Customer Churn Prediction

by
Bernice Abban, Kwabena Aboagye Dougan

***Abstract* — Customer churn is one of the main problems in the telco, banking, food processing industries and the likes. Several studies have shown that attracting new customers is much more expensive than retaining existing ones. Therefore, companies are focusing on developing accurate and reliable predictive models to identify potential customers that will churn in the near future. The aim of this project is to monitor customer activities in real-time, segment customers based on an array of character traits, predict future customer attritions, identify the main reasons for the churns and find anomalies in customer activities. The proposed methodology for predicting customer attritions covers several phases: understanding the business; selection, analysis and data processing; implementing various algorithms for classification; evaluation of the classifiers and choosing the best one for prediction.**

***Keywords* — prediction, churn, decision trees.**

INTRODUCTION

Customer attrition is an important issue that is often associated with the life cycle of the industry. When the industry is in a growth phase of its life cycle, sales are increasing exponentially and the number of new customers largely outnumbers the number of churners. On the other side, companies in a mature phase of in their life cycle, set their focus on reducing the rate of customer churn. (L. Miguel APM, 2001)

The main reasons that cause customer churn are divided into two groups: accidental and intentional. Accidental churn happens when the circumstances are changing so prevents the customers from using the services in the future. Examples of accidental churn are economic circumstances that make services too expensive for the customer. Intentional churn occurs when the customers choose to switch to another company that provides similar services. This type of churn is the one that most companies are trying to prevent. An example of intentional churn is better offers from competition, more advanced services and better price for the same service. (J. Hadden, et al, 2006)

Establishing a system for managing the customer churn is vital. There are two basic approaches for managing customer churn: directed and undirected. In undirected approach, companies rely on superior product and mass advertising to increase loyalty to the brand and to retain customers. In direct approach, companies rely on identifying customers who are likely to churn, and then to adapt their requirements to prevent from churning. (G. S. Linoff, 2011)

In the recent years, churn prediction is becoming very important issue in many industries around the globe. (S-Y. Hung, D., 2006). In order to deal with this problem, the customer operators must recognize these customers before they churn. Therefore, developing a unique classifier that will predict future churns is vital. This classifier must be able to recognize users who have a tendency to churn in the near future, so the customer service providers will be able to react promptly with appropriate discounts and promotions. The most frequently used techniques for this purpose are learning algorithms for classification, like decision trees, logistics regression, *k*-nearest neighbors, Naïve Bayes, neural networks (H. Jiawei et al, 2011)

METHODOLOGY

In order to find a possible solution to the problem of churn prediction i.e. successfully apply a machine learning technique to the available data, one needs a deep understanding of the business rules of the company and their specificity. Such knowledge enables the selection of attributes suitable for the problem at hand. The quality of the data can further be improved by subjecting it to preprocessing. Once a

final dataset is derived, the classification algorithms can be successfully trained and their performances correspondingly evaluated. In the following subsections, we present the identified phases in our methodology.

Business Understanding
In this phase, the focus is set on understanding the project objectives and requirements from the telecommunications business perspective. The aim of the churn prediction is to identify the properties that make a customer churn in order to prevent it and retain the customer. To enable this, we consider customers that churned and analyze their data over a period while they still used the services of the telecommunications company.

Data Understanding
For the purpose of this project, we intend taking data from either a telco, bank or any other service provider in Ghana. We intend anonymizing the data (we only care about the user's dynamics data, not their personal data). The obtained data covers 28 months period from 01.01.2012 to 30.04.2014 (approximately 34 million records). Additional data for the customer complaints is included in the dataset since it is a strong indicator for customer dissatisfaction,

Data Pre-processing
The data pre-processing tasks include careful selection of data attributes and records. Because we are more likely to deal with incomplete and noisy data, some additional data cleaning and transformation will be performed.

Data Selection
We would first identify and extract the most relevant attributes for the project.

- **Demographic attributes:** contain the primary features of the customer such as sex, age, nationality, place of residence, etc.
- **Contract attributes:** contain the attributes associated with the customer contract for a particular service such as type of service, date of conclusion of the contract, price of the service etc.
- **Customer behavior attributes:** describe the customer activities.

Data Cleaning
The presence of noise, unknown or empty values, outliers and invalid values may negatively affect the performance of the machine learning algorithm by using the raw data. The purpose of data cleaning is to reduce the number of inconsistent values, remove noise and incomplete entries and attributes. Since our dataset is sufficiently big, we removed all potentially problematic tuples.

Data Transformation
Data transformation techniques can significantly improve the overall performance of the attrition prediction, which we have seen while experimenting with potential transformations.

Feature Selection
Features selection refers to the process of selecting a subset of relevant attributes of a set of attributes. This reduces the number of input attributes to the learning algorithm, thereby significantly reducing time and resources required to train the algorithm.

Machine learning approaches for attrition prediction
There are many techniques that have been proposed for customer churn prediction. In our approach, we will analyze four machine learning algorithms: random forest classifier, *k*-nearest neighbors' algorithm and logistic regression classifier.

- Logistic regression is a special case of a linear classifier. When applied to a classification problem, it predicts the class using binary dependent variables instead of continuous.

$$\text{logistic}(\eta) = \frac{1}{1 + exp(-\eta)}$$

And it looks like this:

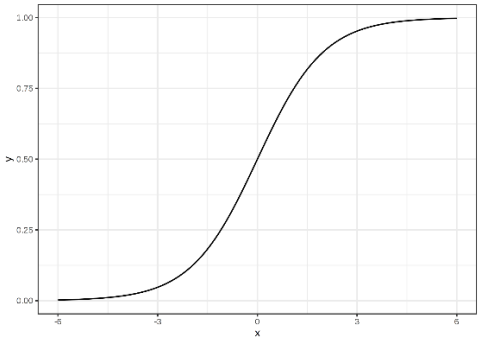

Fig. The logistic function. Its outputs numbers between 0 and 1. At input 0, it outputs 0.5

- *K*-nearest neighbors algorithm compares a test tuple with trained tuples that are similar to it (learning by analogy). The trained tuples are described by *n* attributes (a point in *n*-dimensional space). For a new tuple the algorithm k-nearest-neighbors searches the space for *k* trained tuples that are closet to the

unknown tuple. Most common class of *k* nearest.

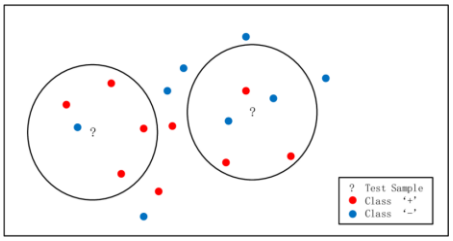

Fig. An example of KNN classification task with k = 5

- Random forest classifier
  Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

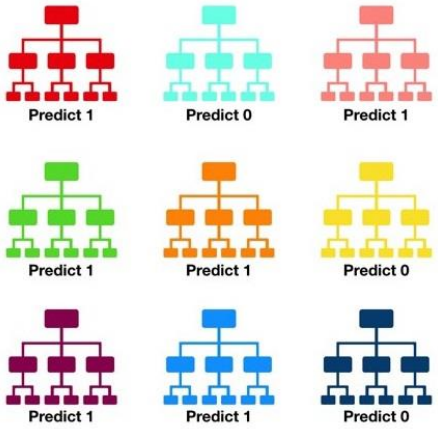

Fig. Visualization of a Random Forest Model Making a Prediction

CONCLUSION

Most industries in the recent years are subjects of major changes and from a fast-growing industry has come to a state of saturation accompanied with strong competitive market. Customers starve for better services and prices, while their requirements are extremely complex and difficult to understand. In order to cope with this problem, we have the following objective: finding the influential factors for the customer churn, as well as building a classifier for predicting the customer churn.

In this documentation we have explained the methodology for building a classifier models that will predict the customer churn. We have carefully extracted the customers' behavior within a one-year period, resulting in dataset containing 22461 customers. The results from this study show that predicting the customer churn can be successful with high accuracy. The classification models derived from *k*-nearest neighbors and logistic regression have an accuracy of over 75%. The highest accuracy is achieved with logistic regression with 80.6% accuracy. The disadvantage of this classifier is its execution time and the need for the vast memory resources. The models based on decision trees also shows high accuracy, while *k*-nearest neighbors show weaker results than other algorithms. Both in terms of execution time and the necessary resources, decision trees are superior to other algorithms. Also, the main advantage of decision trees is their understandable detection of knowledge, that can be easily displayed to the user.

As a result of this AI project and the extracted knowledge, the operator will be able to accurately predict its customers' behavior, and will be able to direct their policies towards customers and their retention. At the same time, the results could lead to cost savings and better building of the company's budget.

REFERENCES

L. Miguel APM. "Measuring the impact of data mining on churn management." *Internet Research*, vol. 11, no. 5, pp. 375–387, 2001.

J. Hadden, et al. "Churn prediction: Does technology matter." *International Journal of Intelligent Technology*, vol. 1, no. 2, pp. 104–110, 2006.

G. S. Linoff, and M. J.A. Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011.

S-Y. Hung, D. C. Yen, and H.-Y. Wang. "Applying data mining to telecom churn management." *Expert Systems with Applications*, vol. 31, no. 3, pp. 515–524, 2006.