User Behavior Based Clustering and a Decision Tree Model for Predicting Customer Insolvency in Telecommunication Business.

Sunu Mary Abraham[1]

ABSTRACT

This research project deals with a telecommunication application that had an objective of building a prediction model to predict solvent and insolvent customers in telecommunication business. This focuses on two main data mining techniques, clustering of the customer base to identify the significant characteristics of insolvent customers using an unsupervised model and classification of the customers as solvent and insolvent using a supervised learning method. This classification model can be then be used to predict insolvent customers much earlier than it is done today, so that the company can take preventive measures to reduce the losses. The results of the project show that the model built in this research is a useful tool in the decision making process.

Keywords: Data Mining, KDD, Classification, Decision Tree, Clustering, Customer insolvency

1. INTRODUCTION

Data Mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules that may be used to make useful predictions. Innovative organizations worldwide are already using data mining to recognize high-valued customers, to reconfigure their product offerings to increase sales, to minimize losses due to error or fraud etc. Data Mining is applicable to any kind of information repository, whether be it relational databases, data warehouses, transactional databases, spatial databases, multimedia databases or other advanced database systems.

The telecommunication market is rapidly expanding and highly competitive especially in mobile services. The mobile companies generate and store tremendous amount of data. The amount of data is so huge that manual analysis of the data is difficult [8]. This creates a great demand for data mining to extract useful information buried within these data sets. One of the major concerns that affect the company's investment and profitability is bad debts, which might arise due to delay in payment with or without the intention to fraud the company. Customers who make delay in payment with/without the intention to defraud the company can be termed as "insolvent" customers. Telecommunication Companies take precautions against such customers who use the provided services without paying their bills. However, the practice of continuing the service, even after the customer becomes insolvent, for a specific period, increases the losses, and unpaid bills end up in the account of uncollectible bills. In order to help the company to identify such customers well in advance and to take preventive actions to reduce losses, a model capable of early prediction of insolvent customers can be used.

1.1 Objectives

1. Cluster analysis of the customer base to build an unsupervised model for the identification of the most significant characteristics of insolvent customers.

[1]Lecturer, Rajagiri School of Computer Science, Rajagiri College of Social sciences, Kalamassery, Kochi, Kerala. Pin - 683104. E-mail: sunumary@rediffmail.com

2. To build a supervised model for classifying solvent and insolvent customers.

3. To predict insolvent customers using the supervised classifier model.

2. Research Methodology

The research followed a typical Knowledge Discovery in Data (KDD) framework, where data mining is the core in the whole process. The research went through all the nine steps of Knowledge Discovery in Data (KDD) .The research design was both exploratory and descriptive. According to Fayyad et al. the nine steps of Knowledge discovery from data (KDD) are: [2][6]

1. Learning the application domain
2. Creating target data set: Data Selection
3. Data Cleaning and preprocessing
4. Data reduction and transformation
5. Choosing functions of data mining
6. Choosing the mining algorithms
7. Data Mining
8. Pattern Evaluation and knowledge presentation
9. Use of discovered knowledge

2.1 Learning the application domain

The study was conducted in the mobile sector of a major telecommunication company in Kerala for post-paid customers. Customers use their mobile phones for a period of one month, called the billing period. The bill is prepared and issued by the 2nd of the next month. Customers receive their bills approximately 10 days later. The due date is approximately three weeks after the date of issue. The company waits for a week after the due date and if the bill is still not paid takes actions against the insolvent customers. The company temporarily disconnects the phone one week after payment due date if not paid. In the next months, the company sends bill to the customer requesting the amount owed. If the customer arranges to pay the bill, the mobile connection is reestablished with a reconnection fee. But if the customer does not pay the dues after disconnection within 60 days from the statement date of the bills, will be permanently closed. The measures that the company takes against customer insolvency come quite late and these customers take advantage of the service for quite a long time before the service is denied.

This research is conducted on the assumption that calling habits and phone usage changes considerably during a critical period before and after termination of the billing period for insolvent customers. Calling habits combined with previous payment patterns can be used for the prediction of customer insolvency so that early detection of these patterns helps the company to take preventive actions to reduce losses. This problem is similar to other fraud detection problems where fraudsters behave differently from normal customers. Thus this data mining application relies on deviation detection [9]. Customers can become insolvent deliberately or due to financial factors beyond their will. Both types of customers should be detected by the prediction model.

Presently, the company employs a system for monitoring international calls, so that if there is a large deviation in the duration and number of international calls made from a particular number, the company will take appropriate action.

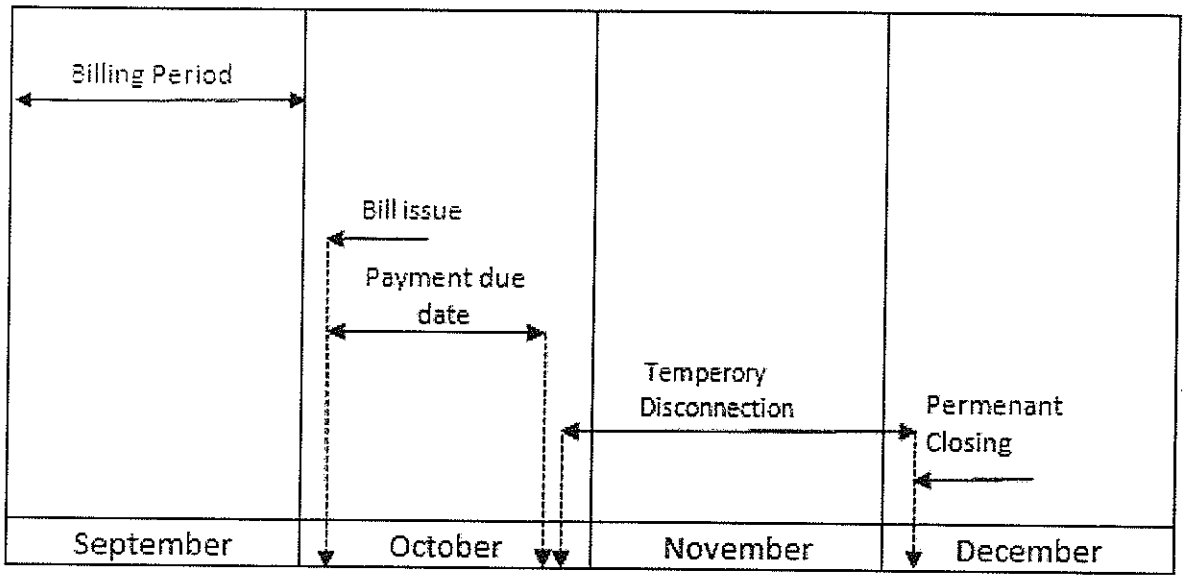

Figure 1 : Billing Process

2.2 Creating target data set: Data Selection

Data for the research came from different sources (data bases) in the organization. The raw data used in the research corresponds to the following departments:

- Systems department
  - Billing data
  - Payment details.
  - Reports of phone disconnections due to failure of payments.
  - Reports of reconnections after payment.
  - Reports of permanent closing of connections
- Switching Center
  - Call detail records

From a total of 17 Lakh company's mobile customers in Kerala 13 Lakh connections are prepaid. Since there is no problem of insolvency in prepaid connections, it is excluded from the sample data set for the study. Out of the 4 Lakh Postpaid customers, the billing & payment data of 4241 customers and the corresponding call detail records are collected for a period of three months. The billing data covered from July 2008 to September 2008 and the call detail records spanned from August 2008 to October 2008. Here, behavior of the customers who became insolvent in October is analyzed. The details of the same set of customers (both solvent and insolvent) in October are taken for the other two months also. This is to study how the normal customers and the customers who have a tendency to become insolvent behave in the past two months before they become insolvent.

2.3. Data Cleaning and preprocessing

To ensure the data to be in a consistent format, missing values and noise were dealt with. Only customers using a specific plan was selected from the original sample since 99% of the mobile users were its subscribers. Another reason for sticking on to the same plan is that different plans have different rentals and schemes of charging calls. This variation may affect the final result.

Also irrelevant records of those customers who were already insolvent for the past few months in the given data set were eliminated from the billing data. Incoming call records, SMS details were removed from the sample CDR data set as only out going call details shows the calling pattern/traffic usage of the user. This will reduce the size of the data set and makes the CDR data more manageable.

Synchronization of Data was done for the following reasons:

- To get the same number of samples (same phone accounts) for all the three months.
- Asynchronous data records (phone accounts) from billing and the corresponding CDR were eliminated.
- To synchronize the payment dates since payment date field for each customer record showed the last month payment date in the sample
- To synchronize the phone accounts that did not appear in the records of payments. This can be got by examining the details of reconnection if it got disconnected by not paying the bill. Otherwise the customer is considered as insolvent. After the elimination of irrelevant data sets and synchronization, data set was reduced to 3641 customers.

2.4 Data reduction and transformation

Data reduction and transformation techniques applied in this research includes dimension reduction (irrelevant or redundant attributes are removed), aggregation, generalization and attribute construction. Attributes not significant to the study were filtered out.

Table 1 Number of fields selected for each type of information

Type of Information	Total Fields	Selected Fields
Billing Information	23	7
Payment Information	3	3
Traffic Usage Information from CDR	43	5
Total	68	15

From a total of 68 fields, from the different types of information system shown in table 1, only 15 fields were selected for the study.

Within the three month study period, information regarding the call transactions are aggregated by two week periods. It was assumed that insolvent customers show a deviation in their calling pattern right after the termination of the billing period (i.e. September 31st, in this context). In the next two weeks, such customers start deviating from the normal calling behavior. The deviation will be maximum during the critical period of one week after the due date after which the phone will be temporarily disconnected.

The traffic usage data was aggregated for five two-week periods (4 two-week periods in the first two months and one two-week period in the third month) because the selected data set consists of customers who become insolvent by the 4th week of October. The insolvent customers need to be predicted while there is still time to take preventive actions. If the insolvent customers can be identified by the end of the first two-week period of October (ie by the 5th two-week period of the study), the traffic usage of the customer for the next two weeks can be controlled thereby reducing the loss of revenue for the company. The information aggregated from CDR include average duration per two week period, average count of

calls per two week period, average count of calls to different numbers per two week period etc.

A new attribute, Period of Delay was added to the structure of billing and payment records. Insolvent customers were identified from the disconnection details given from the systems department and the payment details. A new attribute was added labeling each of the customer as either 'solvent' or 'insolvent'.

2.5 Choosing functions of data mining and mining algorithms

The segmentation of customers according to their billing amount, payment behavior and calling patterns, may be viewed as a clustering problem.

For applying the clustering technique, the data set was further reduced to include only the customers with payment delay greater than zero. This is to identify the characteristics of customers who makes delay in payment and finally becomes insolvent. 753 customers come under this category. Thus, from the above dataset, the unsupervised model was built only with the records of customers who tend to become insolvent. The unsupervised model was developed with k-means clustering algorithm. Following variables were identified to be used in the model using the data within the 3-month study period:

Table 2 Variables selected for Clustering

Sl.no	Variables	Units
1	Average billing value	(in Rs.)
2	Average period of delay	(in days)
3	Average Due	(in Rs.)
4	Average Duration	(in seconds)

Main advantages of K-means are its computational efficiency and its simplicity to understand the results. It groups data using a top-down approach as it starts with a predefined number of clusters and assigns observations to them. This method is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is O (nkt), where n is the total number of objects, k is the number of clusters and t is the number of iterations (normally, k<<n and t<<n). The method often terminates at local optimum[7].

The problem of predicting customer insolvency may be viewed as a classification problem. The distribution of customers was very uneven in the sample data set with 97.12% solvent customers and 2.88% insolvent customers. The number of insolvencies in the data set was very small as very few insolvency cases arise in every given billing period.

As discussed by Weiss and Provost (2001)[10], classiûcation problems with these characteristics are particularly diûcult to solve. Therefore, a new dataset had to be created speciûcally for the data mining function. This is a recommended procedure in similar problems(Dasakalaki, 2006)[2] when there is a minority class to be classified in the classification problem. In this new dataset the distribution of customers between the two classes was altered to approximately 91% of solvent customer and 9% of insolvent customers.

The new distribution between the classes was achieved by eliminating all the records with total due less than Rs.400. Thus, for applying the classification techniques, the data set of 3641 customers was further reduced by eliminating customers whose total due to the company is less than Rs. 400. Now the total number of samples has reduced to 609 .This is used as train and test data set for the classification process.

The elimination of these data did not affect the research since the interest was on detecting patterns of customers with high total due amount. Also the percentage of insolvency was less in low billing values.

Table 3 Percentage of insolvency

Total due (in Rs.)	Total No. of Customers	Solvent Customers	Insolvent Customers	%age of insolvency
<=200	1273	1258	15	1.18
201-300	1231	1219	12	0.97
301-400	528	506	22	4.16
>=400	609	553	56	9.19
Total	3641	3536	105	2.88

The above table makes it clear that insolvency is higher with high dues. ID3 algorithm was chosen to make the decision tree for classifying solvent and insolvent customers. Following variables were identified to be used in the model using the data within the 3 month study period:

Table 4: Variables Identified for Classification

Sl.no	Variables	Units
1	Average billing value	(in Rs.)
2	Average period of delay	(in days)
3	Total Due	(in Rs.)
4	Average duration of calls	(in sec)
5	Average count of calls	(in numbers)
6	Average count of different number of calls	(in numbers)
7	Status of the customer	Solvent/Insolvent

The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node[7]. In order to select the attribute that is most useful for classifying a given

sets, information gain, a measure in information theory is used. Entropy, another measure is used in ID3 algorithm and in many other decision tree construction algorithm. The entropy of a dataset can be considered to be how disordered it is. The entropy is related to information i.e higher the entropy, or uncertainty, of some data, the more information is required in order to completely describe that data. In building a decision tree, the entropy of the dataset is decreased until leaf nodes are reached at which point the subset that is left is pure, or has zero entropy and represents instances all of one class (all instances have the same value for the target attribute).

2.6 Pattern Evaluation and Knowledge Representation

The first objective of the research was to segment the customer base according to their billing amount, calling pattern and period of delay. The clustering was done using the k-means algorithm. The clustering analysis identified 3 different clusters according to the differences in the customer's behavior pattern.

Cluster 1: This cluster can be termed as low bill- low delay that is characterized by subscribers with low billing values, low period of delay and low traffic usage. Such customers do not fraud the company.

Cluster 2 : This cluster can be termed as low bill- high delay that is characterized by subscribers with low billing value, high payment delay and low traffic usage. This is due to the low bill amount and the customers don't consider the low amount seriously to be paid in time and hence the delay. Also, these customers do not have the intension to fraud the company by delaying the bill.

Cluster 3 : This cluster can be termed as high bill- high delay that is characterized by subscribers with high billing value, high payment delay and high traffic usage. It was found that most of the insolvent customers belong to this group. The customers grouped in this cluster thus have a tendency to become insolvent.

The following table and chart shows the centroids of each cluster for the variables used and population percentage of each cluster.

Table 5 Centroids of Resultant Clusters

Variables	Low bill -low delay	Lowbill -high delay	High bill -high delay
Average Total due (Rs.)	377.6143	494.2075	1212.3152
Average Billing (Rs.)	327.695	361.2964	1311.2347
Period of Delay(Days)	6.4	10.9434	14.5
Average Duration(sec)	6113.3954	6392.3954	21,124.2375
Population	51%	31%	18%

The most significant characteristics were identified as high billing value, high total due, high period of delay and high traffic usage. The second objective was to build a classification model for classifying solvent and insolvent customers using supervised learning. Variables used for building the classifier are average billing value, average total due, average period of delay of the three months, average duration, average count of calls and average count of calls to different numbers in the five two week period.

The classification model was built using the ID3 algorithm for constructing the decision tree. The discretized data was used as the data set. Out of the 609 records selected for the classification process, 340 records were used as the training data set to train the algorithm and rest of the 269 records were used to test the classifier model. The training set contained 35 insolvent cases and test data set contained 21 insolvency cases.

The decision tree consists of 15 leaves and 7 nodes in a three level hierarchy.

From the rules generated from the decision tree, 2 types of insolvent customers can be identified

? Customers who become insolvent due to financial reasons. Such type of customers is not deliberately trying to fraud the company.(11/35 of insolvent customers)

? Insolvent Customers who deliberately try to fraud the company by misusing the services provided

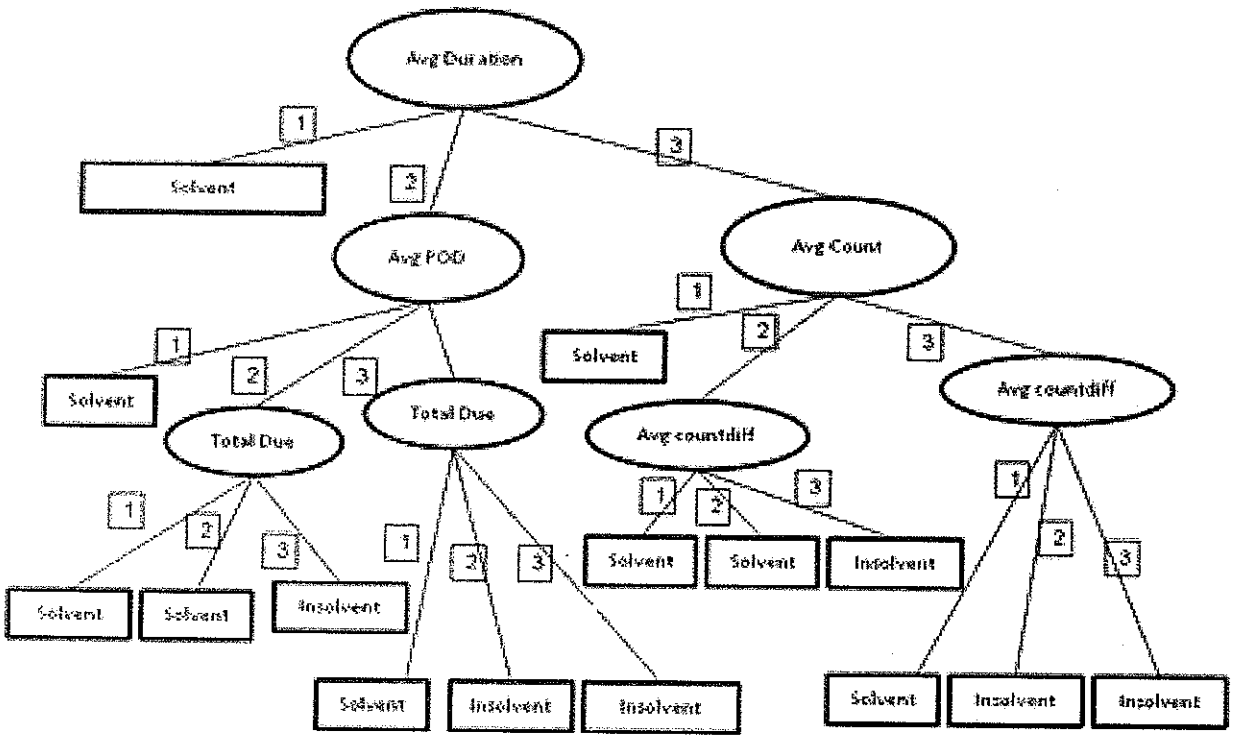

Figure 2. Decision Tree generated using ID3 Algorithm

with the intention of not paying the bill (24/35 of insolvent customers).

This shows that there is a deviation in the behavior of insolvent customers when compared to solvent customers in terms of their calling pattern, period of delay, total due owed to the company etc.

The company can make decisions on the customers according to the way in which the customers become insolvent. The service given to the insolvent customers can be controlled or blocked in the next two weeks before the due date of payment when the insolvent customers will have a tendency to misuse the phone.

The third objective was to use the classifier model built to be used for predicting customer insolvency. The decision tree model was traversed with the test data set of 269 cases with 21 insolvent customers. 257 cases were classified correctly and 12 cases were classified incorrectly. 19 out of 21 insolvent cases were correctly classified as insolvent.

Predictive accuracy of the model can be calculated as the percentage of test samples that are correctly classified i.e. here 95% have been correctly classified.

This shows that the decision tree model is an effective method for classifying solvent and insolvent customers in this context.

3. Conclusion

This research project was an attempt to use data mining techniques for predicting customer insolvency in the telecommunication sector. The results obtained are considered significant. The research study involved a real life application problem. Here two kinds of models were developed. One unsupervised clustering model for identifying the significant characteristics of insolvent customers and a supervised classification model for insolvency prediction. Data was collected from different sources of the department for 4841 customers for a period of 3 months. The data was cleaned, preprocessed, reduced and transformed for extracting relevant information for clustering and classification.

The clustering model allowed the company to understand different group behaviors who made delay in payments and accordingly take actions. The knowledge extracted from the clustering model helped to identify the significant characteristics of insolvent customers which formed a particular cluster.

The supervised classification model was built on a training data set. This model allowed predicting customer insolvency well in advance so that the action measures can be taken against the insolvent customers. 95% prediction accuracy was achieved employing the decision tree classification model in the research. Overall performance of the system should be considered good, since the accuracy of the predictions compare well with other reported problems. This model also identified two types of customers- customers becoming insolvent due to financial problems and those customers becoming insolvent deliberately by misusing the service provided with the intention of not paying the bill. This study has contributed to use data mining as a tool for efficiency in detecting fraud in similar service sectors especially classification techniques and clustering techniques.

References

[1] Burge, P., & Shawe-Taylor, J. (2001):. 'An unsupervised neural network approach to profiling the behavior of mobile phone users for use in fraud detection', Journal of Parallel and Distributed Computing,(61), 915-925.

[2] Daskalaki S, Kopanas I, Goudara M, Avouris N (2003): 'Data mining for decision support on

customer insolvency in telecommunications business', *European Journal of Operational Research* (145),pp.239-255

[3] Daskalaki Sophia, Kopanas Ioannis, Avouris Nikolaos: 'Evaluation of classifiers for an uneven class distribution problem1'(2006): A draft of a manuscript accepted for publication in Applied Artificial Intelligence.

[4] Daskalaki Sophia, Kopanas Ioannis, Avouris Nikolaos(2003): 'Machine Learning Techniques for Prediction of Rare Events in a Business Environment'

[5] Estévez, P.A., Held, C.M., Perez, C.A.(2006):'Subscription Fraud Prevention in Telecommunications Using Fuzzy Rules and Neural Networks', Expert Systems with Applications, Vol. 31, No. 2

[6] Fayyad U.M, Piatetsky-Shapiro G, & Smyth P. (1996): 'From data mining to knowledge discovery in databases', AI Magazine, 17(3), pp. 37-54.

[7] Jaiwei Han; Micheline kamber; Data mining Concepts and Techniques; Morgan Kaufmann Publishers

[8] Pinheiro Carlos Andre R.; .Evsukoff Alexandre G; Ebecken Nelson F.F (2006): 'Revenue Recovering with Insolvency Prevention on a Brazilian Telecom Operator', *SIGKDD Explorations,* Volume 8, Issue (1).

[9] Weiss Gary M. (2005). 'Data Mining in Telecommunications'. In O. Maimon and L. Rokach (.eds), Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, pp.1189-1201.

[10] Weiss Gary. M (2004). 'Mining with Rarity: A Unifying Framework', SIGKDD Explorations, 6(1):pp.7

Authors Biography

Ms. Sunu Mary Abraham is currently working as lecturer in Rajagiri School of Computer Science, Kochi. She holds an MPhil and a Post Graduate degree in Computer Science from Bharathiar University. She has publications in National Journals and her research and teaching interests are in the field of Data Mining, Data Warehousing and Microprocessor Systems.