "*Emotions for sale: Cigarette advertising and women's psychosocial needs*. " Tobacco Control, 14 (2), 127–135, 2005
[2] DiFranza, J. R., Eddy, J. J., Brown, L. F., Ryan, J. L., & Bogojavlensky, A. "*Tobacco acquisition and cigarette brand selection among youth. Tobacco Control"*, 3 (4), 33, 1994
[3] Philip D, Donald S, "*Quantifying Brand Loyalty: Evidence from the Cigarette Market*", NBER Working Paper Series, NBER Working Paper No.28690, JEL No. I12, April 2021
[4] John Daves, "Cigarette brand loyalty and purchase patterns: An examination using US consumer panel data", Journal of Business Reserach, 67 (2014) 1933 - 1943
[5] Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, "Complexity and Application of Tobacco Manufacturer Pricing Game considering Market Segments", Hindawi, Article ID 6701286, 11 pages, Volume 2017