Defection Detection

Using Online Activity Profiles to Predict ISP Customer Vulnerability

Nandini Raghavan
DoubleClick R&D
450, W. 33rd Street
New York, NY 10001
nraghavan@doubleclick.net

Robert M. Bell
AT&T Labs Research
180 Park Ave, P.O. Box 971
Florham Park, NJ 07932
rbell@research.att.com

Matthias Schonlau
RAND
P.O. Box 2138, Santa Monica,
CA 90407-2138
matt@rand.org

ABSTRACT

We describe efforts to develop evolving statistical profiles of online behavior patterns for about 2 million users of an internet service provider (ISP). Profiles can be used for a variety of applications in customer relationship management, here we use profiles to predict customer defection. Logistic regression, using features of the profiles and tenure with the service to predict short-term defection, succeeds in identifying a 20 percent subgroup responsible for 50 percent of the customer loss in the following week. Challenges met include: complex mappings across databases at different scales, developing diagnostics for assessing data quality, kneading the data into a form amenable for statistical analysis, while taking into account data storage issues and computational efficiencies of different environments.

Categories and Subject Descriptors

G.3 [**Mathematics and Computing**]: Probability and Statistics

Keywords

Data Mining, Internet Service Provider, Visualization

1. INTRODUCTION

In this paper we describe efforts to develop statistical profiles of customers of an internet service provider (ISP), based on the users' online activity patterns. The profiles reveal a rich diversity in usage patterns, suggesting that they can be culled and used as features in a variety of applications in customer relationship management, from targeted one-to-one marketing campaigns to focused customer services to enhanced strategies for customer acquisition and retention.

The application we discuss here concerns predicting *short-term* defection. We conjecture that statistical models that incorporate *recent* information about *individual* users' online usage patterns will effectively identify customers vulnerable to defection. We believe that these features capture intangibles such as value and convenience to the customer more effectively than gross demographic variables, which are traditionally used for predicting defection.

In the next section we give some background on WorldNet and the business imperative for constructing profiles. In Section 3, we describe our efforts to create user profiles from online activity logs. In Section 4, we describe an application where we extract features from these profiles to build statistical models to predict customer defection. In Section 5 we describe the research challenges we faced in implementing an analysis of this nature for an ISP population of a few million customers and conclude with a discussion of further issues that need to be addressed in this ongoing effort.

2. BACKGROUND

WorldNet, AT&T's internet service business, has a customer base of about 2 million dial-up customers who pay a monthly fee to gain access to the Internet. Its business model is based on generating revenues through monthly subscription fees. WorldNet offers a variety of subscription plans, the prices of which depend upon the number of free hours offered each month.

WorldNet competes with a host of other ISPs for market-share, including the industry heavyweight, AOL, which has about 20 million customers and whose business model also relies heavily on subscriptions. In Fall 1998, NetZero launched a free Internet service that relies on targeted advertising and e-commerce to generate revenues. Its registered user base grew to about 2 million by Fall 1999 and to about 3 million by early 2000. The astounding growth of NetZero prompted a flurry of free offers by other service providers and has forced several ISPs to re-evaluate their business strategies.

The landscape of the competitive marketplace is changing radically. The common wisdom is that understanding online activities is crucial to successful business models. Customer relationship management for internet services requires real-time personalization and sophisticated customer-centric analytical capabilities.

3. PROFILES

In this section we describe activity profiles of *individual* users as well as *aggregate* usage profiles of groups of customers at modem banks.

3.1 Activity Profiles of Customers

WorldNet users exhibit a rich array of activity patterns. The 2 million customers together account for about 3 million sessions a day. Of these, about 25 percent do not log in even once a week. Those who do, log in once or twice a day on average, but such statistics can be very deceptive. The plots in Figure 1 and 2 illustrate the diversity in observed session-usage patterns for eight customers. Each frame shows the sessions for one customer during a 28-day period. Each session is indicated by a "l—", marking the session start in the 24-hour local time on the x-axis and the date, starting with Sunday the 1-st, on the y-axis. The length of the line corresponds to the duration of the session and colors can be used to indicate weekday vs. weekend (not shown here). Customer 1 is typical, having 1-2 sessions a day of varying length, usually on weekends and evenings. Customers 2-4 have about the same number of sessions, but behave very differently. Customer 2 is typically logged in for very short periods. Customers 3 and 4 have much longer sessions, concentrated during the workday and evening, respectively, the latter stretching into the wee hours of the morning. The patterns for customers 5 and 6 shift around mid-month, signaling changes in activity that would require updates to their profiles. In particular, the type of sudden drop in use by customer 6 may portend a decision to drop WorldNet. Customer 7 has frequent short sessions, consistent with frequent checking for email or some other web-based service. Finally, customer 8 is almost always logged on. These plots differentiate users on several "features" of online behavior.

(i) the number of sessions: customers 1-6 vs. customers 7,8.

(ii) the length of their sessions: customers 2,6,7 vs. customers 1,5 vs. customers 3,4,8.

(iii) the timing of their sessions: customers 2,3,7 (mostly business hours) vs. customer 1,4 (non-business hours only).

We can further differentiate by the intensity of the generated traffic (which is not reflected in the plots here). Features reflecting changes in usage patterns from one week to another can be extracted as well. By doing this on an ongoing basis, re-weighting the statistics with a weighted average of usage in the past several weeks, we can create evolving profiles or signatures of customers.

Similar statistics can be culled about email usage. About two-thirds of WorldNet customers do not use email. Those who do are more likely to retain the service, perhaps because notifying friends about a new email address can be cumbersome. Thus email usage may be a useful indicator in customer vulnerability models.

Another feature of potential importance to customers is the congestion encountered when trying to log in. Dial platform calls to WorldNet are piped through modem banks placed at various points in the network. Each modem bank has a capacity that limits the number of simultaneous calls. When demand exceeds capacity, congestion occurs. Intuitively, a customer encountering repeated busy signals is vulnerable to defecting since they are not enjoying the service that they are paying for.

3.2 Traffic Profiles for Modem Banks

In order to characterize the traffic at a given modem bank, we processed dial-in times and session durations to calculate the total number of active sessions during a given time interval. The plots in Figure 3 illustrate traffic patterns at four modem banks during a 12-day period. The x-axis depicts the time during a 24-hour period. The y-axis represents the average number of active sessions at that time during the 12-day period. In general there are strong diurnal patterns in traffic flows, which can be observed throughout WorldNet's territory. Peak usage occurs during the late evening hours from about 7-10pm, with other modes occurring in the mid-morning hours between 10-12 and again in the late afternoon from 3-5pm. Congestion is reflected in the flat-top curves manifest in the upper left plot. Some modem banks tend to be congested most of the day. The congestion level at a modem bank was determined as follows:

(i) create the traffic profile for each of WorldNet's modem banks (as in the plots above).

(ii) calculate the average area under each profile for 24 hour periods.

Thus the feature that we used to capture congestion is the area under the traffic profile curve. We used clustering techniques to evaluate the discriminating power of several statistics for differentiating among traffic profiles and arrived at this particular statistic. We then created a master table of traffic congestion for all the modem banks.

4. AN APPLICATION OF PROFILING: PREDICTING ISP CUSTOMER DEFECTION

The popularity of the "free service" model exacerbated a prevailing problem that ISPs were trying to combat, i.e., customer defection. ISPs face high costs in acquiring new customers. Since the prevailing view is that a customer saved is worth more than a customer earned, there is an imperative to identify strategies that could make vulnerable customers stay.

We define defection as a *customer initiated action* to terminate service. The process by which a customer decides to defect is a complex one. Among the factors that contribute to this decision are quality of service (or lack thereof) and the value that the customer ascribes to the service. Customers can also be removed from the service for a variety of reasons such as fraudulent usage and non-payment of charges. Such customers, who are deemed undesirable, fall outside the scope of our analysis.

In this section we present models to predict short-term defection. Our goal was to study the feasibility of predicting defection in the following week, based on an individual's usage data (and other predictors) during the current week. In practice, the appropriate time-frame for such an analysis may depend upon business realities. The sooner we react, the more difficult it is to predict. However, the longer we wait, the lesser the chances that we will be able to persuade customers to stay. Thus, a week seemed a reasonable trade-off.

The issue of whether we can predict short-term loss based on a customer's recent usage raises a number of interesting

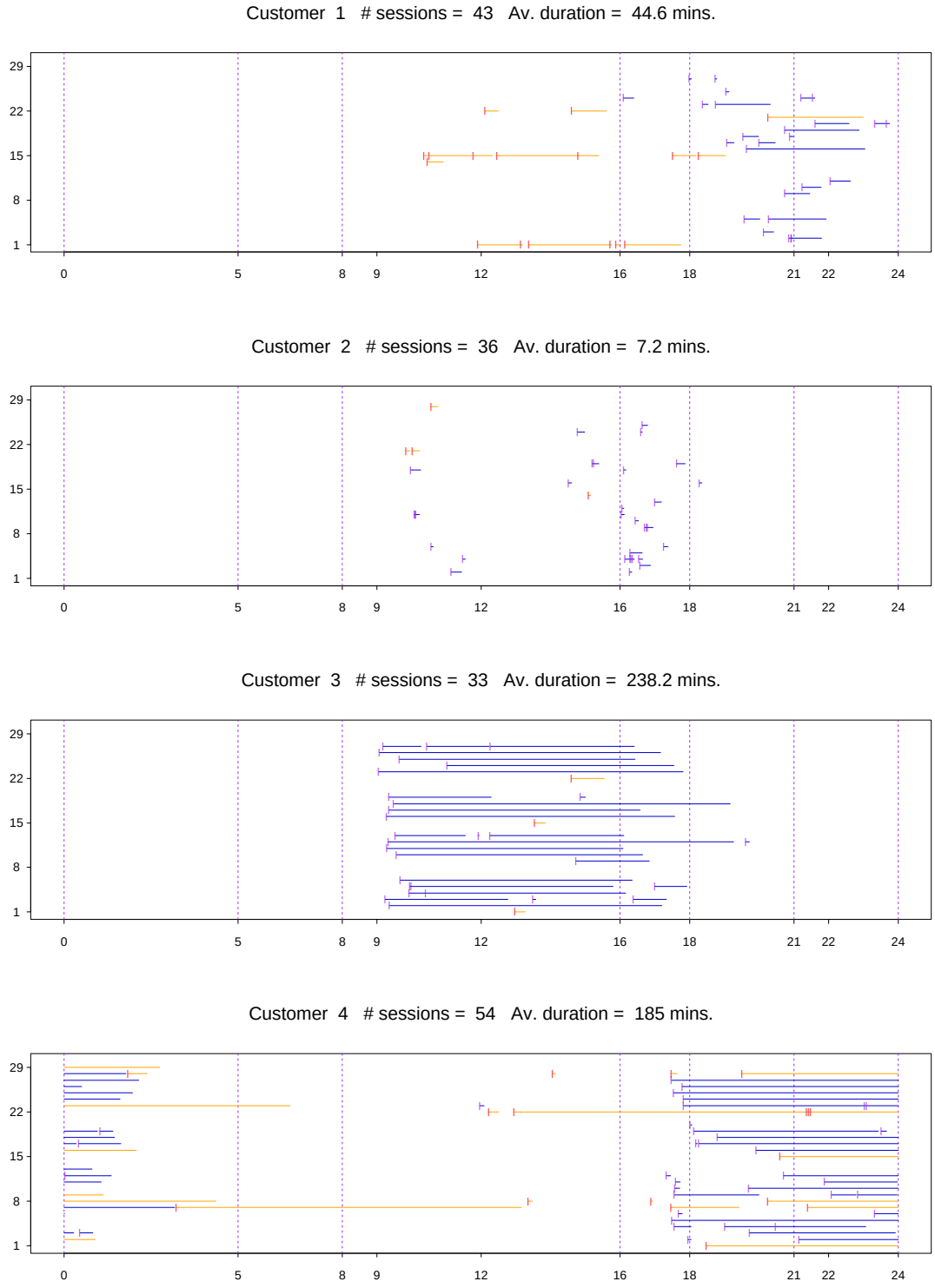

Figure 1: Illustration of customer usage patterns: 1

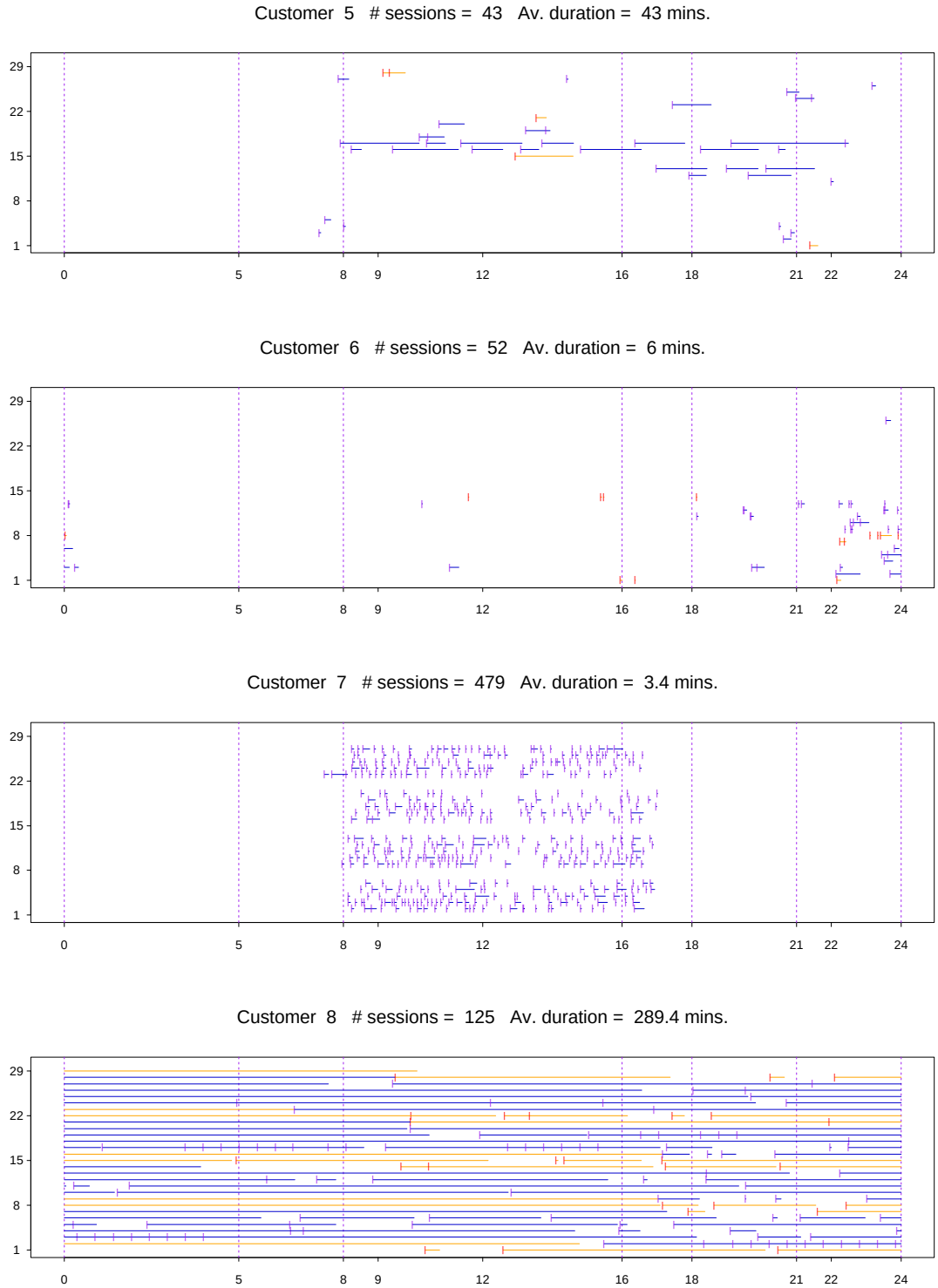

Figure 2: Illustration of customer usage patterns: 2

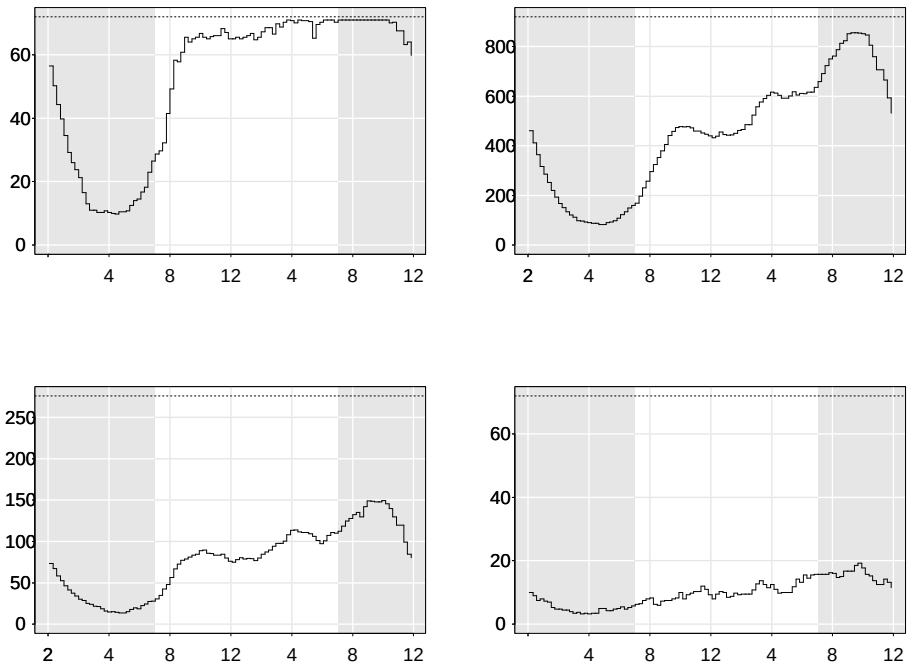

Figure 3: Traffic at 4 modem banks averaged over 12 weekdays

questions: Is recent usage of the service at all useful in predicting whether a customer is likely to stay? Does trouble accessing the service affect retention? Are customers more vulnerable when the billing cycle comes around? Do different predictors come into play at different stages of a person's tenure or for different subscription plans? Does the model change with calendar time? Can we construct a profile of customers at greatest risk of defecting?

4.1 The Data Streams

Data are received and recorded on a continuing basis at various points in the network, then aggregated and sent out daily to several sites including our source.

The Registration database: maintains account information including the subscription plan, service creation and termination dates, reason for dropping service etc. There are about 43 fields. The database is updated daily.

The Session database: maintains session level statistics such as time of call, duration of session, total bytes transacted, origin (modem bank) of call and a variety of statistics related to network traffic. There are approximately 39 fields. Feeds are appended daily to this database.

The Email database: maintains email statistics including date and time a person logs in to the email servers, number and size of the messages received and sent etc. There are approximately 10 fields. Separate databases are maintained for incoming and outgoing email. Feeds are appended daily to these databases.

The registration and session tables are linked through a common account identifier. The email and registration tables are linked through the email address. The email table and the session table are linked only indirectly through the registration table. Since an account is allowed several email addresses, only the primary email address appears in the registration table. A separate registration table of account information is maintained for subsidiary email addresses.

4.2 Feature Selection

A large part of the effort in modeling is directed towards identifying the appropriate variables, and subsequent processing and experimentation to determine the most appropriate form for use in modeling.

Features based on Usage Profiles: The data from session and email logs as well as records for customers who had no activity were extracted, then aggregated to create weekly records for each customer. Anticipating later modeling activities, weeks are defined relative to the date that the customer joined WorldNet, so that they do not necessarily correspond to calendar weeks. The weekly records were then processed to create the following variables: whether a person defected the following week (`status`), indicator of session activity in week (`Ind(session)`), number of sessions (`session.n`), duration of sessions (`duration`), dominant modem bank (`modem`), bytes in and out (`bytes.in`, `bytes.out`), indicator of email activity in week (`Ind(email)`), number of email sessions (`email.n`), number of email messages sent (`emessages.n`).

Re-parameterization can be used to derive meaningful com-

posite features from individual features such as average session length (`session.length`), defined by `duration/session.n` and session intensity (`session.intensity`), defined by `(bytes.in + bytes.out) /duration`.

As Figures 1 and 2 suggest, statistics describing ISP usage tend to be extremely long-tailed (skewed). Skewness can cause pathologies in the statistical analysis. Transformations can reduce problems of skewness. For example, Figure 4 shows histograms of `session.n` and `session.length` before and after transforming. These variables had to be truncated and then "logged" since the range was extremely large.

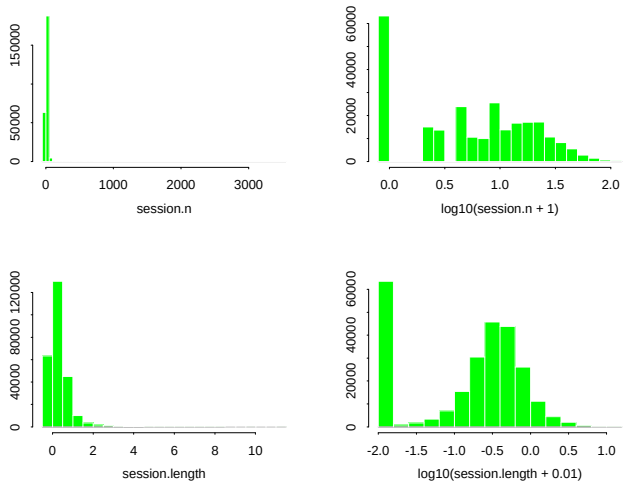

Figure 4: Usage variables before and after transformations

Features based on Equipment Profiles: To determine the congestion a specific user encountered during a given week required mapping `modem` for the customer for a given week to the master table containing traffic congestion for all the modem banks, which was described in Section 3.2. This yields a feature (`congestion`) which is used in the model for predicting defection.

Features based on Tenure: Figure 5, which shows customer defection by tenure, indicates that customers are far more vulnerable to defection when they are first trying out the service. Thus tenure (`tenure`) is one of the features used in the model. The plot also reveals that defection rates spike every 4-5 weeks. These spikes seem to be related to an end-of-month phenomenon coinciding with triggering of a new month's charges. Exact billing dates are hard to come by since a large proportion of customers have their fees automatically charged to their credit cards. However this feature can be captured vicariously by an end-of-month indicator (`Ind(end-of-month)`).

Subscription Plans: The subscription plans (`plan`) range from $10 for 10 free hours ("hourly" plan) to $20 for 150 free hours ("standard" plan) to $22 for unlimited usage ("unlimited" plan). Recently, WorldNet has also started offering a "bundled" plan where AT&T's long-distance customers get a $5 monthly discount off the standard plan when they sign up for WorldNet. We include `plan` in the prediction model using a set of indicator variables, which describe the contrast between the "standard" plan and the rest.

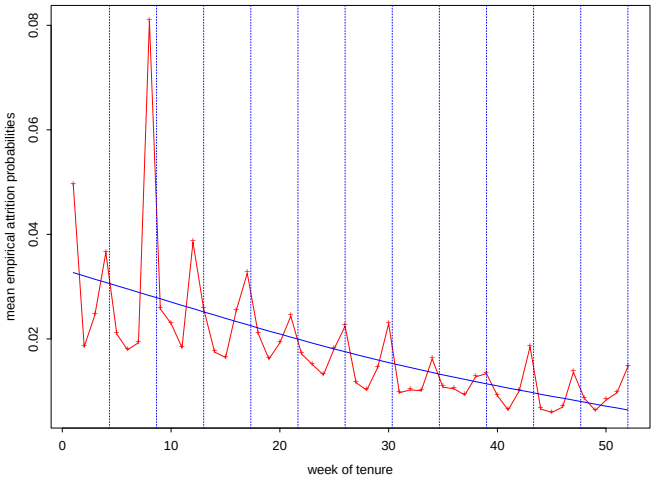

Figure 5: Empirical defection rates by tenure.

In summary, activity patterns such as session and email usage serve as proxies for perceived value and modem congestion for quality of service. In addition we incorporate information about subscription plans, tenure and billing cycle. Traditional studies use demographic predictors but our preliminary studies revealed that demographics were useful only in the absence of specific usage information about a customer. We find them to be poor proxies for individual features. Features we would like to incorporate but do not have access to, are customer care information (which could add further dimensions to perceived quality of service) and external factors such as competition.

4.3 A Model for Vulnerability

The process of deciding on a model is a well-recognized challenge in data mining. We studied several models, but for the discussion here we will focus on the following one to illustrate the main results of our analysis.

We consider a cross-section of WorldNet during the first week of November, 1999. The data consists of weekly records of customers with tenure between 1 and 52 weeks during this period. We used a one-third subset of this data for the modeling and made predictions on an independent one-third sample. Our preliminary studies used classification trees to study the effects of the selected features on the probability of defection. Here we report results from a logistic regression analysis of the data. The model we discuss can be

represented by the pseudo-code:

$$\begin{aligned} \text{logit}(\texttt{status}) \;\sim\; & \texttt{tenure} + \texttt{Ind(end-of-month)} + \texttt{plan} \\ & + \texttt{Ind(session)} + \texttt{log10(session.n+1)} \\ & + \texttt{log10(session.length +0.01)} \\ & + \texttt{Ind(email)} + \texttt{log10(email.n+1)} \\ & + \texttt{log10(emessages.n+1)} \\ & + \texttt{congestion}. \end{aligned} \quad (1)$$

Table 1 shows the results of the analysis. According to this, the variables most useful for predicting retention are (i) `tenure`, (ii) `Ind(end-of-month)`, (iii) `plan`, (iv) `Ind(session)`, (v) `session.length` and (vi)`email.n`. Predicted probabilities of defection for the test dataset are shown in Figure 6. The defection rate for the entire sample was 0.018, and is indicated by the solid vertical line in the plot. Note that there is a three-fold or greater increase in the odds of defection between an average customer and a high-risk one.

Feature	Coeff.	Std.Error	t-value
`Intercept`	-4.203	0.145	-28.888
`tenure`	-0.027	0.001	-22.477
`Ind(end-of-month)`	0.587	0.031	18.693
`Plan: 10hrs`	-0.025	0.026	-0.977
`Plan: Std2`	-0.028	0.018	-1.55
`Plan: Unlmtd.`	0.078	0.008	9.384
`Plan: Bundled`	-0.141	0.023	-6.166
`Ind(session)`	-0.587	0.097	-6.078
`log10(session.n + 1)`	0.087	0.064	1.359
`log10(sess.length+0.01)`	-0.397	0.061	-6.462
`Ind(email.active)`	-0.116	0.118	-0.981
`log10(email.n + 1)`	-1.462	0.361	-4.047
`log10(emessages.n + 1)`	0.272	0.157	1.729
`congestion`	-0.002	0.002	-1.110

Table 1: Results of Logistic Regression Analysis for Predicting Defection

The plots in Figure 7 illustrate some interesting aspects of the subtle differences in the way session and email usage affect retention. The drop in the curve between 0 and 1 session and the relative flatness subsequently suggest that the `Ind(session)` picks up most of the signal in that predictor. In contrast, it does seem to matter how many email sessions a person has. This was, of course, also reflected in the corresponding coefficients in Table 1.

The plots in Figure 8 give some assessment of the goodness-of-fit of the model to the data. In Figure 8(a) the x-axis refers to the ten risk groups that customer were binned into. Customers were ranked and assigned to the groups according to their model-predicted probabilities of defection (lowest decile to highest decile). The steep increase in attrition probabilities for the higher decile groups indicates that the model picks up on defectors in the sample. The plot in Figure 8(b) elaborates on this. The plot shows the decomposition of defectors into the ten risk groups. In particular, we see that 50 percent of the people who defected were in the two highest-risk deciles. Both the plots show good agreement between the empirical and predicted probabilities in

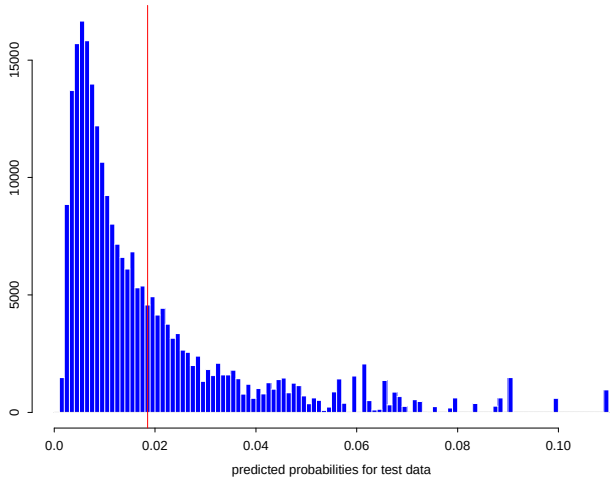

Figure 6: Model predicted probabilities of customer defection for a validation sample.

the different risk groups. This indicates that customers determined by the model to be most at risk were indeed the ones who actually defected.

4.4 Probing the Model

One aspect of model checking is to do diagnostics to determine whether we have missed important predictors or interactions between predictors. We explored interactions of usage features with plan and with tenure. Such an analysis is straightforward for plans, but much harder for tenure.

4.4.1 Interactions of Usage Variables with Plan

We repeated the logistic regression analysis in equation (1) for each subscription plan separately and compared corresponding coefficients across plans. We found that there were not significant differences among the coefficients, suggesting that the model did not vary by plan.

4.4.2 Interactions of Usage Variables with Tenure

In order to understand the effect of tenure on the various predictors, we studied a much larger sample. The dataset consisted of activity logs for all customers during some point in the three month period covering September, 1999 through November, 1999. We extracted weekly records for all eligible customers of WorldNet during this period and created 52 weekly datasets. Thus for instance, Week 1 contained the weekly record of all customers who were in their first week with WorldNet during this three-month period. A single customer might have appeared in as many as 13 of the datasets. We restricted our analysis to 52 weeks because behavior is well-stabilized by this time.

The plots in Figure 9 illustrate how the effects of some of these predictors varied with tenure. Figure 9(a), for instance, shows that the coefficient for the "bundled" plan has a positive effect on retention but that this effect is pretty constant over time, except for the end-of-month fluctuations. Similarly, Figure 9(c) shows a consistently negative relationship between `email.n` and defection, and no trend

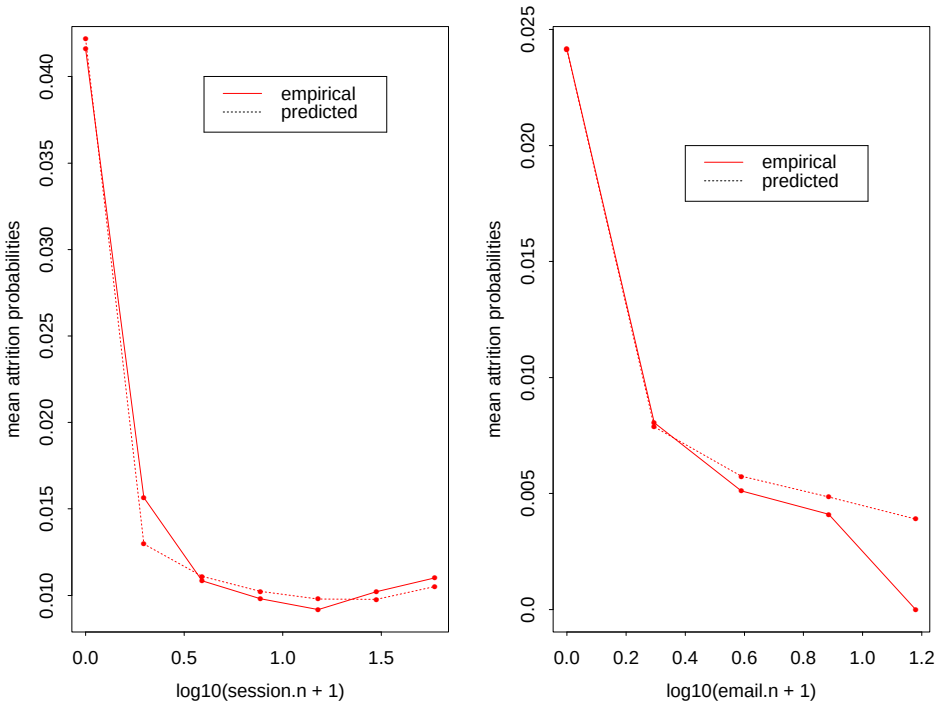

Figure 7: Probabilities of defection vs. predictors.

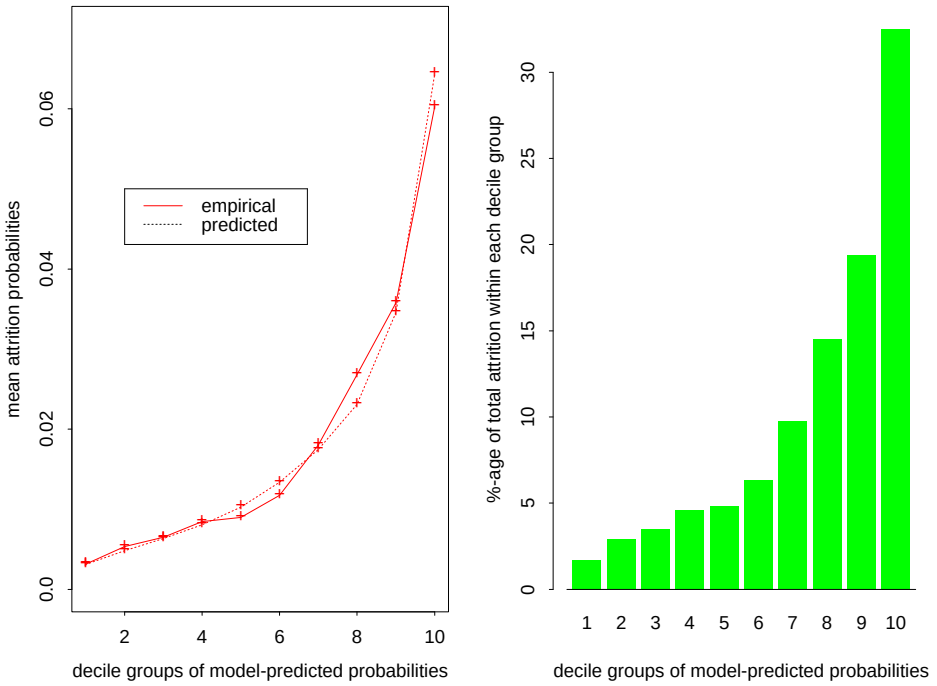

Figure 8: Assessment of goodness-of-fit of the model to validation sample.

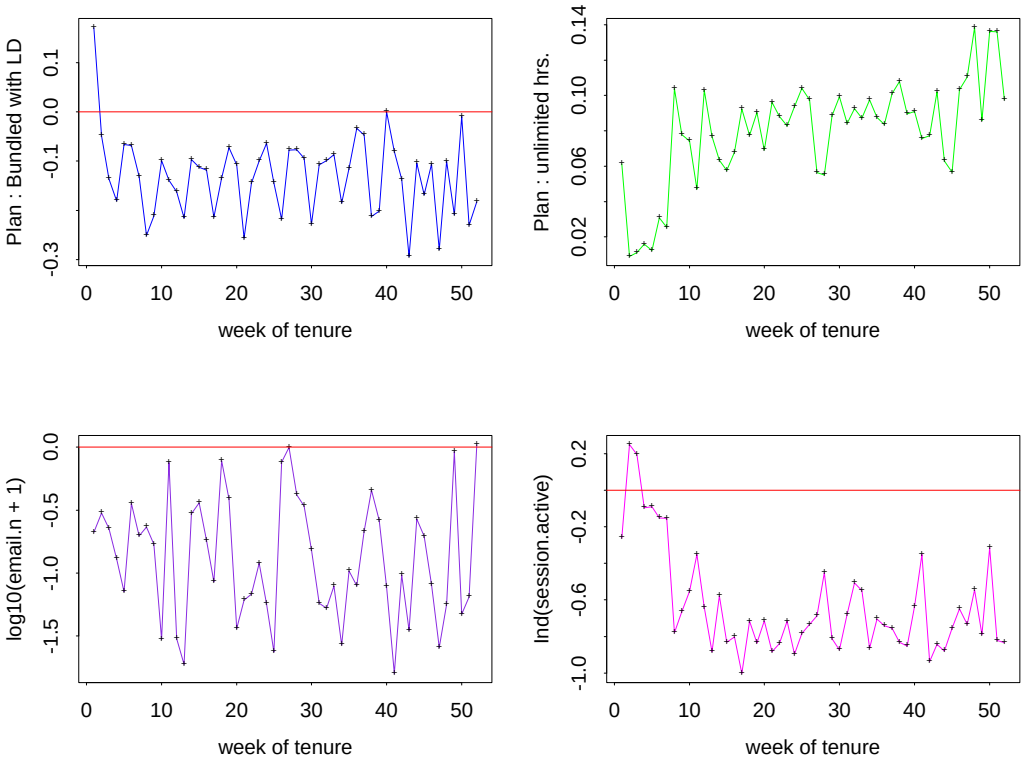

Figure 9: Logistic regression coefficients as a function of tenure.

with tenure. In contrast, Figure 9(b) shows that coefficients for the "unlimited" plan tend to *increase* with tenure. Similarly, the negative relationship with `Ind(session)` shows up after about 8 weeks (Figure 9(d)).

It is important to note that such non-linear trends, as exist in Figures 9(b) and (d), cannot be readily modeled as a logistic regression model with interaction terms.

4.5 Profiling the High-risk Group

In practice we would not only like to be able to rank customers on their vulnerability, but also identify attributes that result in increased vulnerability. The plots in Figure 10 indicate what these might be. These plots show histograms for four features used in the logistic regression model. The "red" portion of each bar highlights the topmost decile of model predicted risk groups. Figure 10(a) shows that users of the "unlimited" plan had an inordinately large proportion of people in this group. Figure 10(b) shows that people early in their tenure with WorldNet are most at risk, and in fact, if a person has stayed on for 3-4 months, they are quite unlikely to fall in the high-risk group. Figure 10(c) starkly suggests that most users in the highest-risk group, do not, in fact, use the service at all. Figure 10(d) echoes the same message about email usage.

In a nutshell, according to the model, people at highest risk to leave are those who have recently signed up for the service, opted for the unlimited plan, haven't got started on email and in fact, aren't using the service at all. This suggests clear strategies for intervention, such as getting people hooked onto email, anticipating usage requirements and proactively advising users to change to a subscription plan more in line with their usage.

5. CHALLENGES

Our work to create statistical profiles of customer activity identified several features that distinguish not only the frequency of use but also how and when customers use WorldNet. The work also highlights the need for dynamic profiles that are sensitive to changes in user behavior. We anticipate that the profiles will have many applications, from describing how the service is used now, to how it might evolve for groups of customers with specific needs.

In particular, logistic regression, using tenure and other features of the profiles to predict short-term defection, succeeds in identifying a 20 percent subgroup responsible for 50 percent of the customer loss in the next week. Diagnostics play an important part in validating the model structure, the coefficients, and the resulting predictions. For example, although analysis generally confirmed that the predictive power of most usage variables were stable with respect to tenure, a few relationships seemed to change after the first few weeks.

Data visualization through graphics proved very valuable in all stages of the analysis: the development of the profiles,

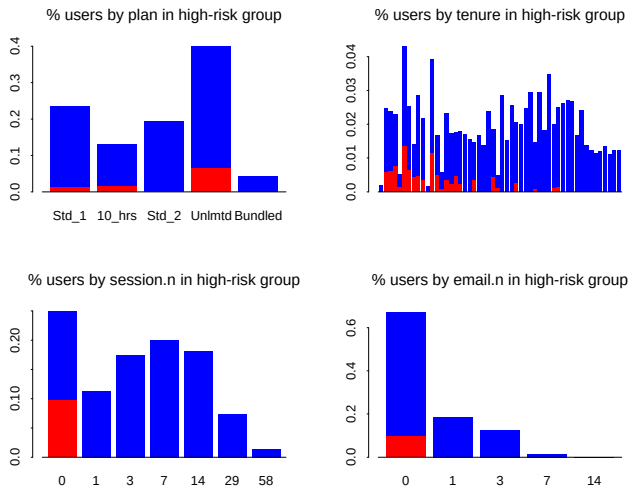

Figure 10: Distribution of the high-risk group in the sample.

feature selection for predicting defection, and model diagnostics.

We overcame many challenges in these analyses, most stemming from the large volumes of data that record information at different granularities. Databases, such as activity logs for sessions or other transactions were generally not designed with statistical analysis in mind. Consequently, creation of an analysis data set required complex linkages (i.e., joins) across multiple databases that often differed in terms of the basic unit and time frame.

Our analysis required creating weekly records for each of about 2 million users, where "weekly" refers to each week of an individual users' tenure with WorldNet. A person starting on Oct. 15, 1999 would be in week 3 for an analysis on Nov. 1, 1999; whereas a person starting on Oct. 1, 1999 would fall in week 5 for the same analysis. Thus, there are two dimensions of time under consideration, "calendar time" and "tenure".

Algorithms for aggregating individual session records to create weekly statistics for each customer required careful consideration of computational efficiencies of different computing environments and data storage capabilities. In practice, we found that many of the required computations were best performed outside the database system.

Data quality is a critical concern for data of this volume and complexity. Data stream in from a variety of different sources. Despite being automated, data can exhibit an assortment of inconsistencies and errors, causing problems when the data streams are mapped and aggregated into customer records. Examples of this are feeds missing partially or altogether for certain periods, "ghost" records from people who have already defected, parallel streams of data having inconsistent records, fields not being consistent with field specifications, etc. To automate the data processing, diagnostics that anticipate and identify such problems need to be in place.

As with all data relating to telecommunications traffic, considerable care needs to be taken when mining online data to ensure that customers' privacy is not violated. AT&T WorldNet has a comprehensive data-use policy to protect privacy. In a nutshell, the data are not sold to third parties. Internally, the data are used for operations, customer care, and marketing. The customer has the right to opt out of marketing uses. The policy can be found at `http://www.att.net/privacy`.

The ultimate goal of the analyses described in this paper is to develop real-time systems for creating statistical profiles of user behavior and applications of the profiles. To be useful to the business, the profiles need to become available while they are still fresh. This need for dynamic analysis creates new challenges to successful implementation. The systems should be able to update existing profiles as new data arrive, rather than recreate profiles from scratch. Even automated data do not always arrive on time (some of Tuesday's data me be available on Wednesday, but some may not show up until the following Monday), therefore the systems need to be flexible enough to handle various contingencies. Finally, data quality checks become critical, because there is not time to carefully analyze a static dataset for inconsistencies. Real-time diagnostics need to alert analysts to potential data errors, some of which may have never occurred previously.

We have taken a large step towards combining statistical challenges with data management challenges. When dealing with massive datasets and messy data often some of the accomplishments appear easier than they actually are. Nonetheless graphical solutions and simple statistical models have already made the endeavor worthwhile. They have led to a deeper understanding of the breadth of customer behavior and quantified key variables - and in the process led to a considerable improvement of our data feeds.

6. ACKNOWLEDGEMENTS

Nandini Raghavan's research was supported in part by NSF grant DMS-9700867 to the National Institute of Statistical Sciences Matthias Schonlau's research was supported in part by NSF grants DMS-9700867 and DMS-9208758 to the National Institute of Statistical Sciences. Alan Karr's research was supported in part by NSF grant DMS-00867 to the National Institute of Statistical Sciences.

7. ADDITIONAL AUTHORS

Daryl Pregibon (AT&T Labs Research, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932, email: `daryl@research.att.org`) and

Alan F. Karr (National Institute of Statistical Sciences,19 T. W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006 email: `karr@rand.org`).