Churn Prediction in New Users of Yahoo! Answers

Gideon Dror, Dan Pelleg, Oleg Rokhlenko, Idan Szpektor
Yahoo! Research, Haifa, Israel
{gideondr, dpelleg, olegro, idan}@yahoo-inc.com

ABSTRACT

One of the important targets of community-based question answering (CQA) services, such as Yahoo! Answers, Quora and Baidu Zhidao, is to maintain and even increase the number of active answerers, that is the users who provide answers to open questions. The reasoning is that they are the engine behind satisfied askers, which is the overall goal behind CQA. Yet, this task is not an easy one. Indeed, our empirical observation shows that many users provide just one or two answers and then leave.

In this work we try to detect answerers that are about to quit, a task known as churn prediction, but unlike prior work, we focus on new users. To address the task of churn prediction in new users, we extract a variety of features to model the behavior of Yahoo! Answers users over the first week of their activity, including personal information, rate of activity, and social interaction with other users. Several classifiers trained on the data show that there is a statistically significant signal for discriminating between users who are likely to churn and those who are not. A detailed feature analysis shows that the two most important signals are the total number of answers given by the user, closely related to the motivation of the user, and attributes related to the amount of recognition given to the user, measured in counts of best answers, thumbs up and positive responses by the asker.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]:

General Terms

Algorithms, Experimentation

Keywords

churn prediction, community question answering, online user behavior

1. INTRODUCTION

While search engines achieved remarkable results in the last two decades, there are still types of information needs that are difficult to answer through traditional search. These include advice requests, opinion seeking and social requests. Some examples of such question are:

- "*my tamale dough is kinda sticky and soft and won't float, what should I do?*"
- "*if legalizing gay marriage does lead to a slippery slope, how will that affect you?*"
- "*am I pretty (here's a link to my photos . . .)?*".

Community-based question answering (CQA) services were introduced exactly for meeting these very personal needs. Services like Yahoo! Answers, Quora, Baidu Zhidao and Naver Ji-Sik-In allow any user to ask any question and any user to answer any of the open questions in the system. One of the important goals of CQA services is to maintain and even increase the size of their active answerer community. After all, these are the users who provide the answers to arriving questions and thus make the whole participation in the service, that is asking a question from the asker's side, worthwhile. In this work we focus on retaining already active users. Specifically, we would like to identify users who are about to quit, a task known as churn prediction [12, 15, 5, 10, 13, 18, 9]. Success in this task will open new possibilities for the site, for example focusing efforts on these users to continue using the service.

The main reason for retaining users is that the cost of acquiring new customers is higher than the cost of keeping existing customers [15]. In CQA services, it is hard to convince users to start answering questions, or even visit the web-service for the first time. Thus, it is important to keep users who already actively answer questions in the system. Yet, there are differences between classic industries with known customer attrition problems, such as telecommunications and banking, and CQA services. One such difference is that once a user subscribed to a mobile service or a credit card company, she will remain a subscriber for a while (several weeks or more), due to the amount of energy put into the subscription process itself. In CQA services however, it is very easy to subscribe and start answering questions. Indeed, the user answering activity graph (Figure 1) obeys the power law distribution, which means that many answerers contribute one or two answers and leave the system never to return. Following this observation, in this work we target the prediction of churn in new users, since converting even a small fraction of soon-to-be churners into active answerers will substantially increase the size of the answerer community.

In this paper we focus our analysis on Yahoo! Answers, which is the one of the largest CQA sites, with hundreds of

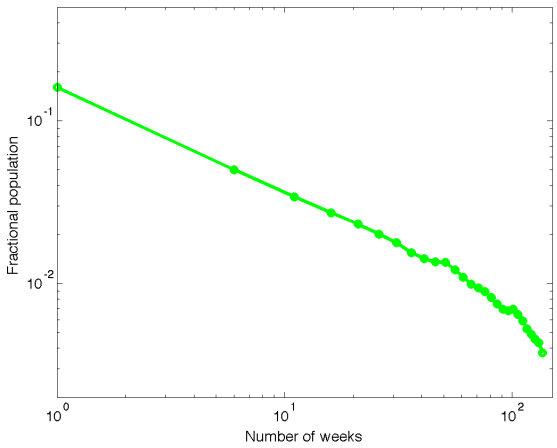

Figure 1: Histogram of the length of participation period per user measured in week (log-log scale).

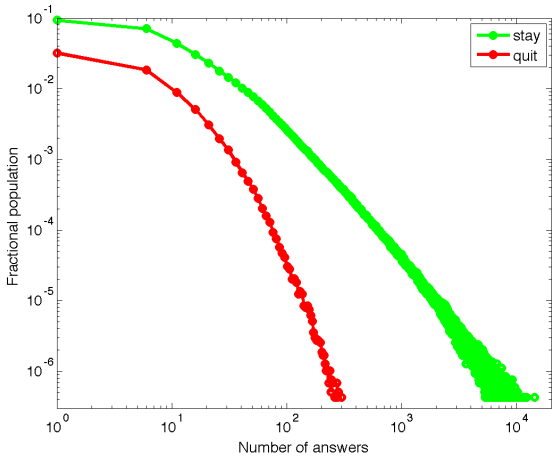

Figure 2: Histogram of the number of answers per user (log-log scale), for the users who left on the first week ("quit"), vs. the users who continued past the first week ("stay").

millions of questions and over a billion answers over several years of site activity. To address the task of churn prediction in new users, we collect a variety of features that model the behavior of users over the first week of their "life" as answerers. These features include personal information, such as age and gender, activity modeling, such as answering time and answering rate, and social relationships with other users, as well as asker-answer communication. We show that each of these aspects contribute to better churn prediction in Yahoo! Answers. To the best of our knowledge this is the first work to study churn prediction in CQA sites, as well as the first work to study churn prediction in new users.

2. BACKGROUND

2.1 Yahoo! Answers

Yahoo! Answers is a question-centric CQA site. Askers post new questions and assign them to categories selected from a predefined taxonomy, such as *Sports* > *Golf*. A question consists of a *title*, a short summary of the question, and a *body*, containing a detailed description of the question. The posted question remains "open" for four days and can be answered by any signed-in user. The asker may choose a best answer while the question is open, and even optionally provide feedback for the best answer in the form of a one to five rating and a textual message. If the asker did not make a choice for best answer, the community votes for it. Once a best answer is chosen, the question considered "resolved".

In case a question is not answered while "open" it is "deleted" from the site. A question or answer may also be deleted if it is detected as spam or if it was found offensive or otherwise not following the site's rules. On top of this, a question may be deleted by the asker herself, for example if a very personal question was asked and the asker is not willing for the question to be publicly available once a suitable answer, which satisfied his/her need, arrived.

2.2 Related Work

The prediction of churn has been extensively studied for over over a decade. Most research on user retention revolved around the wireless telecommunication industry [6, 12, 15, 7, 4, 5, 14, 16, 18]. Yet, other domains were studied as well, including banking [11, 13], grocery retail [2], pay TV [3], online gaming [10], P2P networks [17] and social networks [8, 9].

The main approach for churn prediction is to construct a set of features for each user and train a classifier or regressor for the task [6, 12, 15, 7, 4, 14, 18, 11, 13, 2, 3, 10, 17, 8, 9]. Some features are related to the service and are independent of the user, such as call quality, billing, customer service and pricing [12, 15], and may be utilized to model a prior on churn likelihood in the service at a given time. To distinguish between users, user-dependent features are introduced. These may include information about the users independent of the service, such as age, gender and salary [4, 13]. Other features measure user activity within the specific service, such as minutes-of-use, frequency-of-use and past renewals of service in telecommunications [4, 15], the number of transactions in banking [13] and session length and inter-arrival time in P2P networks [17]. These features are time-dependent and are captured per specific time-frames. A different type of user behavior was studied in [7], analyzing the differences in complaint and repair calls between churners and non-churners. Finally, induced social features were investigated by [9], including in and out degrees of the user node in the social network based on replies in a discussion board, popularity, closeness and betweenness centralities etc.

A complementary approach to modeling independent user behavior is to model the effect of social ties between users. The main hypothesis in this approach is that users that leave a service may influence other users, with whom they have social relationships, to leave as well. [5] model such influence as diffusion processes of churn in the social graphs, showing improvement in churn prediction. [10] combine individual player engagement with diffusion-based influence, where both positive (retention) and negative (churn) influence are taken into account. [16] propose that users leave

in groups, and detect dense social group in which leaders of the group may cause the whole group to leave the provider.

Past work on churn prediction focused on active users, with at least several weeks of documented activity. However, in community-based question answering, many users are early churns, that is, they provide one or two answers and leave without returning to the CQA site, as shown in Figure 1 for Yahoo! Answers. Some of these users are just anecdotal visitors, that do not intend to continue using the service. Still, other users, who planned on using it, decide to give up on the service very early for some reason. If even a small fraction of these early quitters could be convinced to continue and use the service, the number of active users will increase substantially. Hence, unlike prior work, in this paper we focus on predicting churn in new users, specifically within their first week of activity.

3. EXPERIMENTAL SETUP

To collect the data, we studied the answer-posting activity of users on Yahoo! Answers between April and December of 2010. For each user, we aggregate their activity by weeks. We then discard her activity from the second and subsequent weeks, except for a single bit which signifies their existence. In other words, for each user we have the tally of activity over the first week, and an additional label for them not being active beyond that week (churners), or else being active for some additional time (non-churners).

Four types of features were extracted from the activity of the user during this seven-day period: features describing the questions the user answered to, e.g. categories assigned to these questions; features associated to the answers posted by the user, e.g. mean length of the answers measured in characters and in words; features related to the feedback to the work of the user, e.g. the number of thumbs up or thumbs down to her answers; and demographic features of the user such as gender or zip code. The demographic information was taken as is, unverified, from the details supplied by the user. Table 1 details the 64 features extracted together with their types. In total our dataset comprised of $20,000$ examples, consisting of $10,944$ churners and $9,056$ non-churners.

Table 1: Features by type

Question features (33 total)

- *cat_xxx*: of questions answered in each of the 27 top-level YA categories. e.g. "cat_396546089".
- *qlenw*: the average length of the questions in words
- *qlenc*: the average length of the questions in characters
- *deleted_count*: the number of questions the user answered that were deleted
- *deleted_avg*: the fraction of questions the user answered that were deleted
- *q_num_answers_avg*: the average of number of answers, taken over the questions this user answered.
- *nstars_avg*: the average number of stars for the questions answered.

Answer features (10 total)

- *nanswers*: the number of answers submitted
- *anslenc*: the average answer's length in characters
- *anslenw*: the average answer's length in words
- *ansnumurls*: the average number of URLs per answer
- *frac_active*: the fraction of time intervals in which user questions were posted
- *yday*: day in year (1-365)
- *month*: month (1-12)
- *mday*: day of month (1-31)
- *wday*: day of week (1-7)
- *hour*: hour of day (1-24)

Gratification related features (18 total)

- *nbest*: the total number of best answers of the user posted within the first week
- *nbest_voting*: same as *nbest* but only best answers selected by the community
- *nbest_asker*: same as *nbest* but only best answers selected by the asker
- *nbest_awarded*: the total number of best answers awarded to the user within the first week.
- *nbest_voting_awarded*: same as *nbest_awarded* but only best answers selected by the community
- *nbest_asker_awarded*: same as *nbest_awarded* but only best answers selected by the asker
- *nans_before_best*: the number of answers submitted before any best-answer events happening
- *nans_after_best*: the number of answers submitted after the first best-answer
- *thumbs_up*: the number of thumbs up on answers submitted in interval
- *thumbs_down*: the number of thumbs down on answers submitted in interval
- *thumbs_up_down*: the ratio of (1+thumbs_up) to (1+thumbs_down)
- *rating_avg*: the average of rating by asker (only if best answer was chosen by asker)
- *resplenc*: the average asker's response length in characters
- *resplenw*: the average asker's response length in words
- *gratitude_count*: the number of responses with at least one "thanks" statement[1]
- *gratitude_avg*: gratitude_count divided by nbest_asker
- *first_best_asker*: binary indicator for the first answer of the user winning a best answer by asker
- *first_best_voting*: binary indicator for the first answer of the user winning a best answer by voting

Answerer demographic features (3 total)

- *gender*: The gender of the user (as given by the user)
- *age*: The age of the user
- *zip*: The zip-code of the user

4. RESULTS

Using the dataset we tested prediction performance using several classifiers: Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest and K-nearest Neighbors (KNN). In all cases we used conservative hyperparameter setting, with no attempts to tune them: Logistic regression and SVM were trained using stochastic gradient descent, and for the decision tree we used the J48 decision tree learning algorithm pruned such that each leaf contains at least 100 instances. Random Forest was trained with 50 trees and 8 features per split. For KNN we used $K = 10$. Since the different features have very different distributions and scales, we transformed all features to a zero mean and standard deviation one, for the metric based methods (Logistic Regression, SVM and KNN).

Table 2 details the Error rate, the Area Under ROC (AUC) and the F_1 performance measure of predicting churning. The table shows 10-fold cross validation performance. We also give the performance measures for a majority classifier, which is a coarse baseline classifier that assigns every test example to the majority class. Although Random Forest gives consistently superior results it is only slightly better than the Logistic Regression classifier, which is much simpler faster to train and easier to interpret.

Figure 3 depicts the Receiver Operator Characteristic (ROC) of the output of the Random Forest classifier. It shows the True Positive rate as a function of the False Positive rate. It is clear that the ROC curve is significantly higher than that random assignment line, represented by the straight

Classifier	Error	AUC	F_1
Majority	0.453	0.5	0.623
Naive Bayes	0.376	0.705	0.730
Logistic Regression	0.309	0.754	0.737
SVM	0.327	0.65	0.75
Decision Tree	0.319	0.727	0.722
Random Forest	**0.306**	**0.758**	**0.755**
KNN	0.346	0.705	0.685

Table 2: Classifiers' performance

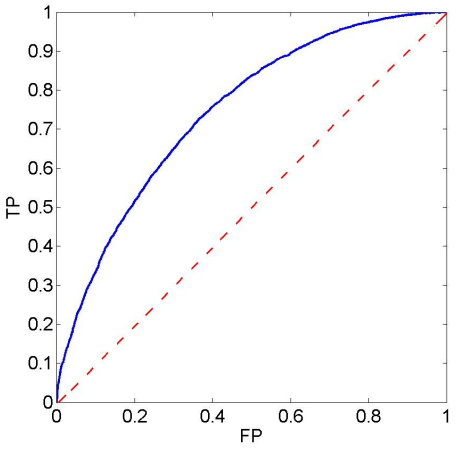

Figure 3: Receiver Operator Curve (ROC) of the Random Forest outputs.

dashed line. Indeed, using the Mann-Whitney statistic [1] the standard deviation of the estimated Area Under ROC curve (AUC) is not greater than $\sigma_{max} = 0.00225$, hence is significantly higher than 0.5.

To get an insight as to which features are more informative for churn prediction, we calculated for each feature the Information Gain for predicting the target variable, churning. Table 2 lists the top contributing features sorted by the value of their Information Gain. We also listed the sign of the correlation between the feature and the target to give a notion of how it affects churning: a positive sign means that larger values of the feature are associated with higher tendency to churn, and vice versa. The table uncovers several trends in the data, some of which obvious while others less so: First, the most informative feature is the total number of answers the user posted. Namely, users who post more answers are much less likely to churn, as suggested by Figure 2. The effect of the feature frac_active, follows, basically, the same reasoning. Second, several temporal features (month, day of year and day in month) all with positive correlations with churning. This means, that in the period of time the data describes, churning becomes more prevalent with time. This pattern was independently verified. Third, features which are related to gratitude (number of best answers, number of best answers by asker, number of best answers by voting, average rating, number of thumbs up) all negatively correlated with churning. Namely the more gratitude a user receives, the less likely she is to churn. We notice that the two features that describe the number of answers before first best answer, and after last best answer are also negatively

Feature	Info-gain	Tendency
nanswers	0.0765	−
frac_active	0.0749	−
yday	0.0734	+
nans_before_best	0.0668	−
month	0.0631	+
mday	0.0549	+
nbest	0.0498	−
nstars_avg	0.0420	+
q_num_answers_avg	0.0382	+
qlenw	0.0353	+
thumbsup	0.0345	−
nans_after_best	0.0335	−
nbest_voting	0.0334	−
nbest_asker	0.0329	−
rating_avg	0.0288	−

Table 3: The 15 most informative features ranked by their information gain with the target. The *Tendency* of a feature is the sign of the correlation with the target.

correlated with churning. Clearly, the number of thumbs up, thumbs down, number of best answers are all positively correlated with the total number of the user posted.

To isolate the effect of gratitude features from the effect of the total number of answers we split users into disjoint sets each characterized by the number of answers posted by the users. For each set we plot the churning rate as a function of the the thumbs-up and thumbs-down rates, namely the fraction of the users' answers that were thumbed up.

In Figure 4 we grouped the users by the number of answers they posted. For each group, we plot the fraction of churners as a function of their thumbs-up and thumbs-down rates. We plot not only an estimate for the churning rate, but also the confidence interval to this estimate calculated with confidence level $\alpha = 0.05$. The lines of Figure 4 (a) that correspond to users with one or two answers exhibit a small but significant decrease in churning rate with thumbs-up rate. An opposite trend is observed in Figure 4 (b) where more thumbs down lead to increased churning. For users with three, four, five and six answers the trend is not as visually clear. But fitting each line with a linear function using the least squares method results in positive slopes for all cases of Figure 4 (a) and negative slopes for Figure 4 (b). This confirms the intuitive hypothesis that a thumbs-up feedback reduces churning albeit in a minute way, wheras a thumbs-down feedback increases churning.

Last, and quite surprising, are question related features (average number of stars, average lengths, average number of answers) which are all positively correlated with churning. This suggests that users that are involved in more popular content are more likely to churn. A possible explanation would be that the longer and more interesting questions tend to attract many users. But of those intially lured to the question, the propensity to answer is higher among "newbies". Possibly, the veterans feel they have a lower chance of winning a best answer, or that their voice will be crowded out, and therefore refrain from answering.

5. CONCLUSIONS

Contributing users are the lifeblood of CQA sites. And

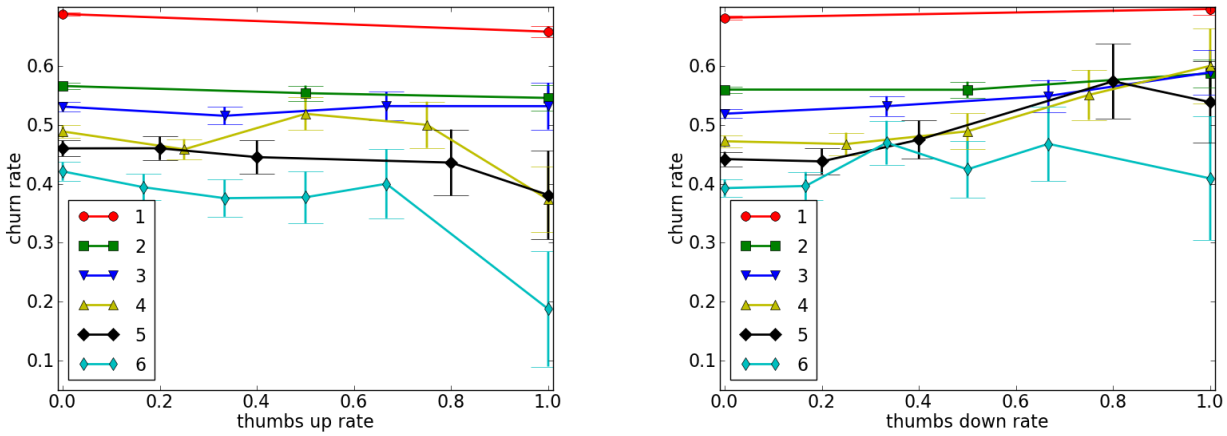

Figure 4: Churn rate as a function of thumbs-up (a) and thumbs-down (b) rates. The legend indicates the sub-group of users, chosen by number of answers posted.

among the contributors, the ones who define the spirit and etiquette of the site are the continuous contributors. These are the users who define the site's collective memory and pass it on. Therefore, a large number of "drive-by" answerers are not as good as a smaller number of users who keep coming back. In identifying users likely to churn early, this work makes a first step in turning this observation into potential operational changes.

Above, we show that the task of identifying potential one-time contributors is achievable to a reasonable degree. The obvious next step is propose changes that will make a site more "sticky". Some of these could be applied to the entire user base, like showing similar questions to the one just answered, to encourage a user to answer again. Others could be targeted just at likely churners, like sending some kind of "we want you back" message. The space of possible actions is large, and it's also true that some actions may cause a backlash and should not be attempted. We feel that the best course of action could probably be answered by usage data and methods similar to the ones in this work, however further development is outside the scope of the current paper.

We also note that analysis of this kind inevitably brings up the nature-versus-nurture debate. What this means here is that repeat contributors may be "born" or "made". If they are born, then what brings them back is some innate quality, which is missing in the early churners, and nothing could be done to change it. Conversely, if they are made, then perhaps there is a strategy to boost participation of the churners, and effort should be made to find it. We acknowledge the issue, however the materials and methods used in this work cannot support a decisive answer in this case. The truth is probably somewhere in the middle, but we leave the task of determining the degree of inherent answer-posting of each user to future work.

6. REFERENCES

[1] Z. Birnbaum and O. M. Klose. Bounds for the variance of the mann-whitney statistic. *Annals of Mathematical Statistics*, 28(4):933–945, 1957.

[2] W. Buckinx and D. Van Den Poel. Customer base analysis: Partial defection of behaviorally-loyal clients in a non-contractual fmcg retail setting. *European Journal Of Operational Research*, 164:252–268, 2003.

[3] J. Burez and D. Vandenpoel. Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 32:277–288, 2007.

[4] K. Coussement and D. Van Den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1):313–327, 2008.

[5] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *EDBT'08*, pages 668–677, 2008.

[6] P. Datta, B. Masand, D. R. Mani, and B. Li. Automated cellular modeling and prediction on a large scale. *Artif. Intell. Rev.*, 14:485–502, 2000.

[7] J. Hadden, A. Tiwari, R. Roy, and D. Ruta. Churn prediction using complaints data. In *Proceedings of world academy of science, engineering, and technology*, volume 13, pages 158–163, 2006.

[8] M. Karnstedt, T. Hennessy, J. Chan, P. Basuchowdhuri, C. Hayes, and T. Strufe. Churn in social networks. *Handbook of Social Network Technologies and Applications (Springer)*, 2010.

[9] M. Karnstedt, M. Rowe, J. Chan, H. Alani, and C. Hayes. The effect of user features on churn in social networks. In *Proceedings of the third ACM/ICA Web Science Conference*, 2011.

[10] J. Kawale, A. Pal, and J. Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *IEEE International Conference on Computational*

Science and Engineering, pages 423–428, 2009.

[11] D. A. Kumar and V. Ravi. Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1, 2008.

[12] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, and H. Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Trans Neural Netw*, 11(3):690–696, 2000.

[13] T. Mutanen, S. Nousiainen, and J. Ahola. Customer churn prediction –a case study in retail banking. In *Proceedings of the 2010 conference on Data Mining for Business Applications*, pages 77–83, 2010.

[14] P. C. Pendharkar. Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. *Expert Syst. Appl.*, 36, 2009.

[15] C. ping Wei and I. tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems With Applications*, 23:103–112, 2002.

[16] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SIAM International Conference on Data Mining*, pages 732–741, 2010.

[17] D. Stutzbach and R. Rejaie. Understanding churn in peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, IMC '06, pages 189–202, 2006.

[18] W. Verbeke, D. Martens, C. Mues, and B. Baesens. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems With Applications*, 38:2354–2364, 2011.