Using Data Mining to Combat Infrastructure Inefficiencies: The Case of Predicting Non-payment for Ethiopian Telecom

Mariye Yigzaw[1], Shawndra Hill[2], Anita Banser[3], Lemma Lessa[4]
Department of Information Science[1,4], Operations and Information Management[1,2]
Addis Ababa University, Ethiopia[14], University of Pennsylvania, Philadelphia, PA[2,3]
myfa1992@yahoo.com [1], shawndra@wharton.upenn.edu [2], bansera@seas.upenn.edu [3], lemma.lessa@gmail.com[4]

Abstract

Data mining and machine learning technologies for business applications have evolved over the past two decades, and are regularly applied in contemporary organizations to everything from manufacturing to online advertising in fields ranging from health care to motor racing. Unfortunately, data mining techniques are not applied as often to problems in the developing world. Despite the fact that some industries, such as banks, airlines, courts, and telecommunications firms, necessitate data storage as part of their business process. We argue that data mining could be used to reduce infrastructure inefficiencies, which is one of the largest problems faced by Africa. We demonstrate that we can potentially reduce the infrastructure inefficiency of the Ethiopian telecommunications industry by ranking customers according to their likelihood of nonpayment using a data mining approach.

Introduction

Information is the life-blood of contemporary organizations, particularly those conducting business. Business organizations use information to maximize their market share, gain competitive advantage, minimize costs, and increase revenue. Modern-day competition among business organizations requires the right information, at the right time, in the right place, and to the right person. Indeed, information forms the cornerstone of any new development.

In service-oriented businesses such as telecommunications (telecom), where customer contact is very frequent and the number of transactions recorded is huge, information must be extracted from very large datasets. Thus, data processing has become a very costly effort. Our ability to generate and collect data has increased in the last several decades, due to various contributing factors (Han and Kamber 2001). Similarly, the database sizes have significantly increased into terabytes of data. Within these data masses lies hidden information of strategic importance. The prolific data collected and stored in numerous large databases far exceeds our ability to analyze without powerful tools (Hill et al. 2006). Data mining tools and techniques play a fundamental role in this process.

Data mining is useful in any field/business where there are large quantities of data from which to extract meaningful patterns and rules (Berry and Linoff 2000). Data mining has attracted worldwide attention in recent years due to the availability of huge amounts of electronic data and the imminent need to turn such data into useful knowledge (Han and Kamber 2001). Information gained from data mining techniques can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration (Han and Kamber 2001). Data mining can also help to reduce company inefficiencies by making good predictions about business outcomes.

The applications of data mining techniques to telecommunications are numerous; they include predicting which customers are likely to default on payments, identifying telecommunication patterns, catching fraudulent activities, improving service quality and resource utilization, and facilitating multidimensional data analysis to improve understanding of customer behavior (Berry and Linoff 2004). Among the many reasons that data mining is especially valuable to telecom firms, the following are relevant to the present study (Mattison 1997):

- **Data intensity:** Telecommunications is one of the most data-intensive industries, the main product being the call or connection. In an environment with so much raw data, data mining paradigms are ideal for monitoring network performance, issuing bills, and network planning and optimization.
- **Analysis dependency:** Telecom firms are dependent on raw, abstract data to generate bills and measure network effectiveness. Data mining is the best method for processing such large datasets.
- **Historical precedent:** The telecom industry has a rich history of data management innovations: an ideal environment for developing powerful mining tools.

Data mining can be used to protect telecom operator revenues due to fraud or customer insolvency (Estevez, Held, and Perez 2006). Such techniques were previously applied to a Brazilian telecom operator to build pattern recognition models for insolvent customer behaviors, based on previous identification of non-payment events (Pinheiro, Evsukoff, and Ebecken 2006). The business target in this previous study was loss reduction by preventing and minimizing the

effects of bad debt events within the residential fixed lines. There are separate data mining applications for fraudulent behaviors and insolvency behaviors, since insolvent customers do not necessarily intend to defraud the company (Pinheiro, Evsukoff, and Ebecken 2006). Separate data mining approaches are also required for fraud prevention (i.e., measures to avoid fraud before it occurs) and fraud detection (i.e., methods to quickly identify fraud after it occurs) (Estevez, Held, and Perez 2006).

A recent report by the World Bank and Africa Infrastructure Country Diagnostic (AICD) cited infrastructure inefficiencies (extendable to those of telecom firms) as a major challenge facing African nations:

> Africa's power and water utilities present very high levels of inefficiency in terms of undercollection of revenues and distribution losses. Utilities typically collect only 70 to 90 percent of billed revenues, and experience distribution losses that can easily be twice as high as technical best practice. According to household surveys, around 40 percent of those connected to utility services do not appear to be paying for them, and the share rises to 65 percent for a significant minority of countries. It is not unusual for the revenues lost as a result of these inefficiencies to exceed the current turnover of the utilities by several multiples.

To our knowledge, very few firms and government agencies in Ethiopia are currently utilizing extant technologies to reduce inefficiencies. In some industries, data are made available daily, yet the firms fail to take advantage of data mining technology. Telecom-focused data mining research has been conducted in Ethiopia. Although not fully related to billing and cancellation, graduating students of Addis Ababa University and the College of Telecommunications and Information Technology (CTIT) have conducted different studies on the application of data mining for improved customer relationship management (CRM). Specifically, the following studies were conducted with regard to the application of data mining techniques to support the activities of the Ethiopian Telecommunications Corporation (ETC):

- Pattern Extraction from Telephone Line Fault Dataset Using Data Mining Techniques (Tamene 2006);
- Data Mining Techniques on Fraud Detection in Telecommunications Networks Using Self-organizing Map (Berhanu 2006);
- Application of Data Mining Techniques to Support Customer Relationship Management (CRM) at Ethiopian Telecommunication Corporation (Fekadu 2004);
- Application of Data Mining Techniques to Customer Classification and Clustering: The Case of Ethiopian Telecommunications Corporation (ETC), Fixed Line (Mulugeta 2009);
- Application of Data Mining Technology to Support Customer Insolvency Prediction at Ethiopian Telecommunication Corporation, Post Paid Mobile Phone Users (Gashaw 2004);
- Data Mining Application in Supporting Fraud Detection on Mobile Communication: The Case of Ethio-Mobile (Jember 2005);
- Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Service (Gebremeskel 2006).

However, there are no published studies of the ETC's CRM with regards to customer complaints or refusal to pay bills, and the resulting actions of cancellation and charging. Therefore, our short case study will focus on this task. We used very simple techniques to rank ETC's customers by their likelihood to experience phone cancellation due to bill non-payment. We built models on attributes we constructed based on usage and change of usage. We found that change in calling behavior is a strong indicator of future non-payment. Prediction and early prevention of default could potentially save the ETC significant money.

Testbed

The ETC has served as the sole, government-owned telecom operator in Ethiopia since the introduction of telecommunications in the country around 1894. Telephone (both wireline and wireless), Internet (dialup and broadband), mobile (pre-paid and post-paid), and other value-added services are among the major telecom services provided by the corporation (ETC, 2008). For the year 2001 (EFY), ETC generated 5.7 billion birr in revenue and 2.48 billion birr in net profit, and provided services for 902,955 fixed, 4,051,703 mobile, and 74,557 Internet subscribers (ETC 2009).

The ETC uses a database system called "USHACOM" to maintain its transactions. Valuable hidden information could be generated if this huge dataset were to be analyzed from different dimensions. Such information could, for instance, help the ETC to identify the most profitable customer types, tailor its services and market strategies, and identify fraudulent customers or those most likely to leave the service. Since it would be impossible to manually analyze such a vast dataset, data mining must be considered.

Each month, thousands of telephone subscribers in Ethiopia complain or refuse to pay their phone bills, resulting in losses totaling millions of dollars. For the year ending 2000 (EFY), the total balance of uncollected bills totaled over 592 million birr, comprising 18% of the ETC's total revenue (3.3 billion birr) (ETC, 2009). Subscriber complaints are attributable to ETC's inefficiency at providing call details to customers, a problem that is currently resolved by the ETC's providing call detail information upon request. While subscribers complain that they are frequently asked by ETC to pay exaggerated bills for services that they did not use, ETC reports that customers frequently complain about paying for what they have actually used. This issue has been a major source of customer dissatisfaction and lost revenue.

Each month, the ETC disconnects and cancels the services of a huge number of subscribers, many of them to be charged in court. The subscribers' refusal or payment, the resulting service cancellation, and the court process bring substantial financial loss to the ETC in the following ways:

1. The ETC repeatedly pays tax for uncollected revenue (bills) each year until the case is settled (either by writing off from the receivables balance or by forced payment of subscribers through legal measures);
2. The ETC must assign many employees to follow these cases at court;
3. The phone lines could remain idle for long periods; and
4. The ETC loses goodwill.

Thus, if the ETC can predict who will likely complain or refuse to pay his/her phone bills and when, this knowledge could be used to implement measures to reduce customer complaints and revenue loss. The ETC could then take preventive/proactive actions. We propose that we can build a quality predictive model using data mining techniques to:

1. Detect possible fraudulent behaviors (watching call usage and restricting/limiting access);
2. Identify areas/locations with unreliable subscribers;
3. Implement *alarming* devices for customers making exceptional calls; and
4. Devise complaint-reduction measures.

Data

The dataset is derived from the ETC's customer and billing record database, and covers the billing period of February 2008 through August 2009. It contains 535,391 customer records, each with several independent variables, including region, start date, and all local and international billing numbers for the entire 18-month period. The dependent variable is the status, which indicates whether or not a line is currently active or inactive.

Variable Descriptions. Each record in the dataset represents a customer, who has been completely anonymized, with the following attributes:

- **Phone code:** A unique identifier derived from the original customer number to ensure privacy;
- **Customer category:** Customer type (e.g., residential, big business, government, international organization, religious, etc.);
- **Start date:** Date that the phone service was turned on;
- **Stop date:** Date that the phone service was turned off. For customers with active lines, this field is blank;
- **Region:** Region of reporting media from which the line originates;
- **Payment mode:** Method of payment (i.e., cash/credit).
- **Reason:** Reason why the phone line was turned off (e.g., payment due, customer requested, GSM move). For active lines, this variable is set as NONE.
- **Status:** The variable we are trying to predict, which indicates the current status of the line, which is either active or inactive (i.e., cancelled, pending, or temporarily disconnected).

Along with the variables listed above, each record also has 18-month variables representing billing numbers for local calls and for international calls.

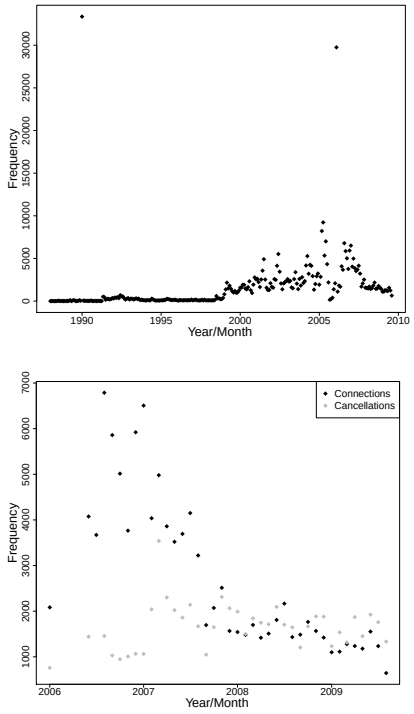

Figure 1: Number of connections by month over the span of time our dataset covers (top) and both connections and cancellations for the past three years of service (bottom).

Descriptive statistics. The variable distributions shed some light on how the data is structured and on potential issues with the dataset. Start distributions plotted by month show a marked increase in phone line connections from 1998 through 2009 (Fig.1, top). While we expect an increase in phone line connections with the introduction of better technology and mobile phones, this sudden increase might indicate potential data collection problems prior to 1998. Also, the high occurrence of January 1990 as a start date confirms faulty data collection.

Stop date distributions, like start date distributions, have unusual spikes during certain months (e.g., March 2007) and little or no activity for others (e.g., January-May 2006) (Fig. 1, right). Again, we assume that these abnormalities might indicate dirty data and/or faulty data collection.

Data limitations. Data is collected and aggregated at a month-to-month level; thus, any potential benefit or information we obtain from examining data at a more granular level (e.g., call-by-call data) is lost. In addition, any errors or biases that might have been introduced in the aggregation phase are transferred and hidden in the data.

There are several missing fields in the dataset. For example, about 9% of start dates are registered as January 1990, which we believe indicates improper or incomplete data collection. Missing data in fields is limiting in general, but it is particularly problematic for a dataset that has few independent variables.

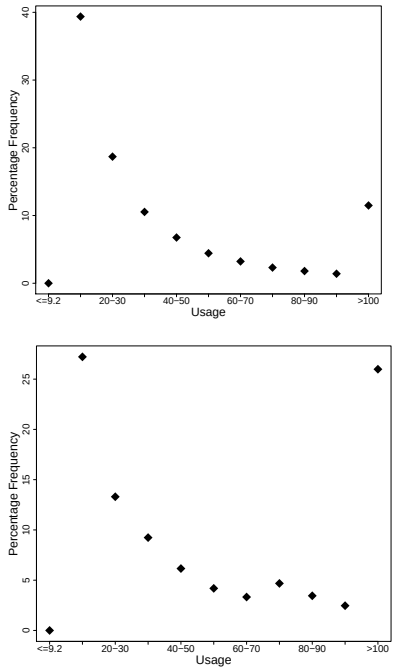

Figure 2: Distribution of usage for active (top) and cancelled phones (bottom). Recently cancelled users were more likely to be high usage individuals.

In addition to the abnormalities mentioned above, there are a few noteworthy procedural steps taken by the ETC before a line is cancelled. These steps are not immediately apparent in the data, but are critical for model construction and data interpretation. There is about a two-month lag period between usage and billing. Therefore, customers get the bill for their current month's usage two months later. If a dispute in billing arises and or if the customer is unable to pay the bill, phone service is temporarily halted until payment is collected. During this period, the customer continues to be charged the standard monthly charge (i.e., 9.2 birr for residential customers, 19.5 birr for business customers). The phone can remain in this state for another two months or more before the line is finally cancelled. Because of this billing setup, or infrastructure inefficiency, data from the two months' prior to termination are very predictive of cancellation. Therefore, when constructing our models, we do not use this information, which would artificially exaggerate the efficacy of our models.

Figure 2 shows the distribution of usage in birr. On average, there were more lines with high usage in May ($>$100 birr) that were cancelled compared to those that were not cancelled. This could indicate either high usage or billing errors.

Method

Preprocessing

Active telephone records were preprocessed under different categories, and a sample was selected so that cancelled

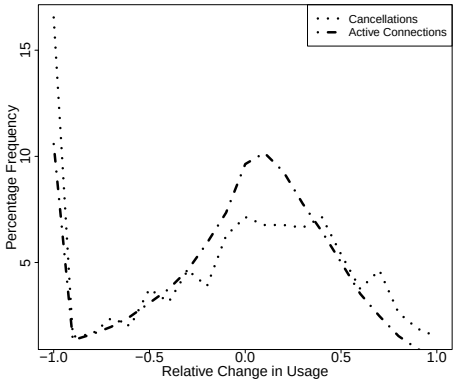

Figure 3: Relative change in usage in May. Users who were recently cancelled were more likely to change their behavior mode than those that were not cancelled. Pattern holds for different months going backward.

phones were equally represented in the data set. Since we want to predict cancellation due to payment default, the 45,396 records were cleaned; as a result, 15,405 entries (bill data from February 2008 to July 2009) were available to be used in the model. Many phones were cancelled in the months prior to February. Thus, the 15,405 entries that were cancelled after April 2008 were selected as having sufficient consumption/billing data.

Based on our understanding of the bill collection process, the following steps were taken to ensure that were not using information that would make our model look deceptively good at predicting cancellation in August:

(1) Our dataset was restricted to users who had $>$ 9.2 usage in May. Users who sustain a 9.2 rent charge (the minimum fee) are most likely already having problems with their phone lines. Because we are trying to predict users whose accounts are currently active but might become inactive in the future, we exclude those whom we know already have problems with their lines. This restriction reduced the dataset to a balanced sample of 1624 users (812 active and 812 cancelled in August).

(2) We removed usage data for June 2009 and July 2009, because the ETC turns off phones for nonpayment months before the phone is cancelled. Furthermore, usage data for June and July may not be available in August.

We then construct the following attributes:

Difference and percentage difference in usage for April and May 2009. It is our understanding that most phone cancellations are a result of disputed bills. We therefore assume that if a user deviates from his/her typical pattern of usage, this might cause a dispute or indicate that the user is no longer interested in keeping the phone line. We plot the distribution of relative change in Figure 3. Note that people who have cancelled phones in August are more likely to have a significant change in their calling behavior between April and May (the month that their phone would first be restricted).

Years since start. and *months since start.* We constructed this attribute by counting the number of years (or months)

since the start date. The intuition behind these attributes is that a phone line that was connected for a longer period would probably be less likely to be cancelled.

International call. We constructed the following attributes using international call usage numbers: The number of months for which a user made international calls before May, the average international call usage before May, the total international call usage before May, and a binary attribute indicating whether or not a user made any international calls before May.

Other attributes. To get a sense of how much a customer used their phone line before potentially encountering problems, we calculated the total and average usages from May backward. To get a sense of the number of users whose phones were billed only the standard monthly bill in the past (which could indicate prior phone suspensions), we calculated the number of months for which usage was > 9.2 from April backward. We looked at April backward because the population on which we built our model included only users who had > 9.2 usage in May

We restricted our dataset to users whose USAGE-3 MONTH values were > 9.2 (the equivalent of restricting the August cancellation data set to users whose May 2009 usage was > 9.2).

Prediction

We used classification techniques such as decision trees, nave Bayes, and logistic regression to predict nonpayment, but due to space limitations, only report our results from the decision tree models. The data described above required a significant amount of cleaning. Our target was set to be all customers who cancelled service in August 2009. The sample on which we built our models was all customers who stopped service in August 2009, plus a random sample of 1992 users who had not stopped service. This was done to obtain a sample that was balanced on the target variable. We removed attributes from which the targets were derived, such as the stop date and reason code. The reported AUC values are the average AUC over 10 independent test sets selected by performing 10-fold cross validation.

Attribute	AUC
Origiinal attributes	0.65
Plus diff	0.66
Plus diff percentage	0.66
Plus years since start	0.66
Plus months since start	0.66
Plus number of months with usage $<= 9.2$	0.67
Plus average usage	0.68
Plus int usage	0.64

Table 1: August 2009 cancellation prediction for residential customers - Results of ranking from different attribute subsets.

Results

Our goal is to rank consumers by their likelihood of non-payment. We evaluate our ranking by the area under the ROC curve (AUC), and compared decision tree models built on subsets of our target attributes. We found that location was the strongest predictor of non-payment; however, billing changes were also significant predictors.

The results of our model show that we can rank customers by their likelihood of non-payment. An AUC of > 0.5 means that we are ranking better than random (i.e. baseline), given that the firm currently has no preventative measures in place. Table 1 shows the results of AUC using decision trees built on different attribute subsets.

Our constructed attributes enabled better ranking than using just the attributes given by the ETC. Note that if we use the change in usage alone, we obtain an AUC of 0.61. We believe that with fine-grained and cleaner data, we would be able to perform even better. In addition, we might use dimensionality reduction techniques to identify the key components in our feature set for prediction as well as identify potentially useful new attributes. Although here we only report results for residential customers, we performed a similar analysis for business and government lines. The sample was smaller, but our ability to rank was about the same in terms of AUC on a balanced subset of about 400 records.

Discussion

We are not the first to propose applying data mining technologies to business data in the developing world. However, despite the availability of technologies and the enormous costs created by inefficiencies, business firms in developing nations do not appear to use methods to predict the likelihood of non-payment. The purpose of this case study was to show how simple techniques may be used on existing data to help such companies combat inefficiencies. Despite very limited and noisy data on service cancellations, we were still able to make predictions that would save the ETC significant money - despite using a simple approach with open source software. Our results provide evidence that this approach is of value, and could be used in other domains. In addition, we should that change in behavior or bill consumption is a key indicator of future non-payment.

It becomes evident through this study that firms, individuals, and government entities may be able to repurpose their data for prediction to reduce costs or increase revenues. To do so, these firms must realize the potential impact that data mining analytics could have on their bottom line, and then take measures to store high quality data. We argue that being able to translate data into information stands to yield tremendous results for the developing world. Thus, data collection from firms, governments, and others should be seen as a priority to move business, communities, and countries forward.

We believe many patterns could have been extracted if the ETC's database had been accessed without restriction on some attributes, and if the records in it had been complete. Surprisingly, the ETC's databases contained no detailed customer profiles (identity, previous payment cancellation, phone use history, and call details) for fixed telephone users in Addis Ababa. Additionally, many of the records were left blank. Although these databases contain the records of phones cancelled back to 2005, they do not contain the associated bill consumption data that far back.

At one time, phone companies in the developed world did not store this data either. Therefore, the first step in eliminating inefficiencies will be to change the culture in the ETC and other entities so that they see value in repurposing existing data for development. We regard this as a tremendous opportunity.

Acknowledgements

We would like to thank the Ethiopian Telecommunications Corporation for generously letting us use their firm as a testbed.

References

Berhanu, H. 2006. *data mining techniques on fraud detection in telecommunications networks using self-organizing map,*. Ph.D. Dissertation, CTIT.

Berry, M., and Linoff, G. 2000. *Mastering Data Mining: the art and science of customer relationship management*. John Wiley and Sons, Inc.

Berry, M., and Linoff, G. 2004. *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*. Indianapolis: Wiley Publishing, Inc, second edition.

Estevez, P. A.; Held, C. M.; and Perez, C. A. 2006. Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications* 31(2):337–344.

2009.

Fekadu, M. 2004. *Application of Data Mining techniques to support customer relationship Management (CRM) at Ethiopian Telecommunication Corporation (ETC)*. Ph.D. Dissertation, Addis Ababa University.

Gashaw, M. 2004. *Application of Data Mining Technology to Support Customer Insolvency Prediction at Ethiopian Telecommunication Corporation*. Ph.D. Dissertation, Addis Ababa University.

Gebremeskel, G. 2006. *Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Service*. Ph.D. Dissertation, Addis Ababa University.

Han, J., and Kamber, M. 2001. *Data Mining: Concepts and Techniques*. Academic Press.

Hill, S.; Agarwal, D.; Bell, R.; and Volinsky, C. 2006. Building an effective representation for dynamic networks. *Journal of Computational and Graphical Statistics* 15(3):584–608.

Jember, G. 2005. *Data Mining Application in Supporting Fraud Detection on Mobile Communication: The case of Ethio-Mobile*. Ph.D. Dissertation, Addis Ababa University.

Mattison, R. 1997. *Data Warehousing and Data Mining for Telecommunications*. Artech Houde, Inc.

Mulugeta, A. 2009. *Application of Data Mining Techniques to Customer Classification and Clustering: The Case of Ethiopian telecommunications Corporation (ETC), Fixed Line*. Ph.D. Dissertation, CTIT.

Pinheiro, C. A.; Evsukoff, A. G.; and Ebecken, N. F. 2006. Revenue recovering with insolvency prevention on a brazilian telecom operator;. *SIGKDD Explorations* 8(1):65–70.

Tamene, F. 2006. *Pattern Extraction from Telephone Line Fault Dataset Using Data Mining Techniques*. Ph.D. Dissertation, CTIT.