Finding The Best Statistical Model To Predict Customer Defection In Telecommunication Retail Setting

Nkululeko Ngcongo

February 11, 2014

University of Witwatersrand

Supervisor: Prof. David Lubinsky

Student Number: 576639

A Research Report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Mathematical Statistics

Candidate's Declaration

I, Nkululeko Ngcongo, declare that this Theses is my own, unaided work. It is being submitted for the Degree of Masters of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Nkululeko Ngcongo
11 February 2014

Abstract

In this study we examine the question of which statistical models work well in predicting customer defection in the retail mobile telecommunication industry. For each of the two data sets that were used (mobile call pattern and billing, and time taken to churn data), four statistical models were fitted and compared namely; artificial neural networks, decision trees, logistic regression and support vector machines. The artificial neural network model proved to be superior than the other three models when fitted on both data sets. This model gave the best area under the receiver operating characteristic curve (0.93 for call pattern data and 0.88 for billing and time taken to churn data), highest lift at 10 per cent of the population (7.01 for call pattern data and 2.12 for billing and time taken to churn data) and lowest misclassification rate (0.04 for call pattern data and 0.19 for billing and time taken to churn data). The logistic regression model under performed the other models when fitted to call pattern data and came out as third when fitted to billing and time taken to churn data whereby they outperformed the decision tree model. Support vector machine came out as the second best model for billing and time taken to churn data and third when fitted to call pattern data. Decision tree model performed well when fitted to call pattern data and worst when fitted to billing and time taken to churn data The study showed that in the retail mobile telecommunication industry, companies can increase revenue streams and competitive advantage by using data mining techniques to predict customers that are likely to churn. The next step for the business is to embark on retention programs to use these methods to reduce churners.

Dedication

This thesis is dedicated to my family, friends and all the under privileged children trying to strive in the ghetto.

Acknowledgments

I would like to thank my supervisor Professor David Lubinsky for dedicating his time and guiding me with my work. I would like to send my sincere thanks to my family for being supportive and understanding all the time. Another great thanks goes to all my friends especially Njabulo Ngcongo, Sivuyile Mgobhozi, John Mukombewrana and Nompumelelo Zama for their support and assistance. A great thanks also goes to the University of California and Data Mining Inc. for their data sets.

Lastly, I would like to thank GOD for making this possible.

Contents

List of Figures

List of Tables

Abbreviations

ANN = Artificial Neural Networks

SVM = Support Vector Machines

Data set 1 = Call pattern churn data

Data set 2 = Billing and time taken to default data

RBF = Radial basis function

SMOTE = Synthetic minority over sampling technique

ROC = Receiver operating characteristic

AUC = Area under the curve

AIC = Akaike information criteria

BIC = Bayes information criteria

SBC = Schwarz Bayesian criteria

1 Introduction

1.1 Background

Statistical data mining is the process of extracting data from different data sources and manipulating the data in order to produce meaningful information that can be used by management to make decisions. Data mining is an 'emerging' field in statistics since technology has allowed us to store large amounts of data to be analysed so that companies, governments and other organizations can make informed decisions. Statistical data mining techniques can be applied to many social science fields [Chow, 2002, Kvam and Sokol, 2004, Crang, 2002, Philip et al., 2011, Mazzocchi, 2007, Juahiainen, 2012]. In this research, we concentrate on using statistical data mining techniques in the marketing field. Marketing departments around the world have huge databases with customer's demographic and behavioural details. They no longer need to rely on gut feel, rather they can use statistics in order to make informed decisions. In the case where the industry has reached saturation the market becomes a churn market and it is difficult and expensive to recruit new customers [Friedman, 1997]. In order for a business to survive fierce competition where churn rates are high, it must rely on statistical data mining techniques to predict churners. Statistical data mining has played an important role in market research in recent years [Imhoff, 2001].

In the retail mobile telecommunication setting, customer relationship management is a very important aspect of the business. Customers have a fixed contract with a known expiry date or termination date. Not all customers will be satisfied with the service they receive and this will lead to customers not renewing their contracts or terminating them earlier than expected. There are various factors that will lead to this, for example:

- Bad service
- Better offers by competitors
- Network inefficiencies

There are also some exogenous factors that one cannot account for that can lead to customer defection, for example:

- Deceased or emigrated customers
- Financial situation where by a customer loses employment and decides to terminate the contract

- Fraudulent contracts that need to be terminated
- Natural disasters

Because retention efforts are expensive, it makes sense to look at retention initiatives for only high value customers. A high value customer may be determined based on the following factors:

- Their 'age on book' is at least more than the initial contract period (excluding new customers)
- They have never missed any of their monthly instalments
- They have participated in a customer satisfaction survey or other study
- They have at least one of the top of the range products
- They must have at least renewed their contract once
- They have not opted out of marketing initiatives

1.2 Statistical problem: finding the best model

The main research question that we address is which statistical technique predicts with accuracy the 'high value' customers that are likely to defect in the retail mobile telecommunication setting. In this problem of predicting customer defection, we are not highly concerned about time taken to defect but mainly concerned about detecting a type of customer profile that is likely to defect. The aim is to predict defection or termination of the service by customers and to also understand the type of statistical techniques that are most successful in predicting customer defection in this setting. This will enable us to classify with a certain probability whether customers are likely to defect or not, based on their historical data.

The retail mobile telecommunication setting is highly competitive therefore, it is easy for a customer not to renew his or her contract. If no new high value customers are recruited as the old ones that churned, then there will be a significant decrease in profit margins. This will lead to business insolvency.

2 Statistical Theory

2.1 Models to be used

The following standard data mining classification models were used in this research to predict churn:

- Artificial neural networks
- Decision trees
- Linear support vector machines
- Logistic regression

The motivation behind using these models is their simplicity and it is fairly easy to interpret the results. We want to find out which model is the most suitable for dealing with retail mobile telecommunication data. Table 1 shows the basics of the four models. Yang and Chiu argued that artificial neural network models are a black box and that the weights of the neurons are uninterpretable. This is a big disadvantage compared to the other three models [WSE, 2006].

Table 1: Model Comparison

Model	**Decision trees**	**Logistic regression**	**Support vector machines**	**Artificial neural networks**
Loss Function	Confusion Matrix	Log Loss	Hinge Loss	Log
High Dimensional Feature	Linear Kernel	Gaussian Kernel	Polynomial	Hyperbolic Tangent
Works Well With	Continuous and Binary	Binary	Continuous	Continuous and Binary
Over fitting	Pruning	Cross Validation	L^2 Norm	Early Stopping

In the remainder of this section we will introduced each modelling technique.

2.1.1 Decision Trees

The basic idea of decision tree models is that for a given training sample $d \subset D$, where D is the entire data set containing $X_i, \quad \forall i = 1, 2, \cdots, n$ individuals with k attributes and $n >> k$, you want to divide d based on the k^{th} attribute and the class $j, \quad \forall j = 1, \cdots, f$ you wish to predict such that you have unique trees with unique individuals [Kamber and Han, 2006]. The

class j is the response variable which can be binary or has multiple states. Suppose that the class variable that you wish to predict is the likelihood that a customer will terminate his/her cell phone contract with a certain service provider (good = not terminate, bad = terminate). The training sample d will be used to build the tree and the model derived from d will be used to classify the X_i in the test sample $T = D - d$. Using the test sample you can also check the model accuracy by checking how many individuals you have correctly classified. The model will enable you to classify new data points entering the system as to whether they will terminate or not.

The decision tree technique is widely used in the data mining industry and is well known for its simplicity. To decide which variable to split on, many functions have been suggested. The most common are GINI index, entropy and information. When a node p is split into l partitions, the quality of the split is given by

$$GINI_{split} = \sum_{j=1}^{l} P(k_j)GINI(p/k) \tag{2.1}$$

where k is the attribute used to split into class j and GINI index at node p is

$$GINI(p) = 1 - \sum_{j=1}^{f}(Prob(j/k))^2$$

where $Prob(j/k)$ is the probability of class j at node p. A pure node is reached if $GINI(p/k) = 0$ and a best split is the variable with the lowest GINI index [Linoff and Berry, 2004].

The entropy of a random variable $X_i, i = 1, 2, \cdots, n$ is

$$entropy(a_1, a_2, \cdots, a_j) = -a_1 \log a_1 - a_2 \log a_2 - \cdots, -a_j \log a_j$$

which is

$$= -\sum a_j log(a_j), \forall j = 1, \cdots, n$$

where a_j is the probability that X_i belongs to a class j.

Let d be the training data set, j be the class that you want to predict (customer terminates contract or not) and k the data attributes and entropy function $g(x)$ then the information gain is:

$$Info(d, k) = g(x) - \sum_{i} \frac{|Y_i/kattribute|}{|d|} g(Y_i \epsilon d/k)$$

The information gain has a huge disadvantage when it comes to splitting data with distinct or unique values as these carry the highest information in the data set (This means that the data will be split first by this variable thus showing it as the most significant variable). The k attribute with the highest information gain will be used to split the data [Kamber and Han, 2006]. Proust suggested that one can also split on the G squared statistics which works out to be twice the size of entropy, that is $G^2 = 2 * entropy$ [Proust, 2012].

2.1.2 Logistic Regression

The second classification technique that was considered was logistic regression. The basic idea is that you have a data set of n distinct individuals with $X_i, \forall i = 1, 2, \cdots, n$ and you want to predict each individual belonging to a certain class j (terminate phone contract = bad or not terminate = good, say) with a certain probability. Let $j = class$ where $j = 1$ if good and 0 if bad and let $\mathbf{X} = X_1, \cdots, X_n$ be the observed data set variables then

$$P(j = Class|X = X_i) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)} \tag{2.2}$$

is the probability of belonging to a certain class [Friedman et al., 2008]. It must be noted that the exponent part is the normal multivariate linear equation where you can have dummy variables, indicators or interaction terms. Not all attributes for X_i points will be significant in predicting the membership of a class j, you can therefore select the attributes that are significant in predicting class j. This may be done by a forward or backward selection method. After fitting the logistic model the contribution or significance of each selected attribute to the model can be determined by likelihood ratio test or the Wald statistic and other methods [Cios et al., 2007].

The odds ratio is used to measure the association between response and predictor variables, that is, the probability of occurring versus not occurring. The odds ratio is widely used in Bio-statistics for evaluating association and relative risk of a certain factor for the groups being studied [Raygoza, 2009]. Suppose two groups are being studied (control and treatment group) and let $\bar{T}$ be the probability of an event in the treatment group and C be the probability of an event in the control group then the odd ratio is:

$$OR = \frac{\frac{\bar{T}}{1-\bar{T}}}{\frac{C}{1-C}} \tag{2.3}$$

If $OR = 1$ then the odds of the event being studied are equally likely to occur in both groups that is the probability of each event occurring is half. This lead to the following equation

$$P(j = Class|X = X_i) = \frac{OR}{1 + OR}$$

which means that

$$OR = \frac{P(j = Class|Y = Y_i)}{1 - P(j = Class|X = X_i)}$$

The odds ratio can be studied for each variable in the logistic regression and this measures the contribution of that variable to the regression equation.

2.1.3 Support Vector Machines

The third technique considered was the support vector machines (SVM). The basic idea is that you have a training sample d and data points $X_i, i = 1, 2, \cdots, n$ and you want to divide the data set into j regions (j may be the class variable). The data will be divided by a set of hyper planes into j regions [Mirowski et al., 2008]. The support vectors are the data points that are closest to the plane that divides the data into j sub regions. You may have quite a large number of the hyper planes that divide the data into j sub region but what you really want is a hyper plane (line) that maximizes the region between the support vectors [Friedman et al., 2008]. This is because maximising the region between the support vectors decreases the likelihood of misclassifying new data points.

Figure 1: Plane separating the data points

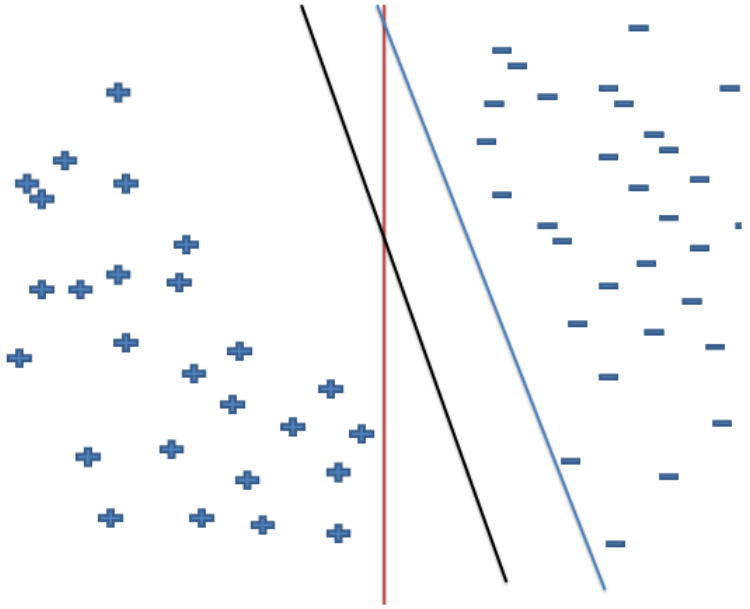


Figure 1 shows three planes that can separate the positive and negative data points with zero misclassification rate [Mirowski et al., 2008]. From this figure, the black plane is the best separator because it gives a bigger margin between the two groups of points. A bigger margin is best because there is a higher chance that if a new data point is imputed it will be classified correctly. Finding the best plane (line) that separates these points is an optimisation problem which can be solved using Lagrangian methods. Sometimes these data points may not be linearly separable.

To define this in a mathematical way, suppose you have a training data set

$$D = [(x^1, y^1), (x^2, y^2), \cdots, (x^l, y^l)], \forall x \epsilon R^n, y \epsilon [-1, 1]$$

where $x^l, \forall i = 1, 2, \cdots, l$ is the vector of individual and attributes, and y^l are regions of belonging for each individual X^n. These points can be separated into -1 or 1 by a hyper plane $< w, x > +b = 0$ where b is the distance from the point to the plane, w the weights vector and $< w, x >$ is the dot product. The separating hyper plane must satisfy

$$y^i[< w, x^i > +b] >= 1, \forall i = 1, 2, \cdots, l \tag{2.4.1}$$

and the distance of x to the hyper plane which is

$$d(w,b;x) = \frac{|<w,x^i>+b|}{||w||} \tag{2.4.2}$$

The optimal hyper plane is the one that minimises $\phi(w) = \frac{1}{2}||w||^2$ and combining this with 2.4.1 and forming a Lagrangian equation with parameter *alpha* it leads to finding a solution of

$$\phi(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i(y^i[<w,x^i>+b]-1) \tag{2.4.3}$$

which satisfies the Karush Kuhn Tucker condition which is first order condition for an optimal value. From 2.4.3 one must find the first partial derivative with respect to b and w and equate to zero for an optimal solution. The solution to the problem is then given by

$$\alpha^{`} = argmin_\alpha \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{j} \alpha_i\alpha_j y_i y_j < x_i, x_j > - \sum_{k=1}^{l}\alpha_k \tag{2.4.4}$$

constrained by $\alpha_i > 0$ and $\sum_{i=1}^{j} \alpha_j y_i = 0$

Assume now that the data is not linearly separable by a hyper plane and suppose now that there is an error $\psi_i, \forall i = 1, 2, \cdots, l$ then the constraint equation 2.4.1 will be modified to

$$y^i[<w,x^i>+b] >= 1 - \psi_i, \forall i = 1, 2, \cdots, l \tag{2.4.5}$$

and the optimal plane is found by w that minimises

$$\phi(w,\alpha) = \frac{1}{2}||w||^2 - C\sum_{i=1}^{l}\psi_i \tag{2.4.6}$$

where C is given subject to constrains. The Lagrangian equation now becomes

$$\phi(w,b,\alpha,\psi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\psi_i - \sum_{i=1}^{l}\alpha_i(y^i[w^T x^i + b] - 1 + \psi_i) - \sum_{j=1}^{l}\beta_i\psi_i \tag{2.4.7}$$

where β and α are Lagrangian multipliers. The equation 2.4.7 is solved in similar fashion to 2.4.3 and the solution is given by

$$\alpha^{`} = argmin_\alpha \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{j} \alpha_i\alpha_j y_i y_j < x_i, x_j > - \sum_{k=1}^{l}\alpha_k \tag{2.4.8}$$

constrained by $0 <= \alpha_i <= C$ and $\sum_{i=1}^{j} \alpha_j y_i = 0$ [Gunn, 1998]

Now that the optimisation problem is solved one needs to know the type of hyper plane to be fitted. When fitting a SVM model one can use kernel functions to map the data into high dimension with the aim of making the data more separable. There are quite a number of kernel functions that are available but we will look at the following kernels:

- Radial Basis Function: $k(x, x') = \exp(-\sigma||x - x'||^2)$
- Polynomial: $k(x, x') = (scale < x, x' > +K)^N$
- Hyperbolic Tangent: $k(x, x') = \tanh(< x, x' > +K)$
- Laplace: $k(x, x') = \exp(-\sigma||x - x'||)$

and the choice of the kernel really depends on the data set. The parameter choices of K, N (degree) and σ also depends on the data set. Furthermore, in R (A statistical analysis software) if these parameters are not given, the program will select the best "parameter" values for you [Karatzoglou et al., 2006].

2.1.4 Artificial Neural Networks

The final model that was used in this research was artificial neural networks (ANN). The reason behind using this approach was that it can fit the data well where linear and other models have proved inadequate. The drawbacks are that this model tends to over fit the data and the fact that it is complex to execute and interpret at times. This data mining classification technique was inspired by biological nervous system architect [Nemati, 2000]. In biology, millions of neurons are interconnected by synapses which carry "information" from one neuron to another. This information is then sent to other neurons as output and the end results are just sensory information (for example: jump).

Data mining construction of a neural network uses almost similar ideology to biology in the sense that you have the following:

- An output vector that passes information
- A "neuron" that processes this information
- A weight for every piece of information entering the neuron

Figure 2: A typical artificial neural network

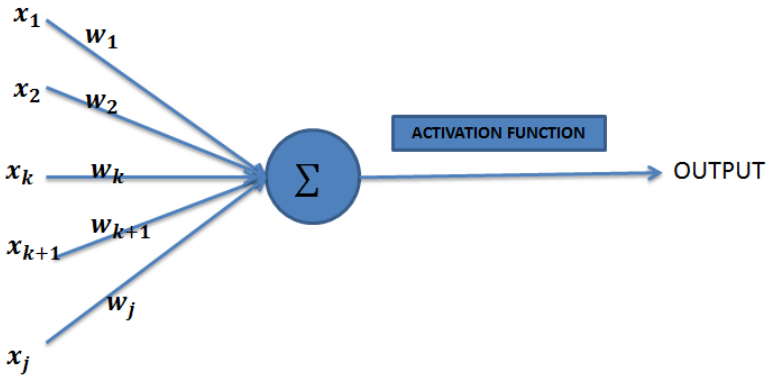


Figure 2 shows a typical ANN where the x_i for $i = 1, 2, \cdots, j$ are the input vectors, the w_i for $i = 1, 2, \cdots, j$ are input vector weights and

$$\Sigma = \sum_{i=1}^{j} w_i x_i$$

is the sum of each weight times the input vector [Cheng and Titterington, 2000]. Let $y_i = \sum_{i=1}^{j} w_i x_i$ be the net input of a neuron then there exists an activation function that gives an output, that is

$$f(y_i) = h(\sum_{i=1}^{j} w_i x_i) \tag{2.5}$$

where $f(y_i)$ is the output from h a sigmoid or linear activation function. The sigmoidal function can be of the form of a hyperbolic tangent, logistic, radial basis function etc. A sigmoidal function is an S shaped curve.

Figure 3: Logistic and hyperbolic tangent sigmoid functions

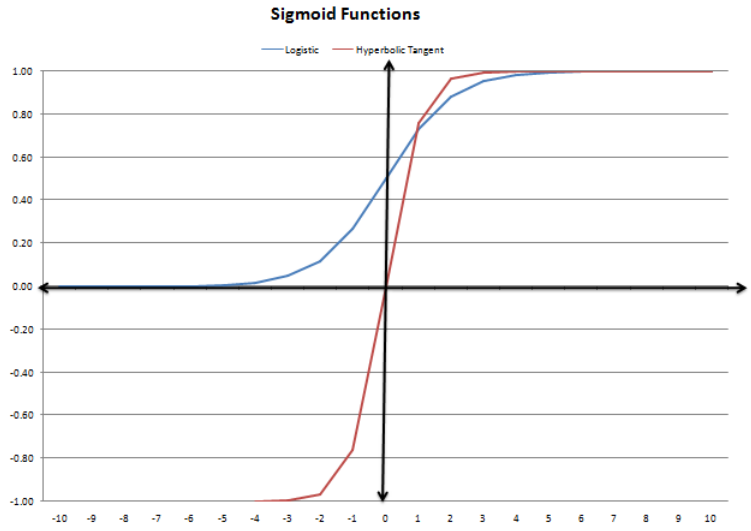


Figure 3 shows logistic (in blue $f(x_i)_{log}$) and hyperbolic tangent (in red $f(x_i)_{tanh}$) sigmoidal functions [Turhan, 1995]. The logistic sigmoid is asymptotic to the lines $f(x_i) = 0$ and $f(x_i) = 1$ while the hyperbolic tangent sigmoid is asymptotic to the lines $f(x_i) = 1$ and $f(x_i) = -1$. The two sigmoid functions are continuous and differentiable on $x_i \epsilon [-\infty, \infty]$ interval [Turhan, 1995]. Furthermore,

$$f(x_i)_{tanh} = 2f(x_i)_{log} - 1 \quad (1)$$

$$= \frac{2}{1 + \exp(-x_i)} - 1 \quad (2)$$

$$= \frac{1 - \exp(-x_i)}{1 + \exp(-x_i)} \quad (3)$$

In this research, we looked at a feed forward artificial neural network and used the hyperbolic tangent as a sigmoidal function. A feed forward neural network has a hidden neural network structure such that the message gets passed from the first neuron to next one but the message is not returned back.

Figure 4: A feed forward neural network with two hidden layers

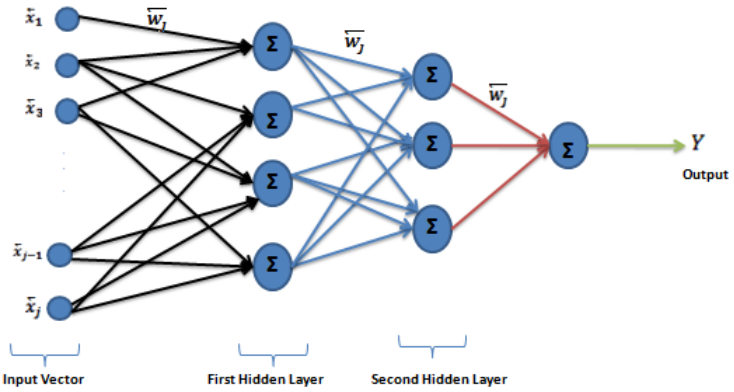


Figure 4 shows a typical feed forward neural network architecture where the w's are the weights of each neuron, Y is the output and the x's are the input vector. When fitting an artificial network model we try to find the unknown weights w_j by minimising the error of the output from the estimated weights. Optimisation techniques such as Back-Propagation, Newton-Raphson and other techniques are used to estimate the w_j. Two problems that may arise from fitting ANN is getting the starting values of the weights to be estimated and over fitting the neural network model. A zero value can be used as a starting point of estimating weights and the early stopping rule in the optimisation technique can be used to avoid over fitting.

2.2 Model evaluation

We evaluate the models using; Bayes Information Criterion (BIC), Receiver Operating Characteristic Curve (ROC), Akaike Information Criteria (AIC), misclassification rates and the lift charts because these are the commonly used evaluation criteria methods.

2.2.1 Bayes and Akaike Information Criterion

The AIC and BIC measure the performance of a statistical model for a data set being analysed. These models depend mostly on the likelihood function and will penalise models with higher numbers of parameters. The main idea is to see which models are over fitting the data amongst the ones that are being compared. These measures are calculated as below:

$$AIC = -2\log(l) + 2k \tag{2.6}$$

and

$$BIC = -2\log(l) + k\log(n) \tag{2.7}$$

where l is the likelihood value, k is the number of parameters and n is the total number of observations. As the model becomes more complicated, k number of parameters used to estimate the model will increase, the AIC and BIC values will increase and the model will be over fitting the data.

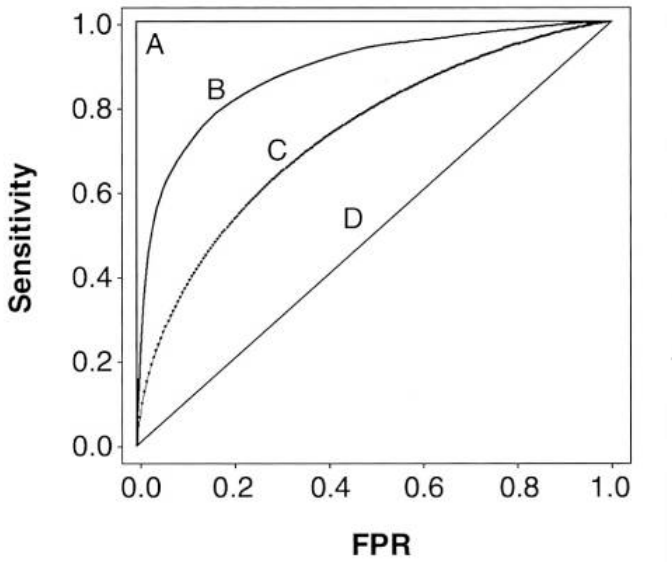
2.2.2 Receiver Operating Characteristic Curve

The ROC curve measures how well the model fits by plotting the false positive and negative fraction and evaluating the area under the fitted curve. Given that we have a class that we want to predict (that is, customer defecting or not) and a given set of data divided into training and test sample. The model is built on the training sample and evaluated on the test sample. For the ROC curve we will be looking at the following:-

- True Positive and Negative Fraction: Predicted to defect in the training sample and actually defected and predicted not to defect and actually not defecting
- False Positive and Negative Fraction: Predicted to defect but does not defect and predicted not to defect but defect.

A best fitting model is the one with the lowest error rate, that is, with low false positive fraction. As a retail mobile telecommunication company you would want to reduce these errors. The ROC curve is then a plot of $sensitivity$ (true positive rate) versus $1-specificity$ (true negative rates/false positive rates). Figure 5 shows a typical ROC curve.

Figure 5: ROC Curve



The 45 degree line ($y = x$ which is labelled D) signifies a worthless model, curve C show that the model is performing better, curve B is better fitting test and curve A is the perfect model [Zou et al., 2007]. The higher the area under the ROC curve the better is the performance of the model. The AUC (area under the ROC curve) is an element of the set $[0, 1]$ [Gatsonis, 2008]

2.2.3 Lift Charts

The idea of the lift chart is that as a marketing firm you do not want to email or SMS all your customers for a promotional offer. Imagine doing this for a base of 1 million customers at a cost of 20 cents and only 500 customers respond to the offer of R10. The cost of sending these SMS's will not be recovered in this case, thus the business will lose a lot of money. The lift chart assists the business in identifying and selecting only the top customers that are likely to respond to the marketing offer rather than using random selection. Measuring a statistical model using the lift curve is done by ranking customers with the highest probability of responding and evaluating the number of the correct predicted customers that actually respond to the campaign at a certain population proportion.

To define this in detail, let S_c be the percentage of customers with highest ranked probability of churning when selected and P_0 be the proportion of customer selected from the whole population of churners and non-churners then

$$Lift(P_0, S_c) = P_0/S_c$$

As the proportion of the population selected increases, the lift value tends to 1 and in fact

$$Lift(100, S_c) = 1$$

and the maximum lift attainable is $1/S_c$ [Kno, 1999]. A lift of a random model is 1 for all P_0 values.

3 Literature Review

The research involved looking at relevant literature and detailed review of sixteen papers by the researcher that focused on churning of customers in industries such as banking, telecommunication and other retail sectors. One of the paper reviewed concentrated on sampling techniques when a class of interest is rare. We have applauded and questioned some of the literature based on their approach toward solving the churn problem.

3.1 Credit Card Churn Forecasting

In this research two data mining techniques were used to build a churn prediction model using credit card data from a Chinese Bank [Nie et al., 2011]. The authors defined data mining as discovering knowledge and patterns from a large data set. They argued that it costs a lot to acquire new customers so it is important to retain existing high value or profitable customers. In the paper they argued that a bank can increase profits by up to 85 per cent by an improvement of 5 per cent in the retention rate. As the economy develops in China, a large number of credit cards have been issued however most of these credit cards were inactive. With an increased competition in the banking sector, it is easier for a customer to exercise their right of switching the product if the current service is not satisfactory.

In this study churn was considered from a customer's initiation point of view, for example

- More favourable competitor pricing
- False information given to customers from acquisition
- Customer expectation not met etc.

and not by customers that churn because of the bank's initiation (for example bad debt). A sample of customers was taken from the database and divided into two time frames. A churner was then defined as a customer with no transaction at a chosen time period t (after) and the customer did make a transaction at a previous time $t-1$ (before). In this paper they used logistic regression and decision trees to predict churn. They also emphasized that these two methods work well in classification problems. The models were validated using percentage of correctly classified, GINI coefficient and ROC curve. They considered two types of errors:-

- Type 1 error: customer did not churn but is classified as a churner
- Type 2 error: customer churned but were classified as a non-churner

The model selection was also based on which model costs the most when selected, that is, the actual currency cost of marketing to the customers that were classified as churners but did not churn. A random sample was selected from a database of 60 million customers from January 2005 to April 2008. The data contained customer's demographic information, transaction information, abnormal card usage and other transactional activities with the bank. The time period was divided into observation period (where the number of total transactions was counted) and the evaluation period (where they check if the customers that were transacting before are still transacting). Out of 135 variables, only 95 variables were included in the final model. This is because some of the variables were found to be correlated (multi-colinearity) and this would have affected the model performance if they were included. Logistic regression and decision tree models were compared; they both showed that the demographic variables were not significant in predicting the churn rate. The activity level variables contributed more to significance in the models than the demographic variables and hence the model with these variables performed better than the model without them (for both models). Logistic regression model performed better than the decision trees and gave less cost on error (decision tree cost = 85283 and logistic regression cost = 80377).

3.2 Data Mining Techniques for the Evaluation of Wireless Churn

The authors of this article start by explaining the fact that the wireless mobile telecommunication industry is very competitive [Ferreira et al., 2004]. As wireless companies grow in numbers customers are faced with wider options to choose from which best satisfy their needs. They explain that there is a battle of advertisement within wireless companies in order to lure customers to change their mind and switch to utilise their services. Churn was defined as abandoning your service provider as a customer and moving to a competitor. Churn is recognised as a crucial issue in consumer business and economics. The author emphasises that predicting churn beforehand can help in retaining high value customers by giving them counter offers and thus saving the business money.

Their dataset came from a wireless carrier in Brazil with a sample of one hundred thousand customers and for a time period of nine months. A churner was defined based on termination of service before the ninth month and this was used as a target variable. From the dataset 1.25 per cent was a monthly churn rate which is very small when trying to model customer churn. The authors overcame this problem of very low churn rate by oversampling. This had an implication on the data and the accuracy of the churn model. The authors used the below variables for predicting churn:

- Billing data (roaming cost, revenue, etc.)
- Customer demographic data (gender, marital, region, etc.)
- Customer relationship data (rate plan, handset age etc.)
- Market data (competitor rates etc.)
- Usage data (airtime, data bundles etc.)

In total there were 37 data attributes (behavioural and demographic variables). These variables were transformed and standardised for modelling purposes. The authors then divided the data into two:

- Simple dataset where no modification was done
- Enhanced dataset where the features were reduced using Least Square Estimation and other methods

Using the feature selection methods, it was found that variables related to the airtime consumption by customers were decisive in defining churn. The two data sets were then standardised. The enhanced data representation had 10 variables while the simple data representation had 20 variables. The data was divided into 70 per cent training set, 20 per cent validation set and 10 per cent test set.

Four models were then run on the data set namely neural networks, decision trees, hierarchical neuro-fuzzy system and genetic algorithm rule evolver. The neural network model had optimal number of hidden layers determined empirically and was trained by back-propagation. The cost of each model was evaluated based on the assumption that 50 per cent of the churners that are offered incentive will be retained, the cost of incentive is 25 dollars, average monthly subscription is 80 dollars and only 20 per cent of those predicted as churners are contactable. Based on a total of two million subscribers for this company, results showed that using a neural network model on enhanced data representation can save the company a large sum of money (44.2 dollars

per client that is likely to churn). The models performed on the enhanced data representation set yielded better results than using a simple data set for all model. Neural network model with fifteen hidden units outperformed the other models.

3.3 Customer Relationship Management at Pay TV

Pay TV is a European company that offers premium channel viewing to subscribers [Burez and den Poel, 2005]. It offers entertainment, news and educational channels to its viewers. Pay TV has a huge database of active customers but in recent years the number of active customers started to decline. It was speculated that the churn was caused by higher fixed cost to customers because it was expensive to maintain Pay TV infrastructure. In this research, they mentioned the following marketing initiatives to try and reduce customer churn

- Give customers free services
- Organising special events to pamper customers
- Survey study on customer satisfaction

In this research, they mention two ways of reducing customer churn. The first of which was an untargeted approach, which is mass marketing to every customer. The second was a targeted marketing approach to customers with a higher probability of churning and provide them with lucrative offers.

Similar to DSTV, if you subscribe to Pay TV you only pay a monthly subscription fee. There are no other charges except for pay per view which was not discussed in this research. The subscription is a twelve months contract by which cancellation before the end of twelve months is not allowed. Customers need to inform Pay TV if they will terminate the contract after twelve months; if this is not done then the contract is automatically renewed. The data was divided into two time buckets that is estimation period (from start of Pay TV to sampling date) and follow up period (a year after the sampling period). Variables that were extracted from the database were:-

- Previous and current subscription
- Demographic (e.g. Age, gender etc.)
- Number of payment reminder notifications to customers

A logistic regression technique was used in this research motivated by its simplicity and because it is widely used in market research. Monthly instalments amounts were used as the class variable. Markov chains were also used and the basic was that customers can move from having product 1 (premium say) to a lower product 2 (say compact). Moving within these two states can influence the probability of churning. Random forests were also used as an additional model. The models used were evaluated by Cumulative lift curves and ROC curve. Random forests outperformed other models and gave the best fit and best cumulative lift curve. Furthermore a field experiment was conducted on the customers with a high probability to churn. Customers were given incentives and response was analysed. It was found that the incentive reduced churn significantly.

3.4 Partial Defection of Loyal Clients

In this research the authors discussed customers partial defection from a Fast Moving Consumer Goods non contractual setting [Buckinx and den Poel, 2004]. In this retail setting customers can change their purchasing behaviour without informing the company about it (for example, in a retail setting where customers do not have loyalty cards). Again, because of high competition in the retail setting it is easy to switch brands. For example, some customers may be price elastic that is, a small increase in price will cause them to switch retailers. They also emphasize looking at customers that are profitable and showing loyal behaviour for retention.

In this research, they looked at two time buckets and looked at behaviour at time 1 and time 2. They then looked at purchasing behaviour in both periods, if there is a change in the negative direction in time 2 then the customer was classified as being partially defected. In this research they used three classification techniques:-

- Logistic regression
- Neural networks
- Random forests

The evaluation criteria used were percentage corrected classified (PCC) and the area under the curve (AUC).

In this study they selected only the behavioural loyal customers for analysis satisfying the following conditions:-

- The frequency of shopping is above average
- Ratio of the standard deviation σ_t of the inter-purchase time to the mean μ_t inter-purchase time is below average

The data chosen for this study contained customer behavioural and demographic attributes. One may argue that most variables that were used in this study were correlated which may have caused bias in the predictions. Random forest outclassed neural networks and logistic regression techniques. The content of this paper is very powerful in the sense that it looks at partial changes in customer behaviour so that corrective initiative can be applied early enough before a customer totally defects.

3.5 Customer Headroom Model

This paper talks about basket analysis in a retail setting in which some baskets were believed to have a missing spend [Shashanka and Giering, 2009]. For example, if a customer usually buys only bread in a store yet it is known from previous experience that bread is associated with butter or milk (say) then there is a possibility that the customer is buying these products from another retailer or the customer does not consume these products at all. If a customer has this property then they can make an initiative to try and cross sell products that are highly associated with the ones that are in the customer's basket.

Customer's transactional data was extracted for all customers who shopped in the sampled time period using their loyalty cards. Log normal distribution of customers total spend and spend in each item was assumed because the data was skewed. Cross shoppers and customers that buy for large communities were excluded from the analysis as they were outliers and will distort the results. Customers spend, frequency, items bought, number of distinct items bought and demographics variables were used to cluster customers into sub-regions. Each sub-region or segment was modelled on its own for an increase in accuracy of the prediction. Singular Value Decomposition was then used to predict customer's potential spend in each subgroup

3.6 Churn Prediction Model

The authors explain how costly it is to recruit new customers in mobile telecommunication retail settings where the service providers are faced with high churn rates. Churn is a highly debatable research area not only in mobile telecommunication but also in other industries [Shaaban et al., 2012]. Data mining techniques have helped service providers to reduce customer churn. The authors defined churners as voluntary and involuntary where by voluntary churn is incidental (unplanned churn) and deliberate (price elasticity, better service and offers). Service providers are concerned with deliberate churn and thus creating a predictive model for this is important. The author mentioned the most frequently used data mining classification techniques with their advantages and disadvantages. These techniques are:

- Decision trees
- Regression analysis
- Neural networks
- Fuzzy logic

The authors sampled 5000 records from a database which was not mentioned and divided it into 80 per cent training and 20 per cent test data set and both train and test data set had a churn rate of 0.2. The data mining and analysis program used by the authors was WEKA. There was a total of 23 variables select from the database and they included demographic, calls and billing data. The authors used decision trees, neural networks and support vector machine for modelling churn and found that neural networks and support vector machine performed better (both 84 per cent model accuracy) than decision tress (78 per cent model accuracy). The authors selected support vector machine as the best model because although the model accuracy rate is the same as neural network model, the support vector machine model is able to pick up more customers that are predicted to churn and they do churn (421 true positives for support vector machines and 403 true positive for neural network model). The authors created three cluster groups of customers (low, medium and high value) based on the 23 variables. We agree with the authors of this paper because:

- It can be clear from the retention program which cluster performs best (more customers are retained)
- Cost can be saved by targeting a cluster that is likely to respond rather that clusters that do not respond

- High value customers can be targeted since they are loyal and profitable to the organisation

3.7 Churn Prediction in the Mobile Telecommunication Industry

In this research Alberts started by explaining why was there a need for predicting customer churn [Alberts, 2006]. In the Netherlands there has been a rapid change in the mobile telecommunication industry, from a growing market to saturation and highly competitive market. Therefore most companies are no longer investing in acquiring new customers they rather invest in retaining the existing ones. It is easier for a customer to switch from one service provider to another because of high competition. The study was carried out for Netherlands Vodafone.

The author used two data mining techniques for predicting churn namely: The Cox survival model and decision trees. These techniques predicted a class of belonging (churner or non-churner) by a certain probability value. In this research the author does not focus on contract customers but only postpaid (prepaid) customers. It is also much easier to predict churn for contract customers because the expiry date of the contract is known. In the research churn was defined as stopping to use the company's services by:

- Voluntary: when the customers switch by choice (say to competitors)
- Involuntary: customers churn because of missed payments or fraud (say)

The proposed research question was the feasibility of modelling churn of prepaid customers using survival and decision tree model. The shortcoming was on how one measures the churn of prepaid customers since there is no specified end date as in a contractual setting. Do survival models have an added value compared to decision tree predictive model? The author defined four states that a prepaid customer can be in:

- Normal use: normal active customers with credit on the prepaid account (1)
- No credit: zero credit in the prepaid account (2)
- Recharge only (3)

- Deactivation: 'churn state' (4)

A customer can move from state 2 and 3 to the normal state after recharging. In general, it takes longer for a prepaid customer to be disconnected in a network. So in many instances prepaid customers churn before they have been disconnected. The paper looked at prepaid customers that have been completely disconnected.

The data was taken from a Vodafone database and was aggregated monthly for each customer. Twenty thousand customers who joined between April and July 2005 were sampled and analysed. In addition the data contained demographic and activity level with Vodafone variables. Some of the selected variables were:

- Number of months since last recharge
- Number of months since last voice mail
- Ratio of incoming call to outgoing calls

The data was manipulated and it was represented as survival data and then Cox Model was fitted. Some customers churned in the sampled period others were censored. Since survival models are not mostly used for classification or prediction, the author used a specific procedure to do this [Ripley and Ripley, 1998]. A hazard function and instantaneous probability was used for this. A predetermined threshold was used and if the hazard function was above this then these customers were churners [Poel and Larivire, 2003]. On using decision trees the data was divided into test and train sample for validating the model. The splitting criteria or variable importance selection that was used was the GINI co-efficient. The problem of over fitting was avoided by pruning the trees that hold the low information. The decision trees outperformed the Cox survival model but the survival model had an advantage over decision trees in that the survival model takes the time aspect into consideration by means of using a baseline. So the author does not only know which customer will defect but also what is the expected time until the customer defect is.

3.8 Analysis of Clustering Technique for Customer Relation Management

This paper reviews different types of clustering techniques used in Customer Relationship Management [Manu, 2012]. Manu defines clustering as creating a group of objects based on their features or attributes in a way that the objects belonging to the same groups are similar and those in different groups are dissimilar. He also mentions that clustering plays a significant role in pattern recognition, text mining, web analytics and customer relationship management. Data mining adds a complexity in the sense that you can have a huge data set with many attributes. The way they defined the components of the clustering task was by using the following steps:

- Pattern Proximity: a distance measure on pairs of patterns (there are various distance measures functions)
- Data Abstraction: extracting a data set
- Cluster Validity Analysis: cluster analysis and validating clusters

In the paper they represented a feature vector of a single data point as

$$X = (X_1, X_2, \cdots, X_p)$$

with p being the dimensions of the space, X is the pattern or vector and the $X's$ are the attributes. The attributes of this feature vector can be qualitative (nominal) or quantitative (continuous or discrete). In the paper they focused on the data with continuous attribute and use Euclidean distance as a measure of similarity ($\sqrt{\sum_{k=1}^{d}(X_{i,k} - X_{j,k})^2}$). Other texts suggest ways of dealing with qualitative data when performing cluster analysis [Linoff and Berry, 2004, Friedman et al., 2008]. The author mentioned the disadvantages of having linearly correlated data when clustering which can distort the distance measure. In such instances one can transform the data using whitening transformation or using the Mahalanobis distance $d_m(x_i, x_j) = (x_i - x_j)\sum^{-1}(x_i, x_j)'$ where x_i, x_j are row vectors and $\sum^{-1}$ is the inverse of the covariance matrix of the $x's$. The author went on to define many clustering techniques with their advantages and disadvantages.

3.9 Churn Prediction in Telecommunications

In this paper which is relevant directly relevant to our story, the authors started by explaining why it is important to maintain customers in a Telecom-

munication retail setting [Idrisa et al., 2012]. If high value customers are lost then the company's revenue will decline significantly. This creates a need to develop a churn probability model that will predict customers that are likely to churn. The authors mentioned that in this setting the dataset has high dimensionality and an imbalanced class distribution. High dimensionality arises from a data set having many behavioural and demographic variables while the imbalance arises from the fact that in general, there are many more non-churners than churners. The imbalance may cause high misclassification rates in the model.

The authors processed the dataset to check for missing values and transforming the nominal values. Below is how the data was processed before applying the classification methods:

- Dataset with useful fields was extracted from the database
- Useless features are removed and the data was reduced in size using principal component analysis.
- Nominal features (70) were transformed to numerical values by grouping into three categories
- Data was further processed by applying Random Under Sampling (RUS) and Particle Swarm Optimisation because churn class rate was low (7.3 per cent)
- Principal Component Analysis, Fisher's Ratio, F-score and minimum redundancy maximum relevance methods were applied for selecting the features to be used in the model.

K nearest neighbour and Random Forest were applied to the datasets in order to predict customers that are likely to churn. These classification techniques were firstly applied to original dataset without any feature selection method applied and then applied to the data set with feature selection methods (four methods). The model performance was evaluated using Area under the Curve (AUC). Random Forest and K Nearest Neighbour performed better when features were selected using minimum redundancy maximum relevance were employed rather than using Principal Component Analysis, Fisher's Ratio and the F-score. The author concluded by stating that using minimum redundancy maximum relevance feature selection and Random Forest model was efficient for predicting churn in the Telecommunication retail setting where the data set is large and high computational costs are involved. The authors complained about the imbalanced class and did enhance the data by

using under or over sampling techniques. These techniques did improve the model performance.

3.10 Turning Telecommunication Call Details to Churn Prediction

A rapid increase in mobile telecommunication service providers has led to high competition [Wei and Chiu, 2002]. In order to survive in such a competitive environment businesses nowadays rely on data mining techniques in order to gain advantage over their competitors. The authors of this article mention that churn management and customer retention is the key in business success in the telecommunication industry. Data mining (information discovery) can be classified into classification, clustering, dependency analysis, data visualisation and text mining as per authors view. In this paper they argued that the use of demographic variables when predicting churn may be misleading because:

- Churn is at customer level rather than contract level as it is common for a customer to have more than one contract
- Often customer databases in mobile telecommunication industry usually don't have substantial demographic information

They analysed churn data for contract customers by using their call pattern changes. They also argued that using call pattern changes (for example, diminishing incoming or outgoing calls) can be used as a signal for churn. The data was taken from a Taiwanese mobile telecommunication provider which has a monthly churn rate of between 1.5 to 2 per cent. The class variable for this analysis was derived from contract end date. The data contained 114,000 customer call records made between October 2000 and January 2001. This data set excluded customers whose contract was terminated based on delinquency. The authors had prior information about the variable that mostly influence churn from the company managements. These variables were:

- Length of subscriber's services
- Payment type (debit order or over the counter)
- Contract type (there are different rates for different contracts)

The call patterns were described based on the three variables:

- Number of minutes for outgoing calls
- Number of outgoing calls made
- Number of distinct people contacted

In the sample data set, a T period was divided into k "sub-regions" in order to evaluate the change in customer patterns. In the data set there are between 1.5 to 2 per cent instances of churn, so the author decided to use multi-classifier class combiner approach. This approach is similar to over-sampling approach in the sense that the small class sample was replicated across different train-test sets while the bigger class was selected at random. A prediction period P was chosen at random from T where churners where defined as having a disconnected status at this period and if the status was active at the end of P then the customers were defined as a non-churner. They also mentioned that there was a retention period R after T and P which allowed the company to offer incentives to keep their customers. They mentioned that data mining techniques are widely used for predicting churn and they used two models (which were not mentioned) on 10 fold cross validation data set. They were mostly concerned about finding the sub-periods where call patterns change and in which prediction period do the models have high accuracy. The model evaluation criteria used were the cumulative lift curves and false alarm rates. The best model gave a lift of 4.68. They also built a model with demographic variables and found that it had a lift of 3.9 which was lower than the lift when no demographic variables are used. It was shown in this research that using behavioural variables for predicting churn is vital and it outperforms the model with demographic variables.

3.11 Churn Prediction Using Complaints Data

In this study the authors explain how valuable it is to maintain existing customers for the business [Hadden et al., 2006]. They also highlight that it is very costly to acquire new customers and with the rise in competition in the telecommunication industry, customers are likely to move to competitors. The authors explained that from past research it has been shown that predicting churn using demographic data is very unstable (Wei and Chiu, 2002). They argued that churn is dependent on customer and not on the contract and so they proposed using call pattern changes. In this paper they took a different approach to this as they predict churn using complaints and repairs data. They used three groups of variables to create the data set namely:

- Provision data: estimations that are made by the company with regards to resolving a complaint or repair
- Complaint data: information about customer complaints
- Repairs data: fault and repair data

We question the authors because they used only 202 customers to train the model with 50 per cent churners and 50 per cent non-churners whilst the test set contained 700 customers with 70 per cent non churners and 30 per cent churners. This data set was very small for training a model and the class ratios were not the same for both train and test sample. The results might be biased and misleading because the model was built on less churners and tested on data with more churners.

The authors used linear regression, regression trees and neural networks to train the data of 202 customers using Matlab and SPSS. The neural network model was performed by back propagation method with different activation functions and in addition a Bayesian neural network was used. The feed forward back propagated neural network using logistic sigmoid gave the best results when a probability threshold of 70 per cent was used for churners. The authors analysed the weights from the 24 variables that were used to develop the model and found that only seven variables were significant. It was not clear on how this variable significance process was done as the authors did not mention full details. The variables that held the most information were:

- Number of engineers arrived on site
- Customer years on book
- Length of repair
- Number of appointments for repair
- Time to resolve a customer query
- If an order has been placed
- Number of times that a specific repair has been done

The authors then used regression trees in order to assess risk of churning which provided an overall accuracy level of 82 per cent. The regression method performed in SPSS gave an overall accuracy of 81 per cent. Bayesian

neural network outperformed the other models for predicting churners and the best performing technology was the regression tree technology.

3.12 Churn Models for Prepaid Customers

The author of this article start by highlighting the importance of Customer Relation Management Department in customer retention [Owczarczuk, 2010]. In retention, the company tries to lure back customers that are likely to defect and in doing so there are cost associated with the process (marketing material) and bonus if the customer is retained. He argued that the retention projects must not target loyal customers as they will continue using the services of the company. We disagree with the author of this paper because neglecting loyal customers will lead to dissatisfaction and thus loyal customers will churn. In instances where the loyal customer base is very small and most profit is generated on the "non-loyal" customers then the author of this paper is correct.

The author worked on predicting churn for prepaid customers rather than contract customer. He argued that it was much simpler to predict churn for contract customers as they have all demographic information about the customers and the exact expiry date of the contract. The author did not want to define churn in a standard terms used in Poland (SIM expiration). This was mainly because if a prepaid customer makes a recharge in month one of sim card purchase then it takes 365 days of non-use for the card to expire. If the customer recharges a month later then the days to expiry (churn) are re-set to 365 days. The author felt like the period was too long and defined churn as having no incoming or outgoing calls in the last six weeks.

The data set was taken from a Polish mobile provider. It contained two years' worth of data (2007 to 2008) and it had 1318 variables (behavioural and demographics). The author used four models for predicting churn namely:

- Logistic regression
- Linear regression
- Fisher linear discriminant analysis
- Decision trees

The idea behind using these models was because of their simplicity and the

ease of interpretation. They mentioned that random forest and support vector machines as the black-box models which are unsuitable for predicting churn. We criticised the author by saying this because he did not had a valid reason as to why these are black-box models. Again we disagree with the author of this paper because these models may be suitable for a different or much more complex data sets than the one used in the study. The author was very cautious when extracting the data from the database because of attribute data type mix and the fact that on a relational database you do not want to accidentally use primary key field in your model. The author sampled 167,595 records and divided it into 51 per cent train, 22 per cent validation and 27 per cent test set.

The author argued that using regression and Fisher discriminant model in a high dimensional vector may lead to wrong conclusions because multi-co-linearity may arise. Also, there may be computational power problem involved. On each variable the t-test was performed and the variables were ranked according to t-score. The top 50 significant variables were used to fit the models. The model performance was obtained from plotting lift curves of each model in the same axis. Logistic regression performed slightly better than the other models. Decision trees were fitted to full data set (1381 variables) and enhanced data set (50 variables) and it gave similar results.

3.13 Mobile Telecommunication Handling in India

India has the second largest telecommunication industry in the world with more than 650 million active customers [Jamwal, 2011]. The author explains that in earlier years (1990's) there were fewer telecommunication service providers and in recent years there are about 17 service providers. This has created a lot of competition and the management in the telecommunication industry are mostly concerned and focused on maintaining existing customers. Our opinion differ with the author of this paper and the management because there is natural churn from death, migration and other so recruiting new customers should also be a priority even if the market is saturated.

The author was motivated to predict churn because the market has a churn rate of 27 per cent per year. This is very high (more than a quarter of customers are lost every year), knowing that it is costly to recruit new customers. The main problem is that it is difficult to predict which customer will churn

and the reasons behind it. Data mining techniques can help us predict churn from the database thus promoting competitive advantage. The author mentions that most organisations lack skills and expertise of data mining and analysis. We agree with the author totally because there is a gap between management and analysts. This gap is because management finds it hard to believe or understand analysts and they may base their decisions on gut feel rather than numbers. The main concern of the author in this research is why do customer churn and who is likely to churn. The author used Chordiant Predictive Analytics Director software to prepare the data and logistic regression and decision trees for modelling churn. The collected data set selected at random had demographic, call details and billing variables of each customer. A total of 15000 customers with a churn rate of 8 per cent were sampled.

From data exploration stage it was found that there was a higher probability of churn for age group of 45 to 48 than the average churn rate and for the customers whose contract are between 25 to 30 months. The customers that paid low monthly fee are more likely to churn and those that are billed less than 190NT in 6 months. Also customers that have less outgoing calls minutes had a higher probability of churning. From these results the author created KPI's (key performance indicators) flags on the database to signal customers that are likely to churn. We criticise the author for not mentioning the model that performed the best. Also the sampled data was too small for churn prediction considering the fact the Indian telecommunication companies have huge databases and an average of 27 per cent churn per year and this may create bias in the models used results.

3.14 Knowledge Discovery on Customer Churn

In this paper they reviewed churn in the retail mobile telecommunication space and used the same data set as in this research (data set 1: call pattern data). The author starts by explaining the importance of customer churn in business nowadays [WSE, 2006]. The business need to focus on getting more knowledgeable about its customers in order to maintain a quality service focus. The study focused on modelling customer churn in a Taiwanese company for prepaid customers who churned voluntary. On the other hand, involuntary churn, that is, customers that churn because of fraud and delinquency were not included in the analysis. Unavoidable churn customers, that is, customers that churn because of death and migration were included

in building the model. This is because the mobile service provider cannot differentiate this with voluntary churn.

The authors used a field test to monitor customers after they had been modelled as churners. This was different to most churn papers cited in this research. Most of them used historic data to predict churn but do not then monitor the customers that have a higher probability to churn in the next time frame. Below are the steps taken in this research:

- Data extraction in the database
- Data transformation and selection of desired variable
- Sampling for modelling
- Modelling and scoring the whole customer database (on SQL)

For model performance the author used hit rate and lift curves. Decision trees and logistic regression were used as classification techniques because of their simplicity and ease of interpretation. The author criticise using the neural networks for predicting churn heavily saying that the one cannot interpret the weights and calling this model a "black box". From the database there were 170 variables selected (containing demographics, billing, usage, call details data etc.) and explored using graphics and chi square. Based on a probability value of 25 per cent (univariate study), variables were reduced to 99. The churn rate in the data set was 0.5 per cent which was very low (but very big when taking into account a database size of more than 1 million records). Due to the low class ratio, the author decided to create bias in train and test data varying the churn rate from 1 to 10 per cent. The best decision tree model was obtained on a churn rate of 2 per cent and a sample size of 375,000. From the list of 5000 customers where a field experiment was conducted a 56 per cent hit rate was obtained. Decision trees outperformed the logistic regression methods. This paper showed that data mining method are applicable even in low churn rates.

3.15 Under-Sampling Approaches for Improving Predictions

The authors explain that the most important thing in classification problem is to improve accuracy in the training data [Yen and Lee, 2009]. It is normal for a data set to have an imbalanced class and when training a model the

majority class will be predicted more accurately than the minority class. A classification technique performs well when the class variable is evenly distributed. The author emphasises that given any data set with a class variable, data mining techniques can be used to train the data and predict the class in the test data set. The authors explained that the process of classification involves the following steps:

- Sample Collection
- Selecting features for training
- Training the data
- Predicting or forecasting the class of the new data set

We felt that the authors of this paper was missing a step of exploratory data analysis because it was not clear in the paper whether this step was included in step two (selecting features for training) or not.

Some authors have suggested techniques like over and under sampling and synthetic minority over sampling technique (SMOTE) to approach the problem of unbalanced class [Chawla et al., 2002]. In over sampling, the instances of the minority class are increased in order to reduce the imbalance. SMOTE regenerates the minority class instances from a sample using the nearest neighbour and create a new sample with more minority instances. The authors explain that generating more of minority class instances without taking into account the majority class can lead over generalisation. Under sampling approach can also be used to reduce the majority class in the data set. In this paper the authors used under sampling based on clustering method in order to overcome the imbalanced class. This was done by clustering the train data into some clusters, say k. In theory the cluster should be dissimilar and so the author evaluated the ratio of the majority class and the minority class in each sample. Based on the authors discretion a cluster with the desired ratio can be used to training the data but with the majority class selected random from each k.

The authors applied the method of under sampling using cluster method to a data set using IBM Intelligence Miner for Data (application) using a neural network classification technique. This under sampling technique was compared to other under sampling technique proposed by other authors on the same data set using neural network model. The data set comes from a 1994/95 United States of America census which contains income data. The

class variable to be predicted was the level of income (binary). There were 30,162 records in the data and the minority class was about 25 per cent. The author used 8 per cent to train the data and 24 per cent to evaluate the performance precision, recall and F-measures. We felt that this was not a very unbalanced data set as one of the cited paper in this research contained about 2 per cent of the minority class which was less than the 25 per cent churn class in this paper [Wei and Chiu, 2002]. In evaluating the performance the under sampling based on clustering method produced good result when compared to the other seven imbalanced class approaches. This method also proved to have a better stability and less run time that the other methods.

3.16 Examining Churn and Loyalty Using Support Vector Machine

The authors start by explaining that the telecommunication industry is amongst the fastest growing industries in the world [Dehghan and Trafalis, 2012]. Companies are offering a wide range of products and because it is hard and expensive to obtain new customers, they rely mainly on maintaining the existing customers. A highly loyal customer is less likely to churn. These types of customers are satisfied with the company's current services and they would like to keep the relationship with the company for longer. The author explained churn and causes of churn for example competitor offering lower prices. From the authors opinion, customer loyalty comes from actively using the service of the same company from a certain period so the decision of whether to churn or not depends on the account length. The authors gave a brief discussion of support vector machines and the uses but we disagree with the authors when they say that support vectors can only separate the data into two classes as they can also be used for multi class problems.

The data comes for this research came from University of California data repository. It contained call details variables and few demographic variables. The authors had to transform some of the variables into integer type as there were many categorical variables in the data. The author used Matlab for the analysis. Firstly, principal component analysis (PCA) was used in order to determine the most significant variables and eight variables accounted for the most variation in the data set, that is, 98 per cent. The authors removed 50 data entries from the training sample and put them into the test data and observed the model performance changes when he did this a 100 times. The

author used support vector machines to model the data. The accuracy plot for the class defect true and false shows that the optimal median accuracy was reached at 57.16 and 57.18 per cent. The optimal value of C was obtained by fine tuning.

We are concerned with the authors approach based on the following reasons:

- The choice of sigma was not mentioned in the study
- Only one model was performed and compared to itself. We think this was not competitive and sound

3.17 Literature Summary

In this section we summarise all the papers that were reviewed in the previous sections. Decision tree technique was the most common technique used. The authors emphasise that the main reason behind using this model was because of its simplicity. Random forests were used in three of the cited paper and they outperformed the other models in all of them. Support vector machines were used in two papers and came as the best model once. Neural network models were used in three papers and came out as the best model twice. The main issues that the authors came across were:

- How to define churners in a non-contractual setting
- How to deal with a rare class when modelling the data

4 Methodology

4.1 Analysis Process

In this chapter we explain the processes, procedures and techniques that were involved in analysing the data. Two data sets from different sources were used in this research, Data set 1 (*Call pattern churn data*) contained mostly information about calls and plan types variables while Data set 2 (*Billing and time taken to churn*) contained information about billing, credit type and contract date variables [Berry and Linoff, 2009b, Blake and Merz, 1998]. Data set 1 comes from University of California data repository and was collected over 51 regional states in United States of America for long distance customers. The sampling date for this data set was not specified in the website. Data set 2 comes from a mobile service provider also in the United States of America. The data contained eight years history of customers in the database sample from a period of May 2000 to August 2008. Both data sets contained few customer demographic variables. The data sets were small in terms of variables (20 for data set 1 and 23 for data set 2) and data set 1 had 3333 records while data set 2 had over 476223 records. Data set 1 contained 14 per cent of customers that churned over the sampled period and data set 2 contained 45 per cent of customer that churned on the sampled period. The data sets were from contractual telecommunication retail settings. Tables A1 and A2 in the appendix show the variables used and a brief description.

The report is structured as follows:

1. Understanding the data sets:
   Getting to understand the data is often the most difficult part in data mining. The data might be polluted, have missing values and contain wrong data types (for example a customer age might be a character field instead of an integer field), so getting it into correct format is vital for modelling. We also explored the relationship between variables in this step.
2. Sampling
   In this stage we split the data into training and test sample by taking random samples from the data set.
3. Analysis and Results, which involves the actual architecture and extensive modelling of the data sets.

4. Model comparison, in this stage we compared model performance and how they differ for the two data sets used in this research.
5. Conclusion and recommendations about the results
6. Summary and future research

4.2 Understanding the data sets

4.2.1 Data Cleaning

- **Data set 1**

The initial data set contained 20 variables and had no missing values so there was no need to use missing values techniques. The class variable "defect" was a binary variable (0 and 1) and converted to a nominal binary variable (true and false). Figures A1, A2 and A3 in the appendix show the distribution of the standardised variables. *State* was not used in model building because there were 51 states where each customer dwells. *Number of voice mail* which showed an undesired distribution (about 70 per cent of customers concentrated to the same point) was later transformed into a binary variable.

- **Data set 2**

This data set was polluted and we had to clean it. The manipulations and cleaning of the data was firstly done in Excel for 476233 records and 26 variables (including 3 derived variables). There were records that had missing customer's *date of birth*. These records were less than 0.1 per cent of the sampled data, so we decided to delete them from the data set. The *initial monthly fee* field also contained some missing value however we decided not to impute mean or median for missing values in this field. The main reason behind this was that the distribution of this variable which was not normal and did not allow this transformation. The distribution of this field is discussed in the data exploration section.

The *age* variable, in years, was derived using *date of birth* and the *sampling date* of the data. The resulting *age*'s were analysed and it was found that some values did not make sense. For example, some customers had a negative *age*, some with *age* more than 110 and others less than 16 years. This meant that there was a possible mistake that was made either during the data capturing or data storing phase. We decided to remove these records and we were left with approximately 4700 records. Time to defect was derived from the data set using contract start date and defect date. Class variable, *churn*, was derived using *defect date* and *account status*. *Account age* was also derived by using *sampling date* and the *contract start date*. After all these manipulation the data was uploaded to SAS JMP for exploration. Figures A5, A6 and A7 show the distribution of the variables in this data set.

4.2.2 Data Exploration

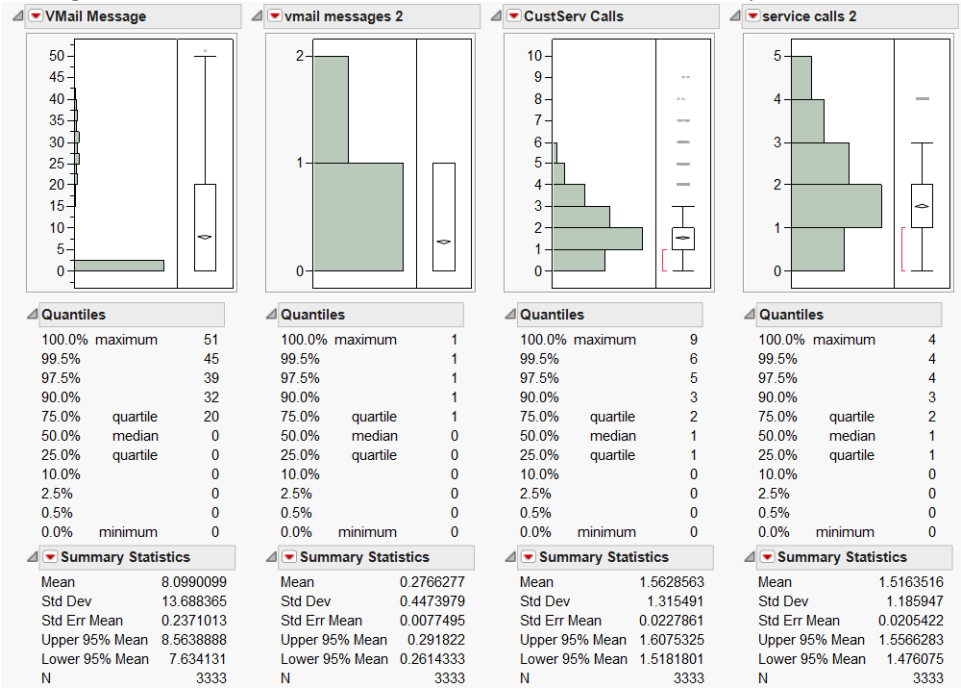
This process is critical in data mining before doing data modelling. It helps to understand the data set and the relationship between the variables at a high level. This process involves knowing the distribution of the dataset so that proper transformations can be done. Data exploration involves mainly descriptive statistics thus summarising the data using box plots, histograms etc.

- **Data set 1**

  We explored the distribution of each variable in the data and decided to transform some of the variables because they were not normally distributed. The *number of service calls* variable was not in a desired distribution and contained a lot of outliers and some values were concentrated at the same point. The *number of service calls* ranged from 0 to 9 (more points were concentrated at 0 and less points concentrated at 9). We then decided to transform this variable into an ordinal variable whereby $0, 1, 2, 3$ and $(4, \cdots 9)$ *service calls* were coded as $0, 1, 2, 3$ and 4. The *number of voice mail messages* variable was not normally distributed since most of the customers had no voice messages and some points were scattered across. There were outliers (range between the mean and the maximum value was large and the standard deviation 13.8 was higher than the mean 8) and this variable ranged from 0 to 51. The mean was not the centre point of the data and the standard

deviation was large. This led us to transform this variable into a binary variable. Zero represented customers that had no voice messages and one for customer that had one or more voice messages. Figure 6 shows the distribution of *service calls* and *number of voice mails* before (*vmail message, custserv calls*) and after (*vmail messages 2, service calls 2*) the transformation. The median was far away from the mean for both variables before transformation and they had a large standard deviation.

Figure 6: Distribution of *service calls* and *number of voice mails*



We then went on to standardise these variables (excluding ordinal and nominal variables) to force them to follow a standard normal distribution thus putting them in the same scale for modelling purposes. We want variables to have an equal contribution in modelling phase in order to arrive at correct conclusions. It was noted that all continuous variables distributions were almost following normal distribution. We then explored the correlation and correlation plots between the variables because, when fitting a model using variables that are correlated, the results and the estimates thereafter are biased and may be misleading. Below are the variables that were found to be perfectly correlated (correlation = 1):

- *Day call charge*: This variable was correlated with the *number of day calls* and the *number of day minutes*.
- *Evening call charge*: This variable was correlated with the *number of evening calls* and *evening minutes*.
- *Night call charge*: This variable was correlated with the *number of night calls* and the *number of night minutes*.

This led us to remove correlated variables (*Day call charge*, *Evening call charge* and *Night call charge*) in the model fitting stage because of redundancy (see Figure A9 in the appendix section for the correlation table). We then went on to find a relationship between each of the remaining variables and the class variable. Below are the variables whose relationship with the class variable stands out more than the other variables (see Figure A4 in the appendix section).

- A logistic bi-variate fit between the class variable and the *number of service call* variable. It was found that there was a strong relationship between the two variables and in fact as *number of service call* variable increased the number of defected customers increased (R-square of 0.1). The Chi-square value of this fit was 204.10 and was significance at 1 percent.
- A logistic bi-variate fit between the class variable and the *number of day minutes* showed that customers with few day minutes are more likely to defect than customers with more day minutes (R-square of 0.1). The Chi-square value of this fit was 221.11 and was significance at 1 percent.
- A logistic bi-variate fit between the class variable and *total international calls* showed that as the *number of international calls* increases the likelihood of defecting decrease. The relationship was not as strong as with the other variable because the R-square value was 0.002 and the Chi-square value was 11.52 and but the fit was significant at 1 per cent.

- **Data set 2**

The distribution of each variable was explored before transforming the data set. *Account age* and *customer age* were transformed by a log function because their distributions were skewed and not normal. After transforming these variables they were then standardise into a normal

scale. Some variables were derived using date variables in the data set as discussed earlier. *Deposit* was converted to a binary variable, that is, all those who paid deposit were represented by one and those who did not pay deposit by zero. *Channel* variable was converted into a binary variable with all the customers that were acquired through indirect marketing represented by zero and those who were acquired through direct marketing by one. *Market* variable was transformed into an ordinal variable with customers belonging to New York market as 1, Chicago market as 2 and Seattle market as 3. The *marker colour* was also transformed into an ordinal variable with unknown colour as 0, red as 1, orange as 2, blue as 3, yellow as 4 and green as 5. *Auto pay* was already in a binary form. We then analysed the correlation of all variables and it was discovered that:

- *Time to defect* had a strong correlation with *account age* ($\rho = 0.588$), however this variable was not used to model the data because of it high correlation level with the class variable, that is, a customer had a *time to defect* value greater than zero if they had churned
- There was also a weaker negative correlation between *account age* and handset *maker* (-0.4562)
- Other variables had partial or no correlation with each other.

Figure 7 shows the correlation values for some variables in the data set

Figure 7: Correlation table for data set two

VARIABLE	DEPOSIT_1	Autopay_1	Market_1	Channel_1	Maker_1	Time_to_Default	Acc_age	INITMONTHLYFEE	AGE
DEPOSIT_1	1.0000	-0.0035	0.0119	0.0064	0.0092	-0.0126	-0.0144	0.0013	-0.0957
Autopay_1	-0.0035	1.0000	0.0222	-0.0491	0.0516	-0.0369	0.0073	0.0632	0.0064
Market_1	0.0119	0.0222	1.0000	-0.1553	-0.0698	0.0834	0.1688	-0.0738	-0.0875
Channel_1	0.0064	-0.0491	-0.1553	1.0000	-0.0110	0.0374	0.0423	-0.0128	0.0922
Maker_1	0.0092	0.0516	-0.0698	-0.0110	1.0000	-0.3072	-0.4502	0.4342	-0.0519
Time_to_Default	-0.0126	-0.0369	0.0834	0.0374	-0.3072	1.0000	0.5892	-0.1818	0.0241
Acc_age	-0.0144	0.0073	0.1688	0.0423	-0.4502	0.5892	1.0000	-0.2507	0.0810
INITMONTHLYFEE	0.0013	0.0632	-0.0738	-0.0128	0.4342	-0.1818	-0.2507	1.0000	0.0143
AGE	-0.0957	0.0064	-0.0875	0.0922	-0.0519	0.0241	0.0810	0.0143	1.0000

We explored the relationship of each variable individually with the class variable by doing a bi-variate study. We found that the relationship of these variables with the class variable stand out the most from the other variables. These variables were *account age*, *initial monthly fee*

and *customer age*. Please see Figure 8 for as logistic fit with the class variable.

- *Account Age*

  In the logistic plot (Figure 8) between the *account age* and *churn* variable (class), it is clear from the chart that as the *account age* increases the likelihood of the customer churning was very low. In business terms, this means that loyal clients (long on book) are less likely to switch to other service providers. The generalised R-Square value was telling us that approximately 0.25 of the variation in this data set was being explained by *account age*. This variable also had a large Chi-square (998.05) and the fit was significant at 1 per cent.

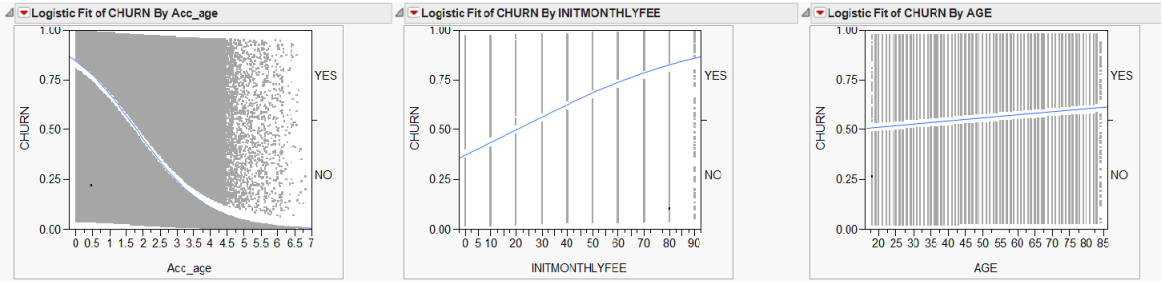
- *Initial Monthly Fee*

  In the logistic plot (Figure 8) between *initial monthly fee* and *churn* variable (class), it is evident that as the customer pay more of *initial monthly fee* the likelihood of defecting increases significantly. The generalised R-Square value shows us that the initial monthly fee variable explains about 0.16 of the variation in the data. In business terms this means customers that pay high initialisation fee are more likely to churn.

- *Customer Age*

  In the logistic plot (Figure 8) between *customer age* and *churn* variable (class), it is very clear that there was little or no change in likelihood of churning as customer age increase or decrease. This meant that *customer age* was not a good predictor of customer churning however it was important to note that customer age does not play any role in defection. Furthermore, this variable had a Chi-square value of 4.84 which was not significant at 1 per cent based on it p-value.

Figure 8: Bi-variate logistic plot for data set 2



4.3 Sampling

4.3.1 Stratifying the data

This section applies to data set 2 in this research. We could have not fitted a model using the whole population, because there were over 470 thousand records for each data set. This was mainly because of computer performance issue. Running a model for such big data set in a low gig ram computer can take long to finish (one day). The support vector machine model was fitted in R using all the records in the data and the model did not finish running. The computer had to be re-booted as R application was not responding. Sampling theory allowed us to develop models in a smaller data set where running these models take less time than using the whole data set. We sampled 1 per cent of the entire data set because we wanted to fit the model that takes less time to run. A simple method of sampling called stratification was used. The sample from data set 1 was stratified according to *state* (place where customers stay) and data set 2 was stratified according to numbering *plan area* (*NPA*) which was an area code that is, part of customer's phone number. This made sure that customers in each area are equally represented in the model. We did not over or under sample the data using the class variable because we felt that this was very biased and our class variable was not highly imbalanced.

4.3.2 Splitting the data

We divided the data sets into two samples (namely training and test sample) for both data sets. On the training sample, we fitted models and applied these models on the test sample in order to evaluate the performance. The split that was used when comparing all the models was 80 per cent training

and 20 per cent test sample. Kohavi suggested that one can use a validation sample and bootstrap methods to enhance the performance of the model from the training sample [Kohavi, 1995]. One can also use k-fold cross validation sample(s) to evaluate model performance and stability.

5 Analysis and results

5.1 Data Set 1 Results

5.1.1 Artificial Neural Networks

We fitted this model on the data set using the SAS JMP application. Firstly, this was done using a 3 hidden units hyperbolic tangent sigmoid neural network architecture. The main reason behind this was to evaluate if, on the standardise data set, the model would perform better than using the un-standardised data set. Secondly, the same neural network architecture model was fitted on the data set without the *number of service calls* variable. This was because we believed that the *number of service calls* variable contained the most information about the data and since the weights and the output from the program do not really tell a complete story about the variable importance, this was a simple way to look at it. Tables 2 and 3 show model performance based on un-standardised data (1) and standardised data (2).

Table 2: Training sample results for standardised and un-standardised data

Training Sample	**Un-standardised Data**		**Standardised Data**	
Measured Metric	**With Service Calls**	**No Service Calls**	**With Service Calls**	**No Service Calls**
R Square	0.588	0.424	0.614	0.546
Mean SE	0.233	0.274	0.232	0.245
Misclassification Rate	0.06	0.06	0.075	0.065
-Log Likelihood	695	908	659	752

Table 3: Test sample results for standardised and unstandardised data

Training Sample	**Unstandardised Data**		**Standardised Data**	
Measured Metric	**With Service Calls**	**No Service Calls**	**With Service Calls**	**No Service Calls**
R Square	0.542	0.435	0.603	0.53
Mean SE	0.239	0.271	0.235	0.24
Misclassification Rate	0.065	0.083	0.081	0.06
-Log Likelihood	379	447	338	384

It was evident from the tables that on both un-standardised and standardised data sets if the number of service calls variable was removed the model performed worse. We say this because there was a decrease in R square value

accompanied with an increase in the negative log likelihood and a slight increase in mean square error. Standardising the data helped us because there was an increase in R square value and a decrease in the negative log likelihood (meaning the model on the un-standardised data was over generalising compared to the un-standardised). It was not clear as to why the misclassification rate for the un-standardised data was better than that of standardised data. We examined this further by running six neural network models on the data set using six different random samples for each model. Each sample was divided into 80 per cent training and 20 per cent test sample. The neural network architecture was a multi-layer perceptron with three hidden units.

Table 4: Neural networks results before transforming the data

Before	**Average Square Error**	**Schwarz Bayesian Criteria**	**Train Misclassification**	**Test Misclassification)**
Model 1	0.179	1 570.962	0.0353	0.0353
Model 2	0.218	1 527.640	0.0552	0.0659
Model 3	0.182	1 226.232	0.0384	0.0422
Model 4	0.172	1 152.413	0.0314	0.0452
Model 5	0.217	1 449.130	0.0657	0.0703
Model 6	0.178	1 210.819	0.0355	0.0422

Table 5: Neural networks results after transforming the data

After	**Average Square Error**	**Schwarz Bayesian Criteria**	**Train Misclassification**	**Test Misclassification)**
Model 1	0.176	1 629.439	0.0329	0.0329
Model 2	0.203	1 468.321	0.0470	0.0444
Model 3	0.188	1 361.707	0.0406	0.0496
Model 4	0.198	1 441.521	0.0435	0.0577
Model 5	0.211	1 585.728	0.0470	0.0563
Model 6	0.200	1 486.565	0.0460	0.0592

Each result was very different and this meant that the model was unstable. This highlighted to us that there was more data exploration required. The *number of voice mail* variable was transformed to a binary variable and also the *number of service calls* was transformed to ordinal variable (discussed earlier). The resulting models were more stable than the previous models. Tables 4 and 5 shows the models result before and after the transformations were done. Although after the data transformation the models had a slightly higher Schwarz Bayesian Criteria (penalty value similar to BIC), the misclassification rates in both training and test data were more stable and the average mean square error did not vary much as with the previous models.

Furthermore, Figures 9 and 10 shows the lift curves for the models before and after transformation. It is evident from the two charts that there was a dispersion in the lift value at 10 per cent of the population before data transformation. After transforming the data this dispersion disappeared and the models seem to have almost the same lift value for all population percentiles.

Figure 9: Lift curves for the six neural networks before data transformation

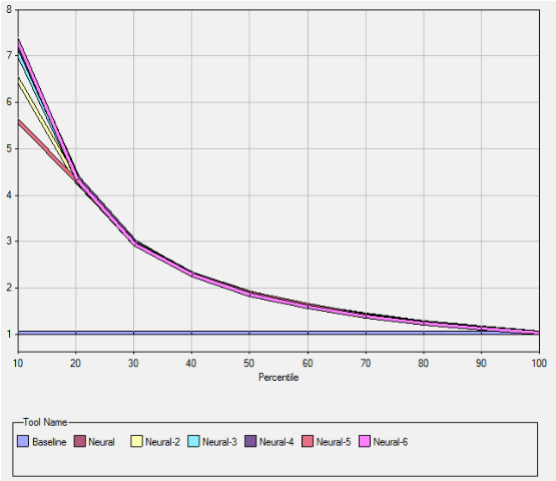
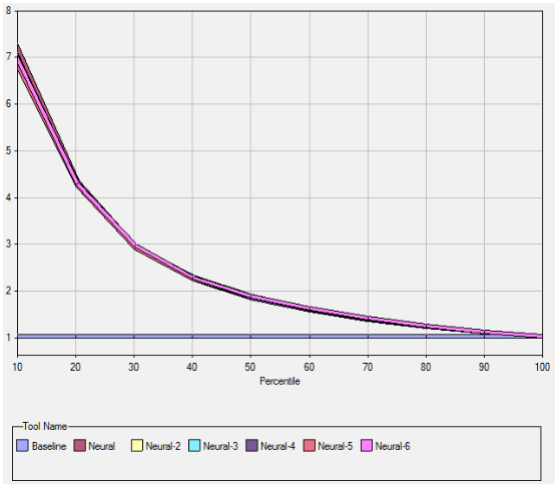


Figure 10: Lift curves for the six neural networks after data transformation



We had a hypothesis that changing the training and test sample ratios would impact the neural network results. This was because we believed that a model with more training instances would outperform the one with fewer instances. In order to test this hypothesis we divided the data into six uneven proportioned data sets as shown in Table 6 with a belief that data set A would perform better than data set F because data set A had more data points to train the model than data set F.

Table 6: Sample Test and Train Ratios

Data Set	**Training Sample (per cent)**	**Test Sample (per cent)**
A	90	10
B	80	20
C	70	30
D	60	40
E	50	50
F	40	60

Tables 7 and 8 show the overview results from the models fitted

Table 7: Train data model performance for data set 1

Train Data	**R Square**	**Misclassification Rate**	**-2Log Likelihood**	**Mean SE**	**True Positive Ratio**
A	0.756	0.035	603	0.179	0.77
B	0.757	0.037	532	0.179	0.78
C	0.712	0.061	540	0.205	0.57
D	0.766	0.036	387	0.177	0.78
E	0.723	0.051	374	0.195	0.64
F	0.721	0.041	299	0.191	0.72

Table 8: Test data model performance for data set 1

Test Data	**R Square**	**Misclassification Rate**	**-2Log Likelihood**	**Mean SE**	**True Positive Ratio**
A	0.72	0.044	74	0.191	0.72
B	0.71	0.048	156	0.198	0.73
C	0.65	0.059	272	0.216	0.61
D	0.73	0.041	290	0.189	0.76
E	0.6	0.0634	510	0.218	0.57
F	0.71	0.042	458	0.188	0.73

The R-square values for all the models did not vary significantly also there were no big variation in the misclassification rate. Data set C had the lowest

true positives for the training data and second lowest from the test data. The results showed that when we decrease the training sample size the negative log-likelihood decreases significantly which meant that the model with less training data had lower AIC and BIC value. Data set A produced similar results as data set F, this meant that neural networks had learned the data using less instances or data points. Thus the hypothesis we had was untrue, that is, in this case training a neural network model with less data yield similar results to the one trained with more data. Furthermore, we went on to confirm this by evaluating the model performance of data set A and F. This was done using ROC and the lift curves. Bearing in mind that an area under the ROC curve greater than 75 per cent meant that the model was performing very well and that the maximum lift we could have had was 7.14. This was because the data set had 14 per cent of customers that actually churned in the sampled period and the maximum lift at 10 per cent of the population was $1/0.14 \approx 7.14$. Figures 11 and 12 show the results obtained for the two data sets:

Figure 11: ROC and lift curves for ANN model data number A

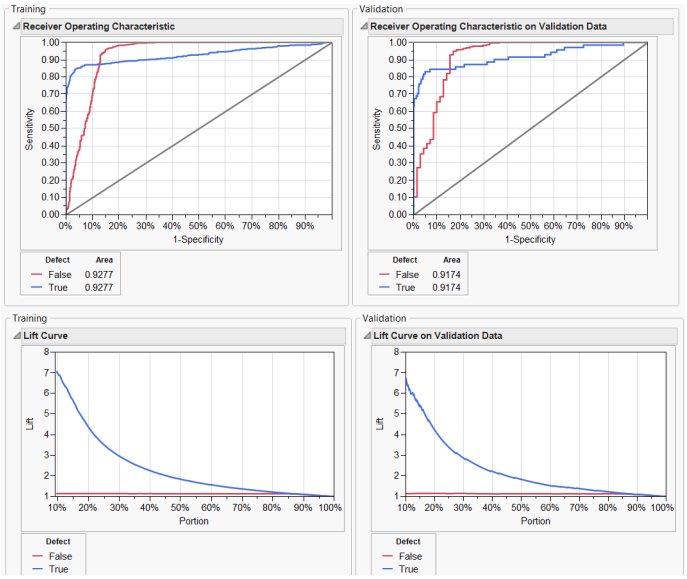
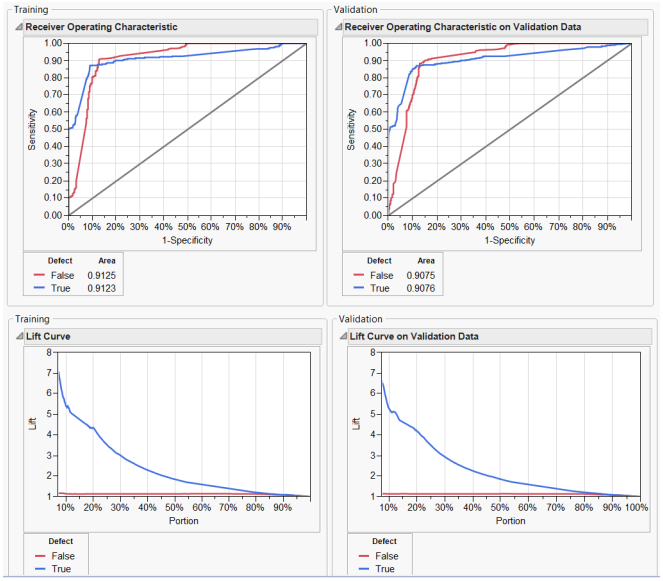


Figure 12: ROC and lift curves for ANN model data number F



The four charts from Figure 11 are results from data set A and the four charts in Figure 12 (two lift and ROC curves) are results from data set F. The area under the curve for the class defect true for data set A was 0.93 for the training and 0.92 for the test sample while it was 0.91 and 0.907 respectively for data set F. Also the lift value at 10 per cent of the population for the four lift curves was approximately 7. This meant that for the two data sets (A and F), we had similarly good results (based on AUC and lift) and they were similar implying that neural networks do not need a lot of training instances to accurately predict new data points.

The weights of the neural network were studied and we realised that they were not telling us anything about variable importance. This is because the results from the model give you weights estimates to each connection of layer within the neural network. These weights when analysed do not give you a clear view of which variables are important although Olden and Jackson have proposed ways of extracting the important variables [Olden and Jackson, 2002]. We then removed all the variables that we thought were significant from the model and evaluated the R-Square, misclassification rate, and the true positive value expecting that the model performance would drop. These variables were selected by looking at a logistic bi-variate plot between the class and explanatory variables and therefore do not take into account the multivariate interdependent to the class variable. Variables that showed a

strong relationship were selected and removed from the model. We removed *international plan*, *number of service calls* and *total day minutes* and the model was as good as random. The R-square value decreased from 0.8 to 0.06, the misclassification rate increased from 0.04 to 0.14 (which is what we would have if no model was used) and the area under the ROC curve for dropped from 0.93 to 0.65. We observed that in this data set these were not the only important variables but there was a higher proportion of the population that when sampled the class variable depends mostly on these three removed variables. This was also the reason why the area under the ROC curve was not 0.5 as in a case of a random model.

Sharma and Panigrahi worked on predicting customer churn using the same data set as in this research [Sharma and Panigrahi, 2011]. They used artificial neural networks techniques to fit the data on a sample of 2427 customers. The authors did not mention the data exploration phase in their analysis and they fit their model in SPSS using all variables but excluding *state* and *customer phone number*. We criticised the authors by doing this as we had found correlated variables in our data exploration phase. Our model performed better than their model (comparing it to data number 2 results). They had an accuracy rate of 0.924 and true positive ratio of 0.663 while our accuracy rate was 0.952 and true positive ratio of 0.73 (using the test data set). It was interesting to see that our models gave the same top three variables as the most important ones (namely *service calls*, *international plan* and *total day minutes*).

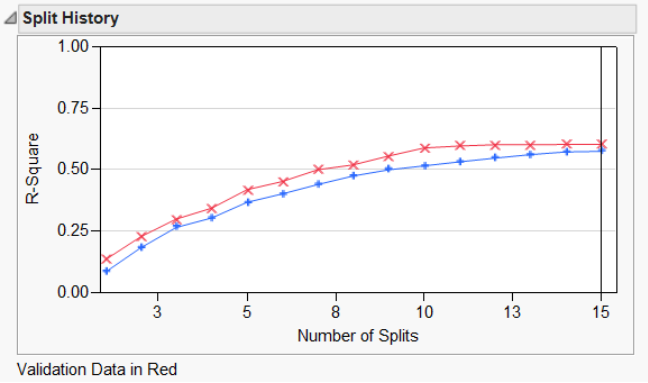
5.1.2 Decision Trees

We then used a decision tree models on the standardised and un-standardised data. This was done because we wanted to check whether standardising the data would have a significant impact in the model results. Firstly the model was fitted on the un-standardised data set split into 80 per cent training and 20 per cent test set. The number of splits was chosen to be ten for both models by evaluating if there was a significant incremental increase in the R-Square value. The AIC for this model was 5152 and the R-Square for the training sample was 0.501 and 0.436 for the test sample. The decision tree model was then applied again to an 80 per cent training and 20 per cent test sample of standardised data set. The optimal number of splits chosen again was ten and the R-Square value for the train and test sample was 0.551 and 0.455 and the AIC value was 4952. This was an improvement from the

previous results based on the R-square value which explained 55.1 per cent of the variation in the data compared to previous R-square of 0.501. We also observed a significant decrease in the model penalty value (AIC, from 5152 to 4952) meaning that the second model was fitting better. The decision tree model performed better on a standardised data set than the un-standardised data set.

We then ran a decision tree model on a newly generated training and test samples (80 and 20 per cent respectively). There were no rules for the minimum or maximum number of splits in the model. The optimal number of split was chosen based on an incremental R-square value after each split. This was done by evaluating if there was a significant increase in the R-square value at each split. The chart in Figure 13 shows the incremental R-Square value from each split.

Figure 13: Number of Decision Tree Splits



It is evident from the chart in Figure 13 that after 14 splits the increment in the R-Square value was small and the optimal value was found to be 0.576 for the training sample and 0.604 for the test sample. We also checked if the model was not over-fitting the data before exploring the results even further. This is because results from an over-fitted model can be misleading and we may make incorrect conclusions about the data. Fivefold cross-validation samples were used and the model was found to be stable with an overall R-Square value of 0.580 and -2log likelihood of 938 which did not vary from the initial model results.

The model misclassification rates for the training and test data sets were 0.054 and 0.048 respectively. These are very good rates because if no model

was used and we classified the data with the majority class (the majority class was 86 per cent) then we would have had a misclassification rate of 0.14. We then evaluated the model confusion matrix and learned that there were about 69.2 per cent of the true positives in the training sample and 73.1 per cent in the test sample. The area under the ROC curve for class defect true was 0.92 and 0.93 for both training and test samples respectively which proved that this model was performing well (A better performing model has an area under the ROC curve more or equal to 0.75). Furthermore, the lift curves were plotted and at 10 per cent of the population with the highest probability of churning, a lift value of 6.5 and 6.7 was obtained for the train and test samples respectively (for class defect true). These two lift values were reasonable high considering the fact that they were close to the maximum attainable lift of 7.14. We investigated variable importance in the model using G-Square statistics which is the same as twice the natural log of entropy. The *total day minutes* gave the highest number of splits (four) and the highest G-square statistic value. Figure 14 shows the number of splits by each variable, the G-Square statistics value and the G-Square plot (bars). The number of splits shows how many nodes were split by that variable, G^2 is the G-Square statistics that measures variable importance and the bars measure the magnitude of the G-Square value. It was evident from the chart that *account length*, *number of evening calls*, *night calls*, *night minutes* and *voice mail messages* were not significant in the model as there were no splits based on these variables.

Figure 14: Decision trees variable importance data set 1

Term	Number of Splits	G^2
Account Length	0	0.00000
Intl_Plan	1	184.03364
VMail_Plan	2	91.93691
vmail_messages_2	0	0.00000
service_calls_2	1	220.76432
Day_Mins	4	380.72336
Day_Calls	1	5.92585
Eve_Mins	4	173.44068
Eve_Calls	0	0.00000
Night_Mins	0	0.00000
Night_Calls	0	0.00000
Intl_Mins	1	87.68550
Intl_Calls	1	143.37650

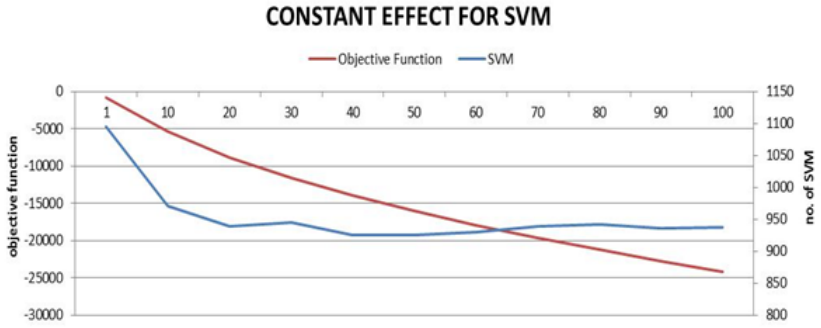
From the decision tree model there were only four pure nodes where the entropy or the G-Square statistics was zero. The minimum number of customers obtained in a final leaf was 20 and they were obtained from the pure node. We then fitted a decision tree model with all the significant variables in the model with a minimum final node size set to 50 customers. The R-Square

values were similar to the previous decision tree model with all variables included for both train and test sample. There was a slight decrease in the AIC value which meant that this model was a better fit than the one with all the variables.

5.1.3 Support Vector Machines

Support vector machines models were applied using R software with KSVM and ROCR packages (used for creating the model and evaluating performance respectively). These models were fitted using three kernel functions namely radial basis, Laplace and polynomial in order to classify the data into two classes (churners and non-churners). The data was divided into 80 per cent training and 20 per cent test sample. We firstly fitted an SVM radial basis kernel model by trying a series of constant values in order to find out which one optimises the model accuracy and minimises the misclassification error. It is important to note that the radial basis and Laplace kernel function depends on sigma and constant values in their construction. Choosing the right sigma (σ) and the right constant (C) for the model will yield good results. At the initial stage, the effect of a constant was crucial and the sigma value was chosen using a built in function in R that automatically finds the best sigma given a constant C. If R did not have this function, we would have chosen sigma using similar method as finding the best constant. Figure 15 shows the objective function value and the number of support vectors for constant values ranging from one to a hundred for a RBF kernel.

Figure 15: Support vector constant effect 1: RBF kernel function



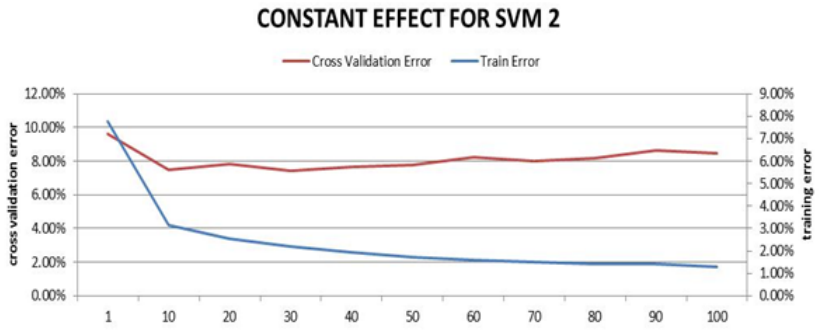
The following phenomenon must be noted from this chart:

- The number of support vectors initially decreases at lower constant values and then stabilises after $C >= 60$ to one hundred
- As the constant increase, the objective function declines significantly.

The chart in Figure 16 plots the error on the training sample (secondary axis) and the cross validation sample (primary axis). The following were noted from the chart:

- As the constant increases, the error on the training sample decreases
- For every increase in the constant, the cross validation sample error slightly increases.
- It can be shown that there is not much change in accuracy of the RBF model with an increase in constant (except for $K < 0$)

Figure 16: Support vector constant effect 2: RBF kernel function



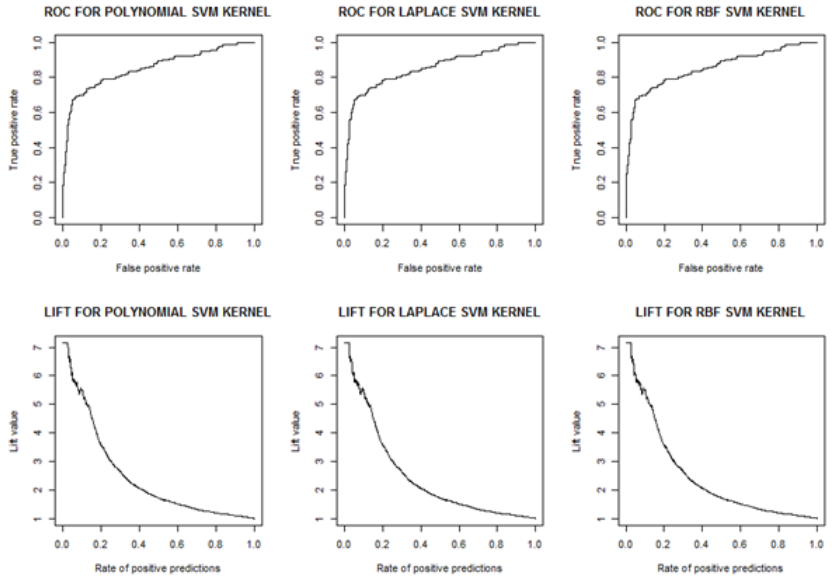
The change in the constant affects mostly the objective function and the training error. In fact, it can be proven that as the constant tends to infinity for the RBF model:-

- the objective function tends to negative infinity
- the training error tends to zero or the neighbourhood of zero
- the cross validation error becomes the same as having no model [Alpaydın, 2004]

On evaluating the three fitted models it was found that Laplace kernel SVM outperformed the polynomial kernel of degree 3 and the radial basis kernel (with best sigma and the constant equalling to 0.05 and 10 respectively). Some of the results are shown in the appendix Figure A8 and Figure 17 shows

the ROC and lift curve. Laplace kernel SVM gave the best training and cross validation error (1.46 and 7.99 per cent respectively), test misclassification rate (8.4 per cent), area under the ROC curve (0.86) and the number of true positive (64 customers).

Figure 17: Support vector machines ROC curve fit for data set 1



We noted the disadvantages of the Laplace kernel SVM and the polynomial model. These two models were over-fitting the data because the variation or the range in the training error compared to the cross validation error (three folds cross validation sample) was large. The range between the two errors (train and cross validation sample) was 0.144 for polynomial kernel of degree three, 0.062 for a Laplace kernel and 0.058 for radial Basis kernel. The RBF SVM seemed to produce more stable results than when the other kernel functions were used. This was then chosen as the best model out of the three.

We then went on to test the strength of the RBF kernel by evaluating the performance of the model when different probability thresholds were used in the test set. The minimum cut off probability threshold was 0.4 and the maximum was 0.95. We noted the following from the results as the cut off probabilities increases:

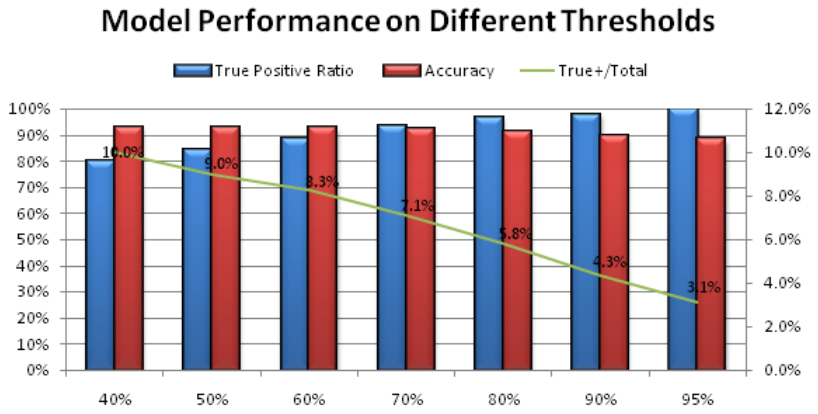
- The true positive ratio increases and at a cut off equalling 0.95 it was equal to 1. These results were disappointing as the number of the

actual true positive was very low, that is, only 31 customers out of a 667 in the test set.

- The model accuracy decreased from 0.93 to 0.89 which we thought was a fairly good level.
- The true negative class increased as the model was classifying most customers to the majority class.

The chart in Figure 18 shows the results from the different probability cut off in details. The red bar is the accuracy rate, the blue bar is the true positive ratio and the line on the secondary axis is the predicted true positive divided by the total sample.

Figure 18: Probability cut off for data set 1 SVM model



Brandusoiu and Toderean used SVM model to predict churn on the same data set [Brandusoiu and Toderean, 2013]. On their data exploration phase, they concluded that the data was "complete" and they were no missing values. They discovered that four *charge* variables were correlated with *minutes* variables and this was similar to our findings. They argued that SVM works well if the class ratio is balanced. They decided to boost the data set by cloning the number of churn class equal yes to be the same as no. They found that a polynomial kernel SVM (of degree 3) outperformed the other three used kernels (sigmoid, RBF and linear) and it gave an accuracy rate of 0.887. In our analysis, polynomial kernel SVM was the worst performing model and it gave the same accuracy rate (0.884) as the latter. Our best performing SVM models (Laplace and RBF) had an accuracy rate of 0.919 and 0.904 respectively thus better than the author's results.

5.1.4 Logistic Regression

The final model that was explored for data set 1 was the logistic regression. This was done using R statistical software. This regression method was performed using a backward selection technique whereby all variables are entered into a model and the non-significant variables are removed one by one. The optimisation method used was the Fisher scoring and the selection criterion was the AIC with a 0.05 level of significance. The model started by putting in all variables and the AIC value was 2538.12 for thirteen variables. After the model had removed four insignificant variables the AIC value had decreased slightly to 2532. We assessed the p-value and the z value and concluded at 5 and 10 per cent level of significance that *account length*, *total number of day calls*, *total evening calls* and *number of night calls* were not significant.

We analysed the coefficient and the z score of each significant variable and found that *international plan factor yes* was the most significant variable (with a z score of 14.88) followed by *total day minutes* (with a z score of 13.01) and *the number of service calls* (with a z score of 12.303). The odds for *international plan factor yes* was 7.188 implying that there was a high association between defecting and having an *international plan*. This reason why this was the case was because 0.42 customers that had international plan defected. The logistic regression equation for the reduced model was

$$\begin{aligned} & \mathrm{f}(defect) \\ = \; & 1/(1+\exp\Big\{2.97-0.71x_1-0.34x_2-0.22x_3-0.25x_4 \\ & +0.17x_5-0.5x_6+1.0x_7-1.9x_8\Big\}) \end{aligned}$$

where x_1 is *total day minutes*, x_2 is the *total evening minutes*, x_3 is the *total night minutes*, $x_4$is the *total international minutes*, x_5 is the *total number of international calls*, x_6 is the *total number of service calls*, x_7 is *voice mail plan* and x_8 the *international plan*.

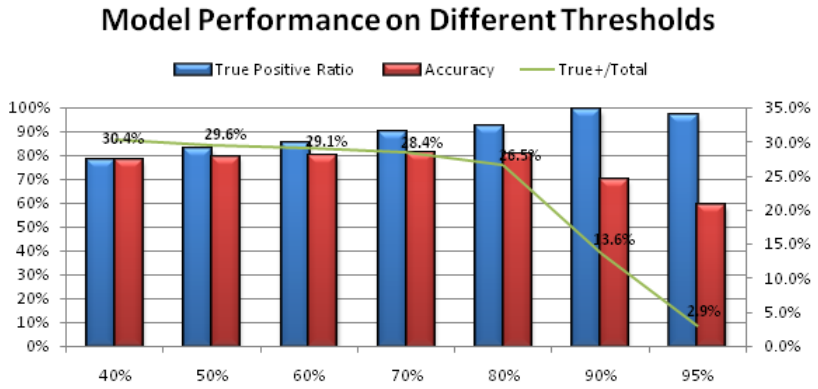
We went on to evaluate the model performance in the test data set. The model had an accuracy value of 0.87 at 0.5 probability value which was exactly the same as for the training sample. The performance was also

evaluated at different probability thresholds and the following were observed as the probability cut off increased:

- The model accuracy level was stable
- The ratio between the true positive and the actual true positive decreased, in fact at probability threshold greater or equal to 0.9 there we no true positives
- The model misclassified the data and put all instances to the majority class at 95 per cent probability threshold.

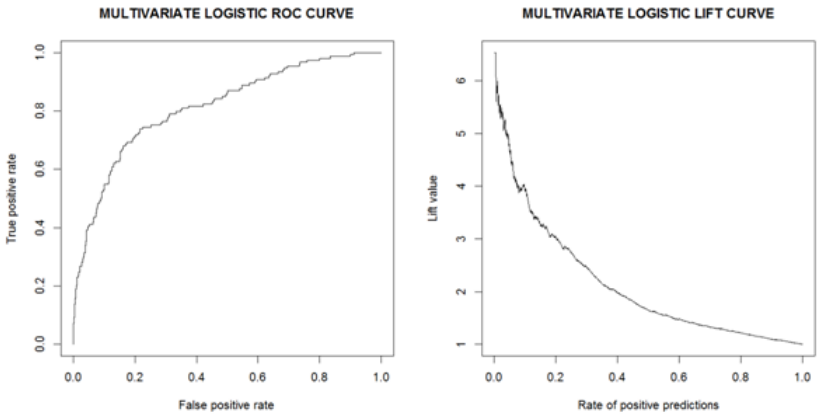
The chart in Figure 19 shows the model performance at different probability threshold values chosen at certain intervals. The red bar is the accuracy rate, the blue bar is the true positive ratio and the line on the secondary axis is the predicted true positive divided by the total sample.

Figure 19: Probability cut off for logistic regression data set 1



A further analysis of the model using ROC and the lift curve showed that the model was not performing well. The area under the ROC curve was 0.81 and the lift value at 10 per cent of the population was 3.6, which was very low considering the fact that a maximum lift attainable was 7.14. A person looking at the results bluntly would say that an average of 87 per cent accuracy on this model (in fact any model) is good, forgetting that if all instances are classified into the same class an accuracy value of 86 per cent would be attained thus this was not a good model. The charts in Figure 20 show the ROC and the lift curves of this model at probability threshold of 0.5.

Figure 20: ROC and lift curve for logistic regression data set 1



5.2 Data Set 2 Results

5.2.1 Artificial Neural Networks

The artificial neural network model was created using nine explanatory variables to predict the class variable churn. We tried to prove the same hypothesis as in data set 1 that the neural network method does not require a lot of training instances to correctly predict the data. The data was split six ways as per table in previous section (Table 6) with varying training and test rates. A feed-forward neural network model with a hyperbolic tangent as a sigmoid function and three hidden units was used to train the data. Tables 9 and 10 show training and test data results based on few model fit metrics:

Table 9: Train data model performance for data set 2

Train Data	R Square	Misclassification Rate	-2Log Likelihood	Mean SE	True Positive Ratio
A	0.57	0.208	1760	0.368	0.565
B	0.58	0.2	1539	0.36	0.66
C	0.56	0.2	1357	0.36	0.65
D	0.585	0.2	1145	0.36	0.69
E	0.568	0.21	981	0.37	0.55
F	0.576	0.2	774	0.365	0.68

Table 10: Test data model performance for data set 2

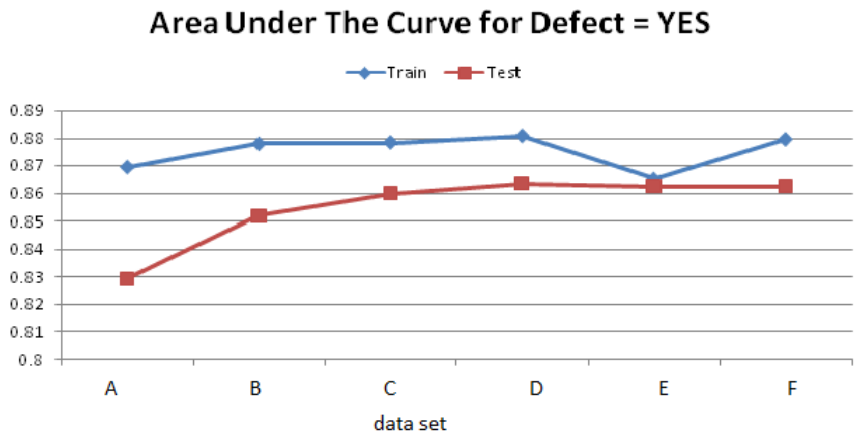
Test Data	R Square	Misclassification Rate	-2Log Likelihood	Mean SE	True Positive Ratio
A	0.467	0.251	226	0.398	0.469
B	0.531	0.221	415	0.387	0.6
C	0.545	0.206	609	0.374	0.63
D	0.548	0.22	808	0.376	0.67
E	0.561	0.21	991	0.37	0.55
F	0.55	0.21	1207	0.374	0.67

The AIC and BIC value increases as sample size increases, that is, the model was penalised for using more observations to fit (based on -2log likelihood). Results from data set D were slightly better than the other data set results because:

- The R-Square value was fairly large for both training and test set (0.59 and 0.55 respectively)
- The model had the highest true positives ratio (0.69 and 0.67 for the training and test set respectively)

Again, these results showed that the neural network model does not need much training information in order to build a good model. The differences in the results in Tables 9 and 10 are minimal for both train and test data. We then went on to evaluate model performance using the receiver operating characteristic curve and the lift curve. The lift curves for all six data sets (for class defect = yes) at ten per cent of the population fluctuated around 2.2 which was a maximum lift that can be obtained ($maxlift = \frac{1}{0.45} \approx 2.22$). This meant that the model in all data sets performed very well (with an exception of data set A). Figure 21 shows the area under the receiver operating characteristic curve for training and test sample.

Figure 21: AUC for ANN models



It is important to note that the scale of the chart in Figure 21 starts at 0.8. There was a slight incremental increase in the AUC value for test data from data set A to data set D which stabilises afterwards. The AUC value for train data stabilises at data set B then drops at data set E and return to "equilibrium" at data set F. This sudden deviation at data set E for the train AUC value was inexplicable unless there were data point(s) that when not sampled in the data set the model performance deteriorates in accuracy.

We then explored what happens when the number of hidden units was increased. Does the neural network model reach a point of stability in accuracy? If so, how many hidden units were needed? Is there any information gain by increasing the number of hidden units? The line chart in Figure 22 shows R-Squared, misclassification and true positive ratio for seven variations of hidden units (from 3 to 21 hidden units) using 80 per cent training and 20 per cent test data.

Figure 22: R-Square for a change in the number of hidden units in ANN model

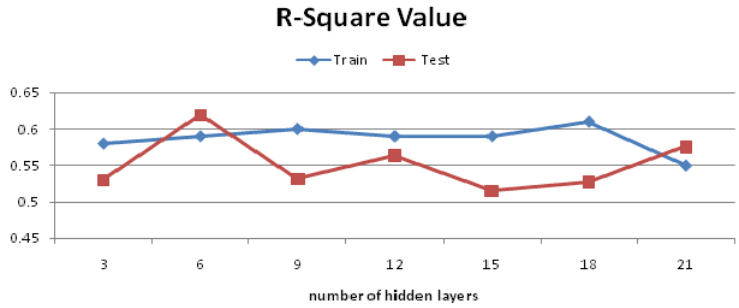
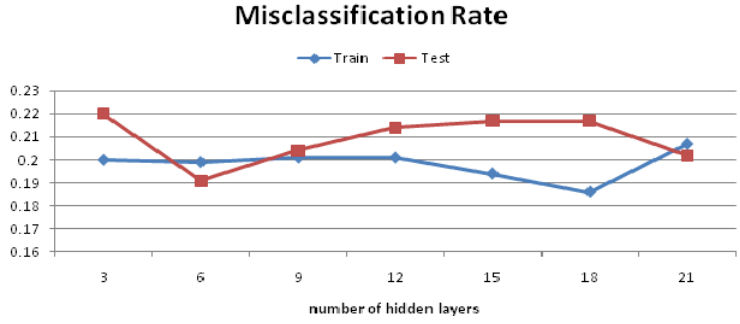


Figure 23: Sensitivity for a change in the number of hidden units in ANN model

Figure 24: Misclassification rates for a change in the number of hidden units in ANN model



The numbers of hidden units were varied by multiples of three up to twenty one hidden units. The feed-forward hyperbolic tangent neural network model

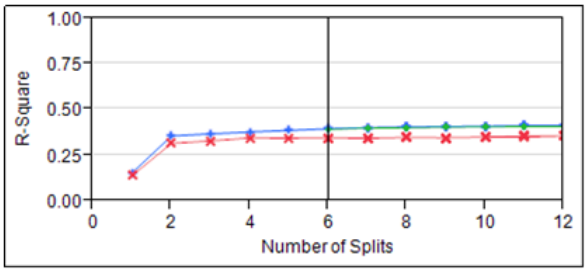
with six hidden units slightly outperforms the other models based on R-Square, Misclassification and True positive rate. The R-Square value drops after eighteen hidden units for the training data set and the true positive rate tends to stabilise after six hidden units. The model with three hidden units was the "worst" performer amongst the seven models based on the three charts (Figures 22, 23 and 24). This model also had the highest AIC and BIC value for the training data set when compared to other models (based on -2log likelihood of 1539).

We then tried to evaluate which variables were important by removing the variables that we thought could be significant using prior information from the data exploration step. Using the six hidden units neural network model we removed *customer account age* and *initial monthly fee*. The misclassification rate increased from 0.19 to 0.22, the R-Square value decrease from 0.6 to 0.45 and the true positive ratio dropped from 0.68 to 0.58. The variable *maker* (hand set colour) was also removed and the R-Square value decreased significantly to 0.1, misclassification rate to 0.39 and the area under the curve for both train and test sample was 0.65.

5.2.2 Decision Trees

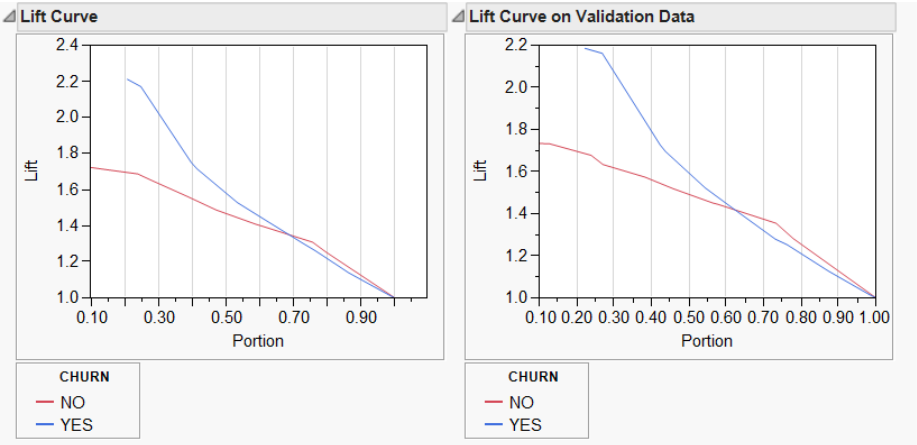
A decision tree model was applied to the second data set using a similar strategy as the one used for data set 1 (with the data set split 80 per cent train and 20 per cent test sample). We stopped splitting the tree by evaluating if there was no significant incremental increase in the R-square value of the model. The results showed that after six splits there was no incremental value on the R-Square, however we made twelve splits and had to prune to six splits as the model was over-fitting. The five folds cross validation sample showed that the overall R-Square was 0.41 and minus twice log likelihood was 3080 which was almost as the same as the model with six splits. The AIC value for the model at six splits was 10169 and the R-square for the training sample was 0.4 and 0.35 for the test sample. The chart in Figure 25 shows the total number of splits for the decision tree model.

Figure 25: Decision trees R-Square value per split for data set 2



The confusion matrix showed 99 per cent of the true negative fraction and 54 per cent of the true positive fraction for the training sample and 98 per cent of true negative fraction and 52 per cent of the true positive instance in the test sample. These were good ratios considering that in the total sample there were 45 per cent instances belonging to the defect class true and 55 per cent belonging to defect class false. The total misclassification rate of the training sample was 20.9 per cent and 22 per cent for the test sample. The model performance was evaluated using the area under the ROC curve which was evaluated to be at 0.85 and 0.83 for training and test sample respectively (for class defect equals true). We then computed the lift curve and found that there was no lift value at ten per cent of the population and that the lift value for the class defect equal true was obtained at 25 per cent of the population which was 2.20 and 2.1 for training and test sample respectively. The reason why there was no lift at lower population proportion of the data was that no customer had a probability of defecting more than 0.75. Figure 26 shows the lift curve for the training sample on the left and the test sample on the right.

Figure 26: Decision trees lift curves for data set 2



We went on to evaluate which variables were important in the model and learned that only three variables were important namely *maker*, *account type* and *account age*. It was noted that *credit class*, *auto pay* and *market* variables become significant if the number of splits were increased. We were surprised as to why the *maker* (colour of the hand set) was the most significant variable. This may be caused by the fact that a certain hand set colour may be

- associated with a certain contract offer
- associated with a high take up of new customers which are in nature having a high probability of defecting
- associated with a certain area where the probability of defecting is high

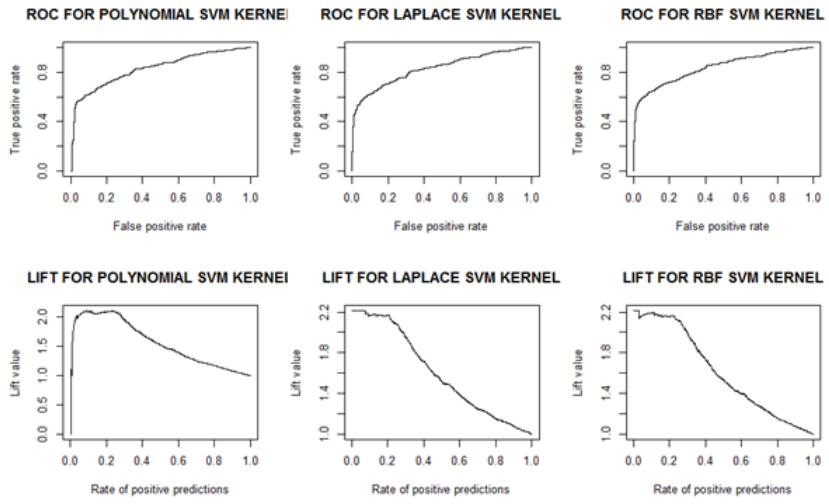
5.2.3 Support Vector Machines

The SVM model was applied using the same strategy as the one used for data set 1. We observed similar results as in data set 1 when looking for an effect of an increase in cost parameter C. The best σ was 0.109 and the best cost parameter was 10 for radial basis and Laplace kernel function. Three kernel SVM models were fitted and compared. The polynomial kernel SVM was of degree three and for all models three folds cross validation samples were used. The model accuracy and results are shown in the appendix Figure A8. The results showed that:

- Polynomial Kernel SVM was the worst performing model with the lowest area under the ROC curve. The lift value (at 10 per cent) for this model was lower than for the three models.
- The Laplace SVM kernel gave the best training error at 0.11 but it misclassification rate (23 per cent) on the test sample was the worst
- The radial basis kernel gave the best results and the model proved to be more stable than the other two model as the range of training (17.8 per cent) and cross validation sample error (20.9 per cent) rate was minimal (range equals 3.1) and less number of support vectors used.

The charts in Figure 27 show the ROC curve and the lift curve for the three support vector machine kernel models

Figure 27: ROC fit for kernel SVM models data set 2



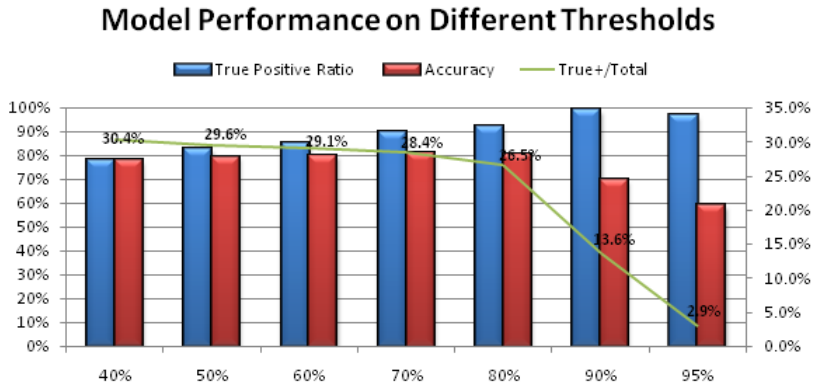
The radial basis kernel SVM was taken to be the best model out of the three. We evaluated the performance for this model at different probability threshold cut off points. We observed the following as the probability threshold increased:

- The accuracy increased until for a threshold greater than 0.8 it decreased sharply
- The true positive ratio increased but the model was classifying more instances into the majority class.

- The true positive total decreased rapidly.

The chart in Figure 28 shows the performance of the model when different probability thresholds are used. The red bar is the accuracy rate, the blue bar is the true positive ratio and the line on the secondary axis is the predicted true positive divided by the total sample.

Figure 28: Probability cut off for data set 2 SVM model



5.2.4 Logistic Regression

We performed the logistic regression on data set 2 in a similar fashion as in data set 1. The Fisher scoring optimisation model stopped after 5 iterations and gave an AIC value of 3719.3. The full model showed that all nine variables that were used in the model were significant but the *credit class factor B* was not significant. It was noted that *credit class factor D*, acquisition *channel* and *customer age* were not significant at one per cent level of significant based on their p-value (0.013, 0.014 and 0.011 respectively). *Market* was found not to be significant when an ANOVA Chi squared test was performed. We ran the variable importance and saw that *account age*, *maker* and *initial monthly fee* were the most significant variables based on their Chi square values.

The odds ratio for *account age* was equal to 4.177 suggesting that a one unit change in *account age* results in 4.177 increase in chances of churning. This also meant that there was a high association between defecting and *account age*. Conversely, *auto pay* had the lowest odds ratio equalling 0.353 which meant that there was a high association between not defecting and this

variable, that is, the chances of defecting per one unit increase as *auto pay* decreases by 0.353. The logistic regression equation for this data was

$$\begin{aligned} & \mathrm{f}(defect) \\ = \; & 1/(1+\exp\Big\{-0.73-0.63x_1-0.42x_2-0.45x_3+1.04x_4 \\ & +0.4x_5-0.3x_6+0.37x_7-1.4x_8+0.17x_9+0.11\mathrm{age}\Big\}) \end{aligned}$$

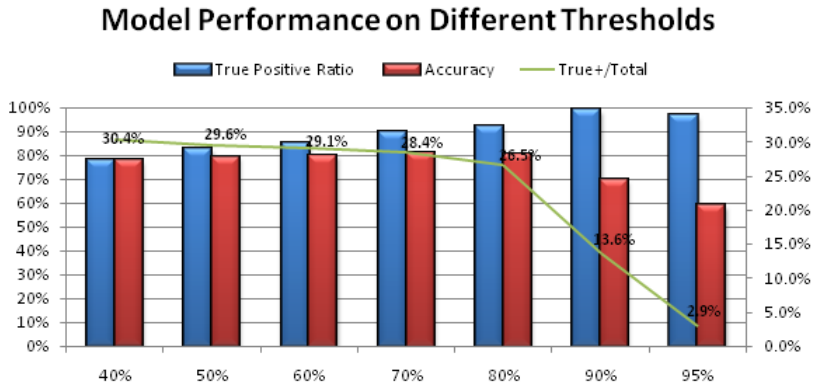
where x_1 and x_2 refer to *credit class factors C and D*, x_3 is *account type*, x_4 is auto pay, x_5 is the *market* x_6 is *channel*, x_7 is the *maker*,x_8 is *account age* and x_9 is *initial monthly fee*.

We then evaluated the model performance on the test set at different probability threshold. These probability thresholds were chosen to be the same as in the case of SVM model. The following were observed as the probability threshold was increased:

- The true positive ratio decreased exponentially and at 0.95 probability threshold the model had 1 true positive out of 954 customers.
- The model accuracy increases from 0.75 to 0.78 and then drops to 0.54 (after probability of 0.8)
- The model classifies most instances to the majority class

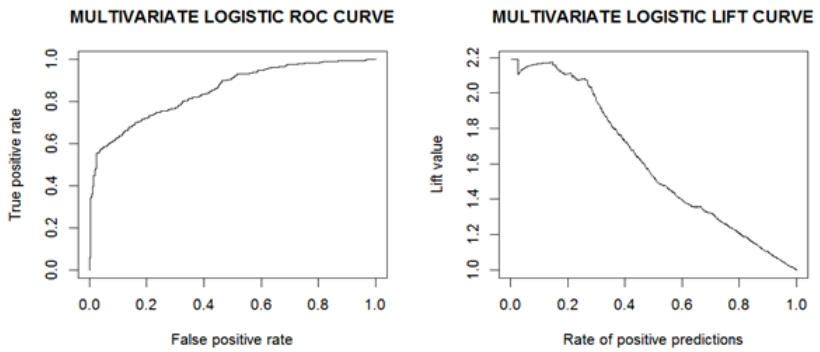
The chart in Figure 29 shows the model performance at different thresholds.The red bar is the accuracy rate, the blue bar is the true positive ratio and the line on the secondary axis is the predicted true positive divided by the total sample.

Figure 29: Probability cut off for logistic regression on data set 2



Also, for the model performance at 0.5 probability value, an ROC and lift curves were computed. The area under the ROC curve was found to be 0.85 and the lift value at 10 per cent of the population was 2.01, which was slightly less than the maximum lift attainable at 10 per cent (2.22). The model performance at 0.5 probability value was fairly good. The two charts in Figure 30 show the ROC and the lift curve for the logistic model at 0.5 probability cut off. The training sample model accuracy at 0.5 probability cut off was 0.755.

Figure 30: ROC and lift curve for logistic regression on data set 2



6 Comparison of Models

In this section we compare the four models based on the area under the ROC curve, lift value, misclassification rate and the true positive rates. Although other measures could have been used to compare these models, we wanted to use only measures that are common in the four models. We could not have used AIC or BIC value as support vector machines model construction does not give log likelihood and probabilities. This was a major drawback for this model. The probability for SVM model was computed using Platt's posterior probability as suggested and improved by Lin and Wang [Lin and Wang, 1999]. The neural network models had probability outputs as a hyperbolic sigmoid function was employed. If we had used a radial basis sigmoid function then the probability would have been computed in similar fashion as in SVM. We found SVM and ANN models to be complex and they both took long to run (10 and 15 minutes respectively). Also, the choice of a constant in SVM model was non-trivial because we had to make sure that the number of support vectors was minimised and also take into account the accuracy of the model. It was much easier to perform decision trees and logistic regression model as they required less computer run time (2-3 minutes for both models).

Rank and score variables were created in order to evaluate which model was the best. This was done because we wanted a robust measure of performance using the four metrics. The score was created by equally weighting all four variables and summing their Z scores and in the case of misclassification value a reciprocal was used as low values were preferred for this variable (other ways could have been used, for example, 1-misclassification rate). The higher the score value the better is the model performance. Rank was created by ranking the models based on each metric that was measured (from 1 to 4). We then took the average rank for the four models with the lowest being the best performing model. Tables 11 and 12 show the model performance for the two data sets. It is evident that the artificial neural network model had outperformed the three other models by a margin for both data sets. This model had the highest score value and the lowest average rank (lower rank preferred than higher rank). The score range of the next best model was almost half the size of the ANN score. The logistic regression was the worst performing model of the four (using data set 1). This model had a negative score (-1.3) and a highest average rank (4, for data set 1). Decision trees outperformed SVM model and was ranked second with a score of 0.52. For data set 2, artificial neural networks outperformed the other three models and had a score value of 0.9 while support vector machine with a radial basis

kernel function came second (score of 0.31 and an average rank equal to 2). The decision tree model was the worst performer with a score value of -0.99 and an average rank of 3.75 (for data set 2).

Table 11: Data set 1 model comparisons

Models for data set 1	**AUC**	**Lift**	**Misclassification rate**	**True positive**	**Score**	**Average Rank**
ARTIFICIAL NEURAL NETWORKS (3 Hidden Units)	0.930	7.010	0.040	0.73	0.810	1.5
DECISION TREES	0.920	6.300	0.048	0.731	0.52	2
SUPPORT VECTOR MACHINES	0.869	5.800	0.093	0.847	-0.020	2.25
LOGISTIC REGRESSION	0.810	3.600	0.130	0.12	-1.310	4

Table 12: Data set 2 model comparisons

Models for data set 2	**AUC**	**Lift**	**Misclassification rate**	**True positive**	**Score**	**Average Rank**
ARTIFICIAL NEURAL NETWORKS (6 Hidden Units)	0.880	2.120	0.190	0.710	0.900	1.5
SUPPORT VECTOR MACHINES	0.838	2.100	0.212	0.859	0.310	2
LOGISTIC REGRESSION	0.850	2.010	0.230	0.740	-0.220	2.75
DECISION TREES	0.830	1.890	0.220	0.540	-0.990	3.75

7 Conclusion and recommendations

The data sets that were used had few demographics variables (2 for data set 1 and 4 for data set 2). They also contained a small numbers of variables (13 for data set 1 and 9 for data set 2) from which models were fitted. Data set 2 was much polluted and it required a lot of cleaning and derivation of new variables. Data set 2 had many variables concerning billing information while data set 1 had more behavioural information. We showed that standardising the data helped as there was an improvement in the R-square value for neural networks and decision tree models on data set 1. All models fitted using data set 2 showed that customer *account age* was a significant indicator of churning while in data set 1 this was not the case. This could mean that in some retail mobile telecommunication settings loyalty is not a big factor that can prevent customers from churning. The customers that churn the most when looking at account age for data set 2 were the new customers. Data set 2 is also suited for survival analysis where we can focus on the expected time until a customer churns rather that finding a type of customer that will churn as in data set 1 [Berry and Linoff, 2009a]. It was much easier to run the models in data set 1 than data set 2 as most variables in data set 1 were binary variables and we had to make sure that there was no problem of linearity.

We recommend artificial neural networks over the three other techniques as they outperformed all of them. This must not be interpreted bluntly, as in some data sets other data mining techniques might be more suited for predicting customer churn because different models can work better when fitting different data sets. This paper has also showed that logistic regression does not work well in instances where the data set has few variables as this requires a more complex statistical model. Furthermore, for industry practice if an artificial neural network model is fitted, we suggest that this model is fitted in conjunction with decision trees model as they were able to extract valuable information from the data concerning variable importance. This can be used to confirm variable importance. It is also up to us to decide which sigmoid function to use, we preferred using a hyperbolic tangent and logistic sigmoid as both these function have a probability output.

8 Summary and Future Research

In the retail mobile telecommunication industry which is very competitive, the likelihood of churning is very high. Using statistical models can help a business retain some of it customers by predicting the ones that are likely to churn and incentivising them. In this research, deployment of artificial neural networks for predicting customer defection in the retail mobile telecommunication industry proved to be helpful. We saw that the artificial neural network technique showed good results but it had some limitation when it comes to variable importance and interpreting the weights. The two data sets had fewer explanatory variables than we would have liked. A data set with more demographic and behavioural variables would have been preferred for this research. It would also be interesting to see how these models will perform in the real world. Furthermore, developing different churn models for high, medium and low value customers might be what is needed for the business as they can minimise cost and maximise profit margins if more retention projects are channelled towards high value customers. A multivariate response model can be built in order to evaluate response rate for all the customers that were incentivised. This is to make sure that the retention projects are working because it does not make economic sense to continue with the retention projects if the response is too low, that is, if the costs of implementing such a projects are not covered by keeping customers that would have defected [Cohen et al., 2006].

References

L.J.S.M Alberts. Churn prediction in mobile telecommunication. September 2006. Online; June 2012.

E. Alpaydın. *The Elements of Statistical Learning Data Mining, Inference and Prediction*. The MIT Press, London, England, second edition, October 2004.

M.J.A. Berry and G.S. Linoff. Customer centric forecasting using survival analysis. `http://www.data-miners.com/companion/sas/forecastingWP_001.pdf`, 2009a. Online; accessed March 2012.

M.J.A. Berry and G.S. Linoff. Customer centric forecasting using survival analysis. 2009b. Online; March 2012.

C. L. Blake and C. J. Merz. *Churn Data Set, UCI Repository of Machine Learning Databases*. University of California, Department of Information and Computer Science, Irvine, 1998. Online; March 2011.

I. Brandusoiu and G. Toderean. Churn prediction in the telecommunications sector using support vector machines. May 2013. Online; June 2013.

W. Buckinx and D. Van den Poel. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research*, 164:1–32, 2004.

J. Burez and D. Van den Poel. Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications Journal*, 32:277–288, 2005.

N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Sythetic minority over sampling technique. *International Journal of Artificial Intelligence Research*, 16:321–357, 2002. Online; June 2013.

B. Cheng and D.M. Titterington. Neural networks: A review from a statistical perspective. *Statistical Sciences*, 9:2–54, Jan 2000.

S.L. Chow. Statistics and it role in psychological research. `http://cogprints.org/2782/1/eolss.pdf`, 2002. Online; accessed March 2012.

K.J. Cios, W. Pedrycz, R.W. Swiniarski, and L.A. Kurgan. *Data Mining, A Knowledge Discovery Approach*. Springer, New York, USA, February 2007.

D. Cohen, C. Gan, H.H.A. Yong, and E. Choong. Customer satisfaction: A study of a bank customer retention in New Zeland. http://www.lincoln.ac.nz/Documents/, March 2006. Online; November 2012.

M. Crang. Quantitative methods: The new orthodoxy. http://dx.doi.org/10.1191/0309132502ph392pr, 2002. Online; January 2013.

A. Dehghan and T.B. Trafalis. Examining churn and loyalty using support vector machine. *Business and Management Research*, 1:153–161, December 2012. Online; June 2013.

J.B. Ferreira, M. Vellasco, M.A. Pacheco, and C.H. Barbosa. Data mining techniques on the evaluation of wireless churn. In *European Symposium on Artificial Neural Networks*, pages 483–488, Bruges, Belgium, 28-30 April 2004. Online; March 2013.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning Data Mining, Inference and Prediction*. Springer, Stanford, Carlifonia USA, second edition, September 2008.

J.H. Friedman. Data mining and statistics: What is the connection? http://www.salfordsystems.com/doc/dmstat.pdf, November 1997. Online; May 2012.

C. Gatsonis. Statistical methods for meta-analysis of diagnostic test accuracy. http://legacy.samsi.info/200809/meta/presentations/diag-test-metan-bib-june08.pdf, June 2008. Online; March 2012.

S.R. Gunn. Support vector machines for classification and regression. http://www.users.ecs.soton.ac.uk/srg/publications, May 1998. Online; April 2013.

J. Hadden, A. Tiwari, R. Roy, and D. Ruta. Churn prediction using complains data. *World Academy of Science Engineering and Technology*, 19:1–6, 2006. Online; May 2013.

A. Idrisa, M. Rizwan, and A. Khan. Churn prediction in telecom using random forest and pso based data balancing in combination with various feature selection strategies. *Computers and Electrical Engineering*, 38: 1808–1819, September 2012.

C. Imhoff. Bouygues telecom. http://www.teradata.com/casestudies/BouyguesTelecomtheIntelligentTelecommunicationsCompany/, March 2001. Online; March 2012.

S. Jamwal. An approach to mobile telecom churn halding in india. *International Journal of Computer Information Systems*, 2:54–58, April 2011. Online; June 2013.

A. Juahiainen. Experimental design. `http://pingpong.ki.se/puplic/pp/`, 2012. Online; January 2013.

M. Kamber and J. Han. *Data Mining: Concepts and Techniques*, chapter 7, pages 383–464. Diane Cerra, San Francisco, USA, second edition, 2006.

A. Karatzoglou, D. Meyer, and K. Hornik. Support vector machines in R. *Journal of Statistical Software*, 15:1–27, April 2006. Online; April 2013.

Estimating campaign benefits and modeling lift, San Diego, Carlifonia, USA, 1999. Knowledge Stream Partners and GTE Laboratories. Online; May 2013.

R. Kohavi. A study of cross validation and bootstrap for accuracy estimation and model selection. *Proccedings of the 14th international joint conference on Artificial Intelligence*, 2:1137–1143, 1995. Online; June 2013.

P.H. Kvam and J. Sokol. Teaching statistics with sports example. `http://www2.isye.gatech.edu/statistics/papers/`, 2004. Online; January 2013.

C.F. Lin and S.D. Wang. Fuzzy support vector machines. *EEE Transactions on NeuralNetworks*, 13:464–471, 1999. Online; June 2013.

G.S. Linoff and M.J.A. Berry. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing, Indianapolis, Indiana USA, second edition, April 2004.

C. Manu. Analysis of clustering technique for crm. *International Journal of Engineering and Management Science*, 3:402–408, July 2012.

M. Mazzocchi. Statistics for marketing and consumer research. `http://consumer.stat.unibo.it/Private/Chap14.ppt`, December 2007. Online; January 2013.

P. Mirowski, S. Chopra, F.J. Huang, and M Mohri. Support vector machines. `http://www.cs.nyu.edu/~yann/2010f-G22-2565-001/diglib/lecture03a-svm-2010.pdf`, June 2008. Online; November 2012.

Dr. H. Nemati. Introduction to data mining using artificial neural networks. `www.uncg.edu/ism/ism611/neuralnet.pdf`, 2000. Online; April 2013.

G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi. Credit card churn forecasting. *Expert System with Applications Journal*, 38:15273–15285, 2011.

J.D. Olden and D.A. Jackson. Illuminating the ""black box": A randomised approach for understanding variable contribution in artificial neural networks. *Ecological Modeling*, 154:135–150, 2002. Online; August 2013.

M. Owczarczuk. Churn models for prepaid customers in the cellular telecommunication industry using large data mart. *Expert Systems with Applications*, 37:4710–4712, 2010. Online; May 2013.

A.A. Philip, A.A Taofiki, and A.A Bidemi. Artificial neural network model for forecasting foreign exchange rates. *World of Computer Science and Information Technology*, 1(3):110–118, 2011.

D. Van Den Poel and B. Larivire. Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 157:196–217, 2003.

M. Proust. *Modeling and Multivariate Methods*. SAS Institute Inc, SAS Campus Drive, Cary, North Carolina 27513, tenth edition, March 2012.

N.P. Raygoza. Effect measures in 2 by 2 tables. `http://www.piit.edu/~super4/34011-35001/`, 2009. Online; January 2013.

B. Ripley and R. Ripley. Neural networks as statistical methods in survival analysis. `http://www.stats.ox.ac.uk/pub/bdr/NNSM.pdf`, 1998. Online; September 2012.

E. Shaaban, Y. Helmy, A. Kherd, and M. Nars. A proposed churn prediction model. *International Journal of Engineering Research and Application*, 2: 693–697, July 2012. Online; June 2013.

A. Sharma and P.K. Panigrahi. A neural network approach for predicting customer churn in cellular network services. *International Journal of Computer Applications*, 27:26–31, August 2011. Online; June 2013.

M. Shashanka and M. Giering. Mining retail data for targeting customers with headroom. *Artificial Intelligence Applications and Inovations III Journal*, 296:347–355, 2009.

Dr. M. Turhan. Neural networks and their supervised training. `rocksolidimages.com/pdf/neural_network.pdf`, 1995. Online; April 2013.

C.P. Wei and I.T. Chiu. Turning telecomunication call details to churn prediction. *Expert Systems with Applications*, 23:103–112, 2002. Online; May 2013.

Knowledge Discovery on Customer Churn, volume 10, Dallas, Texas, USA, November 2006. WSEAS. Online; May 2013.

S.J. Yen and Y.S. Lee. Cluster based under sampling approach for imbalance data. *Expert Systems with Applications*, 36:5718–5727, 2009. Online; June 2013.

K.H. Zou, A. O'Malley, and L. Mauri. Receiver operating characteristic analysis for evaluating diagnostic tests and predictive models. `http://circ.ahajournals.org/content/115/5/654`, 2007. Online; March 2012.

Appendix

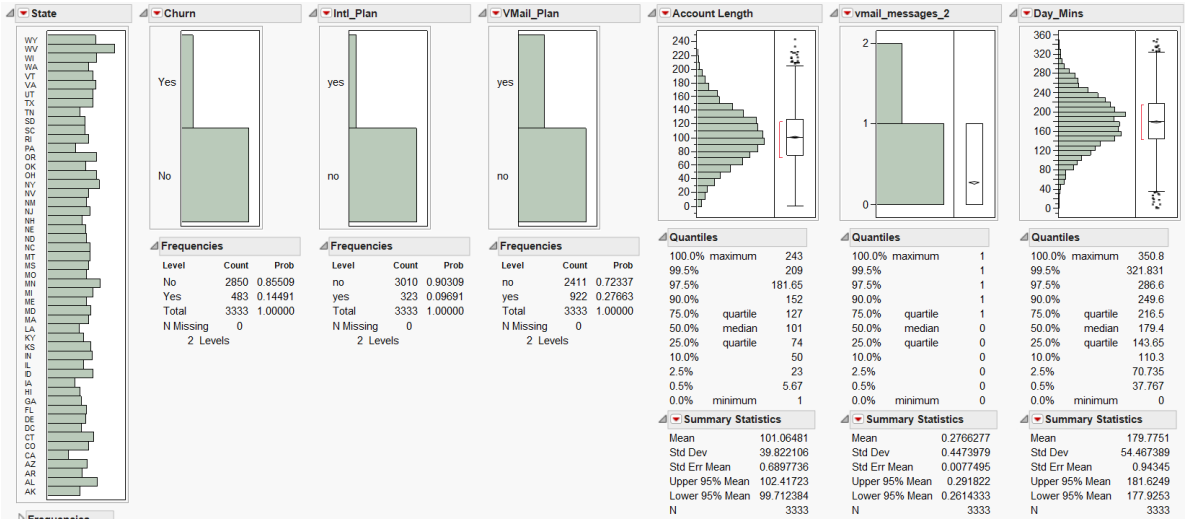
Table A1: Data set 1 variables

Variable	**A brief description**
State	Categorical variable, for the 50 states and the district of Columbia
Account length	Continuous variable for how long account has been active
Area code	Categorical variable, for area of the customer
Phone number	Customer phone number, primary key in the database
International Plan	Binary variable, yes or no
Voice Mail Plan	Binary variable, yes or no
Number of voice mail messages	Integer-valued variable
Total day minutes	Continuous variable for number of minutes customer has used the service during the day
Total day calls	Continuous variable
Total day charge	Continuous variable based on foregoing day calls and minutes
Total evening minutes	Continuous variable for minutes customer has used the service during the evening
Total evening calls	Continuous variable
Total evening charge	Continuous variable based on previous two variables
Total night minutes	Continuous variable for storing minutes the customer has used the service during the night
Total night calls	Continuous variable
Total night charge	Continuous variable based on foregoing night calls and minutes
Total international minutes	Continuous variable for minutes customer has used service to make international calls
Total international calls	Continuous variable
Total international charge	Continuous variable based on foregoing two variables
Number of service call	Continuous variable

Table A2: Data set 2 variables

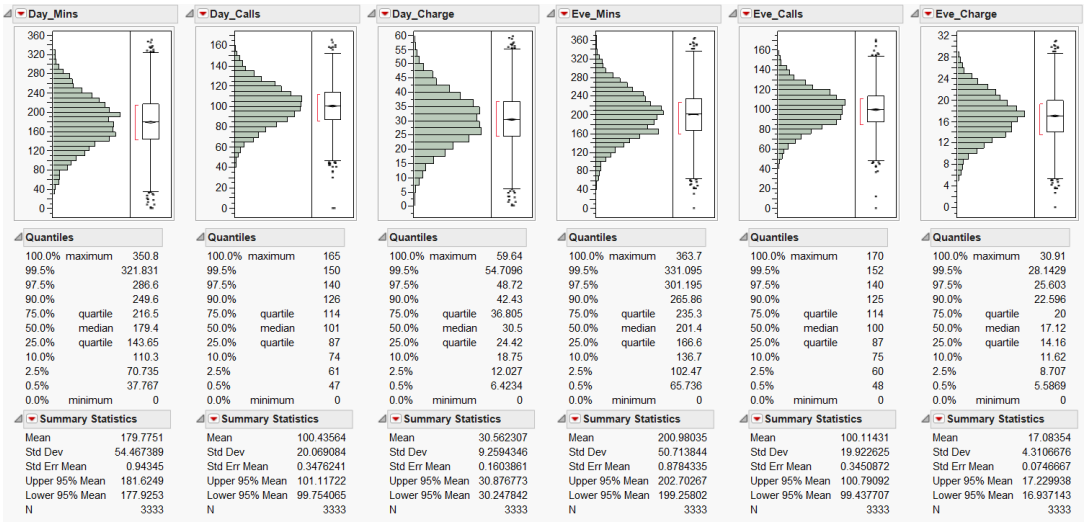
Variable	**A brief description**
Customer ID	Integer, unique client identifier
Npa	Integer, area code one
Nxx	Integer, area code two
Account type	Nominal, account type, standard or premium
Credit class	Ordinal, customer credit class, four classes
Deposit	Continuous, deposit
Auto pay	Binary, auto pay, yes or no
Market	Three market places, Location
Channel	Two acquisition channels, direct or indirect
Dealer code	Categorical, acquisition dealer code
Dealer group	Nominal, acquisition dealer group
Maker	Phone colour, converted from one to six
Start date	Date, start of contract
Stop date	Date, stop of contract
Time to default	Continuous, derived time taken to default
Is active	Binary, account active status, yes or no
Account status	Nominal, account status
Account status dtl	Nominal, account deferred tax status
Effective date	Date, effective contract date
Account Age	Continuous, derived age on book
Initial monthly fee	Continuous, initial monthly fee paid
Contract end date	Date
Birthday	Date, customer date of birth
Age	Continuous, derived customer age
Safety flag	Binary, one or zero
Cut-off date	Date, sample cut-off date

Figure A1: Data set 1 distribution A



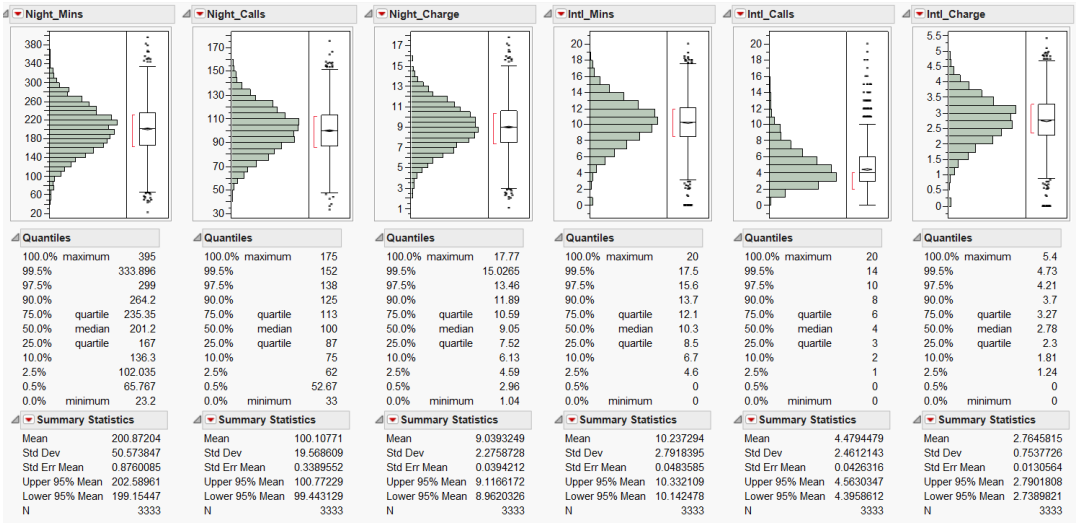
The above chart shows data set 1 variables distribution with descriptive statistics

Figure A2: Data set 1 distribution B



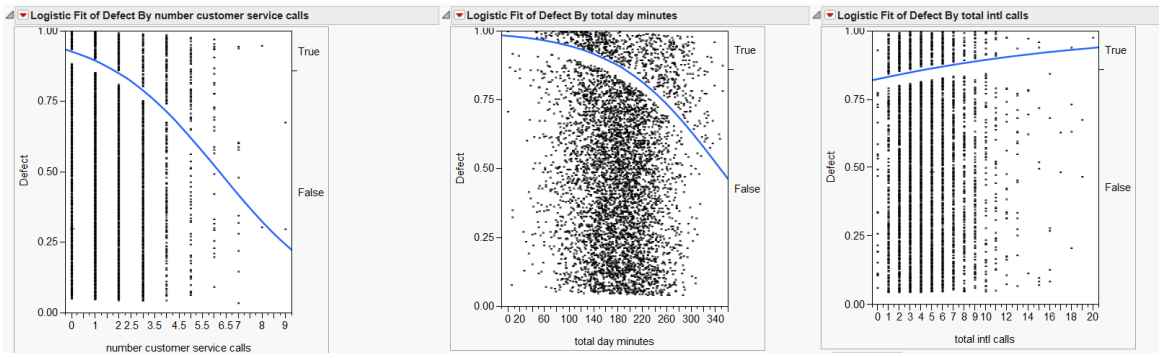
The above chart shows data set 1 variables distribution with descriptive statistics

Figure A3: Data set 1 distribution C



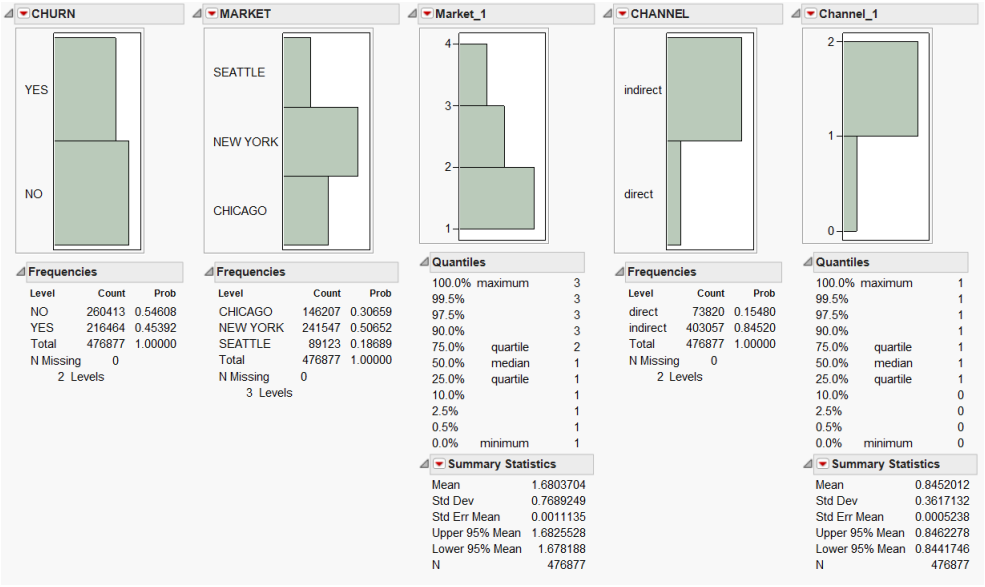
The above chart shows data set 1 variables distribution with descriptive statistics

Figure A4: Data set 1 bi-variate logistic fit



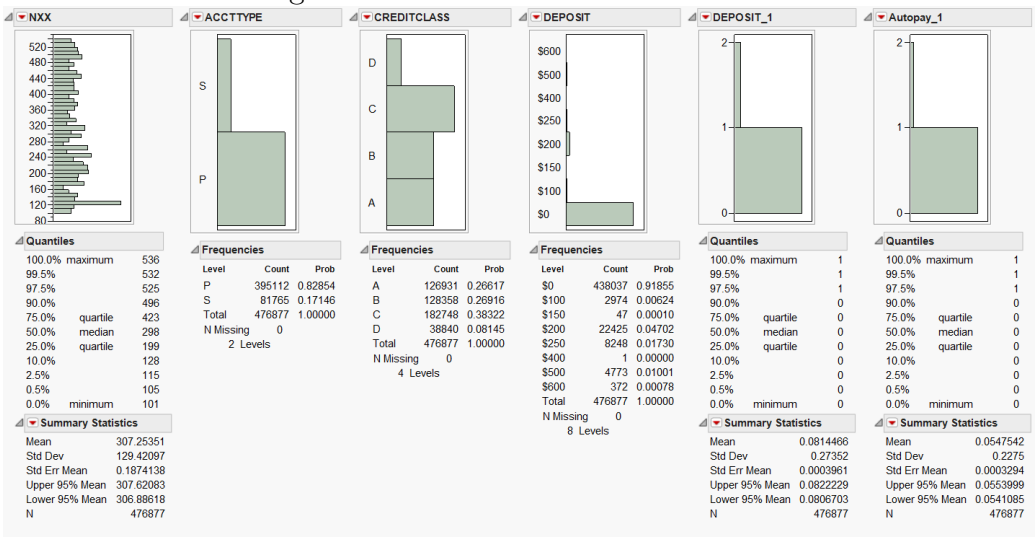
The chart above shows a bi-variate logistic fit between the class variable *churn* and 3 exploratory variables (*number of service calls*, *total day minutes* and *total international calls*)

Figure A5: Data set 2 distributions A



Data set 2 variables distributions with *channel* and *market* transformed to an ordinal variable *channel* and *marker1* respectively.

Figure A6: Data set 2 distributions B



Data set 2 variables distributions with *deposit* transformed to a binary variable *deposit1*

Figure A7: Data set 1 kernel SVM fit

POLY	NO	YES	TOTAL
FALSE	545	79	624
TRUE	10	33	43
TOTAL	555	112	667
Training Error			0.41%
Cross validation Error			14.55%

POLY	NO	YES
FALSE	87.3%	12.7%
TRUE	23.3%	76.7%
AUC		86.9%
Misclasification (Test)		13.34%

LAPLACE	NO	YES	TOTAL
FALSE	547	48	595
TRUE	8	64	72
TOTAL	555	112	667
Training Error			1.46%
Cross validation Error			7.99%

LAPLACE	NO	YES
FALSE	91.9%	8.1%
TRUE	11.1%	88.9%
AUC		86.9%
Misclasification (Test)		8.40%

RBF	NO	YES	TOTAL
FALSE	544	51	595
TRUE	11	61	72
TOTAL	555	112	667
Training Error			2.55%
Cross validation Error			8.36%

RBF	NO	YES
FALSE	91.4%	8.6%
TRUE	15.3%	84.7%
AUC		86.9%
Misclasification (Test)		9.30%

SVM fit results for data set 1 (polynomial of degree 3, Laplace and radial basis kernels)

Figure A8: Data set 2 kernels SVM fit

POLY	FALSE	TRUE	TOTAL
FALSE	482	169	651
TRUE	41	262	303
TOTAL	523	431	954
Training Error			18.2%
Cross Validation Error			22.5%

POLY	NO	YES
FALSE	74.0%	26.0%
TRUE	13.5%	86.5%
AUC		82.8%
Misclass Rate (Test)		22.0%

LAPLACE	FALSE	TRUE	TOTAL
FALSE	460	156	616
TRUE	63	275	338
TOTAL	523	431	954
Training Error			11.9%
Cross Validation Error			21.5%

LAPLACE	NO	YES
FALSE	74.7%	25.3%
TRUE	18.6%	81.4%
AUC		82.9%
Misclass Rate (Test)		23.0%

RBF	FALSE	TRUE	TOTAL
FALSE	478	157	635
TRUE	45	274	319
TOTAL	523	431	954
Training Error			17.8%
Cross Validation Error			20.9%

RBF	NO	YES
FALSE	75.3%	24.7%
TRUE	14.1%	85.9%
AUC		83.80%
Misclass Rate (Test)		21.2%

SVM fit results for data set 2 (polynomial of degree 3, Laplace and radial basis kernels)

Figure A9: Correlation table for data set 1

Variable	vmail_messages_2	Day_Mins	Day_Calls	Day_Charge	Eve_Mins	Eve_Calls	Eve_Charge	Night_Mins	Night_Calls	Night_Charge	Intl_Mins	Intl_Calls	Intl_Charge	service_calls_2
vmail_messages_2	1.0000	-0.0017	-0.0111	-0.0017	0.0215	-0.0064	0.0216	0.0061	0.0156	0.0061	-0.0013	0.0076	-0.0013	-0.0238
Day_Mins	-0.0017	1.0000	0.0068	1.0000	0.0070	0.0158	0.0070	0.0043	0.0230	0.0043	-0.0102	0.0080	-0.0101	-0.0072
Day_Calls	-0.0111	0.0068	1.0000	0.0068	-0.0215	0.0065	-0.0214	0.0229	-0.0196	0.0229	0.0216	0.0046	0.0217	-0.0186
Day_Charge	-0.0017	1.0000	0.0068	1.0000	0.0070	0.0158	0.0070	0.0043	0.0230	0.0043	-0.0102	0.0080	-0.0101	-0.0072
Eve_Mins	0.0215	0.0070	-0.0215	0.0070	1.0000	-0.0114	1.0000	-0.0126	0.0076	-0.0126	-0.0110	0.0025	-0.0111	-0.0146
Eve_Calls	-0.0064	0.0158	0.0065	0.0158	-0.0114	1.0000	-0.0114	-0.0021	0.0077	-0.0021	0.0087	0.0174	0.0087	0.0016
Eve_Charge	0.0216	0.0070	-0.0214	0.0070	1.0000	-0.0114	1.0000	-0.0126	0.0076	-0.0126	-0.0110	0.0025	-0.0111	-0.0146
Night_Mins	0.0061	0.0043	0.0229	0.0043	-0.0126	-0.0021	-0.0126	1.0000	0.0112	1.0000	-0.0152	-0.0124	-0.0152	-0.0098
Night_Calls	0.0156	0.0230	-0.0196	0.0230	0.0076	0.0077	0.0076	0.0112	1.0000	0.0112	-0.0136	0.0003	-0.0136	-0.0156
Night_Charge	0.0061	0.0043	0.0229	0.0043	-0.0126	-0.0021	-0.0126	1.0000	0.0112	1.0000	-0.0152	-0.0123	-0.0152	-0.0098
Intl_Mins	-0.0013	-0.0102	0.0216	-0.0102	-0.0110	0.0087	-0.0110	-0.0152	-0.0136	-0.0152	1.0000	0.0323	1.0000	-0.0112
Intl_Calls	0.0076	0.0080	0.0046	0.0080	0.0025	0.0174	0.0025	-0.0124	0.0003	-0.0123	0.0323	1.0000	0.0324	-0.0136
Intl_Charge	-0.0013	-0.0101	0.0217	-0.0101	-0.0111	0.0087	-0.0111	-0.0152	-0.0136	-0.0152	1.0000	0.0324	1.0000	-0.0112
service_calls_2	-0.0238	-0.0072	-0.0186	-0.0072	-0.0146	0.0016	-0.0146	-0.0098	-0.0156	-0.0098	-0.0112	-0.0136	-0.0112	1.0000

The above table shows the correlation between data set 1 variables. Colours range from dark green (strong positive correlation to dark red (negative correlation). These are the variables that have a positive correlation with each other ($\rho = 1$): *Day charge* and *day minutes*,*evening charge* and *evening minutes*, *night charge* and *night minutes* and *international charge* and *international minutes*