POLITECNICO
MILANO 1863

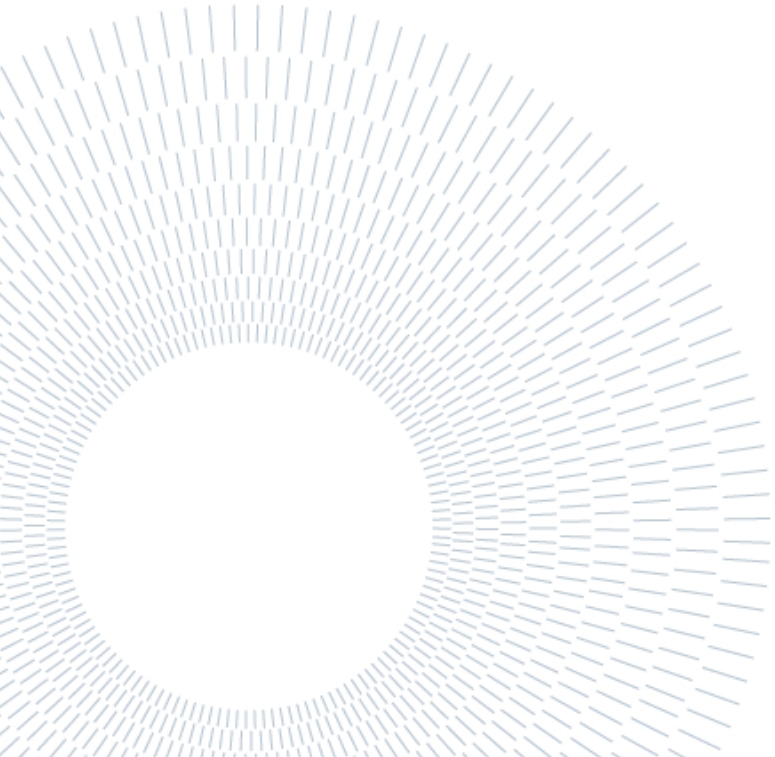
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Assessing partial defection of retail consumers, and the role of private label in its prevention.

TESI DI LAUREA MAGISTRALE IN
MANAGEMENT ENGINEERING
INGEGNERIA GESTIONALE

Authors: **Lorenzo Perego, Mirko Leoni**

Student ID Lorenzo Perego: 968737
Student ID Mirko Leoni: 964918
Supervisor: Lucio Lamberti
Co-supervisor: Emanuele Fedrigolli
Academic Year: 2021-22



Abstract

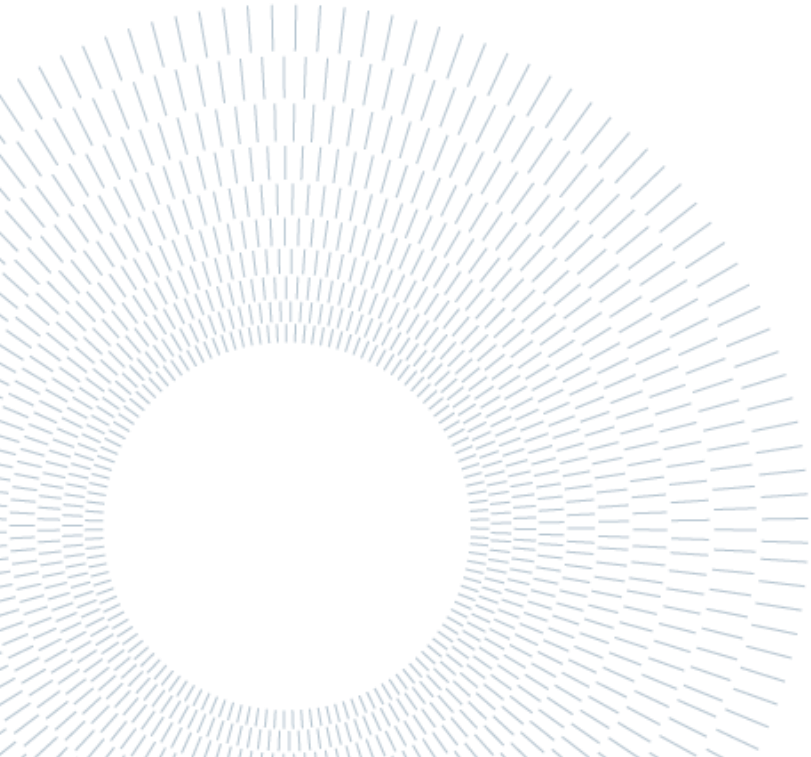
Among the various challenges in the area of Customer Relationship Management, increasing attention is being paid to the issue of customer relationship termination (churn). The final phase of a customer's life is a cause for study with the goal of better understanding it and ultimately preventing it. The phenomenon of churn can occur in two distinct contexts: purchases in which the sales relationship is defined by a supply contract (contractual settings), and purchases in which the sales relationship is not formalized by a supply contract (non-contractual settings). In cases of non-contractual settings, identifying a churn is extremely difficult since a generic customer does not have to withdraw from a contract in order to terminate the relationship with a supplier; therefore, the supplier in this case will not have full visibility into the customer's abandonment; the only thing it will be able to observe is that the customer in question will not visit the stores again. A typical example of a market in which the non-contractual mode is in force is that of fast-moving consumer goods (FMCG). The work presented here finds application precisely in the FMGC sector. In the area of churn prevention, professionals are currently looking for possible factors or behaviours expressed by customers that are able to signal the end of the relationship in an early way. Among these signals, one area that is still much debated concerns the role of private label products in churn prevention, so their ability to retain customers is not yet clear. This thesis aims to study the phenomenon of churn on retailer's customers, and to find an answer about the role that private label products play in increasing customer loyalty. A statistical model will be developed (hidden Markov model) that will be fitted on real data retrieved from a well-known player in the Italian FMGC industry, this is done to evaluate model performance in a real context. Through this model it will be possible to divide the customer base into segments with different churn risk, this classification will be directly generated from the transactional data. Each assignment of a customer to a given class will be given in probabilistic terms, by indicating the probabilities of membership. Finally, the effects of two variables identified as the percentage of private label products divided into two lines: named as "reatiler" and "premium", will be integrated into the model itself via a regression model. The effects of these variables on the result generated by the model will be able to demonstrate the inherent loyalty-building ability of branded products.

Key-words: churn, partial defection, private label, hidden Markov model, HMM, CRM, customer retention, retention rate

Abstract in italiano

Tra le varie sfide nell'ambito del Customer Relationship Management, un'attenzione crescente è posta alla tematica della cessazione della relazione con il cliente (churn). La fase finale della vita di un cliente è motivo di studio con l'obiettivo di meglio comprenderla e, in fine, prevenirla. Il fenomeno del churn può avvenire in due contesti ben distinti: acquisti in cui la relazione di vendita è definita da un contratto di fornitura (contractual settings), e acquisti in cui la relazione di vendita non è formalizzata da un contratto di fornitura (non-contractual settings). In casi di non-contractual settings, identificare un churn risulta estremamente ostico dal momento che un generico cliente non deve recedere da un contratto per terminare la relazione con un fornitore, il fornitore perciò non avrà in questo caso piena visibilità sull'abbandono del cliente; l'unica cosa che potrà osservare sarà che il cliente in questione non ha più visitato i negozi. Un tipico esempio di mercato in cui vige la modalità non contrattuale è quello della grande distribuzione organizzata. In ambito di prevenzione del churn, i professionisti sono attualmente alla ricerca di possibili fattori o comportamenti espressi dai clienti che siano in grado di segnalare in maniera anticipata la fine della relazione. Tra questi segnali un ambito ancora oggi molto dibattuto riguarda il ruolo dei prodotti a marchio nella prevenzione del churn, non quindi è ancora chiara la loro capacità di fidelizzare i clienti. La tesi si propone di studiare il fenomeno del churn su clienti di un supermercato, e di trovare una risposta circa il ruolo che hanno i prodotti a marchio nell'aumentare la fedeltà dei clienti. Verrà sviluppato un modello statistico (hidden Markov model) che sarà fittato su dati di un reale e ben conosciuto distributore italiano operativo nel settore della grande distribuzione, questi dati saranno utili per valutare le performance del modello in un contesto reale. Con questo modello si riuscirà a suddividere la base clienti in segmenti con diverso rischio di abbandono, tale classificazione sarà generata a partire dai dati a transazionali. Ogni assegnazione di un cliente ad una data classe sarà fornita in termini probabilistici, indicando le relative probabilità di appartenenza. In fine, nel modello stesso verranno integrate, tramite un modello di regressione, gli effetti di due variabili identificate come la percentuale di prodotti a marchio suddivisi in due linee soprannominate "retailer" e "premium". Gli effetti di queste variabili sul risultato generato dal modello dimostreranno la capacità di aumentare il grado di fedeltà dei clienti intrinseca dei prodotti a marchio.

Parole chiave: churn, defezione parziale, prodotti a marchio, hidden Markov model, HMM, CRM, customer retention, retention rate



Contents

Executive summary

Introduction

The discipline of customer relationship management involves, in one of its facets, the search for elements and models that will better enable the understanding of the phenomenon of churn/partial defection. This phenomenon can be identified with the final stage of the relationship between a customer and a given company. "Churn," in fact, refers to the instant when the customer stops purchasing the product or service from the company under analysis. In truth, the concept of churn takes on different declinations depending on the type of market in which the transaction takes place. Two types of purchase modes can be distinguished: purchase in which the customer-firm relationship is defined in contractual terms (known in the literature as "contractual settings"), and purchase in which the customer-firm relationship is not defined at the pre-purchase stage through the conclusion of a contract (known in the literature as "non-contractual settings"). The key difference between these two types of purchasing lies in the fact that in contractual settings the company has full visibility into the customer's churn, this is because the customer itself must formally withdraw from the contract to terminate the supply relationship (e.g., telco industry). In contrast, buying in non-contractual settings leaves the company whit no direct signal about the customer's churn: the customer has no type of contract from which it must sever, the customer itself therefore does not expressly issue any signals confirming the churn. One type of industry in which non-contractual setting applies is that of fast-moving consumer goods, the model disclosed here is applied in this contest.

For the reason given above, under non-contractual settings it results of chronic difficulty to stick to the traditional definition of churn, hence it is introduced the concept of "partial defection." Partial defection is alerted when the client under analysis demonstrates a significant reduction in its purchasing habits and enters a state of potential churn.

The churn/partial defection issue is of paramount importance to the business: being able to maintain a healthy relationship with the customer over time ensures prolonged cash flow from the customer. "Over time" is what aligns marketing strategy to firm's value: firm's value is created by present and future cash flows, which are eventually generated by each customer in the firm's customer base, those cash flows must be sustained by consistent marketing actions able to lock-in customers and boost the quality of their relationship.

A better understanding of the churn/partial defection phenomenon, from a managerial point of view, implies a greater ability to implement effective marketing actions aimed at preventing this phenomenon: being able to increase customer retention means increasing the return on the investment made in customer acquisition; each customer acquisition is in fact associated with an acquisition cost, which for the firm represents the investment made in the customer relationship.

In addition, the use of a model that can classify the customer base into risk classes, with relative probability of membership, can make a significant improvement in the addressability of marketing actions and thus a greater yield on marketing efforts.

The most popular method used for segmentation into risk classes is the Recency-Frequency-Monetary analysis. This is usually done imposing thresholds on some specific variables used to describe customer behaviour, typically Monetary and Frequency. In addition, the classification thus obtained is deterministic, in the sense that membership in a given class is not expressed in terms of the probability of belonging to that class.

More in general, the models currently used in churn/partial defection require the classification into risk classes to be provided as input to the model itself: these, in fact, are unable to process directly from the data the necessary classification, which will therefore have to be produced a priori. This is not a problem in the case of contractual settings, in which it is possible to assign a clear flag to churned customer. On the other hand, in non-contractual settings, the non-information on the customer churn, oblige practitioners to decide on how to classify customer.

The solution presented in this paper is able to produce a probabilistic classification of the customer base directly from customer transactional data. Once the model has found the states of the Hidden Markov Chain (HMC), which correspond to the classes

of risk, it is able to compute the transition probabilities between the identified classes. Furthermore, in our model these transition probabilities can be combined with covariates via a multinomial logit model; this allows the overall model to take into account variables that may help in assigning the probabilities of customer membership to classes. In addition, by analysing the coefficients of the multinomial logit model, it is possible to understand what impact each covariate has on the probability of transitions between classes. In particular, by evaluating the impact of these covariates on the probability of transitioning from a higher class to a lower one (the latter with higher churn risk), it is possible to identify churn predictors. In this paper, using the procedure just described, it was found out that the covariates "percentage of private label products per ticket" proved effective as a signal of customer loyalty. Moreover, the issue regarding the loyalty-enhancing potential of private label products is still an open topic of discussion in the literature.

Furthermore, the solution proposed in this paper will be built, and successively tested, using real transactional data provided by an Italian retailer.

Literature review

A literature review has been carried out, in order to assess the current state of the art in the field of churn/partial defection modelling and prevention. The research looked for advancement in the statistical approaches and new findings in the field of churn prediction features, with special regard to the impact on brand loyalty generated by the conscious purchase of private label products.

The literature review was divided into three macro areas of interest:

- **Review on churn/partial defection modelling methods:** this part reviewed the most widely used statistical models for studying the phenomenon of churn/partial defection. This review found that the models discussed in the literature are mainly applicable to contractual settings, where churner clients can be uniquely assigned a label. These tools are primarily supervised learning models. In contrast, for non-contractual settings there is a dearth of models that can analyse the issue of churn/partial defection.

- **Review on hidden Markov models:** this review was conducted with the aim of studying the current and most innovative applications of hidden Markov models, even in areas not necessarily inherent to our own. This study allowed us to increase our knowledge of this innovative type of models, which we then exploited to build the model proposed in this paper. Furthermore, from the literature review it was determined that there is indeed a lack of application of hidden Markov models in the area of churn/partial defection analysis.
- **Review on the role of private label on customer loyalty:** the literature review of the effect that private label purchases have on customer loyalty was done with the aim of unearthing possible predictors to incorporate into our model. Actually, what emerged is that this issue is still open among researchers, and there are conflicting opinions regarding the loyalty-building power of private labels.

Research questions

From the literature review it was possible to came up with five research questions, that our disclosure will try to answer:

I. Develop a fully data-driven classification model that can assign each customer a posterior probability of belonging to each class among those identified. The model should work without any a priori classification.

II. Identify, using our model, churn predictors obtained from the reprocessing of the information available in the transactional dataset, exploiting variables such as type of purchased products.

III. Understand whether private label products have a role in churn prevention and churn prediction.

IV. Compare our model performances with the most traditional and widely adopted RFM models.

V. Provide useful managerial implications for the implementation of our model.

Methodology

A hidden Markov model (HMM) was implemented with the aim of obtaining an innovative and more precise customer classification than those currently used, as well as validating the role that private labels play in customer loyalty.

The development of the model went through several steps, which are report below:

I. **Data Analysis:** the width of the time horizon in which data are available, the choice of variables and the level of aggregation (i.e., choice of the time bucket) can influence the result of model parameters, and for this reason it is necessary to understand which choices are most appropriate.

II. **Model specification and tuning:** in this step, the variables identified in the previous step were selected, and if needed transformed. Several iteration cycles were conducted in order to find the final model. Indeed, this step was an iterative process, in which was involved a lot of trial and learn. The model was then improved by the addition of covariates.

III. **Output analysis:** once found a satisfactory formulation of the model, a thorough interpretation of the results was conducted in order to understand the value of the information obtained as output.

IV. **Model validation:** to evaluate the capabilities of the model, a different dataset with different characteristics and completely new observations was tested. This step was done in order to assess whether the results obtained are consistent and whether the model is able to obtain meaningful states/classes and confirm the covariates impact.

V. **Result discussion and managerial implications:** the last step is to interpret the findings within the managerial context. In addition, the responses obtained to the research questions and possible future improvements will be evaluated.

Data

The data available for the study came from one of the largest FMCG players in the Italian market. The data were collected between August 2020 and September 2022. The dataset contains all transactions made by a pool of customers during the indicated period. The starting aggregation is at the ticket line level (i.e., each row corresponds to a unique product type in each ticket). The following variables are available: Card number, Monetary, EAN code, quantity, ticket date, and product description (in literal form).

Model specification and data management

The goal of the model is to classify customers into segments representing different levels of churn risk and to better understand the presence of possible churn predictors. Model training was done by choosing Monetary and Frequency as continuous emission variables of the HMM. The choice of these variables was aligned with those most widely used in the literature to describe customer behaviour in FMCG sector. Indeed Frequency (number of visits per period) and monetary (€ spend per period) are two effective indicators of relation with the customer (a loyal customer visit often the store and spend an above average amount of money. A "defected" customer never or rarely come to the store and spend less money). In addition, they are two of the three variables that are used to perform the RFM analysis (most used method for segmenting customers in risk classes), and this will allow us later to be able to make a comparison with it. To train the HMM, it was necessary to aggregate the data with to create a customer-specific time series. The aggregation in periods was done with monthly base time, the driver of this choice was the trade-off between:

- Necessity of not to having too long-time buckets that would cause the model to be "slow" in detecting suspicious cases. Churn is a rather rapid phenomenon; too wide time frame can cause the model to not be able to grasp it timely.
- Necessity to not introduce potential doubts between an effective customer churn and an absenteeism period, which may simply be the consequence of a specific buying behaviour.

- Necessity to not over increase the computational difficulty: from a computational point of view short time frames cause poor reading of customer behaviour by increasing the probability of observing periods in which customer never visited the store.

Data cleaning operations were performed to eliminate noise from non-loyal customers as well as those behaviours that are not feasible for a typical customer (i.e. cashiers who swipe their loyalty card resulting as customers with very high frequencies). At the same time as this step, new variables were developed, starting from the information available on the purchased items (feature engineering) to be then used as covariates in the HMM model. This was done with the aim of improving the performances of the model and discovering possible features capable of creating retention in the customer base. The focus was on private label products, these were further disaggregated finding two private label product lines: named in this paper as "retailer" and "premium". For each of them the percentage of products type in the category respect the total product types in the time frame selected was calculated.

This type of covariates engineering deliberately does not give importance to quantity because the main objective is to capture the customer's choice between a private label product and a non-private label product and then to be able to assess any loyalty effects resulting from that choice.

Eventually, the trained HMM model has:

- 3 states representing high, medium, and low classes; corresponding to low risk, medium risk, and high risk of churn/partial defection.
- 2 response variables: Monetary and Frequency through which states/classes are identified.
- 2 covariates (% of each private label category) that impact the probability of transition between states of the system. The meaning of the two identified product lines will be better introduced in the next paragraph about the disclosure of result obtained.

Result disclosure

Impact of covariates on transition matrix

To improve the developed model, the use of covariates was introduced by means of a multinomial logit regression model built on transition probabilities. These covariates do not influence the creation of the HMC states, but only modify the transition probabilities.

The two covariates used, and their significance are listed below:

- **Percentage of "retailer private label products" (perc_pvl_retailer):**

$$\frac{\#\ product\ types\ pvl_{retailer}}{tot.\#\ of\ product\ types}$$

- **Percentage of "Premium private label products" (perc_pvl_premium):**

$$\frac{\#\ product\ types\ pvl_{premium}}{tot.\#\ of\ product\ types}$$

For number of private label product types, it is intended the number of unique EANs referred to private label products, which is then divided by the total number of unique EANs in the ticket.

Effect of perc_pvl_retailer:

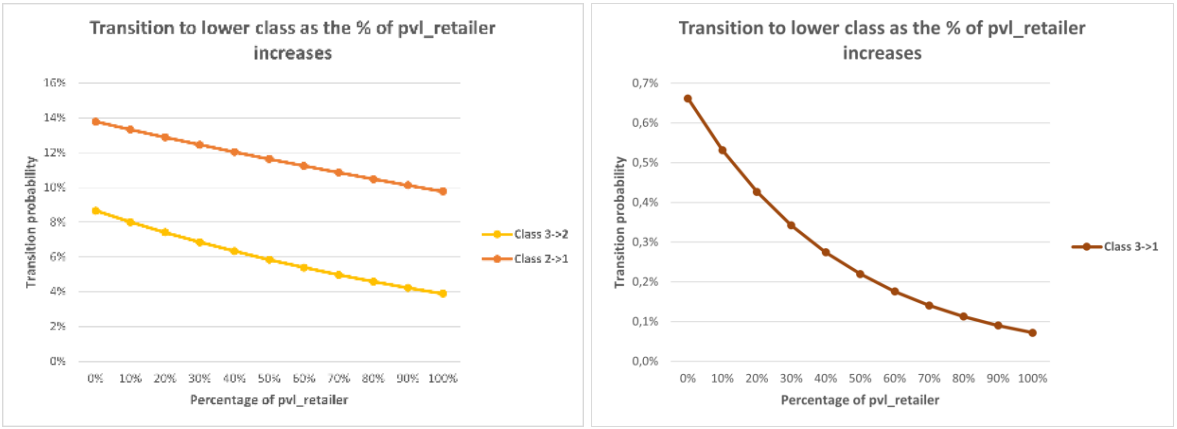


Figure 0.1 - Effect of perc_pvl_retailer on transition probability

The covariate perc_pvl_retailer depends on the number of pvl_retailer products purchased; these products are sold by the retailer itself, which brands them with its logo. The pvl_retailer products are a substitute offering to well-known brand products that the retailer offers to its customers, these products rely heavily on value for money: low prices due to lower promotional and product management expenses, quality guaranteed by the retailer's ability to select its trusted suppliers, know-how derived from its core business. It can be seen from the graphs shown that as the percentage of pvl_retailer increases, there is a corresponding reduction in the probability of transition to a lower status. This means that a customer who is more likely to purchase pvl_retailer products is indeed more loyal to the retailer and has a lower probability of increasing its risk of churn/partial defection.

Effects of perc_pvl_premium:

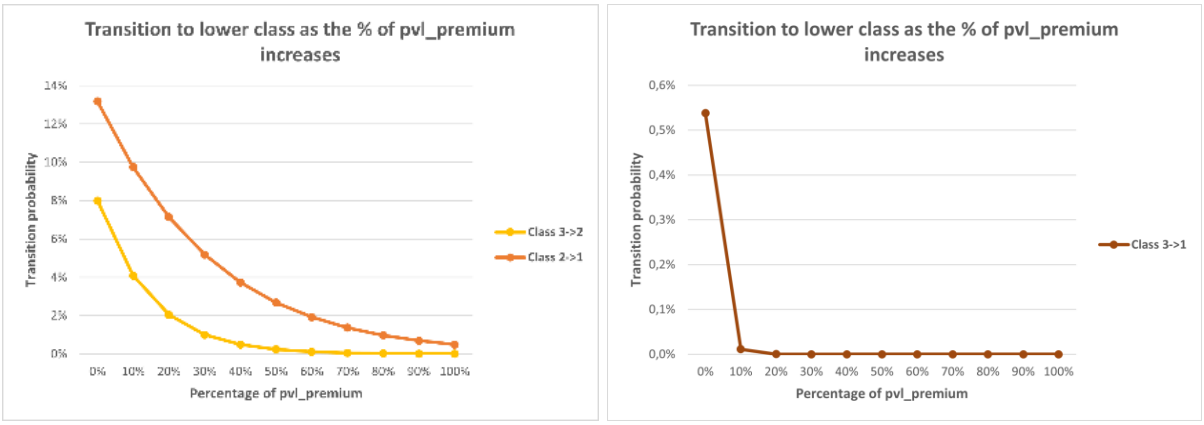


Figure 0.2 - Effect of perc_pvl_premium on transition probability

The perc_pvl_premium covariate represents the percentage of a given customer's purchase in products from the highest-end line offered by the retailer. These are characterized by high prices and uncompromising quality. A customer who is used to buying this type of private label is likely to be a customer who expresses high trust in the retailer and is willing to pay a premium price to have the best product the retailer can offers. The choice of pvl_premium is a conscious one; the purchase is not made for the purpose of savings money (in addition to the higher-than-average price, these products are rarely discounted); in fact, the customer pays more because he or she recognizes the quality of the product and knows that it meets his or her needs.

Among the two covariates so far implemented, this is the one that demonstrates the greatest discriminating power; in fact, the graphs above shows that even a small increase in the percentage of pvl_premium leads to a substantial reduction in the probability of transition to the lower state. This means that a customer inclined to purchase pvl_premium will rarely be a customer with little loyalty to the retailer.

In conclusion, private label retailers and private label premium, can become extremely useful tools in the hands of the marketing manager, who can use them to create marketing campaigns designed to build customer loyalty and boost customer retention.

Performance comparison with traditional models: HMM vs RFM model

RFM model	**Mean Monetary**	**Monetary Variance**
Class 3:	358.8156	207.7375
Class 2:	187.7554	128.9996
Class 1:	43.9602	57.0894
HMM	**Mean Monetary**	**Monetary Variance**
Class 3:	394.431	222.862
Class 2:	153.468	89.901
Class 1:	19.658	23.807

Table 0.1 - Comparison of monetary mean per class RFM vs HMM

RFM model	Mean Frequency
Class 3:	13.156
Class 2:	5.313
Class 1:	1.436
HMM	**Mean Frequency**
Class 3:	13.654
Class 2:	4.702
Class 1:	0.837

Table 0.2 - Comparison of frequency mean per class RFM vs HMM

The tables above compare the mean values of Monetary and Frequency for each of the three risk classes: the values for the HMM model are taken from the output of the model itself, since they correspond by construction to the values of the distributions on which the class/states are fitted. The values for the RFM model, on the other hand, are obtained by calculating mean and variance, over the available history, of all values for each class. The substantial difference in these two models emerges when comparing the mean values for the highest risk class: Class 1 has lower mean values of Monetary and Frequency in the HMM model than the correspondent in the RFM model (Mon = 19.7 vs 44.0, Freq = 0.84 vs 1.44). Having a Class 1 with high mean values of Monetary and Frequency means that the model often includes in the lower-class customers that potentially have performance typical of clients with lower churn/partial defection risk (i.e., not at risk). Eventually, from the comparison of mean values it confirms that the model with HMC succeeds in creating classes that are consistent and coherent with the models currently in use among practitioners, while still being able to more accurately identify customers at high risk of churn/partial defection.

To continue the comparison between HMM and RFM, the transition matrices obtained from the two models were compared. The values contained in these matrices represent the transition probabilities between the individuated classes. In the case of the model with HMC such matrix is obtained as part of the model itself. For the RFM-based model, on the other hand, the transition matrix was calculated by considering for each period the transitions between the three classes.

TP Matrix HMM	Class 1:	Class 2:	Class 3:
Class 1:	80.50 %	19.38 %	0.12 %
Class 2:	13.27 %	83.74 %	2.99 %
Class 3:	0.52 %	7.94 %	91.55 %

Table 0.3 - Transition probability matrix HMM

TP Matrix RFM	Class 1:	Class 2:	Class 3:
Class 1:	77.56 %	19.80 %	2.65 %
Class 2:	35.58 %	47.24 %	17.18 %
Class 3:	5.99 %	20.57 %	73.44 %

Table 0.4 - Transition probability matrix RFM

Looking at the transition matrix of the HMM, it is noticeable that the diagonal (representing the probability of remaining in the same) has always higher probabilities than the respective ones in the transition matrix of the RFM model. This means that the classification done by the HMM turns out to be more consistent. Indeed, this classification is obtained by assigning the probability that the observed emission of Mon and Freq belongs to the fitted distributions. Compared to the RFM model, the HMM is able to eliminate those noisy transitions that are the result of exceeding, even by a negligible delta, the predetermined thresholds, that eventually are even the same for all customers in the given period.

In the RFM model, moreover, it can be observed that Class 2 turns out to be a highly unstable class; the probability of staying in fact corresponds to 47.2% while the probability of changing class (i.e., moving to Class 1 or Class 3) is 52.8%, thus making it more likely to change class than to stay in it. Finally, in both models, Class 3, with the customers at the lowest churn/partial defection risk, is reasonably the one "most closed" to movements to and from Class 1.

Model reaction to topical customer behaviour

Below are shown four typical behaviours that a retailer's customers may engage in. The results produced by our model for each of these customers chosen as archetypes will then be discussed.

- **Stable customer:** during the observation period these clients have a constant behaviour, always staying in the same class or at most making few fluctuations between the highest and middle class (Class 3 - Class 2).
- **Partial defection/announced churn:** those are customers who show a gradual decline in their buying habits that leads them to enter the lowest state, Class 1. Those customers effectively go through a partial defection process, in which their performances decline period after period.
- **Unexpected churn:** for those clients the termination of the relationship is unpredictable. In fact, churn occurs quite suddenly and sometimes even after periods when the client's performance was even improving.
- **Occasional customer:** those customers over the observation time frames have highly variable behaviour. This archetype is indeed very typical one for retailers: in the FMCG field it is common to observe a high level of variability especially if the time horizon of data aggregation is weekly or monthly (as in our study).

Below are reported the graphs showing how the model responds to the four archetypes. For comprehension’s sake a logarithmic scale graph will be provided. Indeed, logarithmic scale is useful to better understand the increase in the posterior probability of belonging to the lowest class.

In fact, it is common for the value of the posterior probabilities of belonging to the lowest class to be extremely low (we are talking orders of magnitude of E-100), especially when the customer is very loyal this belonging to the highest Class 3. Observing the increase in logarithmic scale allows to better grasp interesting signals related to the change in the probability of making a transition from a higher to a lower class (with special regard to Class 1).

Stable customer:

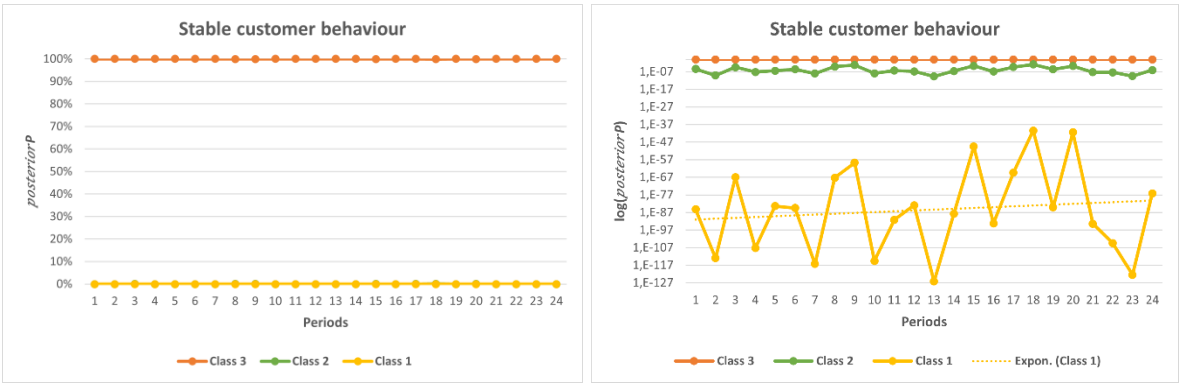


Figure 0.3 - Posterior probability of a stable customer

In these graphs is shown the trend in the posterior probabilities of a stable customer of belonging to each of the three classes. Specifically, the selected customer is assigned to Class 3, and remains stable in the same class for all periods. This means, as shown in the first graph, that the posterior probability of being in Class 3 is constant over time and has a value approximating 1. Instead, in the second graph it is shown the evolution of the posterior probabilities of the same customer, but in logarithmic scale: the value of the posterior probability relative to the lowest class remains almost constant at very negligible values, the trend line is almost horizontal and is composed by extremely small probabilities.

Unexpected churn:

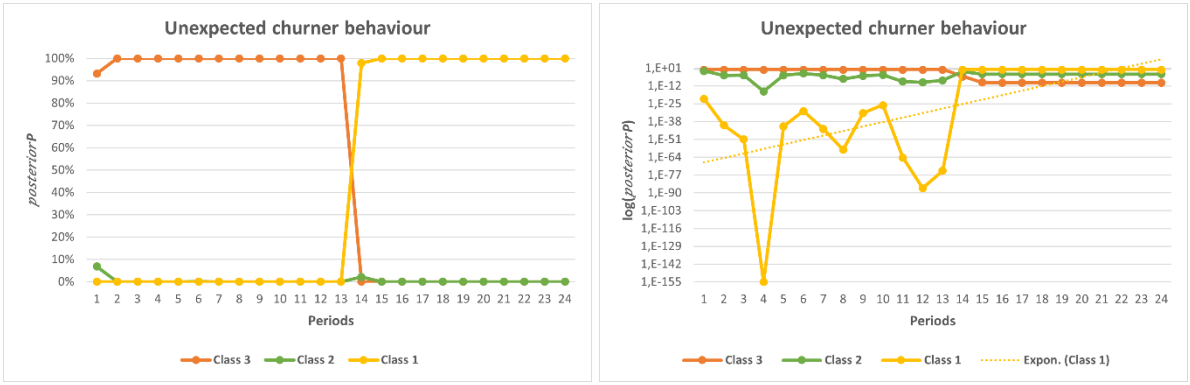


Figure 0.4 - Posterior probability of an unexpected churn customer

In this case the selected client, after a very stable period spent in the highest Class 3, suddenly churns. This behaviour is highly unpredictable and often caused by external factors, over which it is not possible to have any visibility or control. As can be clearly seen in the logarithmic graph, there are no particular signs that would point to a possible churn. On the contrary, in the two periods prior to the churn the value of the posterior probability associated with being in the lower Class 1 was expressing a downward trend, meaning that the customer was indeed doing better than before. This archetype is by far the most difficult to deal with; only the introduction of covariates to the model might help in recognizing this kind of sudden churns.

Partial defection:

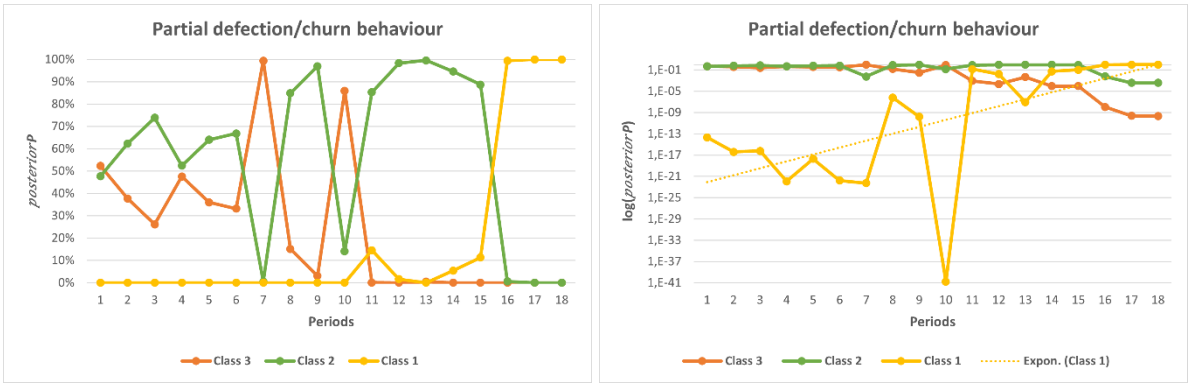


Figure 0.5 - Posterior probability of a partial defection/churn customer

Customers who perform partial defection are those users who after a few stable periods in one of the two high classes (or with small transitions between them) begin to show signs of relationship deterioration. In particular, in the first graph it is shown how the client in the first ten periods is fairly constant, net of some acceptable variability. From period eleven, on the other hand, there is a clear sign of the beginning of a partial defection process: the probability of belonging to the highest class is zeroed out and at the same time the probability of belonging to the lowest class increases significantly, even though the client is actually classified in the middle Class 2. This sudden increase is far more visible in the logarithmic scale graph: transition probability to Class 1 goes from orders of magnitude of E-20 up to orders of magnitude ranging between E-01 and E-05, while transition probabilities of belonging to Class 2 or Class 3 demonstrate a reverse trend (decreasing the probabilistic degree of membership to these two classes). The partial defection is even more evident observing the trend line of the probability of belonging to Class 1. In fact, such trend line is quite pronounced upward, meaning that the associated risk of churn increases constantly during time. This customer represents the typical case in which the use of the Markov model would have allowed for an early warning of customer churn, in the example shown around five months in advance.

Occasional customer:

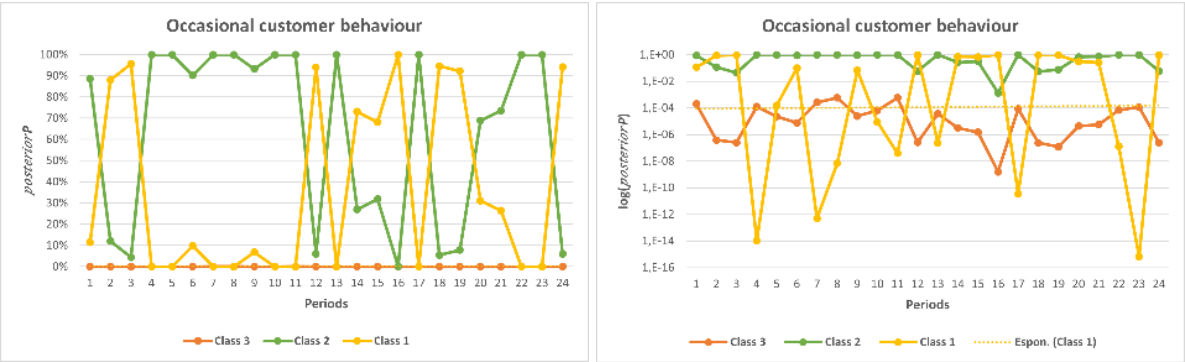


Figure 0.6 - Posterior probability of an occasional customer

Some customer behaviours are characterized by very frequent transitions between classes (often between Class 1 and Class 2).

It can be seen from the posterior probabilities graph that the client in question is never assigned to Class 3 but rather continues to move from Class 2 to Class 1 and vice versa. In these cases, it is not easy to tell whether such a customer is actually a churn/partial defection. However, it is still possible to carry out marketing actions aimed at stabilizing the customer in Class 2 so as to keep his behaviour away from churn.

Model validation

In order to verify the stability of our model, additional transactional data were retrieved from the retailer's management system; in particular, these represent new customers never used before. What we want to evaluate is the behaviour of our model as the data provided change. A new hidden Markov model will be trained with the new data and compared with our Master model.

The new dataset includes 6549 new customers. Like the main dataset this new dataset collects all the receipt lines purchased by new customers.

The main points to be checked are the following:

i. The model, with a fair variety of client, is able to recognize the three risk classes: high risk of churn/partial defection, medium risk and low risk. These classes are not likely to be numerically identical to those in the Master model, indeed the values describing the classes obtained are strictly dependent on the input values on which the HMM is fitted.

ii. Consistency between the Monetary and Frequency values of the classes. The model is expected to be able to recognize that the class with lowest Monetary corresponds to the class with lowest Frequency, and so on.

iii. The lowest class must also be able to accommodate churner customers: the mean value of Monetary and the standard deviation of Monetary must be similar in absolute value in this way we are able to correctly classify even the cases of Monetary close to zero.

iv. The impact of covariates perc_pvl_retailer and perc_pvl_premium must be conserved. The new data should provide the same insights regarding the reduction of risk level as the percentage of private label bought increases. Moreover, perc_pvl_premium must have a higher impact than perc_pvl_retailer especially on the transition to Class 1, the one associated to highest risk of churn/partial defection.

Below are shown the characteristics of the test model classes:

HMM test	Monetary mean	Monetary variance
Class 3:	246.078	144.208
Class 2:	130.675	80.738
Class 1:	26.342	23.584

Table 0.5 - Test model classes: Monetary

HMM test	Frequency λ (mean)
Class 3:	16.5
Class 2:	7.0
Class 1:	2.3

Table 0.6 - Test model classes: Frequency

It can be seen that the states found are consistent and coherent with the requirements listed above.

The trends of transition probabilities as the percentage of private label products increases are shown to be consistent and decreasing for both pvl_retailer and pvl_premium, with higher effect on pvl_premium. This means that the results obtained regarding the covariates are consistent and robust.

Master model:

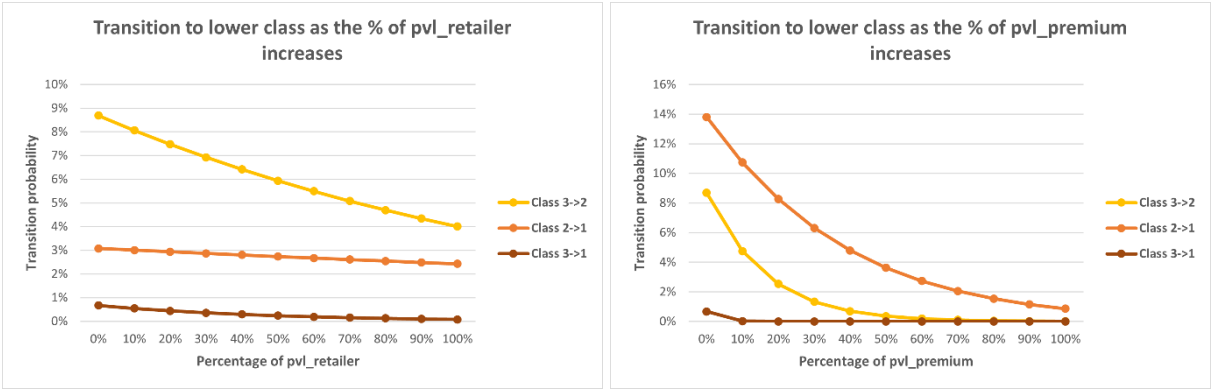


Figure 0.7 - Master model: covariates impact

Test model:

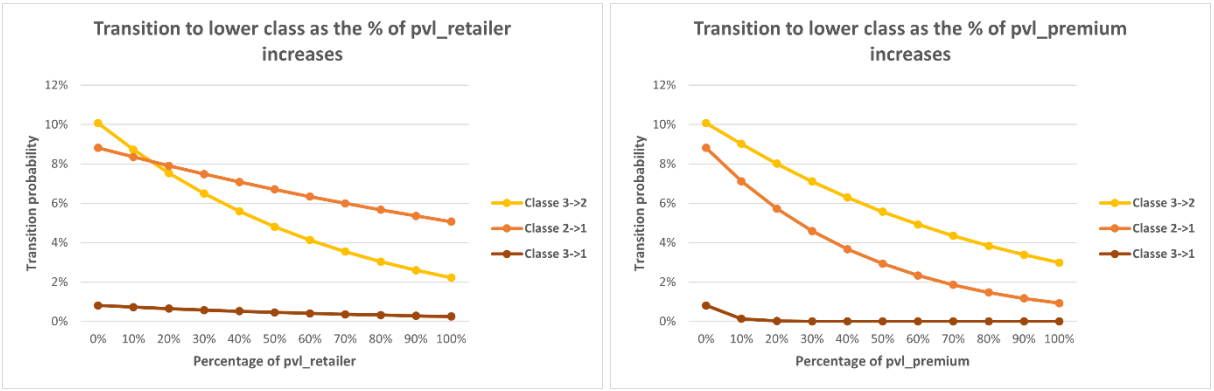


Figure 0.8 - Test model: covariates impact

Conclusion and future improvements

The model developed and discussed in this paper represents a potential starting point for subsequent improvements. Conceptually, the model may find further applications in other non-contractual setting markets.

It should be emphasized that although the proposed model has excellent behaviour and resilience to "noisy" data, since it is still a data analysis tool the latter will perform all the better as the input data are cleaned of spurious data. In such terms, any improvement in the cleaning and outlier detection modes could benefit the performance of the model.

Furthermore, as explained our model uses a normal distribution for fitting the Monetary of the three states of the HMM. From the knowledge acquired from working with retailer's transactional data, the distribution of monetary in customers would appear to be best represented by a gamma-distribution. Such distribution could bring better consistency to the classes and make it easier for the model to create the aforementioned classes of risk. This because often the segments created, even with traditional models, presents distribution of monetary quite left skewed.

A final cue, from which future improvements can be implemented, concerns the integration and testing of new covariates such as product categories (i.e. frozen foods, fruit and vegetables, fresh meat, hygiene products). These, in addition to bringing a quantitative improvement to the classification capability of our model, can generate valuable managerial insights to be used as strategic marketing levers.

Managerial implications

The first managerial implication is due to the greater stability of the identified classes, which means being able to precisely divide the customers into classes that better describes their purchasing behaviour and the associated risk of partial defection/churn. In economic terms, this model can be used to identify a pool of customers which are more likely to churn and abandon the store. Consequently, it is possible to specifically targeting them with retention actions avoiding wasting time and effort on customers who do not present a high risk of defection.

The second managerial implication concerns the results obtained relative to private label products. Thanks to the evidence shown before, we can conclude that retailer private label products are able to induce customer loyalty. This statement is even more valid for premium private label products. With these findings, it is therefore possible to introduce private label products for loyalty marketing campaigns, trying to incentivize clients to purchase and taste these products. Moreover, the model itself provides an answer to the retailer about consumers' perceptions of private labels. In fact, it is reasonable to say that if retailer and premium private labels strengthen the relationship with the customer, it means that the customer himself appreciates them and finds in them adequate value for money. Thus, such products not only provide a higher margin to the retailer but also become a real marketing tool.

Another managerial insight arises from the covariate's construction (which are calculated on monthly bases). This it is such that they are effective alarm bells for spotting the deterioration of customer relationship: if a good customer reduces its purchases in private labels, both retailer and premium, this can be noticed as an increase in the inherent risk of churn. Thanks to this implication the customer deterioration can be spotted in advance (i.e., churn prediction).

The last managerial implication concerns the validation of the model. On the one hand, it is easy to act by varying the type and the number of covariates in input to the model, accordingly to the result pursued for construction of marketing campaigns. On the other, as demonstrated by the model trained on new customers, the model has good flexibility to new data, and consequently resistance to the inherent variability in buying behavior, both in terms of monetary expenditure and frequency. In the end, the test with new customers validates the effectiveness of private labels in increasing customer retention; this shows that the result obtained from the Master model is for all intents and purposes attributable to actual buying behavior that discriminates high-performing customers from customers at high risk of churn/partial defection. From this result, it is possible to create ad hoc marketing campaigns to increase the degree of customer retention, particularly the customers in the high risk of churn class.

A final observation concerns the applications of the model: the model here disclosed was trained and tested on data from a single retailer, however, we have no evidence to affirm that the model is not applicable to data from other retailers active in FMCG sector: the metrics and KPIs used remain applicable, with appropriate adjustments, for any other retailer label. Moreover, the scientific-statistical basis of the model do not depend on the retailer from which the data was retrieved from.

1 Introduction

1.1. General context

In order to create value along customers lifetime the relationship between the firm and customers plays a fundamental role: nowadays to survive the competition and be successful it is required to be master in satisfying customers over time. On the one hand it is crucial for the firm to propose an attractive offer to onboard customer; to this preliminary step usually is associated an acquisition cost, hence the subsequent fundamental step is to create engagement over time, extending the period and the quality of the relationship between the firm and the customer, so a to create positive margin from customers.

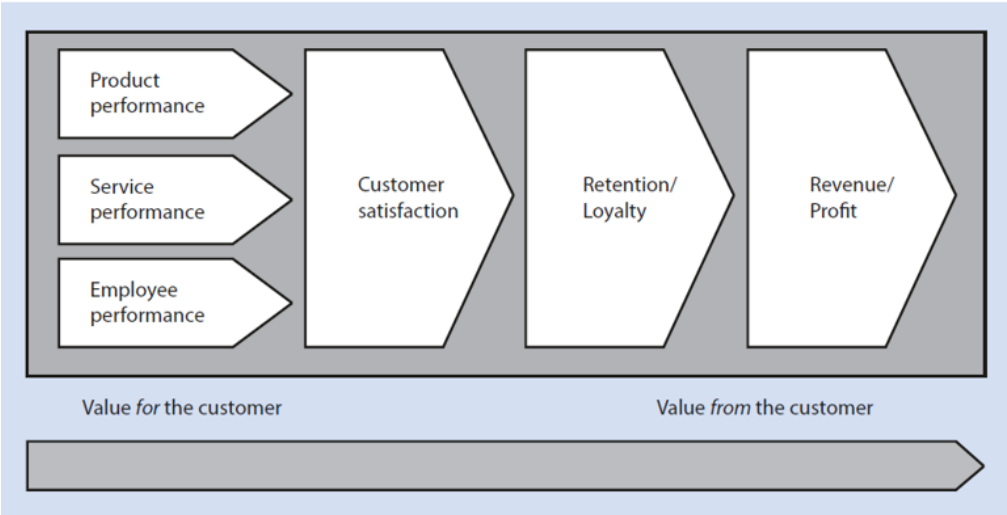


Figure 1.1 - The satisfaction-loyalty-profit chain (source: Anderson & Mittal, 2000)

"Over time" is what aligns marketing strategy to firm's value: firm's value is created by present and future cash flows generated by each customer in the firm's customer base, those cash flows must be sustained by consistent marketing actions able to lock-in customers and boost the quality of their relationship.

The quality of the relationship from the company's perspective is represented by total margin generated, measured by frequency and expense per purchase less the costs generated by the relationship (for instance, calls to customer care, marketing actions, customer acquisition cost, etc.), a good management of customer relationship can improve the overall expense and reduce costs, which generally arise from customer unsatisfaction or from the inability to address the right customer with the most suitable marketing action for him. From the customer point of view quality of the relationship is measured by how much the whole customer experience is satisfactory, going from the product system itself to the complementary services offered (the big challenge here is to find the right balance, since poor satisfaction can have drastic results on customer retention but not with the same magnitude a positive satisfaction boosts retention: not symmetric behaviour).

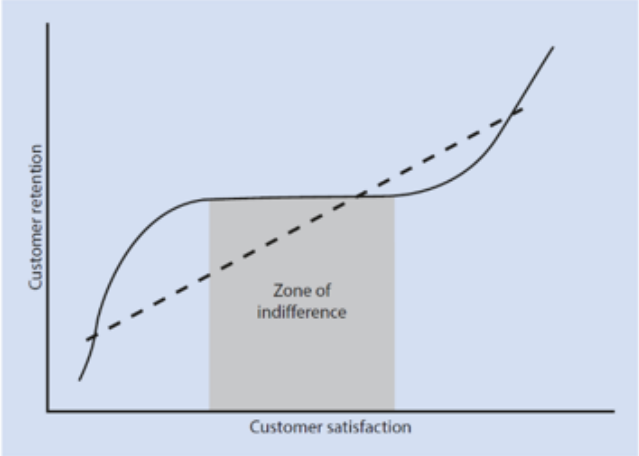


Figure 1.2 - Illustration of the satisfaction-retention link (source: Anderson & Mittal 2000)

Summing up the ideal customer for a company is who spend enough to generate a positive margin and do that in a consistent way during time. Reality is that customers evolve and change during time, and sometimes their relationship with the firm comes to an end, meaning that the customer stop buying from our company. This part of the customer life is covered under the concept of churn. The theme of the churn analysis and churn prevention has recently become one of the most relevant topics of research in the customer relationship management, this is especially true in the food and grocery retail industry: due to the non-contractual nature of food and groceries retail industry, a former customer would leave without directly announcing it to you.

Furthermore, in this industry customers experience low switching costs, having a broad range of options, and being able to easily compare offers and prices. Being tempestive in discovering symptoms of a churning customer, and reasoning behind, can help the firm increasing its customer retention, and accordingly with what we stated before, increase overall customer base value.

1.1.1. Relevant megatrends

In this section we reported some megatrends connected with the topic of churn, customer behaviour and customer relationship management. The following trends are influencing the context in which our topic is centred, they also represent the reasoning why customer relationship is becoming key in more and more industries.

Concept of churn and partial defection is in its essence deeply connected with what drives customer demand: as we stated in different instances, a customer will end its relationship with the company when he/she somehow find out a competitor that can better satisfy his/her needs, or whenever he/she experiences a dissatisfaction from the relationship with a firm. However, what the customer needs and satisfies today may not be what the customer will need or be satisfied with tomorrow, which means that customer behaviour and demands change over time, megatrend are those big topics that influence this change. Moreover, the technological progress and the spread of accessible and low-cost technological infrastructures (such as social networks and internet) deeply influence the way marketing is done and the way customer interacts with companies, giving life to new possibilities both in terms of marketing actions and the way churn or partial defection is monitored and analyzed.

1.1.1.1. Big data and implications for customer relationship management

First thing first we must define what are big data. Referring to the famous 5Vs model retrieved from literature, big data are those data that have at least one out of the following characteristics:

- **Volume:** every day we generate plenty of data. Volume refers to this huge mass of information, which cannot be collected by traditional technologies.

This volume of data is constantly growing, international analysts estimate that data production in 2020 will be 44 times greater than that of 2009. Meaning that for this reasoning it is difficult to set a threshold above which we can speak of Big Data. For now, as a rule of thumb, we identify as big data those data that are larger than 50 Terabytes or volumes of data growing more than 50% annually.

- **Velocity:** data are generated and captured more and more rapidly. This is due to the proliferation of devices equipped with sensors capable of collecting data in real time. What is nowadays challenging companies is the need not only to collect this data, but also to elaborate and analyse it in real time so that business decisions can be made as timely as possible.
- **Variety:** "More isn't just more. More is different." - so wrote Chris Anderson in Wired magazine, it was 2008. Variety refers to the different types of data available, generated from a growing number of heterogeneous sources. Not only transactional and business management systems, but also sensors, social networks, open data. Those can be structure and unstructured, both coming internally or externally the company. Variety also brings more exploitable information.
- **Veracity:** data must be reliable, tell the truth, the underlying idea here is what the one in the field want to avoid: trash-in trash-out. With Big Data, this has become even more challenging: data management technologies change, the speed at which data is collected changes, and the sources increase. However, the quality and integrity of the information remains an indispensable pillar for bringing to life analyses that are useful and reliable.
- **Variability:** much more data, in different formats and from different contexts. The mutability of their aspect is something that needs to be considered as data are interpreted.

Big Data recently has been referred to as the new oil or gold, since it represents an invaluable source of value. Nonetheless, the mere collection data, even if done by leveraging the best technologies available on the market, is not synonymous of information neither of knowledge.

Big data must be treated and transformed into information, and successively this information must be understandably reported to whom it may concern in order to generate an effective knowledge transfer, which eventually will represent a valuable resource to boost performances. Hence it is dutiful to consider another V out of the ones bulleted above: Value.

As we discussed above, analysing and processing large masses of data enable the generation of new knowledge useful for much more informed decisions making. All of this is made possible by technologies that enable management and processing of Big Data in constricted timespan, but also by the spread of innovative algorithms and analysis methodologies that can autonomously extract the information hidden in the data. In the following paragraph we will highlight the main Analytics designs and methodologies; we begin with the four classes of Analytics used in data analytics:

Descriptive analysis: the objective is to describe the current and past situation. This is the case with performance indicators and dashboards to monitor the main business KPIs (they are as a matter of fact the most widely used type of analytics). Moreover, descriptive analysis represents the entry level method in data exploitation, meaning that they are usually the first and easier to be implemented.

- **Predictive analysis:** data are analysed to understand what might happen in the future given current conditions and the knowledge we have of the past. It can be used for budgeting as well as planning purposes.
- **Prescriptive analysis:** the objective is to propose solutions based on the future predictions made. Therefore, they represent a further step after predictive analysis.
- **Automated analysis:** automatic implementation of the proposed solutions after analysing the data.

Big data and analytics are strongly related, since the first one need the latter in order to extract valuable knowledge. Dropping those concepts in the customer relationship management topic, it means being able of enhancing performance, and segmenting customers according to their purchase behaviour boosting business profits. What is more interesting for this dissertation, the concept of Big Data technology enables firms' management to concentrate on customer retention (Wu et al., 2014).

So far, we got that Big Data techniques helps CRM systems to process customer information faster and more smoothly, improving business operations. However, in order to give a valuable argumentation, we also must evaluate implementation cost of those technologies: According to Kunz et al. (2017) Big Data technology requires a great economic effort to be implemented in the most efficient way. In addition, there are high maintenance costs for both software and hardware infrastructure, and this is a major constraint for medium-small organisations. On the other hand, benefits are much greater, in fact BDT is the key to understand customer buying behaviour and from that deploy efficient marketing decisions aimed at increasing customer value, and eventually company value itself. Another great advantage of BDT is the ability of concretely evaluate marketing expenditure and investments, allowing to predict scenario. With transaction data it is possible to predict how decision will impact customers.

There are obviously problems that organisations must face when implementing big data with CRM. Data incompatibility, possible errors, high complexity and security problems are the main ones. Security is a very sensitive issue as it involves a compulsory risk since detailed information on individual customers is needed (privacy issues in the event of a data leak). Moreover, when talking about CRM we must take in mind that masses of data generated by customer exist under the shape of unstructured data, this means further complexity in their management and analysis.

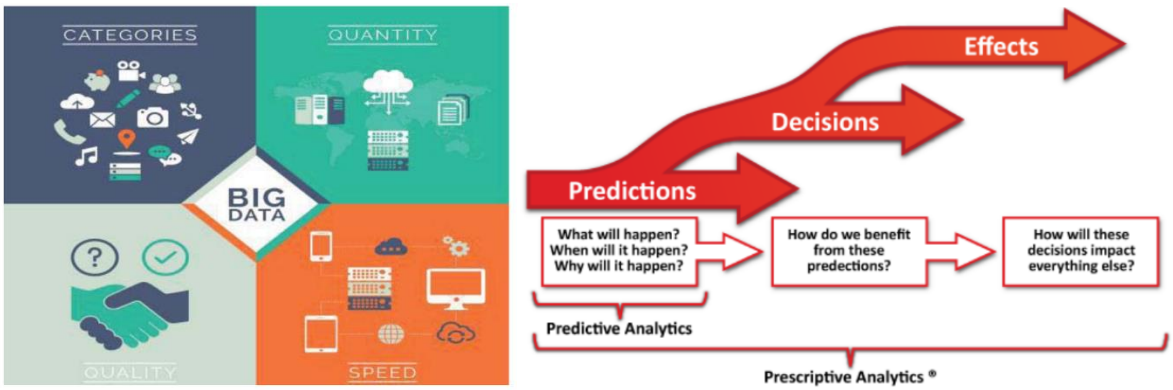


Figure 1.3 - Big-data CRM features - Source: (Stimmel, 2016)

1.1.1.2. Customer centricity and servitization

In recent years, the concept of servitisation or 'service transformation' has developed rapidly. The origin of the terminology dates to a 1988 article by Sandra Vandermerwe and Juan Rada: “Servitisation of business: Adding value by adding services”. Indeed, since the late 1980s, the trend to offer services as an integral part of the product began to develop (a trend that would later be identified with service as the product itself). The main enablers of the transformation of the offer are therefore digital technologies and their ability to collect and analyze data.

It is crucial, when talking about 'service', to be able to capture the customer's desires in order to design the tailor-made experience and make it effective.

This focus on the customer by putting them at the center of every choice is called customer centricity. In CRM, the potential is very high thanks to the use of Big Data as you have a huge amount of data to create experiences built around the customer.

1.1.1.3. Customer experience

Today, consumers give credit to a positive brand experience more than ever. The most important drivers for a good customer journey are speed, convenience, and friendliness. Customers are more attentive to the interactions they can establish with companies and in fact the desire to be part of a community where they can compare and provide feedback has become a key point. Brands who provide customers with consistent and direct communication can foster a sense of community that increase customer loyalty reducing at the same time acquisition costs. There are three key strategies in FMCG: creating one-to-one customer experiences, creating communities that engage and make the customer feel important, and finally innovating products and services by exploiting community feedback and suggestions.

It is important to remember that creating a successful customer experience is not only about engaging the customer but also about being able to reap benefits from this such as reducing time at the checkout stage, collecting useful data and observing the reaction of consumers to the stimuli provided in-store and out-store (cashless payments, real time behaviour observation, audience analytics).

1.1.1.4. Demographic and Behavioural changes challenging customer relationship management

The social, demographic, and cultural changes of recent times have created continuous challenges in CRM. In general, the average age of the population has increased (mainly in developed countries), thus necessitating adjustments to adapt the offering to a higher age group. Adaptations that have also become necessary due to the increasing diversity of customers that has reached individualism. Each customer sees himself as different from the other and wants to be recognized as such.

The digitisation trend also continues to play a key role in sharing news and experiences. Social media have also become powerful communication tools through which brands can be promoted to great effect. The creation of a good social image is of increasing importance.

The possibility of having access to real-time data has extended the horizons of customization like never before. Being able to react instantly to customer requests and behaviour provides a major competitive advantage. Real time data analytics is the type of analysis that consists of analyzing big data at the exact moment it enters the system, giving the company an immediate visualization and understanding of the data.

This ability also makes it possible to respond to the desire for exclusivity and authenticity that is sought even in an everyday action such as shopping at the supermarket.

One of the most important drivers for customer decision-making has become time; differentiation with competitors is often not only guaranteed by the experience itself but rather by the timeframe to create it and get it to the customer. Also, in the area of customer retention, companies must act faster and faster to retain customers. Using real-time analysis techniques is a good place to start in order to better manage one's customer base by evaluating the effects of decisions made by the company's management.

Connected with the desire to reduce time is the complete digitization of checkout operations. It is now common to find self-service payment zones in large chains where the customer himself performs the checkout operations with the aid of technology.

This procedure reduces queue time within the shop, increasing the performance of the shop and at the same time making the customer experience more valuable by not having to waste too much time in line to pay for their purchases. It is therefore clear that going to the supermarket has become a fully-fledged interactive experience where every single detail must be studied to meet the needs of each customer.

A final trend that has grown to such an extent that it has become one of the main drivers of choice between store chains concerns the focus on health and, more generally, on healthy and organic products. The number of customers willing to pay a premium price for certified and more controlled products has increased. At the same time, companies have realised that care for the environment and the ability to take initiatives in this direction can be the differentiator from competitors. The number of individuals who are sensitive to this issue is widening considerably and good corporate communication can lead to excellent business results.

1.2. Customer Relationship Management

About four decades ago, the marketing discipline witnessed a grand paradigmatic shift from focusing on products to focusing on customers that paved the way for the concept of customer relationship management (CRM) The advent of the Internet and the information technologies (IT) has provided firms with opportunities to connect better with the customers, respond directly to their queries, customize solutions, and maintain better relationship with them. As such, in the early days, many considered CRM a technology-based customer solution.

CRM is the set of activities involving selection of most profitable customers, shaping of the interactions between them and the company, with the purpose of improving productivity and maximizing their profitability range for our company. The CRM definition covers a wide variety of subjects that concentrate on business-focused activities. Information and communication technologies (ICTs) components integrated into a CRM strategy enable the automation of those business processes aimed at building and maintaining profitable and sustainable customer relationships.

1.2.1. CRM perspectives

- **Strategic CRM:** focus on creating customer-centric business culture where one directs decisions towards optimizing customer value.
- **Operational CRM:** its main task is to automatically collect data about customers at various touchpoints, as well as aligning and managing the workflow across the sales, marketing, and services divisions.
- **Analytical CRM:** it consists in the application of data warehouse, data mining, and text mining for solution of complex business problems. Analytical CRM uses technology to process and analyse customer-related data that organizations can use to implement marking strategies aimed at increasing customer satisfaction and organization value.
- **Collaborative CRM:** here actions and technologies are implemented with the idea of pursuing better collaboration between the company and customers.

1.2.2. Evolution of CRM

This section describes the stages of development of CRM, from the 1990s until today. It has grown from a tactical marketing tool to a strategic element in all marketing decisions. The growth of the Internet also has increased the adoption rate of CRM in many industries. In the beginning the activities that lather on were covered under the name of CRM were divided among two different products: 1) Sales force automation (SFA) that addressed presales functions, such us maintain data from customers, generate leads and placing sales orders. 2) Customer service support (CSS), including all after-sales activities, such as help desk, call centre and service support. Later, in a second evolution stage CRM took a much more integrated approach, with the aim of connecting Enterprise resource planning processes to the customer, hence incorporating the SFA and CSS. During the early 2000 lots of companies realized the effectiveness of using a strategy-oriented CRM adoption, in this way going from a cost-control perspective to a revenue enhancing one. In the same time internet related services spread generating new needs in customers and new challenges for companies. CRM boosted, thanks to internet technologies, and took a furthermore integrated approach also developing the possibility to interface with systems used by suppliers and partners. Successively, the advent and the diffusion of social media together with

the technological advance, evolved CRM toward a social CRM perspective characterized by the engagement of the customer through the integration of the web 2.0 and social media and using data driven insights to optimize the overall customer experience. Companies encourage active customer participation online,

while they use software applications to track real time social data. This information enables companies to offer relevant content and personalized messages to specific customers and to improve the customer experience at each touchpoint along the customer journey. Additionally, the combination of data across different social media platforms allows companies to determine the customer value not only based on profitability but also based on their online behaviour in terms of referrals, knowledge dispersion and influencing other members of the social media community.

1.2.3. CRM and artificial intelligence

Nowadays CRM is strictly connected to artificial intelligence (AI) practices. Due to the IoT and other technological advancements data production and collection is skyrocketing. CRM has had to evolve to capture and analyse this growing stream of information and data points about customers. Accordingly, many organizations have implemented AI in CRM through machine-based learning algorithms to analyse large CRM data sets. AI is capable of analysing trillions of data and offering concrete solutions in millisecond timescales. The outcome of this symbiosis can potentially result in an extremely enhanced customer experience: in fact, being able to understand the singular customer's behaviour can lead to the company to implement one-to-one marketing actions, for increasing customer satisfaction and remove each kind of frictions. Furthermore, the use of AI on this large customer dataset also can help companies in programming activities (for instance better forecasting of production), hence resulting in leaner processes and cost reduction.

1.3. Concept of churn

1.3.1. Churn in contractual settings

Often a customer-supplier relationship is governed by a contract, taken in many different forms. Nonetheless, the main concept is that there is a clearly identifiable point in time in which termination occurs (sometimes this date is stated in the contract itself). A customer is classified as "churn" when the selling contract comes to an end, because it reached maturity either because the customer for some reason decided to withdraw from it.

At a CRM level, this means being able to have detailed and unambiguous information on the performance of the customer base, moreover it gives a clear event upon which evaluate the effectiveness of implemented marketing actions. It is crucial to remember that it is generally much more costly for a company to acquire new customers than to retain existing ones, this explains why measuring the incidence of churn is relevant for strategic marketing purpose.

It goes that the main indicator derived from the definition of churn is the churn rate, defined as the percentage of customers that stopped using your company's product or service during a certain time frame. For instance, it can be calculated by dividing the number of customers you have lost during that time period (i.e. a quarter) by the number of customers you had at the beginning of that time period. For calculation's sake, it is also useful to introduce the concept of retention rate, which is ratio of the number of customers retained to the number at risk, the latter is usually identified with the whole customer base, since in most of the cases a customer can potentially end its relationship.

1.3.2. Churn in non-contractual setting, and partial defection

Nowadays, customers are subject to a wide range of choices when purchasing a consumer good, this is even more true in the food and grocery industry where they can quite easily compare alternatives, furthermore it is the case that in this market the presence of exit barriers is rare and soft, and in the meantime the number of internal competitors is high and usually for each geographical zone there is a wide variety of stores, each close to the other.

As a matter of fact, some customers are more prone to create a lasting relationship with a particular brand, on the contrary others prefer to change their purchasing behaviour frequently. In the FMCG world, the latter behaviour is facilitated by the almost total lack of switching costs. Indeed, customers usually split their purchases between the most competitive companies, thus the relationship is purely based on convenience expressed in monetary terms, and this is becoming always truer in the Italian context where discounts are quickly eroding market shares of traditional retailers. This means that when a customer experiences a lack of convenience, there is no barrier preventing him/her from terminating his relationship with that store chain and switching to a competitor, which is offering the same type of product (or a substitutive one) at more advantageous conditions.

The only way companies must fight this harmful characteristic of the market is trying to establish a strong relationship with the customer, making in this way switching to a competitor mentally more difficult, given the fact that it is not possible to create physical exit barrier or switching costs (contrariwise in a contractual setting this is possible, for instance specifying in the contract a penalty for contract recission).

Moreover, the lack of a contractual approach renders, as we mentioned earlier, the customer-supplier relationship devoid of any relational obligations. This leads to a total lack of notice when the customer wants to terminate the relationship. In the FMCG world, it is not possible to enter in any kind of contract with the customer, and therefore the concept of Churn as we stated it in the previous paragraph does not apply anymore.

To better describe this situation is introduced the definition of partial defection: when the company experiences a substantial drop in revenue for a given customer, that is called a partial defection. In fact, the customer has not completely broken off the relationship (as it happens in a complete churn), but it is very likely that this will happen in the short term or that the customer is now preferring a competitor. As a matter of fact, what counts is that in both situation the customer has significantly reduced the intensity of its relationship with the retailer under question. Nevertheless, it does not always happen that a customer goes through a partial defection path, sometimes he/she just only stop buying from one moment to the other, those kinds of churns are the most difficult to spot, due to their rapidity and unpredictability essence.

Even if partial defection is a viable alternative for non-contractual setting it does not come with no threats. In fact, sticking to the definition of partial defection we have to identify when a customer reduces his/her purchasing performances; each customer has his/her singular purchasing behaviour (both in quantity and frequency) meaning that it is not possible to define a general rule or a general partial defection behaviour (for instance for a customer that is usually doing groceries for 500€ per month lowering this performance to 250€ per month represents a partial defection, we experience the same kind of partial defection for a customer that usually spend 200€ per month that now spends 100€ per months): it is not possible to perform a partial defection analysis with steady thresholds, notwithstanding which is the historical purchasing behaviour of the customer. This means that the model used to detect a partial defection must somehow take into account the past performance of the customer and compare it to the actual one, in order to classify the customer under observation as a partially defected.

1.3.3. Managerial reasoning behind churn analysis

It is of paramount importance in a first place to define which customer is performing a partial defection and with how much confidence we can classify he/she as partially defected, successively it is extremely helpful to grasp the predictive signs of the deterioration in the customer relationship, so as to act timely and precisely to avoid the ultimate churn. This is explained by the acquisition cost of a customer, it usually is one of the highest costs related to customer during his/her lifetime: every lost customer represents gone money for the company; being able to promptly intercept a customer before his/her churn prevents the loss of that initial investment. Also, understanding the drivers of churn can be extremely useful from a managerial perspective since it allows to better and more precisely direct marketing efforts done to retain customers.

1.4. RFM Segmentation

The best known and most exploited methodology in CRM for analysing customer performance with a data-driven approach is Recency Frequency Monetary analysis (RFM). The objective of this analysis is to segment the customer base in classes able to discriminate the best performing customers ("champions") from those with poor performance ("frozen/lost"). The classes will then be used to fit specific marketing actions, aimed at boosting customer retention and leading customers toward the best class.

1.4.1. RFM Recency version

This type of analysis requires the availability of the so-called "transactional dataset", which is the register containing all the receipt lines purchased by each fidelity card during the opening periods of the stores. Thus, having at our disposal data of purchase done by each fidelity customer in monetary and temporal terms, it is possible to aggregate them in time periods (the choice of the time bucket is data analyst responsibility) and successively derive the following variables:

- **Recency:** defines how long it has been since the customer's last purchase. The lower the value, the "warmer" the customer is, the better. It is computed as the delta in days between last registered purchase and the day of the analysis [days].
- **Frequency:** it defines how often a customer makes a purchase and refers to the total number of visits by the customer during the observation period [visits/period]. It is commonly assumed that the higher is the number of visits done by a customer, the higher is the loyalty towards the company, and thus the better. It is also the case that customers who came more often to the store are the ones that can be retained more easily by offering them in presence marketing actions.
- **Monetary:** defines how much a customer spend in his purchases. For the RFM analysis it is needed the cumulated value along the observation period for each individual customer [€/period]. It therefore reflects the contribution of the single customer to company's revenue.

The higher the monetary value the higher the contribution to the company turnover, and thus the more valuable is the customer in financial terms.

1.4.1.1. How to conduct a Recency – Frequency – Monetary analysis

The first step in conducting an RFM analysis is to identify the time frame of the analysis itself. The length of the analysis period/periods should be consistent with the context in which we are performing the RFM, starting with the type of good sold. For instance, in the case of a food and grocery retailer a consistent time frame generally corresponds to the last three months, this is because food and grocery represent convenience goods. If we were to perform an RFM analysis for a car dealer, for example, the time frame would undoubtedly have to be larger.

Once the time frame is set, the procedure to follow is to assign each client with a discrete score (usually from 1 to 3 or from 1 to 5) for each of the three variables above mentioned. It should be borne in mind that the numerosity of the classes in output from the RFM is directly proportional to the cube of the numerosity of the score scale, for example: one segment will be identified by the customer that score R=1, F=1 and M=1; another segment by those customers that scored R=1, F=1 and M=2; and so on encompassing all possible combinations. It is therefore appropriate to define the numerosity of the score scale consistently with the specific data on which the analysis is to be conducted, always considering that each class created will have to be justified by a recognizable and useful population for the purpose of marketing actions.

This score is assigned using thresholds. There exist different methods to set those threshold, divided into two macro categories: first method consists in established thresholds in a subjective way (empirical method), here the choice of the final value is responsibility of the analyst, this method is easier to implement and more flexible, thus more suitable for small business or for preliminary analysis, it il less time consuming and a good alternative when no much information can be extrapolated from data. On the other hand, the second type of methods are based on the statistical process of distribution analysis: deciles or percentiles are used in order to set thresholds, trying to identify, if it exists, the so-called elbow of the distribution (that intrinsically represents a partitioning between customers). The latter methods are the most rigorous and suitable for a more precise and complete analysis.

Moreover, if the analysis is carried out over several periods, in other words a “rolling analysis”, the use of the statistical methods ensures to take into account components such as trend and seasonality in defining the thresholds, eventually assigning more coherent scores for each RFM variable.

The further step, following the score assignment phase, consists of the class creation and interpretation, in fact at this point the various RFM values are combined in order to create consistent classes for describing the customer base, and finally for the implementation of marketing actions. If, on the other hand, a more synthetic result is desired, a next step can be taken in which a unique score to be assigned to the customer is calculated: this in most cases is done through a simple average or weighted average of the scores obtained along each variable. However, there are more complex statistical methods for obtaining clusters, such as K-means or hierarchical clustering models.

In most cases this version of the RFM is used on the latest available time frame, the underlying idea is to monitor the performances of the customer base in a methodological, thus understanding whether specific marketing actions are needed on specific classes of customers, eventually benefiting from a more directed marketing effort. For this very purpose recency is chosen to be used as a good measure of the "warmth" of a customer. When a customer is recognized to poorly perform in recency, mired and prompt marketing action must be taken to try not to lose the customer permanently, especially if the customer under analysis has always been a good performer. It goes without saying that it makes little sense to use Recency for comparative purposes over time, which instead is the purpose of the dynamic version of RFM analysis.

1.4.2. RFM Regularity version (RegFM)

It is possible to improve the classic RFM analysis, which is static from the temporal point of view, by including in the analysis itself the periods prior to the last one: in this way it is possible to have a perspective on the evolution of the customer over time. This information can be extremely useful in order to validate the marketing actions carried out or being carried out. In addition, performing a rolling analysis makes it possible to identify significant changes in customer behaviour, and ideally find reasoning behind that change. It is evident that with this type of analysis is gradually approaching the study of partial defection.

As discussed in the previous section, it is necessary to replace Recency (which in the case of comparing successive periods loses its meaning) with a variable that in some way describes how the client distributes his expenditures during the period under analysis. The variable generally used in place of Recency is Regularity:

- **Regularity:** it defines how regular is a customer in visiting the store. Regularity is usually measured observing the stability of the interpurchase time over an observation period; interpurchase time is the number of days elapsing between two consecutive visits. The higher the regularity the higher the loyalty towards the company, and thus the better. Interpurchase time stability can in turn be measured in different ways; the most immediate and applicable one, which does not require specific constraints on data, is the average of interpurchase times observed over the period for the client under analysis.

All the other steps of the RegFM are analogous to the ones in the classical RFM, except that they must be repeated for each time frame in which the dataset is divided into.

Moreover, introducing the concept of Regularity makes it possible to compare this value over time, allowing the identification of situations of decline that otherwise would not be possible to spot. In fact, the recency value is not indicative of average behaviour but only tells something about the last purchase.

1.4.3. Use of RFM in churn and partial defection analysis

The very concepts of churn and partial defection are closely related to the measurement of a customer's performance, which in turn is well described by the combo of Recency/Regularity, Frequency and Monetary variables. Therefore, it is straightforward to use the RFM, in both the presented versions (Recency-Regularity), as a valuable tool for churn and partial defection analysis.

As a matter of facts, the output of the RFM model consists of a classification of the customer base into risk classes (for the rolling RFM, the classification is repeated as many times as the number of periods) through which it is possible to identify those customer segments most likely to churn. Moreover, in the rolling RFM analysis it is also possible to identify those customers who did a partial defection, being able to reconstruct a trend in customer performance.

Marketing managers use the segments identified through the RFM to create tailored marketing campaigns, with the aim of 'reactivating' and 're-following' the most at-risk customers, as well as keeping the good performers stable. It is worth emphasising that RFM is not a predictive model, since it analyses the customer's past performance without saying anything about what future performance may be. Eventually, prediction remains analysts' responsibility, who will have to implement subsequent methods, with their own limitations, for estimating the future client's performance.

1.4.4. RFM limits

The RFM analysis, while easy to interpret and implement, has some substantial limitations. Those limitation are disclosed and explained in the following paragraph since they represent a key information for understanding the reasoning behind the necessity for the more comprehensive model that is disclosed in this paper.

It is an analysis based on an a priori choice of thresholds on which the classification will then depend. This results in a lack of robustness, as a change in the thresholds would result in a totally different classification.

Furthermore, the output of the RFM is qualitative, meaning that the model is unable to assign each customer with the probability of belonging to the segment in which it has been classified. This is extremely critical in cases where the customer is on the borderline between two classes: for example, assume an RFM analysis with scores from 1 to 3 (low – high), focusing on the Monetary classification. It might be the case that the Monetary value of a generic customer is around the threshold value that discriminates middle and high class. Assuming that the observed value for the customer in question is slightly below the threshold 2-3, this means that the customer will be classified as an average customer (class 2); such a customer for the RFM model is considered on a par with a class 2 customer whose disbursement in the period slightly exceeds the entry threshold for the aforementioned class. This means that clients who are actually very different (they represent the extremes of the class) are equated, instead a more rigorous analysis would be able to give information about the fact that the first client belongs to class 2 but also is more likely to belong to class 3 than to class 1; the vice versa for the second client given as an example.

As a matter of facts, the main problem from a managerial point of view, is not being able to intervene in the most appropriate manner with customers. Indeed, marketing actions are massively directed at the entire segment considered to be at risk, resulting in a costly outlay. Moreover, as pointed out above, it is not uncommon for a customer to be classified in a non-risk class by a 'statistical rule' when in fact he/she is a risk customer. This last point is in truth extremely critical: although classifying a non-risk customer as a risk customer can result in an unnecessary effort, classifying a risk customer as non-risk may result in the actual loss of that customer, and consequently in the loss of the economic flow generated by him.

1.5. Thesis objective

1.5.1. Structure of the academic research

From the research perspective our plan of actions is modelled upon the following research framework:

- Firstly, we will analyze and review the literature to understand the state of the art in modelling churn/partial defection. Furthermore, we will take advantage of the literature to identify the most common and diffused techniques and tools used in this field. Eventually, at the end of this phase we will come up with the research questions we will try to answer with our research.

It is worth to mention that the order of the following steps is not so strict as it may seem, as a matter of facts, it is true that conceptual model design part comes before the actual writing and implementation part, nonetheless it might be the case that some events and some discoveries done in the implementation part deeply change some aspects of the model's design, the fig. 1.4 well explains the concept from above.

- Second step will be the conceptual design of the model, in which starting from knowledge acquired from literature review we will build up the underlying structure of the model, keeping always in mind to be consistent with the topic and the strategic marketing theory.

- Conceptually the third step is the actual implementation of the algorithm, there we will go through the dataset management part and all the practicalities involved with the drafting of the algorithm itself.
- Last step will consist of the presentation and the discussion of the obtained results. We will highlight the improvements that our model brings to the state of the art in the field of churn management, and we will answer the research questions we posed ourselves in the first phase. Furthermore, we will discuss the main limitation and potential improvements of the model, always maintaining an economic-managerial perspective.

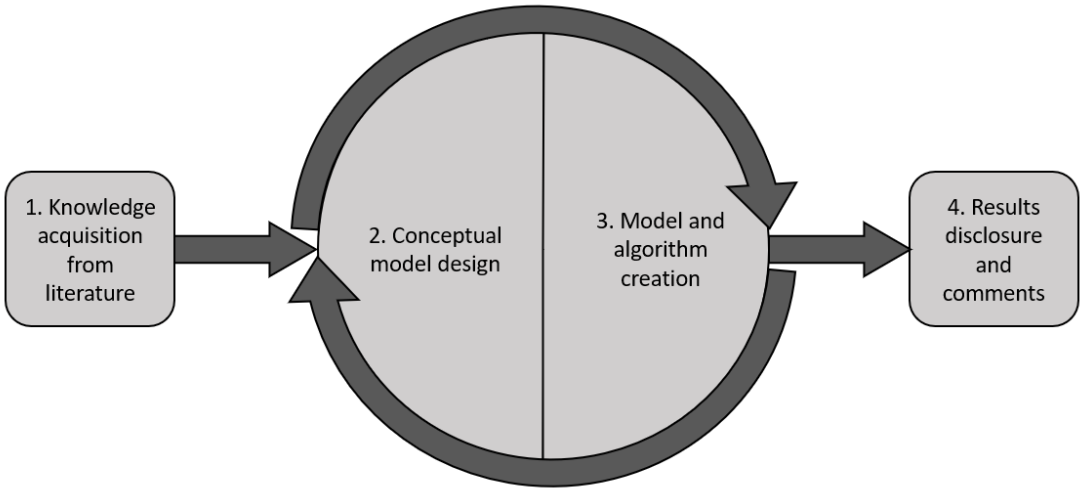


Figure 1.4 - Thesis development framework

Since in this thesis it is also disclosed an on-field application, the creation of the actual algorithm went through a lot of attempts, following the Build-Learn-Measure loop. For simplicity and non-redundancy's sake in this document we will only disclose the final version, and if necessary, the most important ones for a correct understanding of the model.

The chapters of this paper are designed accordingly to the framework above. Keep also in mind that this structure, which sometimes might seem redundant, is done for the sake of simplicity, hoping it results in a clear and pleasant explanation of the model functioning, as well as the functioning of the statistical tools used.

1.5.2. Structure and aim of the model

The model we will present in the next chapters aims to develop in more depth the issue related to the early detection of partial defection/churn, as well as to identify useful insights for the implementation of preventive measures against the latter.

We will use the transactional dataset from the management software of one of the largest Italian companies in the FMCG field. The main steps in which our work will be carried out are as follows: after doing dataset management and feature engineering (creating additional variables useful for later analysis), we will train a three-state Hidden Markov Model. The "response" variables (i.e., the emission emitted by the hidden Markov chain) will be frequency and monetary with the addition of a covariate, the latter will take into account the percentage expenditure on private label products, this was previously created as the number of private label items out of the total number of items purchased. The idea is to provide a probabilistic classification into three levels of deterioration of the customer relationship; with special emphasis on the lowest class, representing customers considered "at high risk of partial defection/churn." The main proposed improvement with our model will consist in the ability to provide a probabilistic estimate of the membership in each class for each customer, thus providing a much greater level of granularity respect to a classical RFM analysis, with greater performances in handling extreme cases laid at the boundary between two consecutive classes. In addition, we will seek to identify useful predictors (covariates in the model) for early detection of partial defections/churns.

2 Literature review

2.1. Literature review methodology

We have done a systematic literature review, in order to collect data in an ordered manner. During this process we tried to critically appraise and synthetize found articles both qualitatively and quantitatively. The article selection process was done on a three steps bases:

I. We used some relevant keyword in order to select a set of potentially insightful articles:

   - **Churn** literature research: *"Customer Relationship Management – CRM – CRM analytics – data mining – big data – churn – partial defection – churn prevention – churn detection – artificial intelligence – relationship marketing – customer loyalty – partial defection analysis – partial defection prevention"*
   - **Hidden Markov models** literature research: *"Hidden Markov models – HMM – mixture model – Markov chains – Markov models – Hidden Markov Chains – HMM fitting – Baum welch algorithm – Forward backword algorithm – Forward algorithm – Expectation maximization – EM – Log likelihood"*
   - **Megatrends** literature research: *"Servitization – Big Data – Big Data Technologies – Customer centricity – Customer experience"*

   Keywords were searched in different online scientific databases, such as: *"Scopus (Elsevier) – Google scholar – Springer Link – Research Gate – ScienceDirect – JSTOR".*

   Furthermore, we referred to some topic-relevant books as well as articles retrieved from reliable websites (such as osservatori.net), where we got a lot of inspiration and knowledge about the argument disclosed in this thesis.

II. Secondly, we decided a selection criterion, in order to filter and discriminate those articles not relevant for our purpose. We decided to carry out a peer-review work, reading abstracts of the papers selected from step one, and keeping only those connecting with the topic of this thesis.

III. Eventually, we integrally read the articles filtered in step two and included relevant arguments and concepts in our dissertation

2.2. Overview on churn literature

In the following section we will provide a summary on the literature related to churn and churn analysis that we reviewed during the research process. It was of paramount importance for us to acquire knowledge about the state of the art of churn detection and prevention techniques. As a matter of facts, before developing the model itself, we wanted to map and understand which statistical tools are the most valid and widespread in this field, so as to ensure that our model is consistent and relevant respect to the state of the art, represented by current research on the topic.

Papers considered for the churn literature review cover the period going from 2011 up to date, all were selected in English language. It is worth to be mentioned that during research phase we came across a lot of documents disclosing the churn in contractual settings, the majority regarding telecommunication industry applications, this is demonstrating that there are research needs for this unconventional type of churn.

As stated before, we faced a scarcity of material in the field of churn in non-contractual settings, this is even more true if we narrow to the research to application in the physical food and grocery retail industry (e-commerce represents most researched topic when dealing with non-contractual settings).

We found that in the literature the modelling of churn is mainly addressed under two different perspectives:

- Enhance predictive performance through increasingly complex models (combining different methodologies), thus focusing on the model's ability of recognizing at risk customers, although not focusing on the specific motivations.

- Understand motivations and drivers that lead customers to terminate their relationship with the company. In this second application, it is then possible to provide management with helpful insights that can foster marketing decision making, hence draw relevant actions aimed at fighting churn.

2.2.1. Techniques for churn analysis from literature used so far

From the analysis we were able to come up with the following summarizing table. In it we reported some of the most relevant papers written by guru authors in this field and the relative publication, it is also reported a brief explanation on the content, giving emphasis on the statistical techniques used.

General overview of literature in churn prediction modelling after 2011:

Authors	Title	Year & Journal	What	Techniques
Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, Bart Baesens	New insights into churn prediction in telecommunication sector: A profit driven data mining approach	2012, European Journal of Operational Research	A profit centric approach to evaluate customer churn prediction models.	C4.5, CART, neural networks, nearest Neighbours, logistic regression, and Support Vector Machine
A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, U. Abbasi	Improved churn prediction in telecommunication industry using data mining techniques.	2014, Applied Soft Computing, Volume 24	A hybrid methodology for boosting performances and extracting influential features in telecommunication.	Decision Tree, Artificial Neural Networks, K-Nearest Neighbours, and Support Vector Machine
M. Clemente-Císcar, S. San Matías, V. Giner-Bosch	A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings	2014, European Journal of Operational Research	A methodology for optimizing the selection of a churn definition in a firm based on profitability criteria.	RFM, Ad hoc function and class
T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.Ch. Chatzisavvas	A comparison of machine learning techniques for customer churn prediction	2015, Simulation Modelling Practice and Theory	Comparison between 5 of the most use techniques for customer churn prediction.	Artificial Neural Networks, Decision Trees, Support Vector Machines, Naïve Bayes classifiers, and Logistic Regression classifiers

Mehdi Mohammadzadeh, Zeinab Zare Hoseini, Hamid Derafshi	A data mining approach for modelling churn behaviour via RFM model in specialized clinics Case study: A public sector hospital in Tehran	2017, Procedia Computer Science	A case study in health-care sector to find potential for loyal customers.	RFML model with customer lifetime value (CLV)
Niels Holtrop, Jaap E. Wieringa, Maarten J. Gijsenberg, Peter C. Verhoef	**No future without the past? Predicting churn in the face of customer privacy**	**2017, International Journal of Research in Marketing**	**A comparison of the principal techniques for modelling churn**	**Logistic regression, classification tree, Hidden Markov Model**
Adnan Amin, Babar Shah, Asad Masood Khattak, Fernando Joaquim Lopes Moreira, Gohar Ali, Alvaro Rocha, Sajid Anwar	Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods	2019, International Journal of Information Management	An evaluation of the most common Customer churn prediction techniques	Naïve Bayes, K-nearest neighbour, Gradient Boosted Tree, Single Rule Induction, and Deep Learner Neural Net
Farid Shirazi, Mahbobeh Mohammadi	A big data analytics model for customer churn prediction in the retiree segment	2019, International Journal of Information Management	Improved Churn prediction model using CRT in retiree segment	Classification and Regression tree
Yixin Li, Bingzhang Hou, Yue Wu, Donglai Zhao, Aoran Xie, Peng Zou	Giant fight: Customer churn prediction in traditional broadcast industry	2021, Journal of Business Research	An application in Chinese broadcast industry where the competition is very high	Logistic regression
Sulim Kim, **Heeseok Lee**	Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees	2022, Procedia Computer Science	Customer churn in influencer commerce Korean Market	Decision Tree
Mai Kiguchi, Waddah Saeed, Imran Medi	Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest	2022, Applied Soft Computing	An in-depth review of related studies to DGBL and related works in churn prediction	Logistic regression, decision tree and random forest.

Table 2.1 - Statistical methods for churn analysis from literature

From the table above it is evident that there are some recurrent data analysis methods and techniques for Churn analysis.

It is important to emphasize that many of the methods we are going to illustrate are well suited to the study of churn situations of contractual settings, while they run into difficulties in our case (non-contractual settings). In fact, all these methods require already classified data to be provided as input for training the model.

This situation is typical for cases of churn in contractual settings, since upon contract termination the customer becomes explicitly for the company a full-fledged churn. It is therefore possible to trace the "churned customer – non-churned customer" information, which is critical for the proper functioning of the models listed above.

On the contrary, for cases of churn in non-contractual settings the information about the churn status of the customer is not directly observable and therefore not available. This gap means that in order to train these models it is necessary to go through a preliminary step in which the analyst discretionally assigning churned or non-churned status to clients in the dataset, thereby introducing substantial bias into the model, given by the churn definition adopted.

We are now going to cover the main models selected from red papers, explaining their usage and functionalities:

2.2.1.1. Ensemble methods

Ensemble machine learning is a machine learning technique based on the construction of multiple hypotheses. The term “ensemble” denotes a set of hypotheses that are processed. In machine learning, the concept of a hypothesis is identified as a decision tree, which is an entity composed of the various attributes ordered in a hierarchical manner. It is therefore an evolution of models based on a single hypothesis and this allows the prediction error to be reduced, as several hypotheses are formulated for each prediction, which are then voted to reach a final prediction.

The most widely used and well-known model in the ensemble field is the Random Forest technique, which is very versatile as it can be used in both forecasting and classification. Random forests derive from the bagging method and consist of constructing independent trees that are then voted on individually to output the prediction of the class to which the observation belongs.

2.2.1.2. Decision Tree approaches

Another very common methodology concerns decision trees, which are mainly used when one wants to segment a population. The power of these tools is their easy comprehension linked to the possibility of using them in both classification and regression. The concept behind decision trees is the subdivision of the space of predictors into well-defined regions through which the class to which a new observation belongs is then identified. In general, the output is not quantitative but qualitative. The main phases for the use of a classification tree are the growth phase or actual construction of the tree (growth) and then the pruning phase (pruning). The main methodologies used in the first phase are the Gini Index and cross entropy while for pruning the misclassification rate is usually used with the aim of minimizing it. This methodology is the basis of the Random Forest technique as a decision tree is the elementary component of the 'forest'.

2.2.1.3. Statistical Classifier

The most widely used model in this category is logistic regression, but Linear/Quadratic Discriminant Analysis (LDA/QDA), Naïve Bayes, Bayesian Networks and Least-Angle Regression are also known.

Logistic regression is part of the General Linear Models and aims to find a causal relationship between the predictors and the output variable, which is of binary type. In logistic regression, probabilities are extrapolated using I logit which is the probability of a positive event divided by the probability of a negative event.

Logit(pi) = 1/(1+ exp(-pi)) and ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k con Logit(pi) = dependent variable e X = independent variable.

The value of Beta, which is a parameter, is usually estimated by means of maximum likelihood estimation (MLE), which consists of testing various values of Beta in several iterations to best fit the logarithmic probability. Finally, the model returns probability values and if they are less than 0.5, the observation will be classified as '0', otherwise it will be classified as “1”. As far as Linear Discriminant Analysis is concerned, it is based on Bayes' theorem. The distribution of predictors through Bayes' theorem is transformed into estimates.

The posterior probability (this estimate) is obtained by combining a priori probabilities with information from the data. At this point, the various observations are classified on the basis of the posterior probability. The main assumption is that each class has a very specific covariance matrix. If one wants to relax this assumption, one speaks of quadratic discriminant analysis. In general, the former (LDA) is used when there are few observations to train the model as the variance problem is more consistent.

2.2.1.4. Nearest Neighbours methods

The most widely used algorithm in this category is the k-Nearest-Neighbour (k-NN), which is basically based on a single parameter, namely K, which is a positive integer that identifies how many 'neighbours' are to be taken into consideration. Given a test observation x, the model goes on to identify the K points closest to x and produces a classification based on the largest class among these K neighbours. It is a non-parametric model and therefore simpler, but at the same time provides no information on which predictors are important as there are no coefficients as in logistic regression.

2.2.1.5. Support Vector Machine (SVM)

These are supervised learning models applicable in both regression and classification domains. The concept behind SVM techniques is the creation of thresholds of non-linear separations between different classes. The functions used can be quadratic or even higher dimensional.

SVMs create a spatial representation in which the training data are divided into classes and separated by as large a space as possible. New observations are assigned by evaluating their position in space. In the case of non-linear classification, the kernel method is used (polynomial or radial kernels are the most used).

2.2.1.6. Neural Networks

Neural networks are a modelling technique similar to non-linear regression. Input variables are transformed and made to interact with a series of hidden layers (hidden layers) from which the final output (prediction) is then obtained.

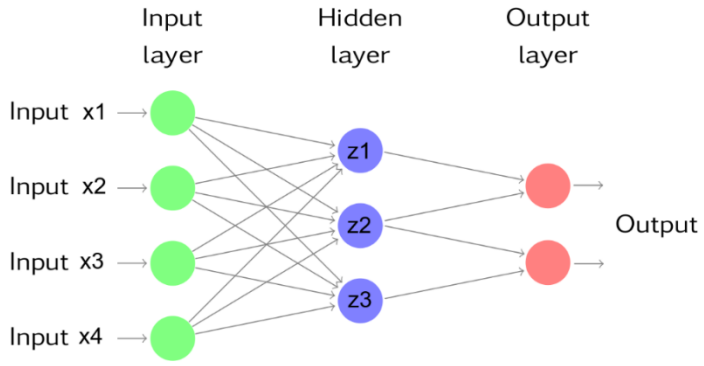


Figure 2.1 - Example of Neural Networks with three layers ("Statistica per Data Science con R", 2019)

As visible from the image, these nodes are part of a sequential structure. This structure is capable of weighting the inputs through the sigmoid in the hidden layer (non-linear function), thus reducing the effect of outliers and making the network more robust. This technique starts with some random 'weight' values that are then refined by exploiting the data; this involves an element of randomness that is mitigated through.

2.2.2. Markov Chains

Markov chains are random processes in which the transition probability at time t between states of the system is dependent only on the state of the system at time t-1. Thus, it is not relevant what happened in the time periods prior to t-1 (i.e., the complete history of the states assumed by the system), this is called Markov property.

The set of all possible states the system could be in constitutes the so-called "state space". From the state space, and the rules governing the relationship between states, it is built the transition matrix, which is a $n \times n$ square matrix (where n = numerosity of the state space) that describes exactly the relationship between the states of the system in terms of the probability of transition from one state to another (or from one state to the same).

The two typical applications of Markov chain are to answer these questions:

- What state the system will be in after *t* periods of time given a given initial state.
- Identification of the path followed by the system knowing the initial and final states and the number of state transitions (i.e., the "steps" taken by the system).

The transition matrix of Markov model is built on the historical data observed from the phenomenon under analysis. For sake of completeness, Markov chains also require the initial state probability distribution, which is the probability a generic observation has of starting in each of the states in the state space.

Markov chain example: simple weather model

For sake of comprehension, hereafter is reported a very basic Markov chain example. The problem here is to guess which will be the weather of tomorrow (next period) using MC.

To make this model work the underlying assumption done are the following:

- The weather of any given period exclusively depends on the weather of the previous period, not on any periods prior to that. Markov assumption:

  $$P[W_t|W_{t-1}, W_{t-2}, \dots] = P[W_t|W_{t-1}] \rightarrow W_t \text{ is the weather in day t}$$

- On each day the weather can be either cloudy, either sunny; no other type of weather condition are feasible. Cloudy and Sunny represents the only two states in which the system could be, the space that contains the two states is called $State\ space\ = \{C, S\}$.
- The transition from one weather condition to the other weather condition is governed by the transition probabilities inscribed in the following transition matrix (which is a square $n \times n$ matrix).

TPM	Sunny	Cloudy
Sunny	0.3	0.7
Cloudy	0.5	0.5

Table 2.2 - Example of transition probability matrix Markov Model

Note that the values on the diagonal of the matrix represent the probabilities that in the next period the system will stay in the same state, for instance the probability of transitioning from the sunny (today) to sunny (tomorrow) is of 0,3, in other words the probability that tomorrow will be sunny again is 30%. Another important characteristic of the transition matrix is that the sum of the probabilities along each line must be equal 1.

This is due to the fact that the system in the next period must be in one of the states of the steady state, no other outcomes are contemplated. Transition matrix can be computed from observed data, in this specific case from the historical weather data we can observe how many sunny days were followed by cloudy days over the total number of sunny days, and so on.

The same transition matrix and the model above described can be explained as the following simple graph, that show in a more immediate way the Markov Chain:

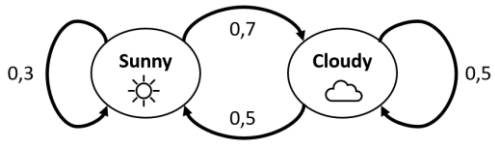


Figure 2.2 - Description of the transition probability with a simple graph

So now, assuming that in the day $W_0 = S$ (i.e., the system starts in sunny state), the probability that in period t_2 the weather conditions will be sunny is computed combining transition probabilities as follow:

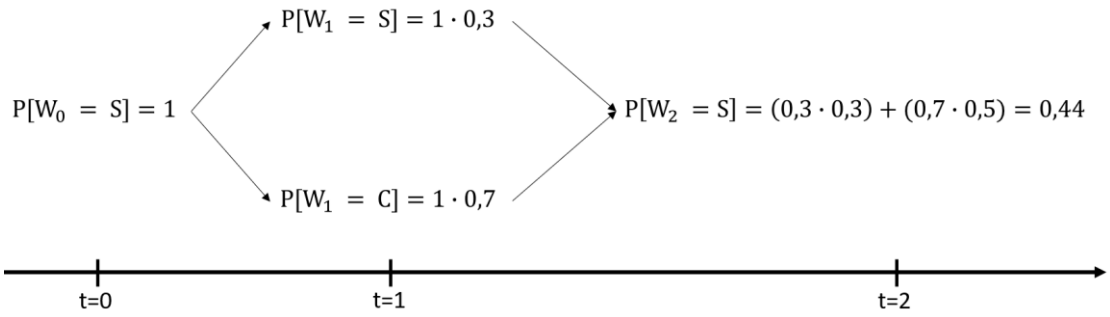


Figure 2.3 - Computation of P[W0 = S] = 1

It is clear that in other to compute the probability in every given the system will be in state Sunny (or Cloud) the only necessary things are, the starting state or the starting probabilities and the transition matrix.

If the starting probability is not known, it can be calculated solving the following system of equations (specific for this example), assuming to start in time zero in sunny state and find convergence in state sunny to infinite (steady state):

$$\begin{cases} \pi_{sunny} \cdot 0{,}3 + \pi_{cloudy} \cdot 0{,}5 = \pi_{sunny} \\ \pi_{sunny} + \pi_{cloudy} = 1 \end{cases}$$

The result is:

$$\begin{cases} \pi_{sunny} = \dfrac{5}{12} = 0{,}42 \\ \pi_{cloudy} = \dfrac{7}{12} = 0{,}58 \end{cases}$$

2.2.3. Markov Chain and RFM in Churn analysis

Markov chains are suitable for churn/partial defection analysis, in fact it is possible to assign a status to each of the segments into which the customer base is classified. Since an a priori classification of the users is necessary, from which the Markov Chain state space is then created, the classes obtained through RFM analysis are often used. From these, the transition matrix of the Markov chain can be calculated, by keeping track for each period of the movements that customers make in the following period, finally updating each individual transition probability. This approach represents an improvement over the simple RFM in that it provides the transition probabilities from one state to another and thus returns information on the "stability" of the segments as well as information on the behaviour of the customer base, being able to identify strange or unexpected behaviour in a more precise manner. From a managerial point of view this information is extremely valuable, since knowing that a customer in class 2 might move to class 1 (a riskier class) with a specific probability allows marketing manager to know more about the degree of risk of the class in question, and by extension the degree of risk of the customer.

On the other hand, this improvement is limited at class level, indeed within the same class will have customers with different degrees of risk, however these differences are not grasped by this symbiotic use of RFM and Markov chains. Another limitation of this solution is linked to the fact that in order to generate a final classification, an a priori classification is used, which introduces bias into the classes created and consequently into the transition matrix obtained, furthermore, a variation of the thresholds established in the calculation step of the RFM (or of the algorithm designated to make an initial classification) corresponds to a variation, even a significant one, in the transition probabilities. This can greatly compromise the stability of the classes and the attribution of a customer to a certain class with the risk of implementing marketing actions on an overly broad customer base or towards the wrong users.

2.2.4. HMM applications from literature

Hidden Markov Models are a variant of Markov Models in which the states, and the Markov chain, are hidden from the observer, who therefore cannot directly observe what state the system is in at a generic instant in time. What can instead be directly observed is the emission that the system produces being in a specific state at the instant in which it is observed. Through these emissions, an attempt is made to extract information about the possible states the system is in. For a more detailed explanation of HMMs, please refer to Chapter 3. Below are some application examples of HMMs with the context to which they are applied. These examples come from the literature review carried out and are extremely useful for better understanding the functioning of Hidden Markov Models.

General overview of literature in HMM application after 2016:

Authors	Title	Year & Journal	What	Techniques
James Lyons, Kuldip K. Paliwal, Abdollah Dehzangi, Rhys Heffernan, Tatsuhiko Tsunoda, Alok Sharma	Protein fold recognition using HMM–HMM alignment and dynamic programming	2016, Journal of Theoretical Biology, Volume 393	Prediction of the novel protein sequence into one of its folds.	HMM, dynamic programming

William N. Robinson, Andrea Aria	Sequential fraud detection for prepaid cards using hidden Markov model divergence	2018, Expert Systems with Applications. Volume 91	Real time fraud detection of credit card.	HMM
Jia Liu, Miyi Duan, Wenfa Li, Xinguang Tian	MMs based masquerade detection for network security on with parallel computing	2020, Computer Communications Volume 156	Detection of anomalous behaviour in network security. Focus on identifying Masquerade attacks	HMM
Chakkarai Sathyaseelan, L Ponoop Prasad Patro, Thenmalarchelvi Rathinavelan	Sequence patterns and HMM profiles to predict proteome wide zinc finger motifs	2022, Pattern Recognition	Prediction proteome wide zinc fingers motifs from a given protein	HMM
Dr. R.K. Srivastava, Digesh Pandey	Speech recognition using HMM and Soft Computing	2022, Materials Today: Proceedings, Volume 51	Manage dubiousness and vulnerabilities in speech signal	HMM, Soft computing

Table 2.3 - Literature in HMM applications after 2016

Summing up, the main fields in which HMMs find use concern speech synthesis, bioinformatics and the study of the human genome, DNA analysis, text and speech recognition, gene family modelling and cybersecurity. From this point of view, the field of application presented in this paper, churn/partial defection analysis, appears to fill an existing gap in the literature.

2.3. Private label and customer retention

The discussion regarding the effects of private label products is currently being debated in the literature, as evidenced by the recency of many papers on the subject.

Private label products are usually manufactured or supplied by third party companies then sold under the brand name of the retailer offering that product.

In other words, the supermarket chain offers a variant to the customer who can choose to buy an equivalent product but branded by the retailer itself. Private label products generally base their strength on an advantageous price-quality ratio for the end consumer, who on the one hand can benefit from a reduced price compared to the competition and on the other hand can benefit from the retailer's know-how in selecting suppliers.

A customer purchasing private label products may therefore be looking for cost leadership, either looking for the best product, relying on his trust in the selective capabilities of the retailer in question (this is generally the case when the customer purchases premium line private label products).

In addition, these products generally provide higher margins for the retailer as there is not that much of a marketing expense component since usually advertising and promotion are usually low; furthermore, from a product management perspective, the retailer has full visibility over all processes along its supply chain, thus providing better information for demand management (manufacturers of certain brands may hide their demand information) and consequently a leaner overall supply chain, which translates into lower costs and better performance.

Another great advantage of offering private label products is that the retailer can use them as marketing leverage, a well reputed private label can strengthen and consolidate the retailer's image.

Private labels usually can be bounded into two categories:

- **Private labels products under the retailer's name:** these are those private labels that include a wide range of products with a brand name that corresponds to the name of the supermarket chain.
- **Premium private label:** these are those private label products with a high positioning that compete directly with the gourmet products of well-known brands. Their name usually is not directly referable to the retailer's name, but of course you can find them only in the specific retailer's stores.

There is still no definitive answer to the question of the effects that the purchase of private labels by a customer has on his/her loyalty. In literature there are various theories according to which the purchase of private labels is only influenced by the desire to save money, but at the same time there are studies which find actual correlations between the purchase of private label products and the level of loyalty.

What will be attempted in this disclosure is to amplify the standard classification in the literature in order to carry out an analytic based assessment on the capacity of the various categories of private labels in customer retention.

2.4. Research needs

The overview conducted on the literature in the FMCG field related to churn revealed several unexplored areas:

The concept of partial defection (an indicator of churn in the noncontractual context) is addressed in a small number of articles and often with an academic/theoretical rather than an application focus. More broadly, the literature is rather rich in those contexts where the concept of churn is well understood and defined and where customer performance is more easily measured given it lower variability, usually in contractual setting price and payment terms are well defined and constant (banking, telecommunications but more broadly membership subscription). Thus, it is needed a complementary disclosure on model and methods to define partial defection from data, possibly giving a level of probabilistic risk associated.

In addition, it was perceived a lack of models with which to interpret the possible signs of a deterioration of the customer's relationship, so as to timely be able to implement marketing actions able to prevent such an issue. This represents a key issue since acquiring a new customer on average involves a greater outlay than retaining one, knowing how to prevent churn is therefore key to the company's long-term success and profit.

A final area on which we have identified our possible contribution concerns private labels and their ability to build customer loyalty. In particular, introducing private label centred covariates in our model we would like to find a more precise answer on this issue, which is still a point without an unambiguous answer.

It therefore seems that this thesis actively brings a valuable contribution to the research: our disclosure tries to bridge the afore mentioned gap by going on to provide practical results through a new model based on Hidden Markov Chains.

2.5. Research questions

The development of the model presented in this paper lays its foundation on some specific questions: finding answers to those questions will guarantee a contribution to

research in the topic of partial defection and customer base classification. The main points we focused on are:

I. Development of a fully data-driven customer purchasing behaviour classification model. The model is expected to output probabilistic segmentation, thus representing an evolution of the more traditional segmentation models (RFMs) that are only able to give deterministic segmentation with analyst-determined thresholds.

II. Identify through the model impactful predictors of partial defection risk, expressed in probabilistic terms. In other words, find out whether there are variables, created by reprocessing the information in the transaction dataset, that are explaining the reasoning behind the drop in performance of a generic customer from the customer base. If this will be the case, we expect that those predictors will improve the predictive power of the model itself, helping it in the classification process.

III. Understanding whether private label products have a link to customer loyalty. This would be very interesting as the debate is still open in the literature and an affirmative answer would allow such products to be used as levers to increase customer retention in marketing campaigns. In particular, a two-lines division of private label products will be analyzed to determine whether and which types of PL products can be exploited as marketing levers or otherwise as indicators of the degree of customer loyalty.

IV. To compare how our model performs compared to more traditional and widely used methods (RFM) and to assess whether there is a real improvement made by using hidden Markov models in the churn analysis domain. The creation of a more stable and precise classifier would be crucial for the managerial implications it would have. Indeed, it would be possible to direct marketing efforts more precisely, especially in monetary terms.

V. Development of managerial implications for the adoption and utilisation in practice of HMM

3 Model development

3.1. Theory and technique behind the model

3.1.1. Markov Model and Hidden Markov Model

We are going to develop a Hidden Markov Model (HMM), which is an evolution of basic Markov Chains. The main feature of an HMM is that the states are not directly observable, hence the Markov chain results hidden to the observer. Nevertheless, what we can observe is the emission that each state of the system will generate; each hidden state emits a continuous emission characterized by a certain probability distribution that is uniquely related to the state itself in which the system is located at any given time "t". Through HMM-type modelling we usually attempt to solve one or more of these three problems:

- Calculate the probability that a given sequence of observations will be observed in the output (forward algorithm).
- Identify which sequence of states best explains a given sequence of observations (Viterbi algorithm).
- Find the probabilities of transition between states given a given sequence of observations. In other words, this activity can be identified as hidden Markov model parameter training (Baum Welch algorithm - Expectation maximization algorithms).

Our effort will be focused on the third problem since given the observations at our disposal we would like to identify the states and find the transition probabilities between them (transition matrix) without classifying the population a priori. Therefore, the goal is to train an HMM model that in output must be able to return the subject's class of membership with the associated probabilities.

The main advantage over basic Markov Chains is to be able to do model training directly on the raw data without having to resort to any a priori classification. We will illustrate in the next chapter 3.1.2 how the training of the HMM is done using Expectation Maximization algorithm, such as the Baum Welch algorithm.

3.1.1.1. Example of Hidden Markov Model: professor's mood

The following section will propose an introductory example of the application of hidden Markov chains, in order to familiarize and better understand how this type of model work. In particular, the following example will propose the solution of a problem of the second type, among those listed above: identifying the sequence of states assumed by the system that best explains the observed emissions, the solution of this problem, in fact, represents the most typical application of hidden Markov models (once this is understood, the solution to the first type of problem will be of immediate understanding). The solution to the third type of question, on the other hand, will be discussed in the next section, with an explanation of expectation maximization algorithms and an explanation of the most popular of these, namely the Baum-Welch.

The problem now to be solved is as follows: every day the professor can be happy or sad (State space = {H,S}) and according to his mood he will be more or less likely to give good grades. The professor wears shirts of three colours: Red, Green and Blue; the choice of shirt colour is made according to his mood that day, according to the following relationship:

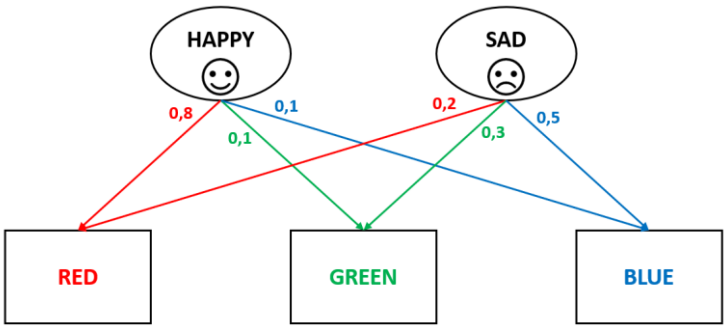


Figure 3.1 - Discrete distribution of emission and hidden states

In this example, for simplicity, the probability distribution of observing a certain color is discrete. By transforming graph from above into a matrix we obtain the emission probability matrix:

TPM	Red	Green	Blue
Happy	0.8	0.1	0.1
Sad	0.2	0.3	0.5

Table 3.1 - Emission probability matrix HMM

It is also well known that the professor's mood transitions from Happy to Sad, and vice versa, in the manner illustrated in the following graph:

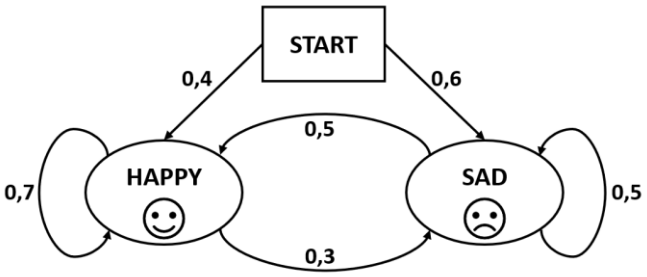


Figure 3.2 - Transition probability graph

From this graph, the following transition matrix and initial probability matrix are derived:

	Happy	Sad
Happy	0.7	0.3
Sad	0.5	0.5

Table 3.2 - Transition probability matrix

Happy	Sad
0.4	0.6

Table 3.3 - Initial probability matrix

During the last 3 days of class, pupils observe the following emissions $\{C_1 = Green, C_2 = Red, C_3 = Blue\}$ and wonder what was the most likely combination of moods the professor had. In symbols:

$$\max_{m_1,m_2,m_3} P[M_1 = m_1, M_2 = m_2, M_3 = m_3]$$

$$m_i = professor's\ mood\ in\ day\ i$$

In general, this probability will depend on the colors $\{C_1, C_2, C_3\}$ observed and by the mood the professor had in the previous days, and it will be so calculated:

$$\begin{bmatrix} P[C_3|C_2, C_1, M_1, M_2, M_3] \times \\ P[C_2|C_1, M_1, M_2, M_3] \times \\ P[C_1|M_1, M_2, M_3] \times \\ P[M_3|M_1, M_2] \times \\ P[M_2|M_1] \times \\ P[M_1] \end{bmatrix}$$

Taking advantage of the Markov chain property and the underlying assumption that the emission only depends on the system state (in this example it means that the colour of the professor's t-shirt only depends on professor's mood of that day), the above probability can be computed in an easier way:

$$\begin{bmatrix} P[C_3|M_3] \times \\ P[C_2|M_2] \times \\ P[C_1|M_1] \times \\ P[M_3|M_2] \times \\ P[M_2|M_1] \times \\ P[M_1|StartP] \end{bmatrix}$$

From emission probability matrix

From transition probability matrix and starting probabilities

The last step is to calculate the probability of each possible combination of states and emission, eventually selecting the greatest out of all. Another possibility, that avoids all manual calculation, is the Apriori algorithm, that require as input the transition matrix, the starting probabilities, the emission probabilities, the observed emission sequence, and produces as output the most likely states combination exploiting the exact same properties explained above.

3.1.2. HMM fitting (EM algorithm - Baum Welch)

BWA allows the estimation of the initial distribution of states, the transition matrix and finally the emission distribution of the hidden Markov model behind the observed phenomenon. This algorithm is part of a larger class of algorithms known as Expectation-Maximization (EM). It is an iterative process in which the forward-backward algorithm is applied to each iteration and at each iteration the algorithm maximizes the expected values of some computed parameters.

The goal of EM is to find a Maximum likelihood Estimate (MLE). This is an algorithm that is able to converge to a local optimum and is therefore affected in an important way by the starting conditions, nevertheless, by giving different "starting points" it is possible to explore different local maxima thus select the best among them.

The following are the main steps of EM both the conventional one and its application to HMMs. The demonstration is taken from the chapter of the paper by Jeffrey W. Miller (2016): "Lecture Notes on Advanced Stochastic Modelling." Duke University, Durham, NC.

3.1.2.1. EM-Baum Welch steps:

• Observed data: $x = (x_1, ..., x_n)$

• Model: $(X, Z) \sim p_\theta (x, z)$. Here, z represents some collection of unobserved variables (for example, in an HMM, $z = (z_1, \ldots, z_n)$ represents the hidden states). EM works best when $p_\theta (x, z)$ is an exponential family.

• Goal: Find $\theta_{MLE} \in \text{argmax}_\theta\, p_\theta(x)$, where $p_\theta (x) = \sum_z p_\theta (x, z)$. We will assume that Z is discrete.

• Algorithm:

1. Initialize θ_1.
2. For k = 1, 2, ... until some convergence criterion is met,

(a) E-step: Compute the function

$$Q(\theta, \theta_k) = E_{\theta k} (\log p_\theta (X, Z) \mid X = x) = \sum_z (\log p_\theta(x, z)) p_{\theta k} (z \mid x).$$

(b) M-step: Solve for $\theta_{k+1} \in \text{argmax}_\theta\, Q(\theta, \theta_k)$.

In an HMM, the parameter θ specifies π, T, and ε_i for each i. Let's suppose that the emission distribution $\varepsilon_i(x)$ belongs to some family of distributions $f_{\phi_i} (x)$ with parameter ϕ_i — for example, if the emission distributions are normal, then we could define $\phi_i = (\mu_i, \sigma^2_i)$ and $\varepsilon_i(x) = f_{\phi_i} (x) = N (x \mid \mu_i, \sigma^2_i)$. Recall that $\pi_i = P(Z_1 = i)$ and $T_{ij} = P(Z_{t+1} = j \mid Z_t = i)$.

With these conventions, the HMM is parameterized by $\theta = (\pi, T, \phi)$, where $\phi = (\phi_1, \ldots, \phi_m)$. Furthermore, we will assume that there are no relationships among π, T, and ϕ. Under these assumptions we have a variation of the E-step and M-step resented before.

(a) E-step:

• We need to compute $Q(\theta, \theta k)$. Let's take a closer look at this to see how we might do it. Recall that:

$$Q(\theta, \theta_k) = E_{\theta k} \log p_\theta(X, Z) \mid X = x.$$

• By the factorization implied by the directed graphical model for an HMM,

$$\log p_\theta(x, z) = \log p_\theta(z_1) + \sum_{t=2}^{n} log\ p_\theta(z_t \mid z_{t-1}) + \sum_{t=1}^{n} log\ p_\theta(x_t \mid z_t)$$

$$= \sum_{i=1}^{m} 1(z_1 = i)\ log\ \pi_i + \sum_{t=2}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m} 1(z_{t-1} = i, z_t = j)\ log\ T_{ij}$$

$$+ \sum_{t=1}^{n}\sum_{i=1}^{m} 1(z_t = i)\ log\ f_{\phi i}(x_t)$$

The only places where z appears in this expression are in the indicator functions, so when we take the expectation of Z given X = x, the expectation moves through and hits only these indicators. Further, the expectation of an indicator function is equal to the probability of the event in the indicator.

For example, $E_{\theta k}\ 1(Z_t = i) \mid X = x = P_{\theta k}(Z_t = i \mid X = x)$. Consequently,

$$Q(\theta, \theta_k) = \sum_{i=1}^{m} P_{\theta_k}(Z1 = i \mid x) \log \pi_i + \sum_{t=2}^{m}\sum_{i=1}^{m}\sum_{j=1}^{m} P_{\vartheta_k}(Z_{t-1} = i, Z_t = j \mid x)$$
$$+\log T_{ij} + \sum_{t=1}^{n}\sum_{i=1}^{m} P_{\theta_k}(Zt = i \mid x) \log f_{\phi i}(x_t).$$

• To simplify the notation, let's define

$$\gamma_{ti} = P_{\theta k}(Z_t = i \mid x)$$

$$\beta_{tij} = P_{\theta k}(Z_{t-1} = i, Z_t = j \mid x).$$

• With this notation, we have

$$Q(\theta, \theta_k) = \sum_{i=1}^{m} \gamma_{1i} \log \pi_i + \sum_{t=2}^{n}\sum_{j=1}^{m} \beta_{tij} \log T_{ij} + \sum_{t=1}^{n}\sum_{i=1}^{m} \gamma_{ti} \log f_{\phi i}(x_t).$$

• For any given θ_k, we can use the forward-backward algorithm to efficiently compute the γ's and β's in order to have an analytical expression for $Q(\theta, \theta_k)$.

(b) M-step:

For the M-step, we need to find a value of θ maximizing $Q(\theta, \theta_k)$.

• First, to maximize with respect to ϕ_i , if the family (f_ϕ) is sufficiently nice (and often it is), we will be able to simply take the gradient with respect to ϕ_i , set it equal to zero, and solve for ϕ_i . In other words, find the value of ϕ_i such that

$$0 = \nabla_{\phi i} Q(\theta, \theta_k) = \sum_{t=1}^{n} \gamma_{ti}\ \nabla_{\phi i} \log f_{\phi i}(x_t).$$

Note that the derivative kills off all the terms in our expression for $P\ Q(\theta, \theta_k)$ except for $\sum_{t=1}^{n} \gamma_{ti} \log f_{\phi i}(x_t)$. The value of ϕ_i satisfying this equation can be thought of as a weighted MLE, in which data point x_t has weight γ_{ti}.

Next, consider $\pi = (\pi_1, ..., \pi_m)$. Things are slightly trickier now since we need to maximize subject to the constraint that $\sum_{i=1}^{m} \pi_i = 1$. Fortunately, we can do this analytically using the method of Lagrange multipliers, as follows. Denoting the Lagrange multiplier by λ, we set the derivative of the Lagrangian equal to zero, apply the constraint, and solve for π:

$$0 = \frac{\partial}{\partial \pi_i} Q(\theta, \theta_k) - \lambda \sum_{j=1}^{m} \pi_i = \frac{\gamma_{1i}}{\pi_i} - \lambda$$

$$\Rightarrow \lambda \pi_i = \gamma_{1i} \Rightarrow \lambda = \lambda \sum_{i=1}^{m} \pi_i = \sum_{i=1}^{m} \gamma_{1i}$$

therefore,

$$\pi_i = \frac{\gamma_{1i}}{\sum_{j=1}^{m} \gamma_{1i}}$$

• Finally, for T, we need to maximize subject to the constraint that the rows sum to 1, in other words, $\sum_{j=1}^{m} T_{ij} = 1$ for each i. As with π, we can do this analytically using Lagrange multipliers. If you work this out, you will get

$$T_{ij} = \frac{\sum_{t=2}^{n} \beta_{tij}}{\sum_{t=2}^{n} \sum_{j=1}^{m} \beta_{tij}} = \frac{\sum_{t=2}^{n} \beta_{tij}}{\sum_{t=1}^{n-1} \gamma_{ti}}$$

Putting all these pieces together, then, the Baum–Welch algorithm proceeds as follows:

1. Randomly initialize π, T, and $\phi = (\phi_1, ..., \phi_m)$.
2. Iteratively repeat the following two steps, until convergence:

   (a) E-step: Compute the γ's and β's using the forward-backward algorithm, given the current values of π, T, ϕ.

   (b) M-step: Update π, T, and ϕ using the formulas above involving the γ's and β's.

3.1.3. HMM with covariates for time-dependent transition matrix

The algorithm we used to fit our Hidden Markov Model, allows to introduce covariates with the aim of better training the transition matrix. This procedure is also feasible using other algorithms, in this paper was used the one proposed by the "DepmixS4" library.

In particular, the library we used, by default calculate the transition probabilities and the initial state probabilities using a multinomial model with an identity link function. Parametrizing by mean of multinomial logistic models allows to include covariates on the initial state and transition probabilities. In this case, each row of the transition matrix is computed by a baseline category logistic multinomial, in other words the parameter for the base category is fixed at zero (as baseline category is chosen the first state). For instance, in our case, which is a 3-states Hidden Markov Chain, this means that first three parameters of the transition model have parameter fixed at zero and the other two are freely estimated (in our case the first three parameters belong to state 1, which corresponds to the highest class). Hence, the transition matrix is built using a multinomial logistic model that compute each transition probability. The multinomial logit has the following formula: $y = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n$, where β_0 is the intercept of the regression and $\{x_1, \ldots, x_1\}$ are the observed values for the covariates.

From the model above illustrated it is possible to compute the probabilities using the following formulas:

$$\Pr(Y_i=1) = \frac{e^{\beta_1 X_i}}{1+\sum_{k=1}^{K-1} e^{\beta_k X_i}}$$

$$\Pr(Y_i=2) = \frac{e^{\beta_2 X_i}}{1+\sum_{k=1}^{K-1} e^{\beta_k X_i}}$$

$$\ldots$$

$$\Pr(Y_i=k) = \frac{e^{\beta_k X_i}}{1+\sum_{j=1}^{K-1} e^{\beta_j X_i}}$$

Power elevation is due to the fact that the model is a multinomial logistic, which therefore uses the logarithm. Power elevation represents the inverse operation. The ratio, on the other hand, is necessary to normalise the values and transform them into probabilities, with a value between 0 and 1.

Fitting the formulae to our model (3 States and 2 covariates) produces:

$$\Pr[Y_i=2] = \frac{e^{\beta_{01}+\beta_{11}\cdot x1+\beta_{21}\cdot x2}}{1+e^{\beta_{01}+\beta_{11}\cdot x1+\beta_{21}\cdot x2}+e^{\beta_{02}+\beta_{12}\cdot x1+\beta_{22}\cdot x2}}$$

$$\Pr[Y_i=3] = \frac{e^{\beta_{02}+\beta_{12}\cdot x1+\beta_{22}\cdot x2}}{1+e^{\beta_{01}+\beta_{11}\cdot x1+\beta_{21}\cdot x2}+e^{\beta_{02}+\beta_{12}\cdot x1+\beta_{22}\cdot x2}}$$

$$\Pr[Y_i=1] = 1 - \Pr[Y_i=2] - \Pr[Y_i=3] = \frac{1}{1+e^{\beta_{01}+\beta_{11}\cdot x1+\beta_{21}\cdot x2}+e^{\beta_{02}+\beta_{12}\cdot x+\beta_{22}\cdot x2}}$$

The sum of the transition probabilities from one state to another (or the same) must be equal one, this is because in the next period the system must be in one of the states in the state space (no other outcomes are possible in Markov chains). The third equation from above is the result of this constrain, in conjunction with the fact that by construction only two parameters out of the three are freely estimated.

3.2. Explanation of the dataset

The available data were downloaded from the CRM software of one of Italy's leading FMCG chains. The total rows amount to 23,616,021, the overall file size is 1.3GB. Each transaction is recorded by customer id, date and individual product purchased; the variables available are shown below:

- Id_store: unique store identifier.
- Complete date: year, month, day, hour, and minute in which each transaction was recorded.
- Ticket number: non-unique ticked number identifier.
- Id_client: unique identifier of customer loyalty card.
- EAN: European Article Number.
- Qty: overall quantity of single product purchased.
- Totsold: overall expense monetary for each item in the transaction ($quantity \cdot price$)
- ECR3: code used by the company in its management systems, is derived from logistics.
- Description: Word description of the product, if available.

Before proceeding with an exploration of the dataset, it was identified which variables available in the dataset were useful and significant for the purposes of the analysis to be conducted.

To concretize the above explanation, four rows were retrieved from the dataset. The table below show the lines corresponding to two transactions belonging to two different customers:

Id_client	Complete date	Totsold	Description
2531794148904	202208191826	2.39	SAIWA ORO GR
2531794148904	202208191826	1.49	CAMPIELLO
265692375804	202205271255	2.99	ARANCIATA FANTA 15CL
265692375804	202205271255	0.75	LATTE UHT TREVAL.500

Table 3.4 - Example retrieved from raw transaction dataset used

3.2.1. Dataset exploration

The data exploration phase was divided into two steps: the first step regarded the analysis of the dataset columns containing product information (description, EAN, ecr code). Next, rows of the transaction dataset were explored in search of possible inconsistencies and outliers, to be handled in the following phase of data handling and cleaning.

Product-exploration: here the first information obtained concerns the fact that 46.31% of the transactions made (i.e., rows of the transactional dataset) report NULL value in the product description column. The product master is found not to be fully complete; this means that not all EANs in the dataset can be traced back to the relative product. This is due to the fact that store managers of each store in the chain are allowed to outsource from local suppliers for certain products, this may be a choice made to meet the demand for typical local products or also a choice driven by the possibility of signing better contracts and sustaining lower logistics costs from local suppliers. This mode of management occurs mainly for fresh products such as fruits and vegetables and some butcher's products. In addition, it happens rather often that product master records are compiled manually by different people, without the use of a common standard. This for instance results in the same brands being abbreviated differently in the product description at discretion of the person populating the product master record.

In terms of the products that can be traced back to a description in the dataset, it can be said that the overall product assortment is rather in line with the one of other FMCG companies. It was also found out the presence of Private label products, identifiable by the name of the distributor cited in the product description. In the same way it was also possible to identify the presence of gourmet private label products.

Transactions-exploration: first thing first, it was noted that some rows were associated to negative values in the tot-sold column, this would mean having products with negative price. In reality, these rows in the transaction dataset are due to the way stores handle offers and reversals; since they are not legally allowed to lower the price of a given product, the ploy used is to add a new ticket line with a negative value, corresponding to the discount or reversal made on the given product.

Continuing the analysis of transactions and customer outliers, a couple of card numbers with a history of excessive transactions were identified; these customer cards were reasonably reconducted as cashiers who pass their personal card for groceries of non-loyal customers, in order to acquire loyalty points. By analyzing the store visit rate, for all customers in the dataset, 88 anomalous badge numbers were identified, which had an unreasonable visit rate in the selected time frame (some badge numbers had a daily frequency including Saturdays and Sundays of -10 charges per day).

In absolute terms, the first and last recorded transactions were made on August 2020 and on September 2022. However, of these months the data are just partial since the first expenditure and the last one does not correspond with the beginning and end of the month. In subsequent modelling stages they will therefore be removed.

3.2.2. Data handling and cleaning (pre-aggregation)

This chapter will present the operations processed on the dataset in order to make it more understandable and manageable for subsequent analysis steps.

First of all, rows with negative monetary values were removed, as well as all rows associated with a fidelity card number identified as belonging to a cashier (i.e., card numbers with an average store visit rate for days of activity greater than 2). This means that for a customer to be classified as a cashier, he or she must make an average of more than two visits per day on all his/her "activity" days, calculated as the sum of the days he or she visits a store.

The fidelity card codes associated with cashiers were removed from the dataset since the customers that will be considered for the development of the model will be only the actual loyal customers.

For each row of the transaction dataset, three Boolean variables were added in order to flag private label products in the respective three lines: retailer and premium (the product lines are exhaustively explained in Chapter 2.3). Membership in a private label line is mutually exclusive, by construction therefore there is no private label product whose sum of the two Boolean variables is different from 1. On the other hand, the case where all of these two variables have the sum of the Boolean values equal to 0 is not ruled out; this in fact identifies products that do not belong to the private label category.

The recognition of private label products was done exploiting the product description column. In detail, a manual selection was made through keywords present in the description (it was not possible to fully automate this procedure because, as described before, not all products of the same brand have the same brand abbreviation hence there was not a single keyword to check. This feature engineering step will then be used to track at the product line level the number of private label items purchased versus the total number of items purchased in each ticket for each customer. It should be specified that the "number of private label items" is not meant in terms of quantity purchased, but rather as the number of ticket lines involving private label products, in the three different categories. For comprehension's sake is done the following example: a customer buys milk of the private label brand in quantity 5 packages (to which will correspond a row in the transactional dataset with quantity equal to 5 and expenditure equal to $5 \cdot milk\ price$) and then 3 packages of cookies of any brand. In this ticket the customer under analysis has purchased one private label product and one non-private label product.

The final step of the transactional database pre-processing was to reconstruct tickets, each ticket is built as the set of items purchased by the same customer in the same visit to the store. In this operation the date along with the fidelity card number were combined to create a unique code to aggregate the initial dataset by ticket. Finally, using the previously constructed features, it was possible to create the three variables related to the total number of private label products per receipt.

Also at this stage, the variable containing the total number of product types per receipt was added. Below it is reported an example of the new dataset obtained from the aggregation by ticket:

Id_client	Complete date	Totsold	Pvl_retailer	Pvl_Premium	Tot_prods
2531794148904	202208191826	18.64	1	0	12
265692375804	202205271255	178.20	2	1	35

Table 3.5 - Rows from dataset aggregated by ticket

3.2.2.1. Covariates choice

This type of covariates engineering deliberately does not give importance to quantity because the main objective is to capture the customer's choice between a private label product and a non-private label product and then to be able to assess any loyalty effects resulting from that choice. Moreover, this choice was made also because the stores' management software calculates the quantity of items purchased by considering each individual item, this is done even for products sold in batches of more than one item, such as water and beverages, each chest is counted as 6 products (quantity = 6). It is therefore clear that quantity would negatively affect by obscuring the value of choice.

3.2.3. Data aggregation choice

After cleaning the dataset and creating the variables needed for further analysis, the mode of temporal aggregation of the data was decided. In fact, the final model is not built on historical data composed by individual receipts; rather, in order to standardize the data across customers, the receipt dataset is further aggregated into periods. It should be noted that churn/partial defection dynamics in FMCG are very rapid, typically in the order of magnitude of a quarter. This means that it is not possible to aggregate data in timeframes that are too extended because of the risk of missing predictive signals on churn/partial defection.

Another constraint in the choice of time bucket arises during the implementation phase: using a time frame that is too small, on the other hand, generates very noisy data; in fact, reducing the period will increase the cases of customer absenteeism. Cases of absenteeism could generate a lot of noise that would risk undermining the analysis, as the customer would turn out to be an absentee (thus with zero Monetary and Frequency for that period) but this may not to be intended as a weakening of the relationship with the customer, rather it could be due to his or her specific buying behaviour. Let's consider for example a time bucket of one week, there may be some loyal customers who buy on a monthly basis, this would generate three out of four null periods that have nothing to do with a potential churn of the customer, who may indeed be extremely loyal. Thus, it is a matter of looking for the right trade-off between the need to have enough visibility to well describe customer behaviour and the need to give the right meaning to absenteeism cases thus avoiding misinterpretation of certain buying behaviour.

Before proceeding to the rationale for the choice of the time bucket, it is worth specifying that once the dataset is aggregated into periods, time series will be constructed for each client from the history of his or her periods of activity.

Below are the aggregation options analyzed and their critical issues:

- **Weekly aggregation:** this is the situation where we go to analyze the extreme of the trade-off at the point where we try to catch a sign of possible deterioration in the relationship between customer and distributor as soon as possible. As explained earlier, the problem with this solution is related to the introduction of a lot of weekly observations with zero values because of customer behaviour variability. Indeed, it is not unusual to observe loyal customers who do not visit the store for a whole week. This option was therefore discarded because, in addition to the increased computational difficulty that prevented the developed model from converging, we felt that even if we had succeeded in finding a solution, it would have been overly influenced by the variability of the purchase behaviour of the customer base in the ways just described.
- **Quarterly aggregation:** this is the opposite extreme of the trade-off respect the previous aggregation choice. In this situation, noise due to customer purchase behaviour is mitigated by the time frame width. The problem that arises with

this choice is that having a two-year dataset available, the maximum available observations would become 24/3 = 8. This, moreover, would only be valid for clients active since the first month of observation to the last. Given the objective of this aggregation, it would result in almost meaningless time series in which many customers would have fewer than 8 observation periods. With eight or fewer observations per customer (with some of them possibly being zero), we do not believe it is possible to adequately train the model, also because the results that would be obtained would not have much meaning relative to the phenomenon under analysis, in which a customer generally churns in even less than three months.

- **Monthly aggregation:** this is the aggregation we decided to use in this study. It is a middle ground between the first two options, and we have seen that it responds well both to the need not to overextend the time frame in order not to miss predictive signals of churn, and to the need not to overly obscure behaviors related to possible churn/partial defection. Moreover, the choice of this time bucket is also optimal from the point of view of the computational complexity provided to the model.

Once the time bucket was chosen, a unique code was generated for the identification of each period. This code was obtained by selecting from the date of each transaction the part of the string containing year and month (example: 20210923104520 unique period code "monthyear" = 202109).

3.2.4. Data handling and cleaning after the aggregation in periods

Once the dataset has been aggregated by month, the subsequent step involved the elimination of the data from the first and last month, 2020-08 and 2022-09, as these months are only partially complete (i.e., not all the days are covered). In addition, outliers defined as customers with more than 90 groceries per month were identified because they are most likely to be cashiers who escaped the cleaning carried out earlier or customers with extremely abnormal attitudes (more than three charges per day, for all days of the month).

Finally, were calculated the percentages of private label products (divided into the two lines retailer and premium) respect to the total products purchased in the month by the customer.

The total number of valid customers at this stage amounts to 19902.

For clarity, it is now reported an example of a client's data from the dataset with the changes and additions made:

Id_customer	Year-month	Totsold	Frequency	%pvl_retailer	%pvl_premium
99999838904	202009	137.4	7	0.42029	0
99999838904	202010	309.37	15	0.225352	0
99999838904	202011	152.12	12	0.296703	0
99999838904	202012	157.26	8	0.210526	0
99999838904	202101	108.61	6	0.271429	0
99999838904	202102	90.62	6	0.27451	0
99999838904	202103	20.22	1	0.25	0
99999838904	202104	132.48	8	0.362319	0
99999838904	202105	129.83	5	0.366667	0
99999838904	202106	129.48	6	0.3	0
99999838904	202107	167.35	8	0.268657	0
99999838904	202108	59.89	4	0.352941	0
99999838904	202109	144.84	7	0.2	0
99999838904	202110	135.97	6	0.253165	0
99999838904	202111	103.05	7	0.227273	0
99999838904	202112	126.59	8	0.30303	0
99999838904	202201	45.82	3	0.285714	0
99999838904	202202	39.25	2	0.47619	0
99999838904	202203	120.09	9	0.285714	0
99999838904	202204	22.93	2	0.052632	0
99999838904	202205	45.18	3	0.25	0

99999838904	202206	80.73	7	0.295455	0
99999838904	202207	122.69	6	0.244444	0
99999838904	202208	69.01	8	0.25	0

Table 3.6 – Ex.1 Customer from the dataset aggregated by month

The example above is retrieved from a customer that has never purchased premium private label products and has been active in all 24 observation periods: in fact, there is one record for each month in the period of analysis.

The dataset just described does not consider periods of absenteeism or more generally periods when a given customer is not active. There are essentially three reasons why a customer might be inactive, and thus have no purchases for a given month: absenteeism, i.e., the customer takes a break from purchases and then resumes in later periods. Churn, i.e., the customer has permanently stopped buying from the retailer under analysis and therefore as of the last month of purchase there will be no more of his transactions. New acquisition, in this case there are no purchases because the person in question is not yet a customer of the retailer, meaning that the first month of activity that will be observed will correspond with the beginning of the relationship with the given customer.

In this table we have an example of a customer with variable and occasional behavior: the first purchase is in 2021-09 and at the same time he makes three months of inactivity corresponding to 2022-11, 2022-05 and 2022-08.

Id_customer	Year-month	Totsold	Frequency	%pvl_economic	%pvl_retailer	%pvl_premium
999891618904	202109	57.04	2	0.038462	0.192308	0.038462
999891618904	202110	49.91	2	0.142857	0.047619	0
999891618904	202112	23.6	1	0.083333	0.25	0
999891618904	202201	35.37	1	0.066667	0.2	0
999891618904	202202	82.95	3	0.037037	0.185185	0
999891618904	202203	58.21	2	0.038462	0.192308	0
999891618904	202204	100.74	3	0.139535	0.162791	0.043424

999891618904	202206	94.06	3	0.105263	0.157895	0
999891618904	202207	49.98	2	0	0.263158	0

Table 3.7 – Ex.2 Customer from the dataset aggregated by month

3.2.5. Time series creation

The data to be given as input to our model should be provided in the form of time series with monthly basis for each customer. However, in order to perform proper training, periods of customer inactivity, absenteeism, and periods without purchases (except for periods when the customer was not actually still a customer) cannot be discarded, since they represent a significant purchasing behavior. For this reasoning, it would be formally wrong not to provide them as training data to the model. Because of this, time series for each customer begins with his or her first available purchase and ends at the last month available in the dataset, 2022-09. In this way, both periods of absenteeism and periods when the customer is already in a state of churn are considered; the only periods of inactivity not considered are, if any, those prior to the first purchase.

Created the time series for each card number, a group of customers was identified whose performance is not usable for the purposes of our analysis; those customers were eventually removed. Indeed, this group is composed of either customers who are too young, meaning that they recently enrolled, or customers who are too infrequently present to be considered as customer at all; cases of this type include customers with too many periods of absenteeism respect the number of periods among the first and last purchase.

In order to ensure that the model learns even from low-performing customers, only customers with fewer than 9 available observation periods between first and last purchase were removed. At the same time, it was imposed another constraint on activity periods: in addition to having at least 9 available periods between first and last purchase, it is necessary that in at least half of them the customer was active (e.g., a customer with 12 periods between the first purchase and the last purchase made may have made a maximum of 5 months of absenteeism).

This concludes the data handling phase carried out before training the model. The total number of time series obtained and usable to train the model amounts to 15840, in other words this means that the model was trained on the data of 15840 customers.

The following is an example of a time series for a client who eventually churned. The same client also had a break in the months preceding the churn.

Id_customer	Date-Year	Totsold	Frequency	%pvl_retailer	%pvl_premium
1013794928904	202009	279.03	6	0.23913	0
1013794928904	202010	242.07	7	0.11215	0
1013794928904	202011	74.57	1	0.194444	0
1013794928904	202012	121.54	3	0.2	0
1013794928904	202101	90.29	2	0.275	0
1013794928904	202102	0	0	0	0
1013794928904	202103	7.85	1	0.2	0
1013794928904	202104	36.57	1	0.0625	0
1013794928904	202105	64.18	2	0	0
1013794928904	202106	13.97	1	0	0
1013794928904	202107	0	0	0	0
1013794928904	202108	0	0	0	0
1013794928904	202109	0	0	0	0
1013794928904	202110	0	0	0	0
1013794928904	202111	0	0	0	0
1013794928904	202112	0	0	0	0
1013794928904	202201	0	0	0	0
1013794928904	202202	0	0	0	0
1013794928904	202203	0	0	0	0
1013794928904	202204	0	0	0	0
1013794928904	202205	0	0	0	0

1013794928904	202206	0	0	0	0
1013794928904	202207	0	0	0	0

Table 3.8 - Customer time serie

3.3. Software choice and libraries

The hidden Markov model was developed using R, a specific object-oriented programming language for statistics. It is free software with "command-line" type interaction; however, there are several graphical tools such as RStudio to make R more user friendly. The version used in this study is R version 4.2.1 - "Funny-Looking Kid." As with many programming languages there are libraries intended as repositories of function sets and data structures implemented for specific purposes that are very useful (in addition to the most basic ones).

The first library we used is "data.table", it was used to structure the data in order to optimize memory usage and speed up computational operations. In addition, a key feature of data table is the embedded advanced query system, which resulted in being extremely useful in querying such a large database.

The second library used is DepmixS4, a recent library (released in 05-12-2021) dealing with hidden Markov model fitting, developed by Ingmar Visser, Maarten Speekenbrink. This package encapsulates an implementation of a general framework for defining and estimating dependent mixed models using R. The main models included are: Markov models, latent/hidden Markov models and latent class/finite mixture distribution models.

3.4. Model implementation

The algorithm introduced in Chapter 3.1.2 was used to train our hidden Markov model. The chosen model will have three states, identified through the Monetary and Frequency emissions of the customer in the given state, while the transition between these states will be affected by the presence of three covariates. The choice of the emissions used to build the states of the model was made taking into consideration what we found in the literature: Monetary and Frequency, in fact, are the two most widely used and most significant variables when to describe the purchasing behaviour of customers in the FMCG sector; it is no coincidence that the most widely used analysis is precisely the Recency-Frequency-Monetary analysis.

The three states identified have the following meaning:

- Low state: corresponding to Class 1, encloses customers with lower Monetary and Frequency values. This state consists of churn/partial defection customers. It therefore represents the final state where the customer is considered lost or at high risk of dropping out.
- Medium state: corresponding to Class 2, these are those customers who have good but not excellent performance. These customers are found to be at no particular risk of churn/partial defection.
- High state: corresponding to Class 3, these are the customers who have high performance and are considered as the most loyal customers. A customer in this class is likely to be considered as "safe" in terms of the possibility of churn/partial defection.

Our model requires as input the time series of customers (where monetary and frequency are present), the number of states the model must have, the distributions of the considered emissions that the model will use to build the classes, and finally the covariates that will affect the probability of transition from one state to another.

As for the distributions, the Gaussian distribution was chosen for Monetary because it allows for well-separated states and guarantees neutrality in creating the states. The Frequency was fitted to a Poisson distribution; in fact, the Frequency is a discrete

variable and thus needs to be fitted with a discrete distribution. In addition, the Poisson distribution is well suited to represent the number of events occurring in a given time frame.

The covariates chosen to be implemented in our model, as described in Section 3.2, are the following:

- Percentage of "retailer private label products" (*perc_pvl_retailer*): this value is computed by period as the tot number of retailer private label items for each grocery, over the total number of items bought in the relative grocery.
- Percentage of "Premium private label products" (*perc_pvl_premium*): this value is computed by period as the tot number of premium private label items for each grocery, over the total number of items bought in the relative grocery.

These 2 covariates are independent of each other by construction, in that a product can belong to one and only one category or at most belong to none (a case in which the product is not a private label product).

4 Results

This chapter will be reserved for the mere presentation of the model output. Considerations, and a further disclosure upon these results will be done in the dedicated Chapter 5.

4.1. Fitted model

Below is the result of the fitting of the hidden Markov model:

4.1.1.1. Initial state probabilities of the model:

	State 1 Class 3	**State 3** Class 2	**State 2** Class 1
Starting probability	19.5%	60.7%	19.8%

Table 4.1 - Initial state probabilities of Master model

4.1.1.2. States of the hidden Markov chain:

	Monetary: *mean*	**Monetary:** *Standard deviation*	**Frequency:** $\log \lambda$	**Frequency:** λ
State 1 Class 3	394.444	222.865	2.614	13.654
State 3 Class 2	153.493	89.893	1.548	4.702
State 2 Class 1	19.663	23.805	-0.177	0.838
	Monetary [€]		Frequency [groceries/month]	

Table 4.2 - States of the HMC from Master model

Response parameters: Monetary = $\mathcal{N}(\mu, \sigma^2)$, Frequency = $Pois(\lambda)$

4.1.2. Covariate: multinomial logit model

The following paragraph contains the logit model trained on the selected covariates pvl_retailer and pvl_premium. The outcome of the multinomial logit model varies depending on the values assumed by the covariates; therefore, for sake of simplicity the results hereafter are reported imposing covariates value to zero (each customer, for each period will have a specific and unique multinomial logit model, as result of the different values of his/her covariates):

$$\boldsymbol{model} = Intercept + \beta\_retailer \cdot x_{retailer} + \beta\_premium \cdot x_{premium}$$

$$\boldsymbol{e^{model}} = e^{Intercept + \beta\_retailer \cdot x_{retailer} + \beta\_premium \cdot x_{premium}}$$

4.1.2.1. Model from State 1 (High):

From St1	Intercept	β_ret	β_prem	x_{ret}	x_{prem}	e^{model}	$P[y_i = St_i]$
$P[y_i = St_2]$	-4.981	-2.791	-36.863	0	0	0.0069	0.62%
$P[y_i = St_3]$	-2.345	-0.860	-6.135	0	0	0.0958	8.69%
$P[y_i = St_1]$	0.000	0.000	0.000	0	0	1.0000	90.69%

Table 4.3 - Multinomial logit model from state 1

4.1.2.2. Model from State 2 (Low):

From St2	Intercept	β_ret	β_prem	x_{ret}	x_{prem}	e^{model}	$P[y_i = St_i]$
$P[y_i = St_2]$	5.965	-0.207	43.155	0	0	389.586	82.53%
$P[y_i = St_3]$	4.400	1.449	45.711	0	0	81.439	17.25%
$P[y_i = St_1]$	0.000	0.000	0.000	0	0	1.000	0.21%

Table 4.4 - Multinomial logit model from state 2

4.1.2.3. Model from State 3 (Medium):

From St3	Intercept	β_ret	β_prem	x_{ret}	x_{prem}	e^{model}	$P[y_i = St_i]$
$P[y_i = St_2]$	1.468	-0.058	-1.214	0	0	4.341	13.48%
$P[y_i = St_3]$	3.291	0.452	1.795	0	0	26.867	83.42%
$P[y_i = St_1]$	0.000	0.000	0.000	0	0	1.000	3.10%

Table 4.5 - Multinomial logit model from state 3

4.1.3. Transition matrix (covariates set to 0)

In the next paragraph there is exposed the transformation of the table above into a transition matrix form (with covariates set to 0): the first line of the transition matrix corresponds to the second model (From St2), the second line corresponds to the third model (From St3) and the third line corresponds to the second model (From St3).

	State 2 Class 1	**State 3** Class 2	**State 1** Class 3
State 2 Class 1	82.53%	17.25%	0.21%
State 3 Class 2	13.48%	83.42%	3.10%
State 1 Class 3	0.62%	8.69%	90.69%

Table 4.6 - Transition matrix from Master model (cov set to 0)

4.2. Model output

For each period and for each customer, the model produces the posterior probabilities that the customer under analysis has of being in each of the 3 states. These posterior probabilities are the result of the observed Monetary and Frequency in the period "t", as well as the values the covariates assumed in that period.

The following represents an example of model output:

Customer ID	Period	State 1	State 2	State 3
1018998848904	202011	1.24E-04	3.77E-03	9.96E-01
1018998848904	202012	4.44E-02	1.43E-14	9.56E-01
1018998848904	202101	3.85E-03	4.61E-08	9.96E-01
1018998848904	202102	9.37E-01	2.06E-36	6.32E-02
1018998848904	202103	5.24E-01	2.16E-14	4.76E-01
1018998848904	202104	3.77E-01	3.90E-17	6.23E-01
1018998848904	202105	2.62E-01	6.22E-17	7.38E-01
1018998848904	202106	4.76E-01	1.14E-22	5.24E-01
1018998848904	202107	3.60E-01	1.93E-18	6.40E-01
1018998848904	202108	3.32E-01	1.70E-22	6.68E-01
1018998848904	202109	9.94E-01	5.60E-23	5.77E-03
1018998848904	202110	1.52E-01	6.35E-07	8.48E-01
1018998848904	202111	3.10E-02	1.57E-10	9.69E-01
1018998848904	202112	8.59E-01	1.57E-41	1.41E-01
1018998848904	202201	8.96E-04	1.46E-01	8.53E-01
1018998848904	202202	2.38E-04	1.64E-02	9.83E-01
1018998848904	202203	4.76E-03	9.00E-08	9.95E-01
1018998848904	202204	9.26E-05	5.48E-02	9.45E-01
1018998848904	202205	9.17E-05	1.13E-01	8.86E-01
1018998848904	202206	1.10E-08	9.94E-01	6.43E-03
1018998848904	202207	2.38E-10	1.00E+00	3.91E-04
1018998848904	202208	2.19E-10	1.00E+00	3.80E-04

Table 4.7 - Example of Master model output

Again, State 1 correspond to class 3 (High), State 2 to class 2 (Low) and State 3 to class 2 (Medium).

4.3. Model validation

In order to validate our model, additional transactional data were retrieved from the retailer's management system; in particular, these represent new customers never used before. What we want to evaluate is the behaviour of our model as the data provided change, in order to ensure that the model is indeed robust and that the results obtained are not fortuitous. In particular, a new hidden Markov model will be trained with the new data, which will then be discussed in absolute terms and compared with our main model.

The new dataset includes 6549 new customers. Like the main dataset this new dataset collects all the receipt lines purchased by new customers.

The main points to be checked are the following:

- The model, with a fair variety of client, is able to recognize the three risk classes: high risk, representing the class with worst performance; medium risk, representing the class with good performance; and low risk representing the class with absolutely best performance and consequently least risk of churn/partial defection. These classes are not likely to be numerically identical to those in the Master model, indeed the values of the classes obtained are strictly dependent on the input values on which the hidden Markov model is fitted.
- Consistency between the Monetary and Frequency values of the classes. The model is expected to be able to recognize that the class with lowest Monetary corresponds to the class with lowest Frequency, and so on.
- The lowest class must also be able to accommodate churner customers: the mean value of Monetary and the standard deviation of Monetary must be similar in absolute value (the values of Monetary are fitted according to an $N(\mu, \sigma^2)$) in this way we are able to correctly classify even the cases of Monetary close to zero.

- The impact of covariates perc_pvl_retailer and perc_pvl_premium must have the same behaviour, in particular the new data should provide the same insights regarding the reduction of risk level as the percentage of private label bought increases. Moreover, perc_pvl_premium must have I higher impact than perc_pvl_retailer especially on the transition to Class 1, the one associated to highest risk of churn/partial defection.

4.3.1.1. Classes found in the Test model

HMM test	Monetary mean	Monetary variance
Class 3:	246.078	144.208
Class 2:	130.675	80.738
Class 1:	26.342	23.584

Table 4.8 - Test model classes: Monetary

HMM test	Frequency λ (mean)
Class 3:	16.5
Class 2:	7.0
Class 1:	2.3

Table 4.9 - Test model classes: Frequency

As exhibited in the above table, it can be seen that the new model demonstrates good behavior, the classes are consistent with what was discussed in the previous section and well distinct from each other.

What has just been described shows that the developed model has a very good ability to adapt even to different data and demonstrates soundness and consistency in its operation. In form the outputs of this model are the same as those of the Master model; therefore, operationally its use is the same as that of the Master model.

In addition, the new model was constructed using the same covariates used in the Master model. This was done to validate the effectiveness of private labels in increasing customer retention. The trends of transition probabilities as the percentage of private label products increases are shown to be consistent and decreasing for both pvl_retailer and pvl_premium. This means that even for new customers, an increase in the purchase of private label products leads to a reduction in the risk of transition to the lower class.

Below are the two graphs comparing the Master model and the model trained with the new data.

Master model:

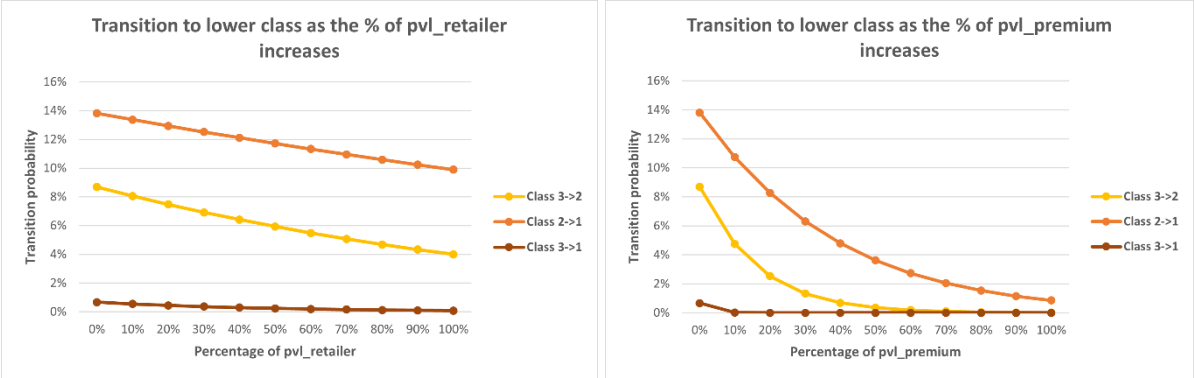


Figure 4.1 - Master model: covariates impact

Test model:

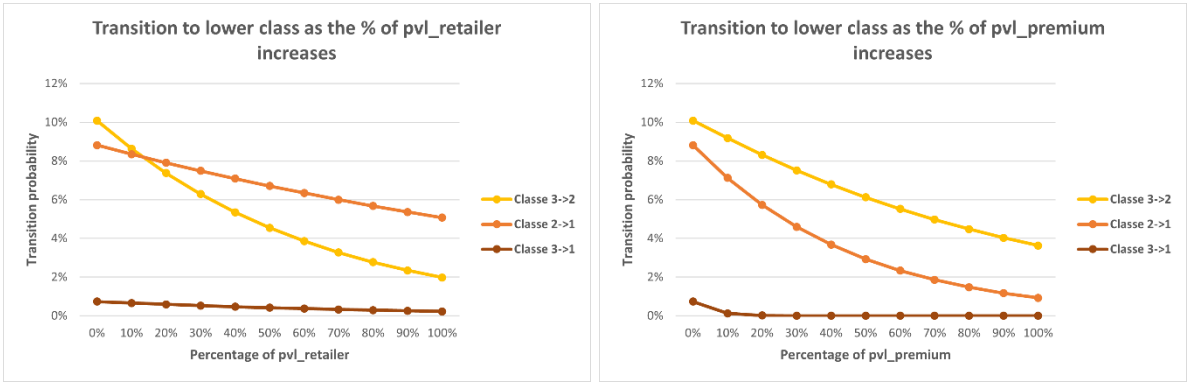


Figure 4.2 - Test model: covariates impact

Finally, the values obtained are not exactly equal between the Master model and the test model: having equal values would mean that the new client data are actually very similar to those used in the Master model, which is deliberately not sought to assess the validity of the model. Instead, what it really matters is the decreasing behaviour demonstrated.

5 Findings and results discussion

5.1. Model tuning

The choice of the number of states, response variables, and covariates was the result of an iterated analysis in which different tests were conducted with combinations of the hyperparameters just listed

The first step involved choosing the response variables, i.e., the emissions from the HMC, and at the same time choosing the most appropriate number of states. As for the response variables, initially only the Monetary was used in order to see if it was sufficient to obtain a good classification. In addition, Monetary was the first response variable tested because of its paramount importance from a managerial point of view being the variable on which we have the final feedback regarding the margin generated by a given customer. The number of states was made to vary from 2 to 6, however, it soon became apparent that having too many states meant, on the one hand, an increasing difficulty in reading the results from a managerial point of view: a classification into many risk classes would be obtained, each of which would then have to be found to have significance from a marketing point of view (ie, the class must be able to be used operationally for marketing campaigns). On the other hand, the model as the number of classes increases decreases the separation between them, increasing the risk of less adequately classifying customers. However, from a model with too few states one would have had the same problem in creating consistent classes that are then viable for marketing operations. Considering this trade-off, the choice fell to a 3-state model, so it was possible to identify three conceptually and managerially significant classes of risk: low risk (Class 3), medium risk (Class 2), and high risk (Class 1). Subsequent tests were conducted to evaluate the performance of different response variables. It was found that monetary alone is not able to satisfactorily learn the variability of customer behaviour (e.g., making an $80 shopping trip in a month is different at the level of customer loyalty than making 4 $20 groceries). For this reason,

tuning was repeated by adding frequency as a response variable, so that we could have an indicator of the number of visits that occurred in the selected time frame. The computational difficulty increases a lot but still the algorithm reached convergence and the 3-state model obtained is more than satisfactory. For completeness, a version with Frequency as the only answer was also tried but the problem turns out to be exactly opposite to that found with the model based only on Monetary (it is not very informative to know that a customer has frequency equal to 4 when then maybe he spent 5€ in the 4 visits to the store). In conjunction with the choice of the response variables, the corresponding distributions were chosen: the Monetary issue was modelled using a continuous Normal distribution, while the Frequency issue was modelled using a discrete Poisson distribution, since this response variable can only take discrete values (it is not possible to make 1.5 expenditures).

Having consolidated the 3-state model with double response variable, we moved on to the analysis on the use of covariates by testing several of them both individually and in combination.

First, covariates related to products were evaluated both from the perspective of their product category (fresh, frozen, fruit and vegetables...) and from the perspective of brand name. In the latter case, the aim is to assess the effect of products that feature retailer's brand directly or indirectly.

Tests with the first type of covariates described did not yield interesting results because, as described in Chapter 3, 46.31% of the transactions in the dataset do not have a valid product description. This means that some products belonging to the chosen product categories are not included in the master data, and thus are not identifiable for the purpose of training the model, which in this case is trained on partial information. However, the following covariates were tested: spending on fresh produce, spending on meats, spending on frozen food, and finally number of products with invalid description versus number of products with valid description. These tests, however, did not produce any interesting results.

Having found out that in the literature the debate regarding the effect of private labels in customer retention is still open; the effect of two categories of private labels was thought to be tested.

These covariates describe customers' buying habits about private label products divided into two lines with positioning in increasing market ranges, as explained in previous chapters. These two variables were tested individually, first pvl_retailer and finally pvl_premium. It was found out that the selected covariates effectively restrain transition from a higher state to a lower one. A more in-depth discussion of this issue can be found in Chapter 5.2.

The utmost step concerning the tuning phase was to simultaneously integrate the two covariates into a single model to get as complete a picture as possible. This was possible because there is no correlation between covariates, and it is therefore possible to get a good and significant multinomial logit.

5.2. Impact of covariate on transition matrix

In order to improve the developed model, the use of covariates was introduced in the manner described in Chapter 3.1.3. These covariates do not influence the creation of the HMM states, but rather make the starting probabilities and transition probabilities dynamic over time with respect to the values assumed by the covariates itself. This means that at each instant of time each client will have an individually fitted model based on the observed covariate and the emission emitted (in terms of Monetary and Frequency), with specific transition matrix and initial states probabilities. The model thus constructed produces in output the posterior probabilities (in the form that was exposed in chapter 4.2 Model output), which indicate for each of the HMM states the probability that in each period the client under analysis will be in each state. Thus, these posterior probabilities are the result of the output and the observed covariate. From a macroscopic perspective, the overall impact of covariates is measured with the variation of the transition probabilities that is generated by the variation of each covariate.

To obtain the final model, several features were created from the information contained in the dataset and then tested as covariates. The underling reasoning behind the feature engineering done was to do educated guesses about which purchase behaviour can tell something about the movements the customer will be more likely to do, in other words the idea was to find relevant covariates that could help the model to better classify customers during time.

In the specific context of churn/partial defection analysis if a covariate proves to be relevant, it could take on the significance of a "churn predictor." In particular, the above statement is valid if the increase of such a covariate correspond to a reduction in the likelihood of falling into a lower class (decrease the transition probability to lower class).

This implication, at the managerial level, represents extremely valuable information for the implementation of anti-churn/partial defection measures. In fact, knowing that a particular behaviour is an expression of a decrease in customer loyalty makes it possible to intervene in a timely and specific manner on the customer in question, proposing marketing actions aimed at avoiding the customer to eventually churn. Furthermore, knowing what characterizes a high-loyalty customer can be a valuable indication to implement at operational marketing level retention action, by stimulating risky customers toward behaviours typical of top clients. This means that if the model shows that a certain behaviour is synonymous of a good loyalty level, high-risk customers can be stimulated to emulate that behaviour in an effort to increase his or her loyalty.

Eventually our Master model uses two covariates. Specifically, these covariates are:

- **Percentage of "retailer private label products" (perc_pvl_retailer):** this value is computed by period as the tot number of Retailer private label products for each grocery, over the total number of products bought in the relative grocery.
- **Percentage of "Premium private label products" (perc_pvl_premium):** this value is computed by period as the tot number of Premium private label products for each grocery, over the total number of products bought in the relative grocery.

For "total number of private label items", it is intended the number of EANs (European Article Number) referred to private label products, which is then divided by the total number to EANs; therefore, no quantity is contemplated.

The covariates above listed are the result of a process of progressive refinements of the final model; once it was verified that the "private lable" element was a valid predictor, a subsequent analysis was conducted to understand what was contained in the "private label" set; this analysis led to a subdivision of the "private label" set into two subgroups, from which the final covariates were then derived. (For further explanation on model tuning refer to Chapter 5.1).

To evaluate the effects of these covariates, individual models were fitted, so that the effect of each covariate can be better understood independently of the others. In the paragraphs to come, the two models will be explained one by one. Consistent with what has been explained in the previous paragraphs, the emphasis will be on the transition from high class to lower class, specifically the transition probabilities that we are going to analyse as the covariate changes are: high Class 3 to middle Class 2, middle Class 2 to low Class 1, and high Class 3 to low Class 1.

5.2.1.1. Model with perc_pvl_retailer:

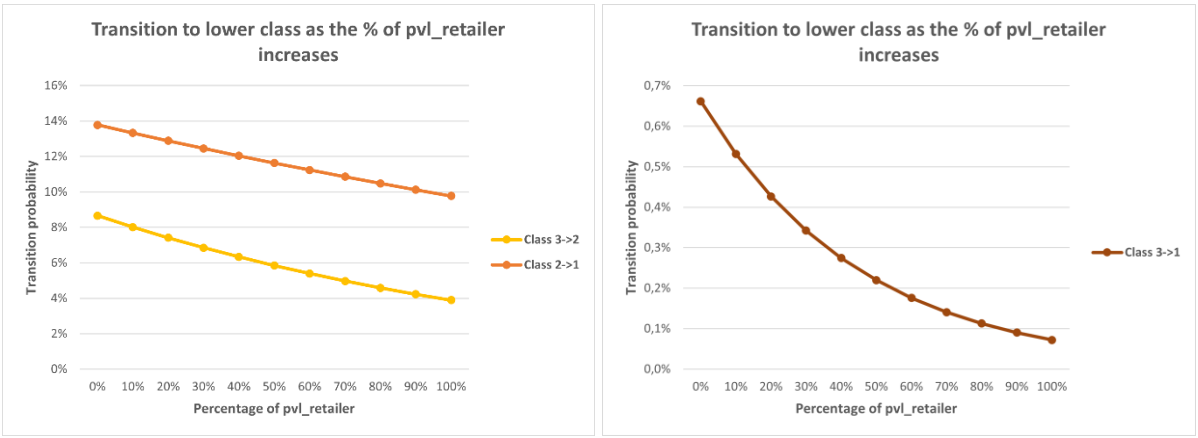


Figure 5.1 - Impact of perc_pvl_retailer

The covariate perc_pvl_retailer depends on the number of pvl_retailer items purchased; these items are for all intents and purposes sold by the retailer branding them with its logo. The pvl_retailers are a substitute offering to well-known brand products that the retailer offers to its customers. These items rely heavily on value for money: lower price than competitors due to lower promotional and product management expenses, quality guaranteed by the retailer's ability to select suppliers. In addition, as is well known, such products guarantee the retailer higher margins than classic products.

A customer inclined to purchase pvl_retailer (thus with a high perc_pvl_retailer) shows trust toward the retailer since these products can be purchased exclusively in the stores of the said chain. Moreover, the explicit and conscious choice of these types of products is itself a signal of brand loyalty. Analysing the output of the model constructed with covariate perc_pvl_retailer, it is observed that as the percentage of pvl_retailer increases there is a reduction in the probability of transition to the lowest state. This demonstrates what was stated earlier, hence a customer who has a high perc_pvl_retailer value in the period considered will be classified by the model as less risky than a customer with a low perc_pvl_retailer value, net of Monetary and Frequency performance.

The following table shows the maximum, minimum, mean and variance of the percentage of pvl_retailer products purchased out of the expenditures of the whole customer base.

	Min	Max	Mean	StDev
perc_pvl_reatailer	0%	100%	11.13%	33.35%

Table 5.1 - Exploration of perc_pvl_retailer

5.2.1.2. Model with perc_pvl_premium:

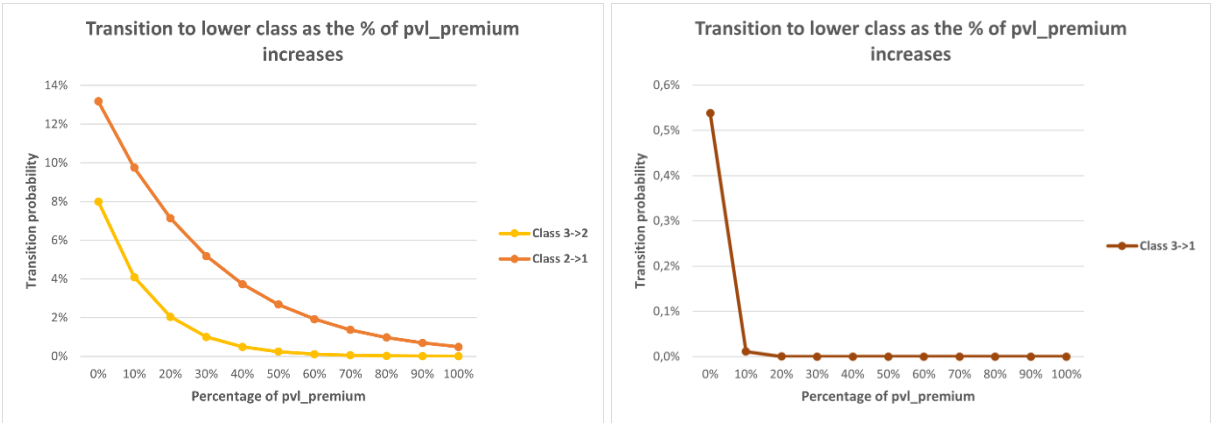


Figure 5.2 - Impact of perc_pvl_premium

The last covariate is perc_pvl_premium, it is related the high-end line offered by the retailer, it is characterized by high prices and uncompromising quality. A customer who is used to buy this type of private label is a customer who expresses high trust in the retailer and is willing to pay a price premium to have the best product that the retailer offers.

The choice of pvl_premium is a conscious one, the purchase is not made with the purpose of saving money (in addition to the higher-than-average price, these products are rarely discounted). Indeed, the customer pays more to the retailer because he recognizes the quality of the product and knows that it meets his needs.

Among the three covariates, this is the one that demonstrates the greatest discriminating power; in fact, in the graphs above it is shown how even a small increase in the percentage of pvl_premium leads to a substantial reduction in the probability of transition to the lower state.

Dwelling on the first graph (transition from Class 3 to Class 2 and transition from Class 2 to Class 1), the curves show a marginal decreasing increase (past the 60% threshold of pvl_premium purchased, the decrease in transition probabilities is marginal), this behaviour from a marketing point of view suggests that soliciting a customer to purchase pvl_premium products can have extremely significant results in terms of retention, however it is good to focus in the 0% - 60% area, stimulating the customer to purchase more than 60% pvl_premium it might result in little reduction of churn risk respect to the marketing effort required.

Moreover, analysing the graph with the transition from Class3 to Class 1, further conclusions can be drawn, in particular it can be seen that between 0% and 10% the probability of transition from high to low undergoes an almost total reduction.

The graph belove shows the zoom on this behaviour in the specific area of interest between 0% and 10%. Already in the neighbourhood of 5% the transition probability is almost zeroed.

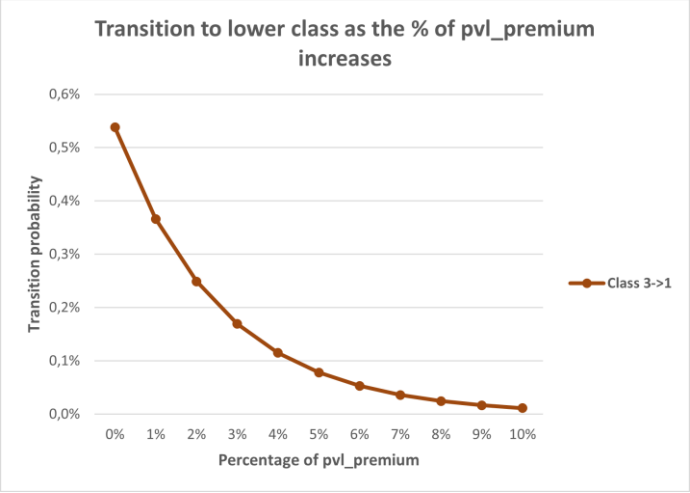


Figure 5.3 - Impact of perc_pvl_premium, focus on Class 3 → 1

It follows that the perc_pvl_premium covariate brings an improvement, albeit a small one, in the classification of sudden churn (i.e., unexpected transitions from high Class 3 to low Class 2).

The following table shows the maximum, minimum, mean and variance of the percentage of pvl_retailer products purchased out of the expenditures of the whole customer base.

	Min	Max	Mean	StDev
perc_pvl_premium	0%	50%	0.11%	3.25%

Table 5.2 - Exploration of perc_pvl_premium

5.3. Comparison between HMM and traditional RFM model

5.3.1.1. Average Monetary and Frequency per class

RFM model	Monetary mean value	Monetary standard dev
Class 3: HIGH	358.8156	207.7375
Class 2: MEDIUM	187.7554	128.9996
Class 1: LOW	43.9602	57.0894
HMM model	Monetary mean value	Monetary standard dev
Class 3: HIGH	394.431	222.862
Class 2: MEDIUM	153.468	89.901
Class 1: LOW	19.658	23.807

Table 5.3 - Monetary per class, RFM vs HMM

RFM model	Frequency mean value
Class 3: HIGH	13,156
Class 2: MEDIUM	5.313
Class 1: LOW	1.436
HMM model	Frequency mean value
Class 3: HIGH	13.654
Class 2: MEDIUM	4.702
Class 1: LOW	0.837

Table 5.4 - Frequency per class, RFM vs HMM

The tables above compare the average Monetary and Frequency values for each of the three risk classes. The values for the HMM model are taken from the output of the model itself, since they correspond by construction to the values of the distributions on which the states are fitted. Values for the RFM model, on the other hand, are calculated by averaging, over the available history, all values for each class.

The RFM analysis was conducted using Frequency, Monetary, and Regularity calculated as the mean interpurchase time for each customer in every period. Thresholds to discriminate between classes were created using thirty-third and sixty-sixth percentiles calculated period by period. In this way these thresholds are dynamic and vary during time. We report for sake of clarity the Monetary, Frequency, and Regularity values used to construct the classes (average values):

RFM thresholds	**Monetary**	**Frequency**	**Regularity**
Class 1: LOW	0-108	0-4	0-3.
Class 2: MEDIUM	108-223	4-7	3-5
Class 3: HIGH	223+	7+	5+

Table 5.5 - RFM thresholds

A preliminary macroscopic analysis highlights the crucial aspect that distinguishes the two models: the composition of the high-risk segment (Class 1) shows significant differences.

For the two best classes, Class 2 and Class 3, on the other hand, the Monetary and Frequency values are quite similar in both models.

The substantial difference in these two models emerges when comparing the average values for the highest-risk class: the HMM model is far better at identifying the high-risk customer; in fact, Class 1 has much lower average values of Monetary and Frequency (less than half) than those of the RFM model. This means having a tool that can classify high-risk cases/churners with more certainty. Taking for example a client with a monthly monetary of 105€, according to the RFM model this would belong to the low Class 1; this is not true for the HMM model where the client would be classified in Class 2 with the highest probability. It is reasonable to think that 105€ per month (think for example of a single person) is not such a low amount to the point of classifying such a client in the worst class. In addition, something even more problematic about the RFM is the partial overlap of statuses that is created especially between low and middle class. This results in uncertainty when classifying between the middle and low classes, which turn out to be precisely the two most critical classes where misclassification can lead to customers loss.

In addition, the comparison of average values confirms that the HMM model demonstrates logical consistency and consistency with the models in use, while still being able to more accurately identify customers at high risk of churn/partial default.

5.3.1.2. Transition matrix analysis

TP_Matrix HMM	Class 1: LOW	Class 2: MED	Class 3: HIGH
Class 1: LOW	80.50%	19.38%	0.12%
Class 2: MED	13.27%	83.74%	2.99%
Class 3: HIGH	0.52%	7.94%	91.55%

Table 5.6 - Transition matrix HMM

TP_Matrix RFM	Class 1: LOW	Class 2: MED	Class 3: HIGH
Class 1: LOW	77.56%	19.80%	2.65%
Class 2: MED	35.58%	47.24%	17.18%
Class 3: HIGH	5.99%	20.57%	73.44%

Table 5.7 - Transition matrix RFM

Looking at the transition matrix of the HMM, it is noticeable that the diagonal (representing the probability of remaining in the same state in the next period) has values that are always larger than those of the generated matrix of the RFM model. This means that in general the classification of the HMM turns out to be more consistent. In particular, compared to the RFM model the HMM eliminates those noisy transitions that are the result of exceeding, even by a negligible delta, the predetermined thresholds.

The transition in the HMM is not simply described by the crossing of a predetermined threshold but rather corresponds to the greater membership, in probabilistic terms, of the client in a new class rather than the previous one. This membership is calculated by how much of the emissions the client produces (in terms of Monetary and Frequency), belongs to the distribution that describes the underlying states (Gaussian in the case of Monetary, and discrete Poisson in the case of Frequency). Hence it is possible to probabilistically define how likely it is that the observed emission comes from each of the state distributions.

Classification in the RFM by thresholds decided a priori makes switching between classes more frequent and less meaningful. In particular, a big difference concerns the most stable class in the model (i.e., the class with higher staying probability), which in the HMM turns out to be the high Class 3 (91.55%) while in the RFM it is the low Class 1 (77.56%).

Furthermore, in the RFM model Class 2 turns out to be an unstable class, meaning that the probability of a class transition in the next period is greater than 50% (staying probability amounts to 47%, probability of transition to another class 53%).

This is a remarkable issue since the state in which misclassifications should be avoided is precisely the average state since the greatest risk is to classify as average a customer who is actually low, thus risking missing the opportunity to act early to avoid churn.

5.4. Model reaction to topical customer behaviours

Below are shown four typical behaviors that a retailer's customers may engage in. The results produced by our model for each of these customers chosen as archetypes will then be discussed.

- **Stable:** during the observation period these clients have a constant behavior, always staying in the same class or at most making few fluctuations between the highest and middle class (Class 3 - Class 2).
- **Partial defection/announced churn:** those are customers who show a gradual decline in their buying habits that leads them to enter the lowest state, Class 1. Those customers effectively go through a partial defection process, in which their performances decline period after period.
- **Unexpected churn:** for those clients the termination of the relationship is unpredictable. In fact, churn occurs quite suddenly and sometimes even after periods when the client's performance was even improving.
- **Variable:** those customers over the observation time frame have highly variable behaviour. This archetype represents the vast majority of the observations, and this should not be surprising since in the FMCG field it is common to observe a high level of variability especially if the time horizon of data aggregation is weekly or monthly (as in our study).

Below are reported the graphs showing how the model responds to the four archetypes. For comprehension's sake a logarithmic scale graph will be provided. Indeed, logarithmic scale is useful to better understand the increase in the posterior probability of belonging to the lowest class. In fact, it is common for the value of the posterior probabilities of belonging to the lowest class to be extremely low (we are talking orders of magnitude of E-100), especially when the customer is very loyal this belonging to the highest Class 3.

Observing the increase in logarithmic scale allows to better grasp interesting signals related to the change in the probability of making a transition from a higher to a lower class (with special regard to Class 1).

5.4.1.1. Stable customers

Figure 5.4 - Stable customer example

In these graphs is shown the trend in the posterior probabilities of a stable customer of belonging to each of the three classes. Specifically, the selected customer is assigned to Class 3, and remains stable in the same class for all periods. This means, as shown in the first graph, that the posterior probability of being in Class 3 is constant over time and has a value approximating 1. Instead, in the second graph it is shown the evolution of the posterior probabilities of the same customer, but in logarithmic scale: the value of the posterior probability relative to the lowest class remains almost constant at very negligible values, the trend line is almost horizontal and is composed by extremely small probabilities.

5.4.1.2. Unexpected churn

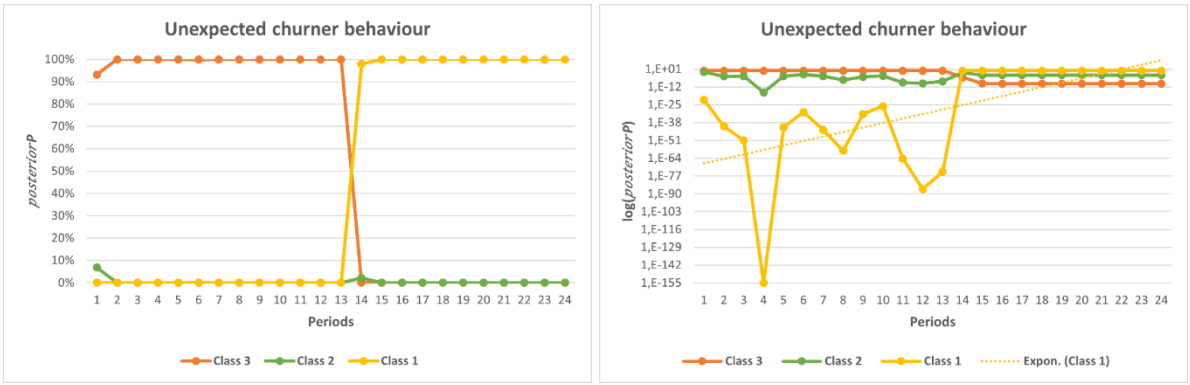


Figure 5.5 - Unexpected churn example

In this case the selected client, after a very stable period spent in the highest Class 3, suddenly churns. This behaviour is highly unpredictable and often caused by external factors, over which it is not possible to have any visibility or control. As can be clearly seen in the logarithmic graph, there are no particular signs that would point to a possible churn. On the contrary, in the two periods prior to the churn the value of the posterior probability associated with being in the lower Class 1 was expressing a downward trend, meaning that the customer was indeed doing better than before. This archetype is by far the most difficult to deal with; only the introduction of covariates to the model might help in recognizing this kind of sudden churns.

5.4.1.3. Partial defection/announced churn customer

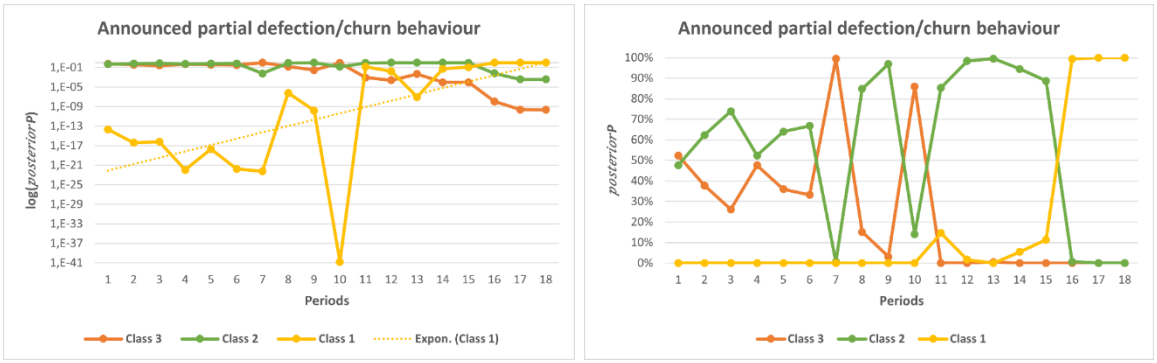


Figure 5.6 - Partial defection/announced churn example

Customers who perform partial defection are those users who after a few stable periods in one of the two high classes (or with small transitions between them) begin to show signs of relationship deterioration. In particular, in the first graph it is shown how the client in the first ten periods is fairly constant, net of some acceptable variability. From period eleven, on the other hand, there is a clear sign of the beginning of a partial defection process: the probability of belonging to the highest class is zeroed out and at the same time the probability of belonging to the lowest class increases significantly, even though the client is actually classified in the middle Class 2. This sudden increase is far more visible in the logarithmic scale graph: transition probability to Class 1 goes from orders of magnitude of E-20 up to orders of magnitude ranging between E-01 and E-05, while transition probabilities of belonging to Class 2 or Class 3 demonstrate a reverse trend (decreasing the probabilistic degree of membership to these two classes).

The partial defection is even more evident observing the trend line of the probability of belonging to Class 1. In fact, such trend line is quite pronounced upward, meaning that the associated risk of churn increases constantly during time. This customer represents the typical case in which the use of our model would have allowed for early identification of customer churn, in the example shown around five months in advance.

5.4.1.4. Occasional customer

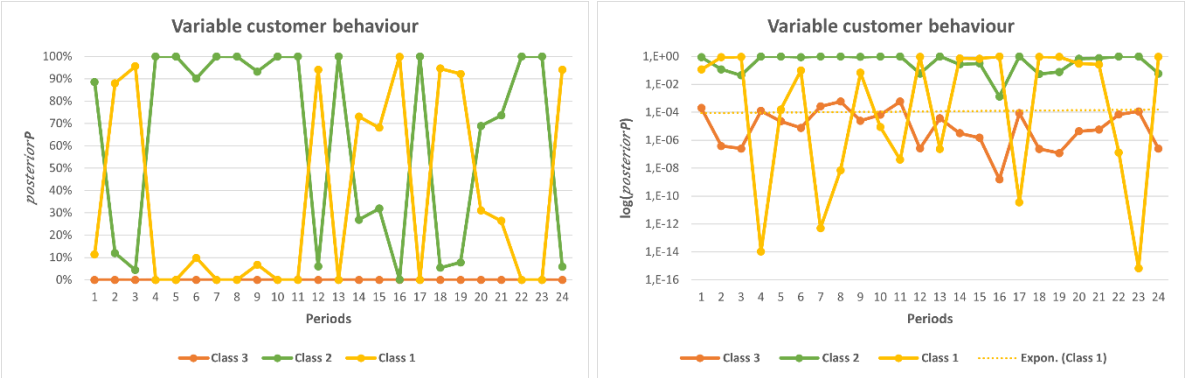


Figure 5.7 - Occasional customer example

Some customer behaviors are characterized by very frequent transitions between classes (often between Class 1 and Class 2). It can be seen from the posterior probabilities graph that the client in question is never assigned to Class 3 but rather continues to move from Class 2 to Class 1 and vice versa. In these cases, it is not easy to tell whether such a customer is actually a churn/partial defection. However, it is still possible to carry out marketing actions aimed at stabilizing the customer in Class 2 so as to keep his behavior away from churn.

6 Conclusions - limitations and research boundaries

6.1. Managerial implications

The first managerial implication is due to the greater stability of the identified classes, which means being able to precisely divide the customers into classes that better describes their purchasing behaviour and the associated risk of partial defection/churn. In economic terms, this model can be used to identify a pool of customers which are more likely to churn and abandon the store. Consequently, it is possible to specifically targeting them with retention actions avoiding wasting time and effort on customers who do not present a high risk of defection.

The second managerial implication concerns the results obtained relative to private label products. Thanks to the evidence shown before, we can conclude that retailer private label products are able to induce customer loyalty. This statement is even more valid for premium private label products. With these findings, it is therefore possible to introduce private label products for loyalty marketing campaigns, trying to incentivize clients to purchase and taste these products. Moreover, the model itself provides an answer to the retailer about consumers' perceptions of private labels. In fact, it is reasonable to say that if retailer and premium private labels strengthen the relationship with the customer, it means that the customer himself appreciates them and finds in them adequate value for money. Thus, such products not only provide a higher margin to the retailer but also become a real marketing tool.

Another managerial insight arises from the covariate's construction (which are calculated on monthly bases). This it is such that they are effective alarm bells for spotting the deterioration of customer relationship: if a good customer reduces its purchases in private labels, both retailer and premium, this can be noticed as an increase in the inherent risk of churn. Thanks to this implication the customer deterioration can be spotted in advance (i.e., churn prediction).

The last managerial implication concerns the validation of the model. On the one hand, it is easy to act by varying the type and the number of covariates in input to the model, accordingly to the result pursued for construction of marketing campaigns. On the other, as demonstrated by the model trained on new customers, the model has good flexibility to new data, and consequently resistance to the inherent variability in buying behavior, both in terms of monetary expenditure and frequency. In the end, the test with new customers validates the effectiveness of private labels in increasing customer retention; this shows that the result obtained from the Master model is for all intents and purposes attributable to actual buying behavior that discriminates high-performing customers from customers at high risk of churn/partial defection. From this result, it is possible to create ad hoc marketing campaigns to increase the degree of customer retention, particularly the customers in the high risk of churn class.

A final observation concerns the applications of the model: the model here disclosed was trained and tested on data from a single retailer, however, we have no evidence to affirm that the model is not applicable to data from other retailers active in FMCG sector: the metrics and KPIs used remain applicable, with appropriate adjustments, for any other retailer label. Moreover, the scientific-statistical basis of the model does not depend on the retailer from which the data was retrieved from.

6.2. Contribution to the literature

The contribution that this thesis makes to research occurs in several ways; first, this disclosure goes to cover an area of the literature that is still rather sparse, as explained in chapter 2. 4 "Research Needs", this model will thus go to position itself in a gap area of research by introducing the application of hidden Markov model as an analytical tool to segment the customer base of an FMCG retailer into levels of churn risk. Furthermore, there are no limitations to extending the use of the model, with due corrections (e.g., by adjusting the width of the periods), to other cases where non-contractual customer relationships exist.

The other contribution to the research concerns the still open discussion on whether or not private labels are synonymous of customer retention; with this model it was possible to demonstrate the usefulness that private labels have in the area of customer retention. In addition, from further analysis on private labels it was also possible to verify how different private label lines affect with different power the customer's risk of churn.

Eventually, research questions had been posed to conduce this study (fully disclosed in Chapter 2.5). those questions arose from the literature review and research needs identified. Below, for completeness of discussion, are the answers to these questions:

- A data-driven model has been developed for segmentation of customers into risk classes; specifically, the model is capable of identifying classes within the customer base based on the customers' own performance. Moreover, this classification is done probabilistically; in fact, the model is able to quantify the degree of membership in a given class.
- Through this model it was possible to identify predictors of partial defection risk, specifically the percentage of private labels purchased per expenditure in any given period, for each type of private label. As presented in Section 5.3, private labels have an inertial effect on the transition to the lowest class, this is especially true for pvl_retailer and pvl_premium type private labels, which represent the retailer's attempt to offer an alternative choice to the competition that is not merely based on cost-effectiveness.

- It was possible to compare the performance of the model with hidden Markov chain, to the model that is currently in most widespread use among practitioners (RFM analysis) and highlight how our model has better performance in classifying more significantly customer with high risk of churn/partial defection (i.e., Class 1 customers). The full disclosure of the comparison is carried out in Chapter 5.2.
- Through the results obtained from our model it was possible to provide managerial insights: the managerial implications are mainly related to improving the addressability of marketing actions, so that the right customer receives the right marketing action, increasing its effectiveness; as well as reducing marketing efforts, thereby increasing the performance of marketing campaigns. The full disclosure of this argument is covered in Chapter 6.1.

6.3. Future improvements of the model

The model developed and discussed represents a potential starting point for subsequent improvements. Conceptually, the model may find other applications in markets governed by customer-firm relationships not defined in contractual terms.

It should be emphasized that although the proposed model has very good behaviour and resilience to "noisy" data, since it is still a data analysis tool the latter will perform all the better when the input data are pre-processed and cleaned of any noise before being used to build the model itself. In these terms, any improvement in how outliers and tellers are cleaned could benefit the final result.

What has been discussed so far mainly concerns the dataset and the context in which our model is made to operate; turning instead to the modelling part there are some tricks to be implemented that can make our model perform better.

First of all, with respect to the Monetary emission, the underlying classes of risk are created by our hidden Markov model using a normal distribution assumption. However, what we have reason to believe, and what we have observed during the development of this model, is that the use of a gamma-type distribution could bring better consistency to the classes and make it easier for the model to create them.

This because often the segments created, even with traditional models such as RFM analysis, presents distribution of monetary quite left skewed, hence gamma distribution might be able to better capture this behaviour. In the DepmixS4 package used to fit the hidden Markov model is not, to date, implemented gamma distributions to describe the system emissions. Unfortunately, our programming skills in R are not enough to make improvements to the source code of the library, consequently we could not test the model with gamma distributions for the Monetary.

A final cue, from which future improvements can be produced, concerns the integration and testing of new covariates: these not only bring a quantitative improvement to the classification ability of our model, but also generate an extremely valuable implication from a managerial and marketing point of view; in fact, allowing for the best understanding on the dynamics of churn/partial defection and thus intervention for the prevention of these.

Bibliography

- Kumar, V., Reinartz, W. "*Customer Relationship Management (2006), third edition*".
- Gold, C., foreword by Tzuo, T. "*Fighting Churn with Data. The science and strategy of customer retention (2020)*".
- Lokuge, S., Sedera, D., Ariyachandra, T., Kumar, S., & Ravi, V. (2020). "*The Next Wave of CRM Innovation: Implications for Research, Teaching, and Practice*". Communications of the Association for Information Systems, 46, pp-pp.
- Algamdi, A., Brika, S.K.M., Laamari, I., Chergui, K. "*CRM system and potential customer loyalty trends: Some evidence of customer engagement*". (2021) International Journal of Advanced and Applied Sciences, 8 (5), pp. 44-52.
- Eva Ascarza, Oded Netzer, Bruce G. S. Hardie (2018) *"Some Customers Would Rather Leave Without Saying Goodbye*". Marketing. Science 37(1):54-77.
- Meena, P., & Sahu, P. (2021). "*Customer relationship management research from 2000 to 2020: An academic literature review and classification*". Vision, 25(2), 136-158.
- Awasthi, P., & Sangle, P. S. (2012). *"Adoption of CRM technology in multichannel environment: A review (2006–2010)"*. Business Process Management Journal, 18(3), 445–471.
- S.Vandermerwe, J.Rada (1988). *"Servitization of business: Adding value by adding services"*. European Management Journal
- Piva, A., 27 March 2019. *"Le 5V dei Big Data: dal Volume al Valore"*. Osservatori.net https://blog.osservatori.net/it_it/le-5v-dei-big-data?hsLang=it-it.

- Arno De Caigny, Kristof Coussement, Koen W. De Bock. *"A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees"*. European Journal of Operational Research, Volume 269, Issue 2, 2018, Pages 760-772.
- Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, 2nd edition, Springer, 2009.
- Ricci, V., *Fitting distributions with R*, 2005
- Jeffrey W. Miller (2016). Lecture Notes on Advanced Stochastic Modeling. Duke University, Durham, NC.
- Dr. R.K. Srivastava, Digesh Pandey, Speech recognition using HMM and Soft Computing 2022, Materials Today: Proceedings, Volume 51.
- Chakkarai Sathyaseelan, L Ponoop Prasad Patro, Thenmalarchelvi Rathinavelan Sequence patterns and HMM profiles to predict proteome wide zinc finger motifs 2022, Pattern Recognition.
- Jia Liu, Miyi Duan, Wenfa Li, Xinguang Tian MMs based masquerade detection for network security on with parallel computing.
- William N. Robinson, Andrea Aria, Sequential fraud detection for prepaid cards using hidden Markov model divergence. 2018, Expert Systems with Applications, volume 91.
- Jorge Florez-Acosta. Do preferences for private labels respond to supermarket loyalty programs? Journal of Economic Behavior and Organization Open Access Volume 188, Pages 183 – 208 August 2021
- Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, Bart Baesens. New insights into churn prediction in telecommunication sector: A profit driven data mining approach 2012, European Journal of Operational Research

- James Lyons, Kuldip K. Paliwal, Abdollah Dehzangi, Rhys Heffernan, Tatsuhiko Tsunoda, Alok Sharma Protein fold recognition using HMM–HMM alignment and dynamic programming. 2016, Journal of Theoretical Biology, Volume 393
- William N. Robinson, Andrea Aria. Sequential fraud detection for prepaid cards using hidden Markov model divergence. 2018, Expert Systems with Applications. Volume 91
- Jia Liu, Miyi Duan, Wenfa Li, Xinguang Tian. MMs based masquerade detection for network security on with parallel computing. 2020, Computer Communications. Volume 156
- Chakkarai Sathyaseelan, L Ponoop Prasad Patro, Thenmalarchelvi Rathinavelan, Sequence patterns and HMM profiles to predict proteome wide zinc finger motifs 2022, Pattern Recognition
- Dr. R.K. Srivastava, Digesh Pandey. Speech recognition using HMM and Soft Computing. 2022, Materials Today: Proceedings,Volume 51
- A. Keramati,R. Jafari-Marandi, M. Aliannejadi,I. Ahmadian,M. Mozaffari,U. Abbasi. Improved churn prediction in telecommunication industry using data mining techniques.2014, Applied Soft Computing, Volume 24
- M. Clemente-Císcar, S. San Matías, V. Giner-Bosch. A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings 2014, European Journal of Operational Research
- T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.Ch. Chatzisavvas. A comparison of machine learning techniques for customer churn prediction 2015, Simulation Modelling Practice and Theory

- Mehdi Mohammadzadeh, Zeinab Zare Hoseini, Hamid Derafshi, A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran 2017, Procedia Computer Science
- Niels Holtrop, Jaap E. Wieringa, Maarten J. Gijsenberg, Peter C. Verhoef. No future without the past? Predicting churn in the face of customer privacy 2017, International Journal of Research in Marketing
- Adnan Amin, Babar Shah, Asad Masood Khattak, Fernando Joaquim Lopes Moreira, Gohar Ali, Alvaro Rocha, Sajid Anwar. Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods 2019, International Journal of Information Management
- Farid Shirazi, Mahbobeh Mohammadi. A big data analytics model for customer churn prediction in the retiree segment 2019, International Journal of Information Management
- Yixin Li, Bingzhang Hou, Yue Wu, Donglai Zhao, Aoran Xie, Peng Zou. Giant fight: Customer churn prediction in traditional broadcast industry 2021, Journal of Business Research
- Sulim Kim, Heeseok Lee Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees 2022, Procedia Computer Science
- Mai Kiguchi, Waddah Saeed, Imran Medi. Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest 2022, Applied Soft Computing

List of Figures

List of Tables

Ringraziamenti

Un grazie particolare al professor Lucio Lamberti, per l'opportunità che ci ha dato di studiare ed approfondire un argomento di ricerca estremamente attuale ed accattivante. Inoltre, questo lavoro non sarebbe stato possibile senza il tempo, la disponibilità e i preziosi consigli forniti dal nostro correlatore ing. Emanuele Fedrigolli.

Una dedica speciale a tutti coloro che hanno reso questi anni speciali, in particolare a Matteo, Luca, Andrea M., Andrea T., Ferdinando e Alessandro, F.P. Bianca; compagni di avventure e risate. Un sentito ringraziamento agli amici di Milano Accademia, casa e sede delle lunghe e numerose giornate di studio.

Un ringraziamento speciale anche ad Arianna e Federica, che ci hanno incoraggiato, supportato e sopportato, specialmente durante le lunghe chiamate per la tesi.

Ultimo, ma non meno importante un grazie dal profondo del cuore ai nostri genitori e alle nostre famiglie, senza i loro insegnamenti, il loro amore e il loro sostegno questo traguardo non sarebbe stato possibile.

Lorenzo e Mirko