Proceedings of 2013 International Conference on
Technology Innovation and Industrial Management
29-31 May 2013, Phuket, Thailand

AN ANALYSIS OF PRICING TELECOMMUNICATIONS NETWORK SERVICES WITH DATA MINING METHODS

Jongsawas Chongwatpol†, NIDA Business School, National Institute of Development Administration, Thailand
email: jongsawas.c@ics.nida.ac.th

ABSTRACT

Research on developing pricing mechanisms for telecommunications service providers has been going on for decades. Many agencies have adopted various pricing schemes to charge their subscribers. However, due to the changes in the economic environment and technological infrastructure, the loss of subscribers is one of the important issues nowadays and these agencies need to adjust their pricing mechanisms to improve retention, to recover the cost of operations, and to maximize profitability. Practically, current pricing mechanisms do not reflect the changes in subscriber behaviors. This study seeks to fill this gap and examines how data mining techniques can help in making telecommunications pricing decisions. Consequently, any telecommunications service providers can evaluate their pricing strategy with respect to the organizational objectives and subscriber satisfaction perspectives.

An in-depth study of a state telecommunications service agency, OneNet - a division of the Oklahoma State Regents for Higher Education, is conducted. OneNet operates as an enterprise-type fund that provide cost-effective, equalized access to advanced network and telecommunications services to educational, governmental, and health care entities. OneNet must recover their costs through billing their subscribers and by justifying appropriations directly from the state legislatures.

Our experiments are based on a data base of 5,000 U.S. domestic subscribers. Many data mining techniques such as stepwise regression model, decision tree, and artificial neural network (ANN) are used to analyze data sets with multiple predictor variables, which include both network and non-network related factors. Our preliminary results show that types of circuits, membership fees, maintenance and repair costs of network-related equipment, and hub locations are the key factors that categorize OneNet's subscribers into four groups. Pricing mechanisms for each group are developed separately based on the identified key factor characteristics. Although we present this research in the context of OneNet, it is equally applicable to other providers of telecommunications services.

Key Words: Data Mining, Pricing, Telecommunications, Neural Network, Decision tree

INTRODUCTION

Research on developing pricing mechanisms for telecommunications service providers has been going on for decades. Many agencies have adopted various pricing schemes to charge their subscribers. However, due to the changes in the economic environment and technological infrastructure, the loss of subscribers is one of the important issues nowadays and these agencies need to adjust their pricing mechanisms to improve retention, to recover the cost of operations, and to maximize profitability.

Practically, current pricing mechanisms do not reflect the changes in subscriber behaviors. This study seeks to fill this gap and examines how data mining techniques can help in making telecommunications pricing decisions. Consequently, any telecommunications service providers can evaluate their pricing strategy with respect to the organizational objectives and subscriber satisfaction perspectives.

An in-depth study of a state telecommunications service agency, OneNet - a division of the Oklahoma State Regents for Higher Education, is conducted. We follow the CRISP-DM model, which is Cross Industry Standard Process for Data Mining. CRISP-DM model is used as a comprehensive data mining methodology and process model for conducting this data mining study by breaking down this data mining project in to six phases: business understanding, data understanding, data preparation, modeling, evaluation, and development. This study is organized as follows. First, related research on pricing telecommunications networks are briefly reviewed in Section 2. In section 3, we take a case study of OneNet, which is considering integrating Data Mining approach to help making pricing decisions. Our research methodology is then discussed in Section 4. Overall results of various predictive modeling and their discussions are presented in Section 5. Conclusion and future research direction are presented in the last section of this paper.

BRIEF LITERATURE REVIEW

Many pricing schemes have been proposed for pricing telecommunications networks. These pricing schemes can be classified into three main categories: cost-based pricing, pricing for best effort services, and pricing with Quality of Service (QoS) guarantees.

Cost-based pricing refers to prices that are directly related to costs. Some of the cost-based pricing models that have been proposed include Fully Distributed Cost (FDC) pricing, Ramsey pricing, and Flat rate pricing. FDC pricing is wildly used as it allocates the total common and shared costs that agency incurs while providing the services to the clients (Courcoubetis and Weber, 2003). Ramsey pricing is a linear pricing scheme that can be used to maximize social welfare and minimize economic misallocation under the constraint of recovering costs. Ramsey prices are sustainable when service providers charge different prices to different customer groups (Berg, 1998). Flat rate pricing is another well-known pricing structure used by service providers. A customer pays a fixed amount for a service at the time the contract is purchased regardless of the actual usage. Customers are charged the average cost of other customers in the same customer group (Courcoubetis and Weber, 2003).

"Best effort" refers to a network service that treats all types of traffic indifferently with no delivery guarantee and with the possibility of traffic loss (Shin et al., 2006). Best effort pricing scheme is employed to overcome the issues of fairness to customers and resource utilization in the case of cost-based pricing when some customers tend to overuse the resources and consequently is resulted in penalizing light users as compared to the heavy one. Usage-based pricing (Li and Wang, 2005, MacKie-Mason and Varian, 1995) is one of the first best-effort pricing schemes introduced to charge the customers for what they actually consume. This pricing scheme can be used to allocate service classes to different uses, to prioritize usage of a congested resource so that customers who value the access the most will get the highest priority, and to recover the costs of providing services. Congestion discount (Keon and Anandalingam, 2005) refers to a pricing approach using price discounts as an incentive to shift demand from congested to uncongested periods in telecommunications systems. Charging flexible contracts (Courcoubetis and Weber, 2003) can benefit both service providers and customers. Customers can vary the amount of bandwidth by changing their contract without the need to predict and reserve maximum resource requirements, while the service providers can provide more services to customers, with or without the need to reserve the resources.

Lastly, Quality of Service (QoS) refers to networks that are capable of providing better service to selected network traffic over various technologies by providing different priorities to different users or data flows, ensuring no traffic loss, and providing timely delivery guarantees (Shin et al., 2006, Guerrero-Ibanez et al., 2010, Keon and Anandalingam, 2005). QoS pricing involves technological enhancements such as Integrated Service (IntServ), Resource Reservation Protocol (RSVP), Multi-Protocol Label Switching (MPLS), and Differentiated Service (DiffServ) architectures (Shin et al., 2006). Thus, QoS pricing schemes introduced in the literature are related to these network architecture issues. For instance, Karsten et al. (1998) have proposed an embedded charging model in the RSVP architecture for an integrated services network. Fankhauser et al. (1998) included the RSVP charging and accounting in the IntServ network. Additionally, Bouras and Sevasti (2005) presented a model for the service provisioning procedure for the deployment of DiffServ-based Service Level Agreements (SLAs) in a bilateral fashion.

This brief literature review of network services pricing schemes is not intended to be comprehensive, but it does illustrate the large number of choices available to internet service providers to select and implement pricing models. However, these pricing schemes do not reflect the changes in customer behaviors, switching from one pricing mechanism to another. Additionally, there are many potentially important factors in both related to network and non-network variables that are neglected in the pricing decision. This study offers a wide range of decision variables based on the pricing evaluation though data mining approach. Consequently, such variables can be included in the current pricing schemes.

Case Example

OneNet, a division of the Oklahoma State Regents for Higher Education, operates as an enterprise-type fund that provide cost-effective, equalized access to advanced network and telecommunications services to educational, governmental, and health care entities. OneNet must recover their costs through billing their subscribers and by justifying appropriations directly from the state legislatures. The main clients served by OneNet are K-12 schools, colleges and universities, career technology centers, courts, libraries, state and federal agencies, and hospitals and clinics. OneNet established a cost-based pricing rate structure by focusing on the allocation of the costs of serving different types of clients and of different bandwidth offerings. Table 1 Presents OneNet's current rate charged to its clients with various bandwidth-based tier charges.

Table 1: Current Rate

No	Name	Mbps	Rate
1	SUD	0	$22.00
2	56K	0.056	$263.00
3	T1	1.5	$514.00
4	T1-OH	1.5	$514.00
5	Ethernet	10	$1,300.00
6	DS3	44.736	$3,510.00
7	Fast Ethernet	100	$2,300.00
8	OC-3	155	*ICB
9	OC-12	622	*ICB
10	Gigabit Ethernet	1000	*ICB

*ICB: Individual Case Basis

The rates OneNet currently charges for its services are quite straightforward, following a simple cost-based pricing model that averages all cost components for all clients at a given bandwidth rate. However, rapid changes in the economic environment, client utilization, and technologies have increased pressures on the network's infrastructure, client connection policies, and the operating budget, raising the question as to whether the current rates are adequate to recover its cost of operation. Another problem is that that current rate does not reflex the changes in i) financial support such as State Appropriations, State Universal Service, and private funding, ii) the operating budgets, or iii)network infra-structure. For instance, if funding support is cut, the revenue generated from these fixed rates will not cover OneNet's cost of operations. In addition, the current rate structure does not reflect the changes in E-Rate, funds from the Universal Service Fund to assist K-12 schools in obtaining affordable telecommunications and internet access. Additionally, other factors such as the high costs of new technology investments and upgrades, the costs to provide additional value-added services, or the cost associated with new bandwidth offerings generate the need to reassess whether OneNet's rates recover its costs of operations.

METHODOLOGY

In this study, we follow the CRISP-DM Model, a popular data mining method as a complete blueprint for this study (Shearer, 2000). After understanding the domain of pricing telecommunication network services and developing the objectives of achieving pricing decision though data mining approach, we begin our analysis by understanding the relevant data source, accessing data quality, and discovering first insights into the data. The next step is toward data preprocessing from the initial raw data to the final dataset, ready for the model development. This preprocessing step takes about 90% of time to clean, transform, construct, and format the relevant data. We then apply analytical data mining techniques to understand and predict rates charged to OneNet's clients. We also need to evaluate and assess the validity and the utility of our developed predictive models before deploying the data mining results into the domain as stated in the objectives of the study. Figure 1 presents the overall CRISP-DM framework of this study.

The first part of this study is to utilize data mining approach to understand the important factors that customers switch from OneNet to other internet service providers. The second part of this study focuses on determining the suitable pricing model for each customer group based on the customer segmentation profile.

Our experiments are based on a data base of 5,000 U.S. domestic subscribers in which 3,708 customers currently subscribed to OneNet's network and 1,292 customers are considered the loss of subscribers. A completed list of variables obtained, which include network and non-related network variables, is given. These variables, for instance, include

- Site ID, type of the organization, circuit speed, business location, equipment type, governing funding support, private funding support, circuit cost, equipment cost, administrative cost, rent expense, equipment maintenance expense, information service, library database subscriptions, membership fees, fiber relocation cost, and intra-agency payment.

Data Preparation

The next step in this analysis is to examine the quality of the data. In order to identify whether any inconsistencies, errors, or extreme values exist in the dataset, frequency distribution, descriptive statistics, and cross-tab analysis are performed. We then assess whether or not the data is complete or has missing values and what variables to be included in the model. Since including such variables with high-missing values in the model or even applying missing value imputation method can lower the quality of our findings, we make an assumption that the model excludes variables with over 50% information missing; see similar methodology from Park and Edington (2001).

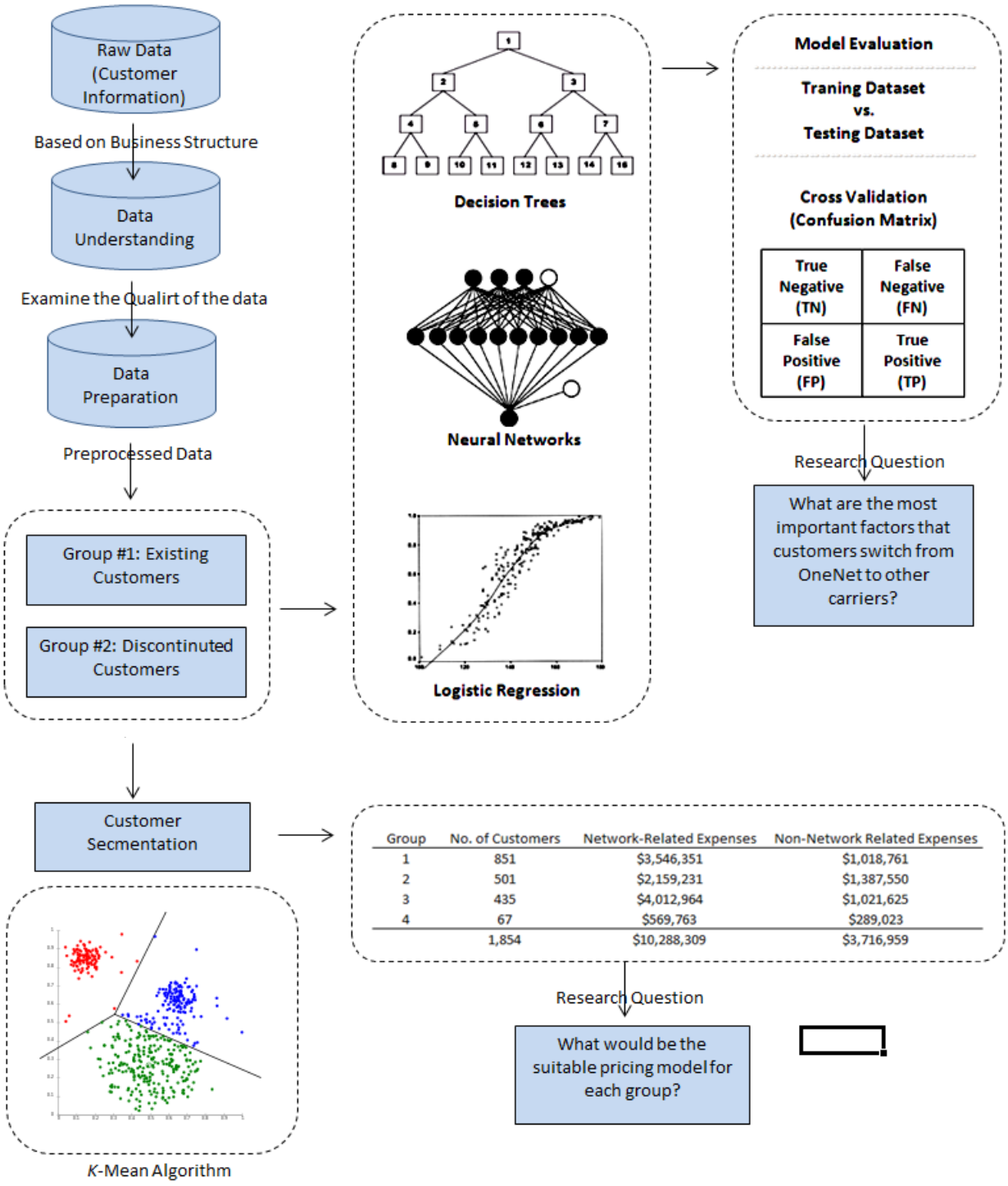


Figure 1: The Overall CRISP-DM Framework

Prediction Model

Many data mining techniques such as logistic regression model, decision tree, and artificial neural network (ANN) are used to analyze data sets with multiple predictor variables, which include both network and non-network related factors.

- Logistic regression is often used to predict an outcome variable that is binary or multi-class dependent variables. It allows the prediction of discrete variables (dependent variables) by a mix of continuous and discrete predictors as the relationship between dependent variables and independent variables is non-linear. It builds the model to predict the odds of its occurrence instead of point estimate event in the traditional linear regression model.
- Decision tree is another data classification and prediction method commonly used due to its intuitive explainability characteristics. Decision tree divides the dataset into multiple groups by evaluating individual data record, which can be described by its attributes. It is also simple and easy to visualize the process of classification where the predicates return discrete values and can be explained by a series of nested if-then-else statements.
- Artificial Neural Network (ANN) is a mathematical and computational model for pattern recognition and data classification through a learning process. It is a biologically inspired analytical technique, simulating biological systems, where learning algorithm indicates how learning takes place and involves adjustments to the synaptic connections between neurons. Data input can be discrete or real valued; meanwhile the output is in a form of vector of values and can be discrete or real valued as well.

For a technical summary including both algorithm and its applications see Jackson, 2002, Turban et al., 2011, and Shearer, 2000)

PRELIMINARY RESULTS

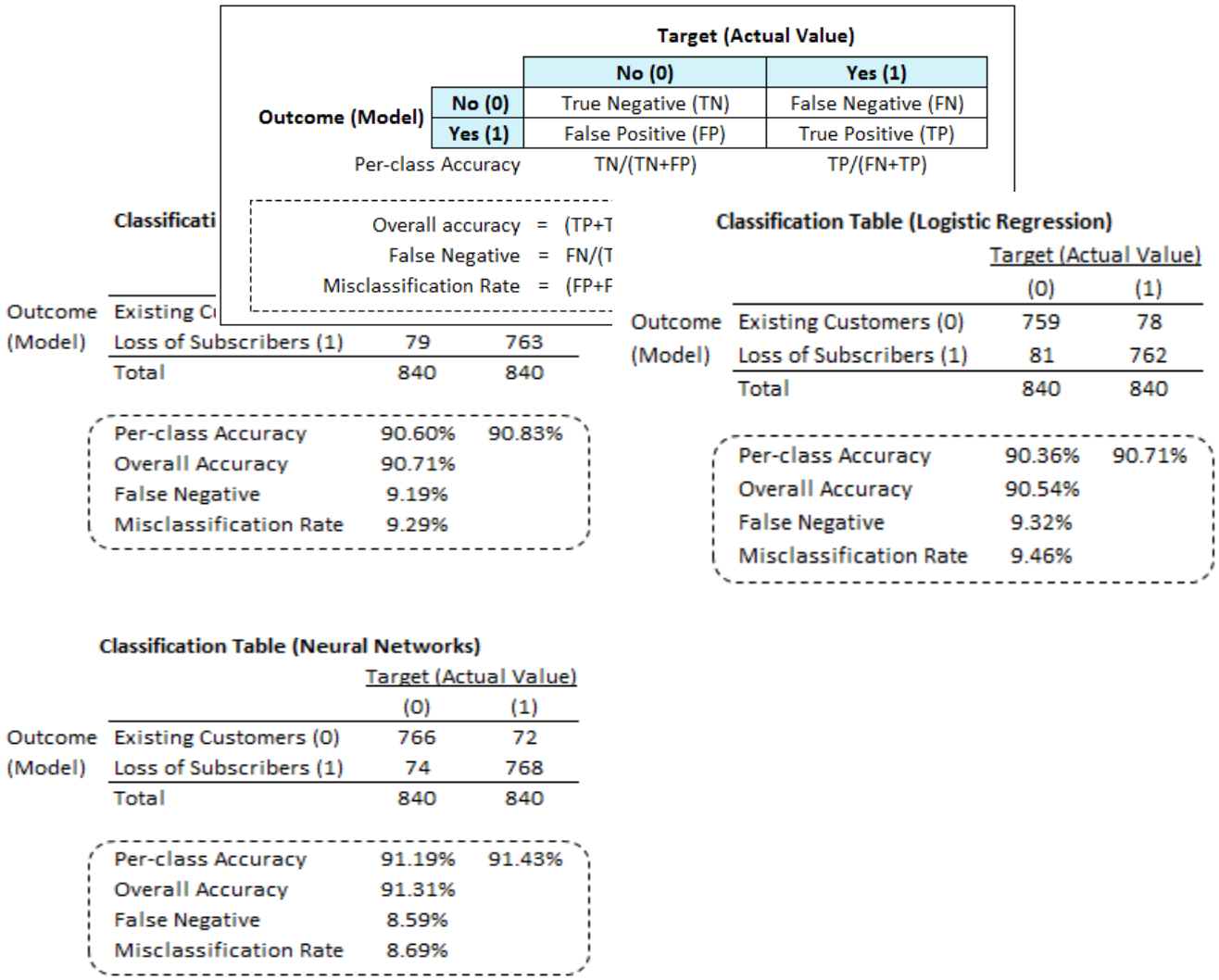
After excluding variables with outliers and high missing values, we first develop predictive models on the original sample dataset, which is composed of 2,400 records (1,200 existing customers and 1,200 losses of subscribers). The binary variable of OneNet's clients (Target = 1 for the loss of subscribers and Outcome = 0 for existing customers) is the output variable of the prediction models. After recoding all categorical input variables, the selected variables are tested whether the association between the input variables and the logit of binary target variable satisfy the linearity assumption. The problematic variables are then transformed to satisfy such assumption. Different models are constructed and compared in order to predict the loss of subscribers. Both training and testing datasets do not differ significantly for any of the variables studied. Training dataset is only used to extract models by the data mining algorithms. Then, those models derived in the training data set are then applied on the testing dataset for the correct discovery of intrusions. In other words, this testing dataset is used to prune the models generated by the data mining process in the training dataset to avoid overfitting and instabilities in the classification accuracy. Statistical analyses are performed using SAS Enterprise Guide 4.3 for data preparation and SAS Enterprise Miner 7.1 for model development and comparison.

We use three different criteria to select the best model on the testing dataset. These criteria include false negative, prediction accuracy, and misclassification rate. False negative (Target = 1 and Outcome = 0) represents the case of an error in the model prediction where model results indicate that hip fracture occurrence is not present, when in reality, there is an incident. The false negative value should be as low as possible. The proportion of cases misclassified is very common in the predictive modeling. However, the observed misclassification rate should be also relatively low for model justification. Lastly, prediction accuracy is evaluated among the three models on the testing dataset. The higher the prediction accuracy rate, the better the model to be selected. The details of performance measures are outlined as follows:

- True Negative (TN): the number of subscribers who are predicted to stay with OneNet and actually are staying with OneNet.
- True Positive (TP): the number of subscribers who are predicted to leave OneNet and actually were moved to other service providers.
- False Negative (FN): the number of subscribers who are predicted to stay with OneNet but actually were moved to other service providers.
- False Positive (FP): the number of subscribers who are predicted to leave OneNet but actually are staying with OneNet.

Figure 2 presents the classification table and the prediction results of the logistic regression, neural networks, and decision trees model. Neural network model produces the best results with overall misclassification rate of 8.69%, followed by the decision trees and logistic regression with misclassification rates of 9.29% and 9.46%, respectively. Neural Network model also has the lowest false negative rate of 8.59% and decision tree model comes out as the runner up with false negative rate of 9.19%. Thus, we select the neural network model as our final model to predict the loss of subscribers.

Our preliminary results show that types of circuits, circuit costs, government funding support, and hub locations are the key factors that lead to the loss of subscriber to OneNet.



		Target (Actual Value)	
		No (0)	Yes (1)
Outcome (Model)	No (0)	True Negative (TN)	False Negative (FN)
	Yes (1)	False Positive (FP)	True Positive (TP)
	Per-class Accuracy	TN/(TN+FP)	TP/(FN+TP)

Overall accuracy = (TP+T
False Negative = FN/(T
Misclassification Rate = (FP+F

Classificati

Outcome (Model)	Existing C		
	Loss of Subscribers (1)	79	763
	Total	840	840

Per-class Accuracy	90.60%	90.83%
Overall Accuracy	90.71%	
False Negative	9.19%	
Misclassification Rate	9.29%	

Classification Table (Logistic Regression)

		Target (Actual Value)	
		(0)	(1)
Outcome (Model)	Existing Customers (0)	759	78
	Loss of Subscribers (1)	81	762
	Total	840	840

Per-class Accuracy	90.36%	90.71%
Overall Accuracy	90.54%	
False Negative	9.32%	
Misclassification Rate	9.46%	

Classification Table (Neural Networks)

		Target (Actual Value)	
		(0)	(1)
Outcome (Model)	Existing Customers (0)	766	72
	Loss of Subscribers (1)	74	768
	Total	840	840

Per-class Accuracy	91.19%	91.43%
Overall Accuracy	91.31%	
False Negative	8.59%	
Misclassification Rate	8.69%	

Figure 2: Prediction Results and Model Comparison on the Testing Dataset

After understanding the causes of loss of subscribers, the next step is to determine the pricing mechanism so that OneNet can use in charging its clients appropriately. We first segment all 3,708 customers into sub-groups that share similar characteristics. The *k*-Means Algorithm is deployed in the clustering process. See Collica 2011, for the details on customer segmentation and clustering using SAS enterprise miner. Table 2 presents the group segmentation after applying the *k*-Means algorithm. Our preliminary results show that types of circuits, membership

fees, maintenance and repair costs of network-related equipment, and hub locations are the key factors that categorize OneNet's subscribers into four groups.

Table 2: Group Segmentation Based on the k-Means Algorithm

Group	No. of Customers	Network-Related Expenses	Non-Network Related Expenses
1	851	$3,546,351	$1,018,761
2	501	$2,159,231	$1,387,550
3	435	$4,012,964	$1,021,625
4	67	$569,763	$289,023
	1,854	$10,288,309	$3,716,959

Linear regression-based pricing is developed for each group. As a non-profit organization, the primary goal is cost recovery; therefore, the values of dependent variables for calibrating models are from the actual cost allocated to individual clients, defined as Rate (R_i) where "*i*" refers to the group members. In this analysis, we split the data at random into two sets (estimation and validation data), build the regression model on the estimation data, apply this regression model on the validation data, and then compare predictive fit and regression estimates between the estimation sample data and validation sample data. We also perform stepwise regression on these independent variables to simplify the model and to determine which variables are among the highest correlation with the dependent variables.

The following equation represented an example of the rate charged to its clients in group #1, where B_i refers to the baseline bandwidth subscription, L_i refers to the Last Mile Costs, and C_i refers to the channelized cost.

$$R_1 = 804.06 + 1.8\ (B_i) - 1.08609(L_i) + 1.853\ (C_i) \quad \text{--------- (1)}$$

The associated independent variables are varied among groups. However, a non-linear regression rate structure can also be appropriated to determine the rates charged to individual customers. For instance, Equation #2 represents the rate charged to member group #2.

$$R_2 = -35.90 + 875\ (B_i^{0.56}) \quad \text{--------- (2)}$$

The rate derived from the non-linear regression is based on the direct variation in bandwidth subscriptions. The higher the bandwidth clients subscribe to, the higher the rate charged to clients. In contrast, the rates derived from the linear regression vary not only depending mainly on the Last Mile, but also on the direct variation of bandwidth subscription and the reverse variation of the client group.

DISCUSSION AND CONCLUSION

With data mining approach, OneNet is able to understand the nature of its customers who switch to other carriers as they can provide the same services with the same or even lower rates. First, the preliminary results show that types of circuits, circuit costs, government funding support, and hub locations are the key factors that lead to the loss of subscriber to OneNet. OneNet can utilize this information to determine what incentives can be adjusted and offered to its clients to improve retention and, meanwhile, recover its cost of operations. For instance, since government funding is one of the key reasons of the loss of subscriber, the current rate charged to its customers should reflect the changes in funding support such as E-Rate, funds from the Universal Service Fund to assist K-12 schools in obtaining affordable telecommunications and internet access.

Secondly, we segment the customers into 4 groups based on customers' characteristics such as types of circuits, membership fees, maintenance and repair costs of network-related equipment, and hub locations. We then develop pricing model that is suitable for each group members. Consequently, OneNet can decrease the perception of unfairness to individual customers who tend to overuse the resources compared to other members in different groups. Linear and non-linear regression models are example of rates charged to customers in groups #1 and 2, respectively.

Note that we present this study as a pilot study to determine whether appropriate data is available, to understand the exploration of data mining approaches, and to develop initial models to determine what factors influence the pricing decision. With only 3,708 data records from an organization, our findings are still limited and the model and the model appears not to be generalizable. Thus, the final paper will include predictive modeling results with larger sample sizes and different organizations so that the finding can be applicable to other providers of telecommunications services.

REFERENCE

1. Berg, S. V. (1998). Basics of rate design: Pricing principles and self-selecting two-part tariffs. Paper presented at the Austrailian Competition and Consumer Commission, Melbourne, Australia.
2. Bouras, C., & Sevasti, A. (2005). Service level agreements for DiffServ-based services' provisioning. Journal of Network and Computer Applications, 28(4), 285-302.
3. Collica, R (2011). Customer Segmentation and Clustering Using SAS Enterprise Miner, Second Edition. Cary, NC: SAS Institute Inc.
4. Courcoubetis, C., & Weber, R. (2003). Pricing Communication Networks: Economics,Technology and Modelling. Hoboken, NJ: John Wiley & Sons Inc.
5. Fankhauser, G., Stiller, B., Christoph, V., & Plattner, B. (1998). Reservation-based Charging in an Integrated Services Network. Paper presented at the 4th INFORMS Telecommunications Conference, Boca Raton, FL,

6. Guerrero-Ibanez, A., Contreras-Castillo, J., Barba, A., & Reyes, A. (2010). A QoS-based dynamic pricing approach for services provisioning in heterogeneous wireless access networks. Pervasive and Mobile Computing, 7(5), 569-583.
7. Jackson, J.(2002), Data mining: a conceptual overview. Communications of the Association for Information Systems, 8: p. 267-296.
8. Karsten, M., Schmitt, J., Wolf, L., & Steinmetz, R. An embedded charging approach for RSVP. In Quality of Service, 1998. (IWQoS 98) 1998 Sixth International Workshop on, 1998 (pp. 91-100)
9. Keon, N., & Anandalingam, G. (2005). A new pricing model for competitive telecommunications services using congestion discounts. INFORMS Journal on Computing, 17(2), 248.
10. Li, K. F., & Wang, J. (2005). A Multi-Objective Internet Pricing Model. NWeSP.
11. MacKie-Mason, J. K., & Varian, H. R. (1995). Some FAQs about usage-based pricing. Computer Networks and ISDN Systems, 28(1-2), 257-265.
12. Park, J. and D.W. Edington (2001), A sequential neural network model for diabetes prediction. Artificial Intelligence in Medicine, 23(3): p. 277-293.
13. Shearer, C. (2000), The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5(4): p. 13-22.
14. Shin, S., F. Cope, R. I., F. Cope, R., & Tucci, J., E. (2006). Internet Pricing: Best Effort versus Quality of Service. Academy of Information and Management Sciences Journal, 9(2), 1.
15. Turban, E., R. Sharda, and D. Delen (2011), Decision Support and Business Intelligence Systems, Pearson