

# Gaining Insight into Customer Churn Prediction using Generalized Additive Neural Networks

David A de Waal and Jan V du Toit

**Abstract**—As the South African telecommunications industry is opened up, competition for customers is increasing. Retaining customers is therefore of critical importance to services providers. Some of the techniques that have been successfully applied to churn prediction include logistic regression, decision trees and neural networks. Support vector machines are a recent addition to the modeller's arsenal of techniques and seems to be the technique of choice in many applications as it promises increased accuracy. However, although improved accuracy in predicting churn is important, gaining insight into reasons for churn might be the ultimate aim, as this allows the service provider to introduce interventions that could increase customer retention and therefore profitability. In this article, a recent development of generalized additive models, namely Generalized Additive Neural Networks is proposed as a technique for modelling churn. It not only promises great accuracy, but also ease of interpretability. Previously, neural networks have been known to offer accurate predictions, but the black box nature of the model offered little insight or explanation as to why a customer was about to churn. This shortcoming of neural networks is overcome by restricting the neural network architecture to a simplified architecture implementing a generalized additive model and the use of partial residual plots that provide a graphical view of the relationships (which could be nonlinear) between the input variables and the target (and therefore also reasons for churn).

**Index Terms**—Customer churn, Neural networks, Partial residual plots

## I. INTRODUCTION

THE South African telecommunications landscape is changing: a second fixed-line operator license has been awarded and Virgin Mobile entered the mobile phone market. Although these two events have not yet shaken up the South African telecommunications market as was expected, the landscape is slowly changing and competition

D. A. de Waal is with the Centre for Business Mathematics and Informatics, Private Bag X6001, North-West University, Potchefstroom, 2520, South Africa (e-mail: Andre.DeWaal@nuw.ac.za, Tel (018) 299-2535, Fax (018) 299-2584).

J. V. du Toit is with the Department of Computer Science and Information Systems, Private Bag X6001, North-West University, Potchefstroom, 2520, South Africa. (e-mail: Tiny.DuToit@nuw.ac.za, Tel (018) 299-2548, Fax (018) 299-2570).

for customers is increasing. Furthermore, the launching of Mobile Number Portability allows customers to keep their mobile phone numbers when changing service provider. Churning (the phenomenon whereby a customer leaves a service provider) might therefore increase and it is vital that service providers retain existing customers and strengthen relationships with customers to remain profitable.

In a recent study [1], the churn rate for U.S. mobile carriers was estimated at between 2% and 3%. Furthermore it costs the mobile carriers between \$400 and \$500 to sign a single customer who typically generates about \$50 in monthly revenue. An amount of \$207 million could have been saved by the top six US mobile carriers if they have better customer retention strategies [2].

If the situation in South Africa is anything like that in the US, South African mobile operators are losing millions of Rand to customers chasing "special deals" and the latest mobile phones. It would therefore be prudent to invest in techniques that may be used to limit customer churn.

Several Data Mining techniques have been successfully applied to customer churn (attrition) prediction [14]; [15]; [16]; [17]. These techniques include logistic regression, tree-based methods and artificial neural networks [3]. Due to the nonlinear nature of the causes of churn [3], it is important that the modelling technique is able to model nonlinear relationships between the independent variables (demographic information, contractual information and call pattern information) and the dependent variable (binary target variable indicating churn or no-churn). Inspection of these nonlinear relationships could then provide the service provider with reasons for churning.

Although neural networks are known for their accurate predictions, they are usually regarded as 'black boxes' and reasons for reaching the predictions are usually absent.

In this article, a special type of neural network, known as a Generalized Additive Neural Network or GANN, is proposed for customer churn prediction. A GANN implements a Generalized Additive Model and exploits partial residual plots as a visual aid in interpreting and understanding complex nonlinear relationships between the independent variables and the dependent variable [4].

The rest of the article is organized as follows. In Section II the GANN architecture is introduced. Section III contains a description of Partial Residual Plots and an example of a complex nonlinear function modeled by a GANN. Automation of the algorithm is described in Section IV. Some of the results obtained with the automated GANN system are presented in Section V. Conclusions are given in Section VI and the article ends with ideas for Future Work.

## II. GENERALIZED ADDITIVE NEURAL NETWORKS

A Generalized Additive Model (GAM) [5]; [6] is defined as the sum of unspecified univariate functions,

$$g_0^{-1}(E(y)) = \alpha + f_1(X_1) + \dots + f_p(X_p) + \varepsilon$$

where  $E(\varepsilon) = 0$  and  $\text{var}(\varepsilon) = \sigma^2$ . A link function,  $g_0^{-1}$ , is used to constrain the range of the response values.

Multilayer perceptrons (MLPs) are the most widely used type of neural network for supervised prediction. Theoretically, MLPs are universal approximators that can model any continuous function [7]. For this reason, MLPs can be used as the univariate functions of GAMs. A MLP that has a single layer with  $h$  hidden neurons has the form

$$g_0^{-1}(E(y)) = w_0 + \sum_{t=1}^h w_t \tanh(w_{ot} + \sum_{j=1}^p w_{jt} x_j).$$

The activation function used for the hidden layers in this case is the hyperbolic tangent function suggested by [8]. This nonlinear regression model has  $h(p+1)+1$  unknown parameters (weights and biases). The parameters are estimated by numerically optimizing some suitable measure of fit to the training data such as the negative log likelihood.

When GAMs are implemented as neural networks, backfitting is unnecessary, since any training method suitable for fitting MLPs can be utilized to simultaneously estimate the parameters of GANN models. As a result, the usual optimization and model complexity issues also apply to GANN models.

The basic architecture for a GANN has a separate MLP with a single hidden layer of  $h$  units for each input variable, given by

$$f_j(x_j) = \sum_{t=0}^h w_{tj} \tanh(w_{otj} + w_{1tj} x_j).$$

The overall bias  $\alpha$  absorbs the individual bias terms. Each individual univariate function has  $3h$  parameters, where  $h$  could vary across inputs. The architecture can be enhanced to include an additional parameter for a direct connection (skip layer)

$$f_j(x_j) = w_{0j} x_j + \sum_{t=1}^h w_{tj} \tanh(w_{otj} + w_{1tj} x_j)$$

so that the Generalized Linear Model is a special case.

The following set of instructions for constructing a GANN interactively [4] takes advantage of their constrained form to simplify optimization and model selection. This methodology guides the modeler in visually deciding on the appropriate complexity of the individual univariate functions.

1. Construct a GANN with one neuron in the hidden layer and a skip layer for each input in the model. In this step the univariate functions are initialized to

$$f_j(x_j) = w_{0j} x_j + w_{1j} \tanh(w_{01j} + w_{11j} x_j).$$

This gives 4 parameters for each input. Binary inputs only have a direct connection.

2. Fit a Generalized Linear Model to give initial estimates of  $\alpha$  and the  $w_{0j}$ .
3. Initialize the remaining 3 parameters in each hidden layer as random values from a normal distribution with mean zero and variance equal to 0.1.
4. Fit the full GANN model.
5. Examine each of the fitted univariate functions overlaid on their partial residuals.

6. Prune (remove neurons) the hidden layers with apparently linear effects and grow (add neurons) the hidden layers where the nonlinear trend appears to be underfitted. If this step is repeated, the final estimates from previous fits can be used as starting values.

The effect of each input on the fitted model can be interpreted by using partial residual plots.

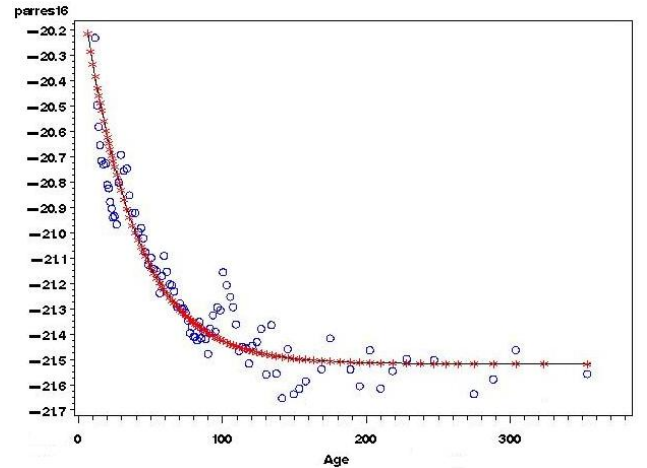
## III. PARTIAL RESIDUAL PLOTS

The visual diagnostics used to aid the model selection process for GANNs are plots of the fitted univariate functions,  $\hat{f}_j(x_j)$ , overlaid on the partial residuals,

$$pr_j = g_0^{-1}(y) - \alpha - \sum_{l \neq j} \hat{f}_l(x_l) = (g_0^{-1}(y) - g_0^{-1}(\hat{y})) + \hat{f}_j(x_j)$$

versus the corresponding  $j$ th input. With partial residuals [9]; [10] the effect of the individual inputs, adjusted for the effect of the other inputs, can be investigated. The  $j$ th partial residual is the deviation between the actual values and that portion of the fitted model that does not involve  $x_j$ .

An example of a partial residual plot showing a nonlinear relationship between the variable Age and the probability of churn is given in Figure 1.



**Figure 1: Nonlinear relationship**

In this example, Age might give an indication of the elapsed time since the customer started using a specific service provider (in days or weeks). From this plot the service provider could infer that it would be rewarding to pay special attention to new customers, as they are the most likely candidates to churn. An intervention could be planned.

In the next section, an automated approach to the construction of GANNs is presented. This new method is objective and relies only on a model selection criterion for model selection. No human interaction is needed for choosing the single best model.

## IV. AUTOMATED CONSTRUCTION OF GANNs

Automation of the interactive construction algorithm is not trivial. The algorithm must operate effectively given the following two sources of uncertainty. First, the univariate functions for the variables are unspecified. Second, the optimal neural network architectures needed to approximate the unknown univariate functions are also unknown.

In the interactive construction algorithm, human judgment is also required to interpret the partial residual plots. But, automating human judgment is not easy. Complex image processing and machine learning techniques could be used. However, the insight that objective model selection criteria can be used to differentiate between “good” and “bad” models makes the construction of an automatic algorithm possible.

The development of an efficient algorithm overcoming the uncertainty just described as well as eliminating the reliance on human judgment is the subject of the rest of this section.

A high level algorithm [11] is given below to make comparison with the interactive algorithm possible.

1. Construct a GANN with a skip layer for each input in the model. In this step the univariate functions are initialized to  $f_j(x_j) = w_{0j}x_j$ . This gives 1 parameter for each input. Binary inputs only have a direct connection.
2. Fit a Generalized Linear Model to give initial estimates of  $\alpha$  and the  $w_{0j}$ .
3. N/A.
4. Fit the full GANN model.
5. Expand the model, that is for each variable in the model, selectively prune the hidden layers and add neurons to the hidden layers. Examine each of the models resulting from a prune or a grow step.
6. Select the best (based on model selection criterion) unexpanded model. Repeat from step 5 until search the space has been exhausted or the time has lapsed. Report the best model and terminate.

To arrive at a working implementation, several additional decisions need to be made. They are described in the following subsections.

#### A. GANN ENCODING

Let each GANN architecture be represented by a string over a finite alphabet. The alphabet consists of ten digits and is given in Table 1.

Symbol	Description
0	No MLP (Input removed)
1	MLP with a skip layer
2	MLP with no skip layer and 1 hidden node
3	MLP with a skip layer and 1 hidden node
4	MLP with no skip layer and 2 hidden nodes
5	MLP with a skip layer and 2 hidden nodes
6	MLP with no skip layer and 3 hidden nodes
7	MLP with a skip layer and 3 hidden nodes
8	MLP with no skip layer and 4 hidden nodes
9	MLP with a skip layer and 4 hidden nodes

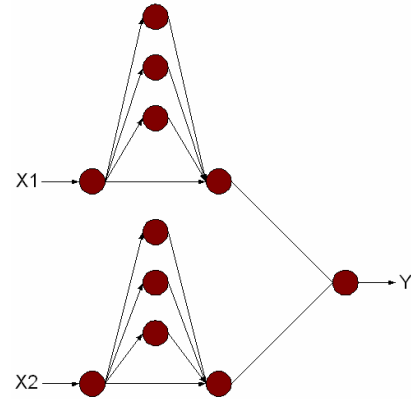
**Table 1: Alphabet**

The restriction to a finite alphabet is not strictly needed, as the model selection criterion will prevent too complex neural networks from being exploited. This restriction however simplifies the implementation as an upper limit is placed on the complexity of the underlying neural network. Also, in the interactive algorithm, a skip layer is always

included. This restriction is relaxed in the automated algorithm.

A well-formed string consists of a finite number of digits from the alphabet. For example, the string 102003313 represents a GANN architecture with nine variables where variables 2, 4, and 5 are deleted, variables 1 and 8 are each represented by a MLP with only a skip layer, variable 3 is represented by a MLP with no skip layer and one node in a hidden layer and variables 6, 7 and 9 are represented by a skip layer and one node in the hidden layer. Every digit in the string represents an architecture for a specific variable (called a sub-architecture).

An example of a GANN with two inputs, represented by the string 77 is given in Figure 2.



**Figure 2: Example GANN**

To achieve specific aims such as the reduction in the size of the search space or limiting the complexity of the univariate functions, further restrictions might be placed on the given alphabet (only certain digits allowed).

#### B. MODEL SELECTION CRITERIA

Two philosophies of model selection, efficient and consistent criteria, are described subsequently.

An assumption commonly made in both regression and time series is that the generating or true model is of infinite dimension, or that the set of candidate models does not contain the true model. The goal is then to choose one model that best approximates the true model from a set of finite-dimensional candidate models. The candidate model closest to the true model is assumed to be the appropriate choice. Examples of efficient criteria are FPE, AIC, AICc, and Cp [12]. The Akaike Information Criterion (AIC) is defined as

$$AIC = -2 \log(L(\hat{\theta} | y)) + 2K$$

where  $\log(L(\hat{\theta} | y))$  denotes the natural logarithm of the likelihood function of the parameter vector  $\theta$ , given the data  $y$  and  $K$  the number of estimable parameters in the approximating model [12]. In the special case of least squares estimation with normally distributed errors and constant variance, the AIC can be expressed as:

$$AIC = n \log(\hat{\sigma}^2) + 2K \quad \text{where } \hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n} \text{ (the MLE of } \sigma^2 \text{)}.$$

$\varepsilon_i$  are the estimated residuals for a particular candidate model, and  $K$  is the total number of estimated regression parameters, including the intercept and  $\sigma^2$ .

Many researchers assume that the true model is included in the set of candidate models and consequently of finite dimension. Under this assumption the goal of model selection is to correctly identify the true model from the list of candidate models. A model selection criterion that identifies the correct model asymptotically with a probability of one is said to be consistent. Examples of consistent criteria are SIC, HQ, and GM [12].

Several forms of the Schwarz Information Criterion (SIC or SBC) have been proposed in the literature. The generic SBC is defined as follows [12]:

$$SBC = -2\log(L(\hat{\theta} | y)) + K \log(n)$$

In the special case of the Gaussian error model, the SBC can be expressed as

$$SBC = n\log(\hat{\sigma}^2) + K \log(n).$$

## V. RESULTS

The automated algorithm has been implemented in SAS SAS® Enterprise Miner™ and applied to various data sets (e.g. Adult and Breast Cancer) with excellent results [13]. That article also contains a comparison with other techniques and some of the comparative results obtained with the Automated GANN system on the Breast Cancer data set and the Adult data set are given in Table 2 and in Table 3.

Method	Validation Error
GANN (auto-system)	0.022
LogitBoost	0.028
Real AdaBoost	0.032
Gentle Adaboost	0.031
Discrete Adaboost	0.032
CART	0.045

**Table 2: Breast Cancer Results**

Method	Error
GANN (auto-system)	13.65
C4.5	15.54
Voted ID3 (0.6)	15.64
T2	16.84
NBTree	14.10
FSS Naive Bayes	14.05

**Table 3: Adult Results**

In all cases the automated Generalized Additive Neural Network approach was found to be superior to other techniques in all respects (misclassification rates, error rates and Gini Coefficients)

The authors are confident that the automated GANN system described in this article should perform in a similar manner on customer churn data. Furthermore, the automated GANN system generates partial residual plots as standard that should provide graphical clues as to the reasons for

churn. The reasons could then be investigated further and the necessary steps taken to limit churn and therefore to retain existing customers.

## VI. CONCLUSION

In this article, Generalized Additive Neural Networks is proposed as a technique to model and predict Customer Churn. Benefits of using this technique are that it has been shown to be very accurate on other data sets [13], that it can be computed automatically and that it provides insight into the relationships between the input variables and the target variable. This insight can then be used by the service provider to plan strategies to limit customer churn and increase profitability.

The developed algorithm is general and can be applied to any classification problem such as propensity modelling and even to build scorecards (application and behavior). Furthermore, although the automated technique is based on neural network technology, some of the problems usually associated with the use of neural networks are removed. Firstly a GANN is not a black box as the modeller can inspect partial residual plots to gain insight into the relationships between the input variables and the target variable. Secondly the optimal GANN architecture is identified automatically (usually this is done by trail and error) which could results in significant time savings.

## VII. FUTURE WORK

The obvious next step would be to apply the proposed technique to a real data set from a service provider in South Africa and report results. This is also the aim of our forthcoming research, but dependent on the cooperation and willingness of a service provider to provide data. As confidentiality of such data is always an issue, a data set from a previously done study exploiting other techniques might also be used.

## ACKNOWLEDGMENT

DA de Waal and JV du Toit thank SAS Institute for providing them with Base SAS® and SAS® Enterprise Miner™ software used in computing all the results presented in this paper. This work forms part of the research done at the North-West University within the TELKOM CoE research programme, funded by TELKOM, GRINTEK TELECOM and THRIP.

## REFERENCES

- [1] Y. Zhao, B. Li, W. Liu, S. Ren, "Customer Churn Prediction Using Improved One-Class Support Vector Machine" in *Advanced Data Mining and Applications*, vol. 3584/2005, Springer, Berlin, pp. 300-306, 2005.
- [2] Duke Teradata, Teradata Center for Customer Relationship Management. Retrieved on: November 7, 2002.
- [3] T. Au, S. Li, G. Ma., "Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques", *Journal of Comparative International Management*, vol. 6, no. 1, June 2003.
- [4] W. J. E. Potts, Generalized additive neural networks, in *Proceedings of the Fifth ACM SIGKDD International*

*Conference on Knowledge Discovery and Data Mining*, pp. 194–200, 1999.

- [5] T. J. Hastie & R. J. Tibshirani, *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London, 1990.
- [6] S. N. Wood, *Generalized Additive Models: An introduction with R*, Texts in Statistical Science, Chapman & Hall/CRC, London, 2006.
- [7] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, United Kingdom, 1996.
- [8] W. J. E. Potts, *Neural Network Modeling Course Notes*, SAS Institute Inc., Cary, NC, 2000.
- [9] M. Ezekiel, “A method for handling curvilinear correlation for any number of variables”, *Journal of the American Statistical Association* 19(148), pp. 431–453, 1924.
- [10] W. A. Larsen & S. J. McCleary, “The use of partial residual plots in regression analysis”, *Technometrics* 14(3), pp. 781–790, 1972.
- [11] J. V. Du Toit, “Automated Construction of Generalized Additive Neural Networks for Predictive Data Mining”, PhD thesis, School for Computer, Statistical and Mathematical Sciences, North-West University, South Africa, 2006.
- [12] K. P. Burnham & D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn, Springer, New York, 2002.
- [13] D. A. de Waal, J. V. du Toit., Generalized Additive Models from a Neural Network Perspective, ‘*Seventh IEEE International Conference on Data Mining – Workshops*’, 2007.
- [14] J.-H. Ahn, S.-P. Han, 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, pp. 552-568, 2006.
- [15] C.-P. Wei, I.-T. Chiu, “Turning telecommunications call details to churn prediction: a data mining approach”, *Expert Systems with Applications*, vol. 23, pp. 103-112, 2002.
- [16] W.-H. Au, K. C. C. Chan, X. Yao, “A Novel Evolutionary Data Mining Algorithm With Applications to Churn Prediction”, *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 6, pp. 532-545, December 2003.
- [17] L.-S. Yang, C. Chiu, “Subscriber Churn Prediction in Telecommunications”, *WSEAS Transactions on Mathematics*, vol. 6, no. 2, pp. 316-323, February 2007.

**J. V. du Toit** is a senior lecturer in the Department of Computer Science and Information Systems, North-West University, South Africa. He obtained his Ph.D. from the North-West University, Potchefstroom Campus, South Africa during 2006 and his fields of interest include Artificial Intelligence and Data Mining.