PREDICTING CUSTOMER CHURN IN MOBILE TELECOMMUNICATION SECTOR USING DATA MINING ANALYSIS APPROACH WITH R

BY

VICTORIA OLUWAKEMI, OKESIPE
MATRIC NO.: 200888

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF IBADAN, IBADAN,
OYO STATE, NIGERIA

March, 2020

OUTLINE

1 INTRODUCTION

2 AIM AND OBJECTIVES

3 LITERATURE REVIEW

4 DATA AND METHODOLOGY
- Data
- Methodology

5 RESULTS AND DISCUSSION

6 CONCLUSIONS AND RECOMMENDATION

7 REFERENCES

Customer Churn

- Customer Churn is a big problem for mobile network operators. Mobile Number Portability (MNP) has in addition magnified this problem for mobile telecommunication companies, although it has rightly empowered the subscribers of these services to switch networks without losing their cellular phone numbers.

Telecom Sector

- The main goal of many telecommunication firms has consequently shifted from building an oversize customer base into keeping customers in-house.In this study, we seek to predict and model customer churn in the telecommunication sector, so as to maximise customer lifetime value for the company.

AIM AND OBJECTIVES

Aim

The main aim of this work is to formulate a churn prediction model which assists telecommunication operators to predict customers who are most likely to switch to a competitor.

Objectives

Our specific objectives are to:

(i) determine the likelihood of an existing customer in a telecommunication company to churn or not; and

(ii) figure out the distinctive attributes of customers who churn.

- In this study, we rely on the operational definition of churn as given by **Alberts (2006)**, churn occurs when a customer has permanently stopped using/recharging his SIM-card as early as possible. It is also the rate of movement of subscriber within a system (GSM operators).
- Thus, churn situation is when a number of customers switch/port their preferences among network providers. Literature reveals the following three types of customer churns **(Lazarov et al., 2007)**:

1. **Active churner (voluntary):** these are customers who want to quit the contract and move to the next provider.
2. **Passive churner (non-voluntary):** when a company discontinues service to a customer.
3. **Rotational churner (silent):** those customers who discontinue the contract without the prior knowledge of both parties (customer and company), where each party (e.g. customer or company) may suddenly terminate the contract without any notification.

LITERATURE REVIEW

- According to **Hamelin, et al., (2010)**, a number of researchers and academics have studied factors that may cause churn. However, it is important to investigate why customers are leaving before selecting the appropriate churn-reduction mechanisms.
- According to **Fox, et al., (2002)**, a better price is the main factor but not the ”prevailing reason”. In fact, as it has been observed, offering a lower price does not necessarily mean higher customer loyalty. They also reveal that customers churn to the competitor who best matches their needs in terms of service features, technology, and service quality.

LITERATURE REVIEW

- The probability that a subscriber will change the actual carrier depends on the satisfaction level reached in addition to factors relating to service attributes, which include call quality, tariff level, handsets, brand image, income and other relevant factors.
- In addition to these service attributes, **Kumar (2007)** has indicated that the transparency level of any company is highly associated with customer satisfaction. When a specific firm adopts transparent marketing strategies, communicates transparent tariffs and makes information available, it increases the customer confidence and satisfaction.

LITERATURE REVIEW

- Burez, et al., (2007) pointed out two types of targeted approaches to managing customer churn; reactive and proactive.
- When a company adopts a reactive approach, it waits until customers ask the company to cancel their service relationship. In this situation, the company will offer the customer an incentive to stay.
- On the other hand, when a company adopts a proactive approach, it tries to identify customers who are likely to churn before they do so. The company then provides special programmes or incentives for these customers to keep the customers from churning. Targeted proactive programmes have potential advantages of having lower incentive costs.

LITERATURE REVIEW (CONT'D)

- However, these systems may be very wasteful if churn predictions are inaccurate because companies are wasting incentives (money) on customers who will not churn. Therefore, it is important to build a customer-churn prediction model as accurately as possible (Burez, et al., 2007; Van den Poel, et al., 2004).
- Sulaimon, et al., (2016) stated that the application of logistic regression to the study of customer churn and retention decision in the Nigerian telecommunication industry falls into proactive methods.

- This makes for a better understanding of the needs of subscribers to predict customer churn and retention in the industry in order to enhance better marketing strategies and provide research driven policy guide for the operators in their quest for the optimal development of the industry.
- Consequently, to retain customers, telecommunication companies however must understand subscribers' and market services, which motivate them to remain with the current service provider and not switch (port) to other competitors who render similar or same service.

Data set (Telcom Customer Churn)

- The Law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome. This implies that random processes are supposed to compensate for whatever happened in the past. This is called the gambler's fallacy or the law of averages.
- In the telecommunication dataset (international) for this study, each row represents a customer, while each column contains customers attributes. The raw data contains 7043 rows (customers) and 21 columns (features). The Churn column is our target, while other columns are used as features to the model taking note of the following details:

Telecommunications Reference Data Set

- Customers who left within the last month - the column is called ”Churn”.
- Services that each customer has signed up for - phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information - how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic info about customers - gender, age range, and if they have partners and dependents.

Use of **R**

- We seek to formulate a model for churn prediction for telecommunication companies utilising two data mining approaches namely: logistic regression and decision trees with the help of **R**. **R** is a rich environment for statistical and mathematical computing and has many capabilities for exploring data, especially when it has to do with prediction.

Practically, data are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. Noisy: containing errors or outliers. Inconsistent: containing discrepancies in codes or names.

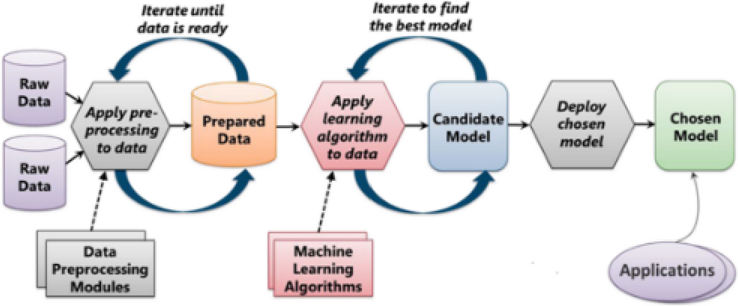

Figure: *The Machine Learning Process*

Phases of Data Preparation

A. **Data Preprocessing:**Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. The steps include:

Step 1: Import the libraries
Step 2: Import the data set
Step 3: Check out the missing values
Step 4: See the Categorical Values
Step 5: Splitting the data set into Training and Test Set
Step 6: Feature Scaling

B. **Data Wrangling:** This includes the preparation of data during interactive data analysis and model building.This step iteratively changes the shape of a dataset until it works well for finding insights or building a good analytic model.

Below is a typical analytical pipeline for building an analytic model:

1. Data Access
2. Data Preprocessing
3. Exploratory Data Analysis (EDA)
4. Model Building
5. Model Validation
6. Model Execution
7. Deployment

Logistic Regression

- **Logistic regression** is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. In this study, the outcome is measured with a dichotomous variable (in which there are only two possible outcomes).
- Here, **the dependent variable,** which is binary or dichotomous, is coded as 1(TRUE, success, Churn) or 0 (FALSE, failure, not-churn)

Logistic Regression

- The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = w_0 + w_1 x_1 + \cdots + w_n x_n \qquad \text{Log(odds)} \tag{1}$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$\text{odds} = \frac{p}{1-p} = \frac{\text{probability of presence of characteristic}}{\text{probability of absence of characteristic}} \tag{2}$$

and

$$logit(p) = log\left(\frac{p}{1-p}\right) \tag{3}$$

Decision Trees

- Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables (James, 2018).
- A decision tree is a tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

Decision Tree Process

- Start by dividing the predictor space into j distinct and non-overlapping regions.
- For every observation that falls in a region, the mean of the response value is predicted in that region.
- Each region is split to minimize the residual sum of squares. To do so, it takes a top-down greedy approach known as recursive binary splitting. Because all observations are in a single region before the first split.
- This approach deploys a greedy method. Because the best split occurs at a particular step, rather than looking ahead and making a split that will result in a better prediction in a future step.

Mathematically, we define the pair of half-planes as:

$$R_1(j,s) = \{X | X_j < s\} \text{ and } R_2(j,s) = \{X | X_j \geq s\} \tag{4}$$

and we seek j and s to minimize:

$$\sum_{i:x_i \in R_1(j,s)} (\hat{y}_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (\hat{y}_i - \hat{y}_{R_2})^2 \tag{5}$$

Logistic Regression Model for Customer Churn

At .05 significance level, derived Logistic Regression Model for Customer Churn is given as:

$$logit(p) = -0.759x_1 - 1.621x_2 + 0.343x_3 + 1.693x_4; \qquad \text{log(odds)} \quad (6)$$

where;

$$\begin{aligned} logit(p) &= log\left(\frac{p}{1-p}\right) \\ x_1 &= \text{ContractOne year} \\ x_2 &= \text{ContractTwo year} \\ x_3 &= \text{PaperlessBillingYes} \\ x_4 &= \text{tenure\_group0-12 Months} \end{aligned}$$

Decision Tree visualization

- For illustration purpose, we are going to use only three variables for plotting Decision Trees. They are: ”Contract”, ”tenure_group” and ”PaperlessBilling”.

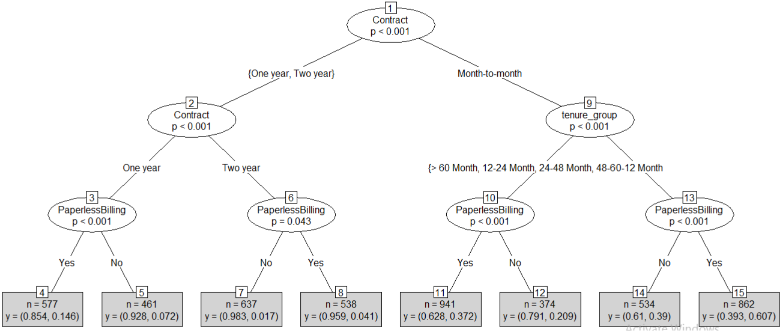

Figure: *Customer Churn Decision Tree*

Discussion

Deep learning neural networks can also be used to predict customer churn as it can capture non-linearity in models which will lead to better model formulations

CONCLUSIONS

- From the results obtained in the previous chapter, we can see that Logistic Regression, and Decision Tree can be used for customer churn analysis.
- Throughout the analysis, we observed that: Features such as Tenure group, Contract, Paperless Billing, Monthly Charges and Internet Service appear to be variables of great importance in customer churn. There does not seem to be a relationship between gender and churn.
- However, Customers in a month-to-month contract, with Paperless Billing and are within 12 months tenure, are more likely to churn. On the other hand, customers with one or two year contract, with longer than 12 months tenure, that are not using Paperless Billing, are less likely to churn.

RECOMMENDATION

- There is a need for Telecommunication service providers to focus on attributes that have been identified as significant to the telecommunication subscriber porting decision. This will ensure that subscribers at risk of porting are given special attention. In order to enhance customer satisfaction and unbroken patronage.
- In addition, the network provider needs to understand the dynamics of subscribers behaviour in the market.
- Therefore, further research in this area can be focused on other areas of Customer Relationship Management (CRM), like customer attraction and customer development.

REFERENCES

Alberts, L. J. (2006). Churn prediction in the mobile telecommunications industry: An application of Survival Analysis in Data Mining, Unpublished Master Thesis, Maastricht University.

Burez, J., & Van den Poel, D. (2007). CRM at a Pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications, 32(2), 277288 `http://dx.doi.org/10.1016/j.eswa.2005.11.037`

Fox, M. & Poje, M. (2002). Churn: are rural telecoms immune? Rural Telecommunications, 21, 1-5.

Hamelin, N., Nassali, A & Harcar, T. (2010). Determining churn drivers in Moroccan telecom sector. Journal of International Business Disciplines, 4(2), 16-34.

James Le (2018). An article on Decision Trees in R. DataCamp COMMUNITY. Retrieved from: `https://bit.ly/2DcNG9B`

Kumar, N. (2007). Gaining momentum towards customer centricity. Retrieved from `http://www.gtnews.com/article/6953.cfm.`

Lazarov, V., Capota, M. (2007). Churn prediction, Bus. Anal. Course TUM Comput. Sci. (2007).

Sulaimon Olanrewaju Adebiyi, Emmanuel Olateju Oyatoye, Bilqis Bolanle Amole (2019). Relevant Drivers for Customers' Churn and Retention Decision in the Nigerian Mobile Telecommunication Industry. Journal of Competitiveness. Vol. 8, Issue 3, pp. 52-67, September 2016. ISSN 1804-171X (Print), ISSN 1804-1728 (Online), DOI:10.7441/joc.2016.03.04

Thank you!