A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry

Carol Anne Hargreaves

***Abstract*—This paper focuses on the application of a machine learning algorithms such as the logistic regression, to firstly, derive insights from the data to identify the factors that drive churn, secondly to identify which customers are highly likely to churn and their probability of churn and thirdly, to develop a retention strategy that reduced churn and maximized the revenue of the company. The final chosen model was the logistic regression model with a high churn prediction accuracy of 75.3%. The top five significant variables for driving churn was "FiberOptic", "MonthtoMonthContract", "DSL", "OneYearContract" and "StreamingMovies". By implementing the retention strategy, the business was able to reduce the churn rate by 40% and more than double their overall profit. By spending $88,000, the retention strategy was able to retain a revenue of $893,908.50, which is ten times more than the amount of money spent on the retention strategy.**

***Index Terms*—Machine learning algorithms, churn, logistic regression, retention strategy, factors, churn reduction, revenue maximization.**

I. Introduction

Churn prediction and management is a rising concern for leading businesses, but it is particularly prominent and critical in the highly competitive telecommunications industry. Reducing customer churn using churn analysis is of utmost importance to telecommunication companies in an ever-competitive and integrated market [1]. Commonly termed as customer churn in the telecommunication industry, it refers to the migration of subscribers from one provider to another. Customer churn can be divided into two broad categories, voluntary and non-voluntary churn. Non-voluntary churn refers to subscribers whose contracts are terminated by the company while voluntary churn can be subdivided into deliberate and incidental churn. In this paper, we will be focusing on reducing the churn rate for deliberate churn subscribers, who choose to migrate their services to other competing companies in search of better rates and services. A two-pronged approach of retaining potential churners and building loyalty within the existing subscribers is considered, since the cost required for acquiring new customers always outweighs the cost used for retaining existing customers to extend their service with the company [2]. Therefore, it is crucial for telecommunication companies to accurately identify customers who intend to churn, and to roll out attractive marketing programmes that cater to the needs of these potential churners, henceforth preventing them from terminating their services with the company. Therefore, fully understanding of the contributing factors that cause customers to churn is critical in effectively managing customer churn. In this study, logical and systematic steps will be taken to achieve the objective of maximizing revenue by reducing churn with the use of machine learning strategies.

In a nutshell, the aim of this paper, is to develop a churn reduction system, while retaining existing valuable customers and maximizing the revenue of the company. This paper is structured into 6 sections, while Section 1 is the introduction, Section 2 gives a brief literature review, Section 3 a brief overview of the analytical methods used, Section 4 the analysis results, Section 5 the retention strategy and its impact, after which Section 6 presents the conclusion.

Manuscript received March 1, 2019; revised October 24, 2019.

Carol Anne Hargreaves is with the Department of Statistics and Applied Probability, National University of Singapore, Singapore (e-mail: stacah@nus.edu.sg).

II. Literature Review

Customer churn identification is a classification problem with two levels, "Churn" or "Do not Churn". However, due to the large dimensionality and possible unbalanced class proportion, it requires a large amount of time and effort to train various models before finding the ideal classifier. A common technique used by [3] is to choose a set of machine learning algorithms – decision tree, logistic regression, neural network and to compare the predictive power of all algorithms before deciding on the best algorithm. While research suggests that the neural network may be more stable and does outperform the decision tree and logistic regression, it does not uncover the pattern in an easily understandable form and is seen as a black box. To achieve a good fit between the model complexity and accuracy, [4] employed the logistic regression and decision tree while [5], used the logistic regression and concluded that customer dissatisfaction, switching cost, demographic and service usage variables affect customer churn.

Other than studying the various types of machine learning models, [6] proposed call behavior-based churn prediction techniques which make use of subscriber call details and contractual information to identify potential churners. [7] introduced a new approach to design targeted retention strategies that maximized the profits of the investment for these strategies. The authors proposed constructing the gain loss matrix which accounts for the gain of retaining targeted customers and the loss/cost incurred.

doi: 10.18178/ijfcc.2019.8.4.550

All of the authors above have successfully demonstrated the different machine learning techniques for accurate churn prediction. It is however, worth noting that a successful churn reduction system should also consider the type of data available and the metric used in designing the target retention strategies. This study focuses on using contractual details to build a logistic regression, decision tree and random forest model and proposes a new metric which considers the retention cost involved in the implementation of the targeted retention strategies.

III. Methods Used

A. Data Cleaning & Data Exploration

The training and test data sets [8] contain 2409 and 1350 observations, respectively. In each data set, there are a total of 21 variables: 3 numerical variables, 17 categorical variables, including the response variable, Churn, and 1 string variable CustomerID. There were a few missing values that made up 0.21% and 0.15% of the training and test data sets respectively, because these missing values were associated with non-churners, these observations were removed. Hence, the training and test data sets reduced to 2404 and 1348 observations respectively.

Bar charts were used to explore the relationships between categorical variables and the churn behavior of customers. The 16 categorical variables in the training data set were plotted against Churn, where Non-churners are indicated by 0 and Churners by 1. As an example, 70.2% of Churners use Fiber Optic and in Fig. 1, we can observe that 65.5% of customers who have Fiber Optic Churn, versus Non-churners, 34.5%. Further, 67.4% of customers who have Month-to-Month contract Churn versus Non-churners 32.6%.

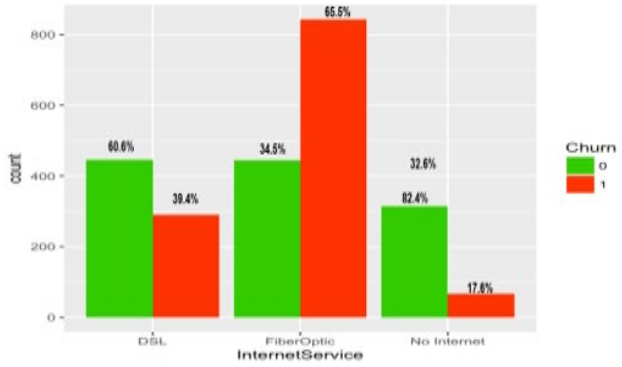

Fig. 1. Churners vs. Non-churners for Fiber Optic.

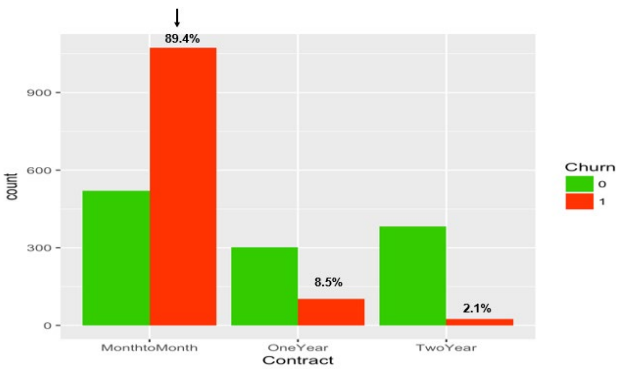

Fig. 2. Churners for type of contract.

89.4% of Churners who subscribe to Month-to-Month contracts compared to 8.5% of Churners who have 1-year contracts and 2.1% of Churners who have a 2-year contract (See Fig. 2). This strongly indicates to us during the data exploration stage that Fiber Optic and Month-to-Month Contracts are likely to be the main drivers of churn and could possibly imply inadequacy in the services provided by the company for Fiber Optic and Month-to-Month contract.

Box plots are plotted for numerical variables ‘Tenure’ and ‘Monthly Charges’.

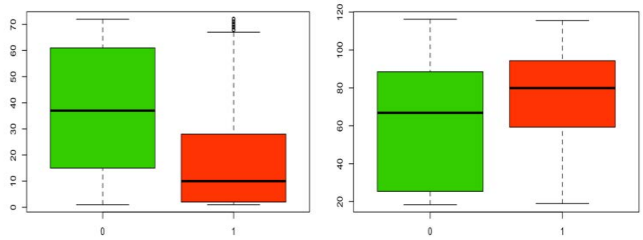

Fig. 3. Tenure and monthly charges as possible factors.

In Fig. 3, the box plot diagrams for the distributions of Non-churners (green) and the Churners (red) and displayed side-by-side for easy comparison. As a brief observation, differences between the median values of Non-churners and Churners for both the numerical variables can be clearly noticed, suggesting that ‘Tenure’ and ‘Monthly Charges’ are significant in influencing Churn. Moreover, even though Churners have shorter tenure, their ‘Monthly Charges’ are higher than Non-Churners. This indicates that Churners have larger purchasing power than Non-Churners and could possibly be high revenue generating customers in the long run. Hence, failure to retain such customers would result in a huge loss in revenue for the company.

B. The Machine Learning Technique: Logistic Regression

The Logistic Regression Model was selected for use in this study because it is the most basic and robust classification algorithm. [4] used two machine learning algorithms; decision trees and the logistic regression to contruct a churn prediction model. The test result graded the logistic regression model ahead of decision trees. Further, the Logistic Regression model was chosen as it requires little running time compared to other complicated machine learning algorithms and its output is also easy to interpret.

As the dependent variable of the Logistic Regression Model is binary, the dependent variable selected was “Churn”. was used to classify the customers as Churners or Non-Churners.

The main assumption for building a Logistic Regression Model is that the independent variables do not have significant multi-collinearity [9]. “Total Charges was highly correlated with “Monthly Charges”, hence was excluded from the final dataset. A total of 20 input variables were used for the Logistic Regression Model and the model was carried out using the backward stepwise regression method. Variables with a p-value greater than 0.05 was deemed as insignificant to Churn and was removed from the model. The process stopped when the model was left with variables with p-values less than 0.05. R programming language was used for the analysis.

IV. Results & Findings

A. Evaluation of the Accuracy of the Logistic Regression Model

Insights about the input variables and their effect on Churn was gained by understanding the signs and size of the coefficients of the Logistic Regression model and the resulting odds ratio. The size of the coefficient indicates the main drivers of Churn. The larger the coefficient, the higher the significance of the variable to Churn. A coefficient with a negative sign indicated that the predictor variable and Churn had an inverse relationship with each other. A positive coefficient indicated that the predictor variable had a positive relationship with Churn. In order to reduce Churn, variables with an inverse relationship with Churn should be encouraged while variables with a positive relationship with Churn should be improved upon. From Fig. 4 below, we can see that the variables "Fiber Optic" (2.79), "Month-to-Month-Contract" (1.54) and DSL (1.50) have the highest positive relationship with Churn, confirming our initial data exploration findings that "Fiber Optic" and "Month-to-Month-Contract" are likely drivers of Churn. So, to reduce Churn, the telecommunication company needs to investigate their fiber optic service as it is likely that the speed of the fiber optic is not as fast as promised. For "Month-to-Month-Contract", the telecommunication company should consider moving customers from "Month-to-Month-Contracts" to 2-year or 1-year contracts as customers with these types of contracts are less likely to churn.

An added advantage of using the Logistic Regression is that it calculates the logarithmic odds of churning. The odds ratio is calculated to study the effect of each variable in affecting the odds of churning [10]. The odds ratio for "Fiber Optic", "Month-to-Month-Contract" and "DSL" were 16.4, 4.7 and 4.5 respectively. This means that holding all other variables constant, customers with "Fiber Optic" are 16.4 times more likely to Churn as to customers without "Fiber Optic". Similarly, for customers using "Month-to-Month-Contract, they are 4.7 times more likely to churn than the customers with other contract types. Lastly, customers with "DSL" are 4.5 times more likely to churn as compared to a customer without "DSL".

```
Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = churn.data6)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2357  -0.7625  -0.1396   0.7429   2.8897

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.776134   0.337026  -5.270 1.36e-07 ***
Dependents           -0.250857   0.122679  -2.045 0.040872 *
tenure               -0.030917   0.003455  -8.947  < 2e-16 ***
MultipleLines1        0.445958   0.141072   3.161 0.001571 **
OnlineSecurity1      -0.321548   0.138286  -2.325 0.020059 *
TechSupport1         -0.473406   0.141140  -3.354 0.000796 ***
StreamingTV1          0.391865   0.155202   2.525 0.011574 *
StreamingMovies1      0.553511   0.154943   3.572 0.000354 ***
PaperlessBilling1     0.273908   0.116542   2.350 0.018758 *
MonthlyCharges       -0.019820   0.008626  -2.298 0.021578 *
FiberOptic1           2.791652   0.494763   5.642 1.68e-08 ***
DSL1                  1.503262   0.265092   5.671 1.42e-08 ***
MonthtoMonthContract1 1.542468   0.256811   6.006 1.90e-09 ***
OneYearContract1      0.838533   0.251240   3.338 0.000845 ***
ElectronicCheck1      0.517351   0.110216   4.694 2.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3332.6  on 2403  degrees of freedom
Residual deviance: 2287.9  on 2389  degrees of freedom
AIC: 2317.9
```

Fig. 4. Logistic regression coefficient summary.

Other than understanding how different variables would affect Churn, there is a need to evaluate the training model to find out its predictive power. This is done by measuring the accuracy of the model using the training dataset. Accuracy measures such as the confusion matrix, recall, precision, specificity, overall accuracy, Receiver Operating Characteristic (ROC) curve, and the Area under the Curve (AUC) were calculated and are shown below. These accuracy measures are used to evaluate the accuracy of all models in this paper.

The confusion matrix measured the ability of the model to classify observations into the correct class. Table I below shows the confusion matrix for the Logistic Regression training model.

From the confusion matrix, 963 and 881 observations are classified correctly as churners and non-churners respectively. However, 323 non-churners are classified as churners while 237 churners are classified as non-churners by the model. TABLE II below shows the "Recall" and "Overall" Accuracy and the "Area under the Curve (AUC)".

TABLE I: Confusion Matrix - Train Model

	Non-Churners(P)	Churners(P)
Non-Churners (A)	881	237
Churners (A)	323	963

TABLE II: AUC, Recall & Overall Accuracy

Recall Accuracy	80.3%
Overall Accuracy	76.7%
AUC	0.767

Overall, the model was able to predict 76.7% of the observations accurately and predicted 80.3% of churners accurately. The Logistic Regression training model fits well with the training dataset as the AUC value is greater than 70%. This indicates that the Logistic Regression model has good predictive power in predicting Churn and therefore, the Trained Model should be fitted using the validation dataset.

The results for the Logistic Regression model on the validation dataset were very good. Table III below shows the confusion matrix for the validation dataset.

The Trained Logistic Regression Model fitted well on the test dataset too. From Table III below, 504 and 511 customers were classified correctly as churners and non-churners respectively. However, 168 non-churners were classified as churners while 165 churners were classified as non- churners by the model. Overall, the model was able to predict 75.3% of the observations accurately and predicted 75.3 % of churners accurately. From the ROC Curve in Fig. 5 below, the Logistic Regression Test model showed good predictive power as it had a high AUC value of 83.9%, which is greater than 70%. This means that the Logistic Regression Model was able to predict churn accurately on unseen data.

TABLE III: Confusion Matrix -Validation Dataset

	Non-Churners(P)	Churners(P)
Non-Churners (A)	511	165
Churners (A)	168	504

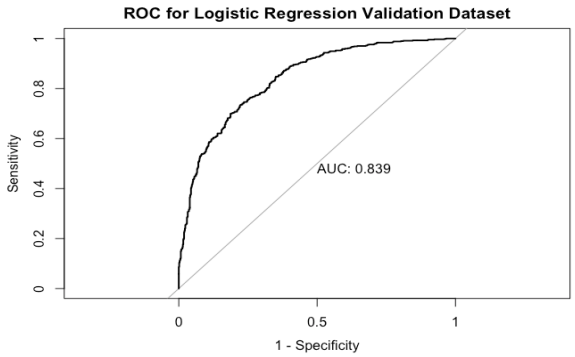

Fig. 5. Receiver operating characteristic curve.

B. The Business Retention Strategy

Using the predictions from the logistic regression model, the business strategy was developed to reduce the churn rate and to maximize the revenue. In order to carry out the business retention strategy, a retention budget was allocated and the predicted churners were split into top 20% revenue customers and middle 60% revenue customers. Using the business formula by [11], the retention budget considered the cost of the retention offer ($120), the cost of contacting the customer ($1) and the administrative cost ($10). The total cost of retention for one predicted churner amounted to $131. Using the predictions from the logistic regression model, the estimated retention cost needed to retain 672 predicted churners was $88,000 rounded to the nearest thousand. As it is not feasible to retain all predicted churners, it was necessary to decide which predicted churners to include and which to exclude from the retention packages. The predicted churners were first ranked in decreasing order of their "TotalCharges" and split into their revenue groups, where the top 20% formed the high revenue group, middle 60% the medium revenue group, and the bottom 20% formed the low revenue group. The revenue contribution from the top 20%, middle 60% and bottom 20% were 59.3%, 39.7% and 1.0% respectively. As the bottom 20% only contributed 1% of the revenue, they were not considered in the retention strategy. 60% of the budget which amounted to $52,800 was allocated to the top 20% group and the remaining 40% which amounted to $35,200 was allocated to the middle 60% group.

The top 20% group consisted of a total of 135 predicted churners, where 90.4% used a "MonthtoMonthContract", and 9.6% used a "OneYearContract". This observation suggested that customers with a "TwoYearContract" may be less likely to churn. Hence, it was decided to get the predicted churners in the top 20% group to sign a "TwoYearContract" as this would allow the company to retain them and earn revenue from them for the next two years. Secondly, 99.3% of the top 20% were using "FiberOptic". Hence, a retention strategy that was related to "FiberOptic" was considered a relevant retention strategy, provided the "FiberOptic" service was investigated and the company could guarantee high quality from the "FiberOptic". The resulting retention offer for the top 20% who were on a "MonthtoMonthContract", was "Free additional 1G of Fiber Optic for the first six months plus 50% off for the additional 1 Gbps for the remaining eighteen months when the predicted churners signed up for a "TwoYearContract". Following the assumption that all targeted predicted churners will accept the retention offer, that is, all 135 predicted churners would be retained, the company was able to retain a large sum of revenue which amounted to $535, 759.20 while spending only $49,455. This retention strategy was able to retain revenue that was worth ten times more, showing its effectiveness.

The middle 60% group consisted of 403 predicted churners, giving the company a total revenue of $358,139.30. As 99.8% of the predicted churners in this group had internet services, we selected a revenue strategy related to internet services. Further, 97.8% of the predicted churners in the middle 60% group had a "MonthtoMonthContract" while 2.2% had a "OneYearContract". It was also noted that "StreamingMovies" was the fifth most significant variable in our Churn Logistic Regression Model. So, we decided to use "StreamingMovies" in our retention offer, to push the predicted churners to a "TwoYearContract". Our retention offer to the predicted churners in the middle 60% who had a "MonthtoMonthContract", we offered them a $90 voucher to stream movies on their television. Similarly, for predicted churners in the middle 60% who had a "OneYearContract", we offered them a $18 voucher to stream movies on their television. The total retention cost for the predicted churners in the middle 60% group was $35,622, and assuming that the predicted churners will accept the retention offer, this busiess strategy would retain all 403 predicted churners and will retain a large sum of revenue which amounted to $358,139.30. This retention strategy was able to reatin a revenue that is worth ten times more than the cost of the retention strategy.

V. Conclusion

Churn is an important problem for the telecommunication companies. Telecommunication companies should focus on retaining existing customers instead of focusing on acquiring new customers given that the cost involved for customer retention is far lower than the cost to acquire a new customer.

The top five significant variables for driving churn was "FiberOptic", "MonthtoMonthContract", "DSL", "OneYearContract" and "StreamingMovies".

By grouping the predicted churners according to their revenue group and crafting a retention strategy based on the significant variables in the Churn Logistic Regression Model, we were able to reduce churn effectively and at the same time maximized the revenue of the company. By spending $88,000, the retention strategy was able to retain a revenue of $893,908.50, which is ten times more than the amount of money spent on the retention strategy.

In conclusion, assuming all churners accepted the retention plan offered to them, the crafted retention strategies were able to reduce the churn rate by five times, from 49.9% in the test dataset to 9.9%, and retain more than ten times the costs that were invested for the retention strategies.

Acknowledgment

I would like to acknowledge and thank, Zhao Wenxin,

Bernice Lim Fang Yuan, Ko Wing Lam, Looi Boon Hon, Chua Wei Kian and Chen ZhiYi Florence for their contributions to this study. I also would like to thank IBM for making the "Teleco Customer Churn dataset" available online for use.

REFERENCES

[1] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications*, vol. 36, no. 3, part 1, pp. 5445-5449, 2008.

[2] J. H. Ahn, S. P. Han, and Y. S. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," *Telecommunications Policy*, vol. 30, no. 10-11, pp. 552-568, 2009.

[3] H. Hwang, T. Jung, and E. Suh, "An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry," *Expert Systems with Applications*, vol. 26, no. 2, pp. 181-188, 2004.

[4] G. Nie, W. Rowe, L. Zhang, Y. Tan, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15273-15285, 2011.

[5] A. Keramati and S. M. Ardabili, "Churn analysis for an Iranian mobile operator," *Telecommunications Policy*, vol. 35, no. 4, pp. 344-356, 2011.

[6] C. Wei and I. Chiu, "Turning telecommunications call details to churn prediction: A data mining approach," *Expert Systems with Applications*, vol. 23, no. 2, pp. 103-112, 2002.

[7] A. Lemmens and S. Gupta, "Managing churn to maximize profits," *SSRN Electronic Journal*, 2017.

[8] IBM Community. (2019). *A Collaborative Community Space for IBM Users*. [Online]. Available: https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set

[9] S. W. Menard and I. NetLibrary, *Applied Logistic Regression Analysis*, Thousand Oaks, Calif: Sage Publications, 2002.

[10] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd edition. Hoboken, N.J: Wiley, 2013.

[11] A. C. Bahnsen, D. Aouada, and B. Ottersten, "A novel cost-sensitive framework for customer churn predictive modeling," *Decision Analytics*, vol. 2, no. 5, 2015.

Carol Anne Hargreaves received PhD in Statistics, University of South Africa, Pretoria, South Africa in 2002. She has over 30 years analytics experience, with leading roles in the pharmaceutical, healthcare, education & FMCG industries. Prof Hargreaves has worked with a variety of leading companies to make businesses more intelligent and profitable.