**Assignment Two**

**MIS784 Marketing Analytics**

**Yash Ahuja**

**Student Id: 219608443**

Table of Contents

[Introduction and problem definition 2](#_Toc52071372)

[Literature review 3](#_Toc52071373)

[Methodology 5](#_Toc52071374)

[Predictive techniques 5](#_Toc52071375)

[Logistic regression 5](#_Toc52071376)

[RFM Model 5](#_Toc52071377)

[Evaluation criteria 6](#_Toc52071378)

[Empirical Study 6](#_Toc52071379)

[Results 7](#_Toc52071380)

[Conclusion and Recommendations 9](#_Toc52071381)

[References 9](#_Toc52071382)

# Introduction and problem definition

Due to the recent advancement in technology over the past few years, extremely competitive market and surplus supply, customers are now considered as the real ruler of the market causing the focus of the companies being shifted from products to customers. As a result of these advancements, maintaining customer relationships has become the most critical issue for companies. However, many companies lose their valuable customers to other competitors which is called as customer churning. This is where the idea of customer churn prediction comes in. The companies observe the buying behaviour of the customers and predict the customer churn (i.e. customers who cease business with the company after some time). The companies then launch targeted marketing campaigns towards those potential churners in order to retain them (Khodabandehlou & Zivari Rahman, 2017). Customer churn can decrease profit, renounce referrals from existing customers and increase customer acquisition costs (Tamaddoni, Stakhovych, & Ewing, 2016). Based on various research studies, it has been suggested that the cost of acquiring a new customer is to the cost of retaining five existing customers. Customer churn in retail industry is higher than any other businesses because of the presence of large number of stores and high competition among them. Therefore, the most crucial application of customer churn prediction is in the retail industry. Various marketing experts have suggested that companies, on average, annually lose around 25 percent of their customers and approximately 25 to 85 percent increase in revenues can occur as a result of 5 percent reduction in customer churn implicating the importance of customer churn prediction (Khodabandehlou & Zivari Rahman, 2017).

The companies predict customer churn through predictive models built on the historical data of the consumers indicating the buying behaviour of the customers. The main goal of this study is to understand which models can be used in customer churn prediction, how they are implemented and investigate their differences and limitations in predicting customer churn. It is also worth noting to observe what are the evaluation criteria of these models and benefits of using these models in terms of revenues or marketing expenditure of companies, which customers to focus on, profitable or all and when should the company stop pursuing customers. It also aims to understand the issues faced by these predictive models such as what aspects should be considered (e.g. variables used) while building these models for churn prediction (Tamaddoni et al., 2016). Moreover, how these models can aid top executives in redesigning the existing strategies of the companies based on data driven decisions for reducing customer churn and what are the drivers of customer churn (Gustafsson, Johnson, & Roos, 2005).

# Literature review

It has been suggested that the simpler methods such as logistic regression and RFM models can be used for customer churn prediction. These models are benchmarked models for customer churn prediction. Logistic regression model is used for predicting the probability of the customer to be a churner or non-churner (Tamaddoni et al., 2016). It makes use of the variables in the same way as linear regression and has same assumptions of data as in the case of linear regression such as no missing values, all variables numeric, no extreme values and all variables independent. The variables used in churn prediction are called predictors such as amount spent, quantity of items, type of products etc and the churn to be predicted is called label. It calculates the probability of each customer to be a churner (1) or non-churner (0). As a result of its simplicity, robustness and regularization to avoid overfitting, it is one of the most favoured method for churn prediction by marketers. RFM model, on the other hand, focuses on three variables such as recency, frequency and monetary implicating the time when last product was purchased, number of products purchased, and the average amount of money spent on each transaction in the observation period. Now, customers are split into equal five groups based on the values of these variables and hence a three-digit code is assigned to each customer indicating their churn probability. The highest and lowest number represent the churn probability based on the coding of the RFM codes. This traditional approach for churn prediction is very simple to use. The performance of both the models (LR and RFM) tends to decrease in case of complex and large datasets (Miguéis, Van den Poel, Camanho, & Falcão e Cunha, 2012)

The evaluation criteria most used for customer churn prediction is cumulative lift which specifies the cumulative percentage of churners in each decile (Tamaddoni et al., 2016). Other studies also claim that higher top-decile lift (i.e. cumulative lift) represents better performance of the model. This criterion helps in customer retention campaigns by providing majority of the churners in top deciles and thus helping in saving marketing expenditure of the company. Since, customer churn prediction is a classification problem of categorising the customers into churners and non-churners. The traditional criterion used for classification model evaluation is confusion matrix which includes accuracy (specifying the correct predictions out of total predictions), misclassification rate (1 – accuracy). This metric has its limitations when the classes are imbalanced. Therefore, sensitivity/recall (the number of positive class members correctly identified), specificity (the number of negative class members correctly identified), precision (the fraction of examples assigned the positive class that belong to the positive class) can be used in cases of class imbalance (Burez & Van den Poel, 2009).

Also, a further extension of confusion matrix is area under the curve (AUC) criteria which specifies the ratio of sensitivity to 1-specificity and should be close to 1 for an ideal classifier and is the most used classification criteria. Moreover, Lift % is also used as an evaluation criterion which represents how much better a classifier performs as compared to random model. Lift % representing the churn rate of each decile divided by overall churn rate. Also, lift analysis measures have certain limitations as they are observed in a certain period and they do not consider the feedback loops for next observation period (De Caigny, Coussement, & De Bock, 2018). Other criterion used for evaluation such as gini coefficient which focuses on less risky segments too instead of riskiest segments only (as in the case of lift% and cumulative lift), this metric considers all scores including less risky and high riskiest segments and thus give an unbiased evaluation of the model performance. As, the model who performs better in identifying riskiest segments (likeliest churners) does not necessarily be good at identifying less risky segments (less likely churners) which can help in identifying which customers to focus on along with profitable customers and when we can stop pursuing them once their gini coefficient becomes low (Lemmens & Croux, 2006). Although, different studies have suggested different criteria for model evaluation and but none of the studies represent which criteria can be generalized for churn prediction or which criteria has more importance in terms of the size of the dataset.

Some studies suggest that models such as ANN is one of the most successful and popular models in churn prediction and produces better results than traditional approaches such as logistic regression and RFM models when the data is complex and large. Also, it has better performance than other models such as SVM, decision trees and ensemble learners also produce optimum results for customer churn prediction (Khodabandehlou & Zivari Rahman, 2017). Whereas, some studies also suggest that ensemble learners produce better results than advanced ANN models on simpler data (Lemmens & Croux, 2006). Whereas, other studies suggest that when evaluating the best modelling technique for churn prediction models is highly dependent on the evaluation criteria used for evaluation (Arisholm, Briand, & Johannessen, 2010). Although, different studies have suggested different models for churn prediction evaluation based on the complexity and size of the data but none of the studies represent which models produce best results for all kinds of datasets and evaluation measures.

Some studies have talked about the effects of variable selection in churn prediction for complex and simple models and determined that complex models perform better with large number of variables and some variables or factors when included in the model play a significant role in model performance implicating the importance of some variables (Tamaddoni et al., 2016). Also, some studies have also provided the generalised description of what type of variables play an important role or must be included while building the model for customer churn prediction as purchase frequency, time and amount in case of RFM models. All the information regarding selection of variables helps in understanding the black box nature of customer churn to some extent but still further detailed implications of variables selected for model building are not provided by any of the studies (Miguéis et al., 2012).

It has also been suggested that customer churn prediction models provide top executives meaningful insights about their customers through evaluation measures which are used by them in designing targeted customer retention campaigns for potential churners. Also, these insights provide them a deep understanding of the buyer behaviour of their customers and thus, the top executives can use this information to alter their strategies as per customer requirements and as a result, companies generate increased revenues. Moreover, various model building practices show that price and quality variables become insignificant when customer satisfaction variable is included in the churn prediction model and the effects of price and quality variable on churn are compensated by the customer satisfaction variable implicating customer satisfaction variable to be the critical drivers for customer churn (Gustafsson et al., 2005).

But none of the studies have discussed in detail about the drivers of low customer satisfaction level which according to majority of them is leading to increased customer churn rate. Moreover, none of the studies incorporated the impact of increased competition on the customer churn while building the customer churn prediction models using simple and advanced methods. These areas along with above mentioned information not provided by each study (i.e. knowledge gaps) should be investigated in order to get the true essence of the consumer behaviour and hence to predict customer churn reliably and accurately.

# Methodology

## Predictive techniques

Predictive modelling techniques used in this study are logistic regression model and RFM model.

### Logistic regression

Since, we are predicting customer churn i.e. the dependent variable is binary, logistic regression model is used to predict the churn probabilities for each customer considering the independent variables. The variables used in churn prediction are called predictors and the customer churn probability outcome (0 being non-churner and 1 being churner) to be predicted is called label. The model is first trained on a large dataset and then the trained model is implemented on the testing data set in order to check the reliability of our model. Firstly, in training, variable selection is done in order to determine which variables will be used in building the model, then the model calculates the logit (i.e. odds) of each customer to be a churner and the corresponding LL calculation (log likelihood) is calculated. The log likelihood is based on summing the probabilities associated with the predicted and actual outcomes of churn. This statistic is analogous to residual sum of squares (SSR) in linear regression implicating how much unexplained information there is after the model has been fitted. The trained model is then implemented on the testing data and the predicted probabilities of each customer to be a churner or non-churner based on the selected predictors are calculated which range from 0 to 1 and an upper limit is set for probability of being a churner (i.e. 0.5) and finally the customer churn outcome of customer being churner (probability 1) or non-churner (probability 0) is calculated for testing data.

Like every other model, logistic regression model also has its limitations as the logistic regression model has various assumptions for the data used in the model building process which includes no missing values, all numeric values and no extreme cases. These assumptions are checked through residual analysis after the model development. The strengths of the logistic regression model include ease of use, robustness whereas the weaknesses of the LR model are its inability to handle complex, non-linear and large datasets.

### RFM Model

The RFM model is applied on the same data used for testing the LR model. Variable selection is the first step of the RFM model. Since, RFM model only focuses on the frequency, recency and monetary aspects of the customer data. The variables selected for building the RFM model only relate to time since the last purchase, quantity of the products purchased by the customer and the average amount spent by the customer on a single transaction in the observation period. The selected variables are then transformed or aggregated in order to meet the requirements of the recency, frequency and monetary codes. After the transformation, the recency code is generated by sorting the time since last purchase variable from lowest to highest, quantity variable from highest to lowest and the amount variable from highest to lowest. All the customers are then grouped into equal five groups 1 to 5 with first group having 1 code and second group having 2 code and so on until the fifth group. The process is done for all the three variables recency, frequency and monetary codes. Finally, RFM score is generated for each customer which is a three-digit code and is sorted from highest to lowest implicating the churn probability of each customer with highest score (555) having highest churn probability and lowest score (111) having lowest churn probability. Based on the RFM scores, the companies can target top portion of the customers for customer retention campaigns.

RFM model also has its limitations as the model does not perform well on complex and large datasets. Moreover, the strength of the RFM model being its simplicity can also become its weakness in real-world scenarios where it is likely that most of the customers of the company have ordered once, spent less and have not purchased in one or more years.

## Evaluation criteria

The evaluation criteria used in the evaluation of logistic regression model is confusion matrix. As known from the literature review, it includes various measures such as accuracy (specifying the correct churner and non-churners out of total predictions), misclassification rate (1 – accuracy). This metric has its limitations when the classes are imbalanced. It also includes sensitivity/recall (the number of churners correctly identified) which is of more relevance as focus is on churners and specificity (the number of non-churners correctly identified). Both measures overcome issue of class imbalance faced by accuracy measure. But these measures are difficult to calculate when the data is complex or large.

Another criterion used for evaluation of logistic regression model and RFM model is lift analysis. This measure helps to evaluate the performance of each model as well as comparison of both the models on the same data set. Lift analysis includes cumulative lift which represent the cumulative percentage also known as cumulative response/concentration of churners in each decile. Higher cumulative lift in top deciles represent better performance of the model as it provides majority of the churners in top deciles and thus helps in saving marketing expenditure of the company by targeting only those deciles. Cumulative lifts for RFM model can also be calculated and graphs can be constructed in order to make the comparison between the LR model and RFM model. Moreover, the lift analysis graph also provides comparison of the LR model and RFM model with the random model. But the disadvantage of lift measures is that these measures are observed in a certain period and they do not consider the feedback loops for next observation period i.e. a reduction in the likelihood of the churn when customers were targeted with the retention campaign.

# Empirical Study

TESCO PLC company’s data consisting of 30,000 records (20,000 training set, 10,000 testing set) in a period from 1 January 2015 to 31 December 2015 provided from TESCO Clubcard is used by LR model in customer churn prediction. The LR model is first trained on the training set and the trained model is then applied to the testing set for reliable prediction of customer churn. All the independent variables namely id, purchase, T.last, T.active, loyalty, service failure, total profit, AP.spent, BH.spent, DL.spent, DY.spent, FV.spent, GM.spent, GR.spent, LQ.spent, MT.spent and Socio.Economic and the dependent variable namely churn are model building variable which are included in the LR model. All the variables used are numeric. The LR model is constructed based on these variables of the training set (20,000 records) and then the constructed model is then applied to the same variable in the testing set (10,000 records) in order to reliably calculate the predicted churn of the testing set. In order to check the reliability of the model, it is necessary to test the trained model on the testing data which includes churn label so that we can check the reliability of the model before deployment by comparing the predicted churn values with the actual churn values.

In order to make the comparison between LR model and RFM model, same data i.e. testing set is used for building the RFM model. The variables used in RFM model include all the variables except loyalty, service failure, total profit and Socio.Economic as these variables have no significance in terms of recency, frequency and monetary codes. The difference of T.active and T.last respectively generates new variable called Time since last purchase (in weeks) and the average of all the total spent variables generates a new variable called average spending as required for calculating recency and monetary codes in RFM. The model built on above mentioned variables generate the RFM score indicating the churn probability of each customer.

# Results

|  |  |  |  |
| --- | --- | --- | --- |
| **CONFUSION MATRIX** | |  |  |
|  |  |  |  |
|  | **Predicted churn** |  |  |
| **Actual churn** | 0 | 1 | **Total** |
| 0 | 4751 | 998 | 5749 |
| 1 | 1120 | 3131 | 4251 |
| **Total** | 5871 | 4129 | 10000 |

Figure 1

|  |  |
| --- | --- |
| **PERFORMANCE OF LOGISTIC REGRESSION MODEL** | |
| **Accuracy rate** | 79% |
| **Misclassification rate** | 21% |
| **Sensitivity rate** | 74% |
| **Specificity rate** | 83% |

Figure 2

|  |  |  |  |
| --- | --- | --- | --- |
| **CUMULATIVE RESPONSE (CONCENTRATION)** | | | |
|  |  |  |  |
| **Number of deciles** | **Logistic  regression** | **RFM Model** | **Random  Model** |
| 0.00 | 0.00 | 0.00 | 0.00 |
| 1.00 | 23.52 | 23.52 | 10.00 |
| 2.00 | 47.05 | 47.05 | 20.00 |
| 3.00 | 70.57 | 70.57 | 30.00 |
| 4.00 | 94.10 | 94.10 | 40.00 |
| 5.00 | 100.00 | 100.00 | 50.00 |
| 6.00 | 100.00 | 100.00 | 60.00 |
| 7.00 | 100.00 | 100.00 | 70.00 |
| 8.00 | 100.00 | 100.00 | 80.00 |
| 9.00 | 100.00 | 100.00 | 90.00 |
| 10.00 | 100.00 | 100.00 | 100.00 |
| Total |  |  |  |

Figure 3

Figure 4

The figure 1 and 2 describes the performance measures of logistic regression model based on confusion matrix. The figure 3 and 4 implicates that the logistic regression model and RFM model have same performance by having approximately 94% of churners in their top 4 deciles. Moreover, both models are performing better than baseline random model which is having only 40% of churners in top 4 deciles.

# Conclusion and Recommendations

Based on the results, the accuracy of the LR model is 79% which means around 79% of the total churn predictions are correct (refer figure 2). However, the data used has slightly imbalanced classes with more records in the non-churning class as compared to churning class. Other measures such as sensitivity of the LR model is 74%, implicating how well the churner class is predicted and hence the model is performing well. However, it can also be seen that specificity rate is higher i.e. 83% indicating the model is predicting non-churner class in a better way as compared to churner class (refer figure 2) but that might not be the case as potential reason for this may be that the same model built on training data which has more non-churner class records, is applied to the testing data (refer figure 1). Moreover, it is evident that both the models LR and RFM have successfully performed in predicting customer churn by indicating majority (approx. 94%) of the churners in the top 4 deciles of the data. Higher cumulative concentration in top deciles is implicating both the models are performing well and their results can be used (refer figure 3 and 4)

The models have provided the segment of the customers that the managers need to target for retention campaign. Moreover, by only focusing on the top 4 deciles for customer retention campaign as majority of the churners are in these deciles, the managers can have reduced marketing expenditure which may ultimately lead to increased customer retention rate.

However, the study illustrated various knowledge gaps in the field of customer churn prediction when it comes to large, complex and non-linear datasets which require further research as it has been found that these benchmarked models LR and RFM do not perform well in such cases whereas advanced models such as ANN, SVM, ensemble learners can handle complex dataset but have their limitations because of increased computational costs and inefficient results with simple and small datasets.

# References

* Arisholm, E., Briand, L. C., & Johannessen, E. B. (2010). A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. *Journal of Systems and Software, 83*(1), 2-17. doi:<https://doi.org/10.1016/j.jss.2009.06.055>
* Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3, Part 1), 4626-4636. doi:<https://doi.org/10.1016/j.eswa.2008.05.027>
* De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760-772. doi:<https://doi.org/10.1016/j.ejor.2018.02.009>
* Gustafsson, A., Johnson, M. D., & Roos, I. (2005). The Effects of Customer Satisfaction, Relationship Commitment Dimensions, and Triggers on Customer Retention. *Journal of Marketing, 69*(4), 210-218. doi:10.1509/jmkg.2005.69.4.210
* Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology, 19*(1/2), 65-93. doi:10.1108/JSIT-10-2016-0061
* Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research, 43*(2), 276-286. doi:10.1509/jmkr.43.2.276
* Miguéis, V. L., Van den Poel, D., Camanho, A. S., & Falcão e Cunha, J. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications, 39*(12), 11250-11256. doi:<https://doi.org/10.1016/j.eswa.2012.03.073>
* Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing Churn Prediction Techniques and Assessing Their Performance:A Contingent Perspective. *Journal of Service Research, 19*(2), 123-141. doi:10.1177/1094670515616376