Customer Churn Prediction Using a Two Step Approach – K-Means and Decision Trees

submitted by

William Amonoo
Gokul Krishna Krishnan
Spoorthi Kondagadapu

5/11/2017

Contents

Abstract

Customer retention is one of the most critical challenges that a maturing and highly competitive telecommunication industry faces. In this paper we use a customer billing and transaction data set to investigate the ability of businesses to predict which customers are likely to churn. We propose a relatively simple two step approach of using K-means clustering analysis and then decision trees to predict customer churn. The K-means clustering algorithm helps us to identify unknown groups of customer profiles based on specific variables within our data set. Consequently customers can be segmented into a number of groups which gives an in-depth knowledge of the customer base. The decision tree is then applied on the groups to predict which segment of customers is likely to churn. Our approach is applied on a data set from the telecommunications industry and our results showed a true positive rate of 73.2%, a false positive rate of 6.2% and an overall prediction accuracy of 83.5%. In other words our model predicts correctly that a customer will churn 73% of the time. 6.5% of the time, it will falsely predict a customer as churning when they have actually not decided to leave. In a business context, a company will want to reduce how much it spends on people who have not decided to churn and rather focus resources on customers likely to churn. Thus any prediction of churning has to yield small false positive values.

1 Introduction

Customer churn refers to when a customer terminates his/her relationship with a company. Churning or otherwise retaining customers is a major concern for companies who have customers that can switch easily to other competitors. Examples include insurance companies, telecommunication companies and credit card issuers. This becomes especially true in liberalized economies where there can be fierce competition.

The axiom among businesses is that it is more expensive to get a new customer than to keep an existing one. Depending on the industry one finds themselves in, acquiring a new customer can range anywhere from 5 to 25 times more expensive than retaining an existing one (Gallo, 2014). Thus retention of customers is one of the most important aspects for any business. Loyal relationships translate into cost savings in businesses in a number of ways. As customers keep buying from a company over time, the operating costs to serve them goes down, they may refer others to the company and will also pay a premium to continue to do business with an organization they are familiar with rather than switch to a competitor (Almana, Aksoy, & Alzahrani, 2014). In the financial services sector, for example, a 5% increase in customer retention leads to an increase of 25% in profit (Reichheld, n.d.). One of the key metrics in understanding whether a business is retaining its customers is customer churn rate. Understanding and preventing why customers churn therefore represents a huge additional potential revenue source for every business.

“Consequently churn management has emerged as a crucial competitive weapon and a foundation for an entire range of customer-focused marketing efforts.” (Richeldi & Perrucci, 2002, p.3). It is an established fact that companies provide specialized offers to customers who have signaled an intention to churn. However the save rate of such measures rarely exceed 30 – 40 % implying that most of the time, attempts to prevent customer churn fail. These efforts fail because most customers are identified and offered alternative value propositions only after they have already made up their minds about the product or service, and may have possibly compared alternative offers and are ready to abandon service. Additionally customers are becoming smart and share information on how to maximize the benefits of a bluffed cancel call. The result has been that it has become more expensive to extend save offers to customers bluffing about their intention to churn and organizations have declining ability to save people who have really decided to churn (PWC, 2011).

Customers generally only signal an intention to churn at the point when they have initiated action towards such as end. For example, a telecommunication company may not know a customer wants to switch until they signal their intention by calling to cancel their account. This presents the challenge of being able to predict what a particular customer will do at a future date; in other words, businesses need to answer when a particular customer will churn. The second dimension to this problem is to understand the reasons leading to the churn. Thus prediction and understanding may represent the most important aspects of managing customer churn.

In the age of big data sets, organizations are now using data mining and machine learning methodologies to perform big data analytics in particular churn detection on their customers as an effective approach to the problem. This addresses the first dimension of churn analysis, the prediction

component. This is the focus of our paper: comparing the performance of two algorithms, specifically K-means clustering and decision trees in the prediction of customer churn.

It must be stated that addressing the first dimension of the problem may not be enough to ensure effective churn management. This is because the predictions do not give the reasons why a customer wants to leave to a competitor and also how it can be prevented. For an effective churn strategy, the root causes need to be known. Only then can at-risk customers be targeted with right offers before they make up their minds to churn. This will usually involve an analysis of the business processes and systems that manage a customer's relationship from sale to service, to cancellation- the complete customer experience pipeline. A detailed root cause analysis can produce the so-called leaky pipe report that pinpoints where along the infrastructure certain information is getting lost or mishandled. Tying each leak down to a specific churn event could make it easy to do a cost-benefit analysis that would fix the leak (Springer et al., 2014).

Another tool that give insights into the root causes of customer churn is the structured call monitoring of cancel / save, sales, customer care/billing and technical support calls. Call monitoring is a powerful tool that can generate knowledge about the key drivers of churn, their nature and frequency. It also leads to statistically sound insight into how problem areas may be contributing to the total churn. Consequently separate statistical models will be required to understand each unique churn driver identified and its treatment. E.g. models can be designed to scoring specific customers subsets (e.g. customers who are about to come out of a contract) and their likelihood to churn due to some particular churn driver (e.g. contract expiry) and their response to treatment (e.g. a campaign targeted at contract renewal) (PWC 2011).

1.1 Objective

The main objective of this paper is to demonstrate the performance of two relatively simple algorithms namely K-means and decision trees in predicting customer churn in the telecommunication industry. As noted earlier, the ability to make accurate predictions about who is likely to churn is an important dimension in a holistic strategy to address customer churn.

2 Literature Review

The telecommunication industry is an interesting case study in the management of customer churn. This includes cable or satellite television providers, both wireless and landline telephone service providers and internet providers. With the liberalization of markets, many countries now allow private carriers to compete with state-run telecommunications companies where they still exist. Also there has been an enormous amount of activity and breakthroughs in computer technologies leading to major telecommunications advances globally. According to Deloitte's 2017 Telecommunications Industry Outlook, the telecom sector continues to be a crucial force for growth, innovation and disruption. However with maturing markets, the provision of telecommunication services has become highly competitive (Miller, 2015). A manifestation of such competition is the problem of churn. Springer et Al., (2014), indicate that churn levels among wireline providers tend to hover around 2 – 2.5% per month translating into an estimated mean of 1.3 million people switching or leaving a service provider with 5 million customers and consequently losing $2 billion worth of revenue each year.

Ahn et al.(2006), discuss customer churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry using customer transaction and billing data. Using a logistic regression model they analyze partial and total defection using the status of a customer (e.g. a customer's status could change from active or using the service on a regular basis to non-use or deciding not to use it temporarily without having churned yet or suspended (being suspended by the service provider) all the way to a total defection. They concluded that call quality-related factors influence customer churn. However customers who were participating in membership card programs were actually more likely to churn. They also discovered that heavy users were also more likely to churn.

According to Li et al. (2003), predicting customer attrition is a challenging work. It requires large amounts of data as well as the specification of the right statistical model. Customer attrition may not be caused by a single factor but rather a multiple of reasons. These could range from a customer no longer needing a service, migrating to another type of service or switching to another competitor for the provision of the same service. More often than not, an organization does not know why a customer leaves and thus predicting attrition by any single cause is not appropriate. Customer churn is also not an additive sum of the attrition of each cause that may have been identified.

As part of their study, they evaluated the performance of models namely a multilayer perceptron model (MLP), logistic regression and tree-based methods in the development of predictive models for customer churn. They found that the tree-based method performed the best among the three methods. This could be attributed to the fact that the tree-based method is nonparametric and less sensitive to data distribution. Also it uses the fewest number of predicting variables which may make the model more stable and less affected by missing data problems.

Missing data is a common problem in marketing research. Logistic regression and MLP can only use cases with no missing data. Thus a lot of information could be lost when missing data is discarded. However the tree method works regardless of whether missing data exists or not. This robustness could be an advantage over the other two methods that have been discussed. However decision trees have

the limitation that they do not perform very well for non-linear relationships. However pruning decision trees can help to improve classification accuracy of the tree. In some cases, constructing a dual step multi-algorithm of a neural network model with tree-based models may give the best results in terms of classification accuracy (Almana et al., 2014).

Sharma and Sachdeva (2017) evaluated the performance of several algorithms in the prediction of customer churn. From their results they proposed a hybrid approach which is an integration of two techniques; the random forest and support vector machine algorithms that feature Artificial Bee Colony (ABC). This use of ABC with two best local and global values boosted the predictive capability of their hybrid algorithm technique compared with eight other algorithms.

3 Methodology

3.1 Data

The data set used in this project is a customer billing and transaction data from a of telecommunication company. It consists of 20 variables with 3333 rows. The types of information that are stored include:

- Customer demographics such as location
- Call statistics such as the length of calls made at different times of the day, the number of local and long distance calls
- Billing information, i.e. the amount each customer is paying for local and long distance calls
- The amount of time the customer has spent talking to customer service
- Products and services used by the customer such as how many voicemail plans they are subscribed to.

A complete description of the data set is found in the appendix.

3.2 K-means Clustering

K-means clustering is a type of unsupervised learning and can be described as an iterative clustering algorithm whose aim is to partition a number of observations into k clusters. The algorithm works iteratively to assign each data point to one of the K clusters (groups) based on the features that the data set possesses. The data points are grouped based on similarities in features. Rather than defining groups before looking at the data, the algorithm allows for the discovery of groups that have not been explicitly labeled in the data. Thus this algorithm is very useful in business cases where the goal is to identify unknown groups such as creating customer profiles based on activity monitoring, segmenting customers by purchase history, etc (Dunham, 2003).

K-means clustering results in the following:

- The centroids of the K clusters. These can be used to label the new data.
- A label for the training data set since each point is now assigned to a single cluster.

To use the algorithm, the value of *K* has to be specified. Generally, there is no method for determining the value of *K.* One may have to run the algorithm with a number of different *K* values and then compare the results or cross-validation. The advantages of the K-means clustering are obvious: It is one of the simplest methods to solve clustering issues and also the technique can be used to simplify large data sets.

3.3 Decision Trees

Tree-based methods are suitable for both regression and classification problems. They usually work by partitioning the predictor space into a number of regions using a set of rules. Each observation is predicted as belonging to the most commonly occurring class of training data in the segment where it belongs. This binary splitting is done recursively to grow the tree. When building a tree, either the Gini index or the cross entropy is usually used to evaluate the quality of a particular split. These two methods

are more sensitive to node purity than misclassification error rate. By aggregating many decision trees and using techniques like bagging, boosting and random forests, the predictive power of tree models can be improved substantially (Gareth, Witten, Hastie & Tibshirani, 2013).

The main advantage of decision trees is the fact that they are very intuitive and easy to understand since they are displayed graphically. They are also relatively easy to explain compared to many of the machine learning methods. These make it particularly suitable for non-technical managers.

3.4 Workflow

The workflow for churn prediction follows a framework of pre-processing of the data, K-means clustering and finally using the output from the clustering to make churn predictions through a decision tree. The first step is to read the dataset and pre-process it. The data is composed of two components, customer transactions and billing data. For ease of operations, these separate data sets were combined into a single excel file.

For churn analysis, the excel file is filtered using the first seven columns which hold customer call transactions. We then output an integer flow variable with a given number value. The value can also be controlled from a quick form. Then we start a loop that increases a variable within a user-defined interval by a certain amount. This is very handy for nodes inside a loop that take a continuous parameter. The current value is accessible via the scope variable loop value. At the end of the loop we collect the results from all loop iterations and then perform K-means clustering that assigns a data vector to exactly one cluster. The algorithm terminates when the cluster assignments do not change anymore. The clustering algorithm uses the Euclidean distance on the selected attributes. Then we filter chunks of rows and now we calculate the distance matrix for each row and then we create a list of collection values, a list of rows with the values of the collection in one column and all other columns given from the original row. After that we calculate the following formula: square root (average (distance * number of rows)). Then we take a list of user-defined rules and try to match them to each row in the input table. If a rule matches, its outcome value is added into a new column. The first matching rule in order of definition determines the outcome. After the loops ends we write the clustered data table into a spreadsheet.

In the decision tree analysis phase, output from the first phase was split into training and testing data set. In in our case we split the training and testing dataset in 80 and 20 percent respectively. Then we input the training dataset into the decision tree model with GINI coefficient and used the trained decision tree to predict customer churn. We then we calculated the AUC and plotted the roc curve for the model. Predictions are evaluated based on confusion matrix and ROC.

4 Results

In the appendix section you can see the table with different classes of variables with labeled clusters. To profile each cluster, we use the following equation calculate (cluster center – Average)/(sd/√(cluster-size)) for each variable. The table below is an example for the variable "account length".

Cluster 0	-2.46
Cluster 1	-24.70
Cluster 2	5.63
Cluster 3	24.36
Cluster 4	-0.55
Cluster 5	-0.40
Average	101.06
Standard Deviation	39.82

With those results we are able to categorize the variables into five categories namely, very low, low, medium, high and very high for all the variables. An extract is shown in the table below.

	Account Length	**Vmail**	**Day Calls**	**Evening Calls**	**Night Calls**	**International Calls**	**Customer Service**
Cluster 0	Low	Very Low	Low	Very High	Very Low	Medium	Low
Cluster 1	Very Low	Very Low	High	Medium	Very High	Low	Low
Cluster 2	Medium	Very Low	Very Low	Very Low	Medium	High	Low
Cluster 3	High	Very Low	Very High	Medium	High	Low	Low
Cluster 4	Medium	Low	Medium	Medium	Low	Medium	Very High
Cluster 5	Medium	Very High	Medium	Medium	Medium	Medium	Low

Now we applied this to the churn table we got the following categorized table.

	Churn
Cluster 0	Low
Cluster 1	Medium
Cluster 2	Low
Cluster 3	High
Cluster 4	Very High
Cluster 5	Low

We can compare the cluster churn to the table with variables such as customer who uses very high customer service are likely not to churn. We are also able to cross reference it with all variables to make some meaningful business decisions.

The confusion matrix below was obtained using the decision tree with 162 observations.

Churn	**Predicted: No (0)**	**Predicted: Yes (1)**
Actual: No (0)	TN = 91	FP = 6
Actual: Yes (1)	FN = 26	TP = 71

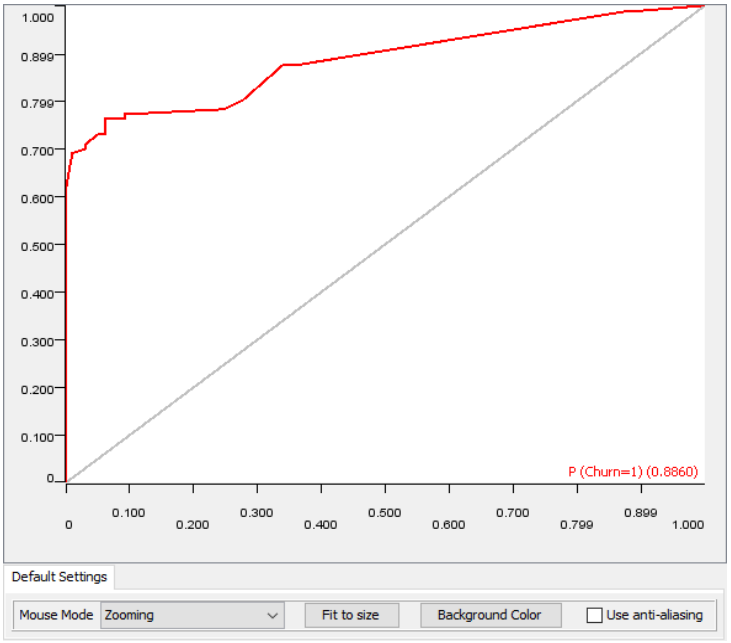
We compute the following metrics from our confusion matrix:

- Accuracy: Overall, how often is the classifier correct?
  (TP+TN)/total = (71+91)/194= 83.5%
- Misclassification / error rate: Overall, how often is it wrong?
  (FP+FN)/total = (6+26)/194 = 16.5%
- True Positive Rate / sensitivity / recall: When it's actually yes, how often does it predict yes?
  TP/actual yes = 71/105 = 73.2%
- False Positive Rate: When it's actually no, how often does it predict yes?
  FP/actual no = 6/97 = 6.2%
- Specificity (Which is equivalent to 1 minus False Positive Rate): When it's actually no, how often does it predict no?
  TN/actual no = 91/97 = 93.8%
- Precision: When it predicts yes, how often is it correct?

TP/predicted yes = 71/77 = 92.2%

- Prevalence: How often does the yes condition actually occur in our sample?
  actual yes/total = 97/162 = 59.9%

The resulting ROC curve with AUC = 0.89 is shown below.



5 Discussion and Conclusion

In this paper we have shown that using the process of K-means clustering and decision tree analysis to predict the customer attrition rate is highly accurate and also it can be a simple use case in real world business as it is easy to explain it to non-technical people with its results being easily interpreted. There are other ways to classify the data such as random forest and ensemble methods which are widely used now but these methods can be hard to explain to the people without any data science skills for them to trust results from such models.

Additionally the unsupervised clustering such as K-means can be used for a wide range of business problems and it can be a starting point in a business to evaluate the performance of such products where metrics or key performance indicators have not yet been developed.

5.1 Further study

This study has only covered the first phase in the management of customer churn, i.e. the prediction phase. As indicated earlier in the paper, it is vital to understand the root causes of churning for efficient management. Thus it will be interesting to integrate the “understanding” phase into the prediction phase so that businesses can have a simple but yet effective means to address customer churn.

6 Appendix

Appendix 1: Description of data used for the study

Variable Name	Description
Account Length	Number of days since the customer created account
VMail Message	Number of messages sent by the customer via vmail
Day Mins	Number of minutes the customer has used the service during the day
Eve Mins	Number of minutes the customer has used the service during the Evening
Night Mins	Number of minutes the customer has used the service during the Night
Intl Mins	Number of minutes the customer has used the service for international calls
CustServ Calls	Number of minutes the customer has called the customer service
Churn	Binary Churn for customer
Int'l Plan	Number of International plans the customer has subscribed form the service
VMail Plan	Number of vmail plans the customer has subscribed form the service
Day Calls	Number of calls the customer has called during the day
Day Charge	Rate for Day calls
Eve Calls	Number of calls the customer has called during the Evening
Eve Charge	Rate for Evening calls
Night Calls	Number of calls the customer has called during the Night
Night Charge	Rate for Night calls
Intl Calls	Number of international calls the customer has called
Intl Charge	Rate for International calls
State	State Code
Area Code	Area Code
Phone	Phone Number of the customer

Appendix 2: clustering information obtained after analysis

Account Length	VMail Message	Day Mins	Eve Mins	Night Mins	Intl Mins	CustServ Calls	Clustername	Clustersize	k	Iteration
100.855	8.271	225.516	218.829	225.392	10.138	1.520	cluster_0	1132	3	0
99.893	8.454	229.373	236.785	211.926	10.163	1.516	cluster_0	854	4	1
89.788	8.628	231.908	242.211	210.985	10.130	1.549	cluster_0	639	5	2
99.887	8.399	240.943	208.498	243.566	10.208	1.527	cluster_0	513	6	3
100.388	7.864	128.761	210.667	214.191	10.292	1.590	cluster_1	1206	3	0
100.573	7.884	122.600	224.181	212.297	10.232	1.587	cluster_1	882	4	1
107.099	8.226	142.369	235.570	235.392	10.333	1.552	cluster_1	706	5	2
107.051	8.436	140.940	219.658	246.584	10.370	1.589	cluster_1	628	6	3
102.128	8.189	189.655	168.798	156.648	10.284	1.579	cluster_2	995	3	0
102.122	8.455	184.650	189.300	140.058	10.342	1.578	cluster_2	759	4	1
115.852	8.091	211.574	190.627	148.536	10.364	1.556	cluster_2	667	5	2
115.528	7.798	219.643	181.718	152.678	10.335	1.549	cluster_2	591	6	3
101.821	7.641	185.806	150.267	232.982	10.223	1.571	cluster_3	838	4	1
100.494	8.000	194.719	148.275	237.692	10.136	1.529	cluster_3	670	5	2
89.305	7.863	176.376	137.142	220.691	10.130	1.581	cluster_3	530	6	3
91.262	7.548	121.301	187.054	169.013	10.214	1.631	cluster_4	651	5	2
101.198	7.344	114.188	191.649	163.693	10.299	1.627	cluster_4	515	6	3
91.448	8.663	189.688	260.569	175.768	10.058	1.504	cluster_5	556	6	3

7 References

Almana, A. M., Aksoy, M. S., & Alzahrani, R. (2014), A survey on data mining techniques in customer churn analysis for telecom industry. *Journal of Engineering Research and Applications, 4(5)*, 165-171.

Almana, A., Aksoy, M., Alzahrani, R., (2014) A survey on data mining techniques in customer churn analysis for telecom industry. *Int. Journal of Engineering Research and Applications, Vol. 4, Issue 5*, 165–171.

Dahiya, K., & Bhatia, S. (2015) Customer churn analysis in telecom industry. *Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference,* 1-6.

Deloitte (2017) Telecommunications Industry Outlook – Interview with Craig Wigginton, retrieved from https://www2.deloitte.com/us/en/pages/technology-media-and-telecommunications/articles/telecommunications-industry-outlook.html.

Gallo, A., (2014) The value of keeping the right customers. *HBS No. 10 (Boston Harvard School Publishing, 2014)* retrieved from https://hbr.org/2014/10/the-value-of-keeping-the-right-customers,

Gao, Y., Zhang, G., Lu, J., & Ma, J. (2014) A bi-level decision model for customer churn analysis. *Computational Intelligence, 30(3)*, 583-599.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007) Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research, 34(10),* 2902-2917.

Han, S.-P, Lee, Y. -S., Ahn, J.-H. (2006) Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy 30*, 552 – 568.

James, G., Witten, D., Hastie, T., Tibshirani, R., (2013) *An introduction to statistical learning with applications in R*, Springer, New York.

Li, S., Au, T., Ma, G. (2003) Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques. *Journal of Comparative International Management, Vol. 6, No 1*, 10-22.

Miller, T., (2015) *Marketing data science – modeling techniques in predictive analytics with R and Python*, Pearson, New Jersey.

Nath, S. V., & Behara, R. S. (2003) Customer churn analysis in the wireless industry: A data mining approach. *In Proceedings-annual meeting of the decision sciences institute*, 505-510.

PWC (2011) Curing customer churn. Retrieved from https://www.pwc.com/us/en/increasing-it-effectiveness/assets/curing-customer-churn.pdf, accessed 5/10/2017

Reichheld, F., (n.d.) Prescription for cutting costs, Bain & Company. Retrieved from http://www.bain.com/Images/BB_Prescription_cutting_costs.pdf.

Richeldi, M., Perrucci, A., (2002) *Churn analysis case study*, Telecom Italia Lab, Contract No. IST-1999 – 11993 (Deliverable No. D17.2). Retrieved from http://www-ai.cs.uni dortmund.de/PublicPublicationFiles/richeldi_perrucci_2002b.pdf

Sharma, R., Sachdeva, R. (2017) Performance Evaluation of Churn Customer Behavior based on Hybrid Algorithm. *International Journal of Computer Applications 159(6)*:14-19.

Springer, T., Kim, C., Debruyne, F., Azzarello, D., Melton, J., (2014) Breaking the back of customer churn, Bain & Company Retrieved from http://www.bain.com/Images/BAIN_BRIEF_Breaking_the_back_of_customer_churn.pdf, accessed 5/10/2017

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015) A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory, 55*, 1-9.