TopChurn: Maximum Entropy Churn Prediction Using Topic Models Over Heterogeneous Signals

Manirupa Das
The Ohio State University
das.65@osu.edu

Micha Elsner
The Ohio State University
elsner.14@osu.edu

Arnab Nandi
The Ohio State University
nandi.9@osu.edu

Rajiv Ramnath
The Ohio State University
ramnath.6@osu.edu

ABSTRACT

With the onset of social media and news aggregators on the Web, the newspaper industry is faced with a declining subscriber base. In order to retain customers both on-line and in print, it is therefore critical to predict and mitigate customer churn. Newspapers typically have heterogeneous sources of valuable data: circulation data, customer subscription information, news content, and search click log data. An ensemble of predictive models over multiple sources faces unique challenges – ascertaining short-term versus long-term effects of features on churn, and determining mutual information properties across multiple data sources. We present TopChurn, a novel system that uses topic models [5, 29, 24] as a means of extracting dominant features from user complaints and Web data for churn prediction. TopChurn uses a maximum entropy-based approach [21] to identify features that are most indicative of subscribers likely to drop subscription within a specified period of time. We conduct temporal analyses to determine long-term versus short-term effects of status changes on subscriber accounts, included in our temporal models of churn; and topic and sentiment analyses on news and clicklogs, included in our Web models of churn. We then validate our insights via experiments over real data from The Columbus Dispatch, a mainstream daily newspaper, and demonstrate that our churn models significantly outperform baselines for various prediction windows.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support, Logistics*; I.2.7 [**Natural Language Processing**]: Text analysis; H.3.5 [**Information Storage And Retrieval**]: Online Information Services

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743053 .

Keywords

churn prediction; maximum entropy; feature engineering; topic modeling; sentiment analysis; predictive modeling; data science; Web Science; exploratory data analysis

1. INTRODUCTION

Customer churn, also known as the attrition rate, is a well-studied problem, especially in the telecommunications industry [3, 6, 26, 1, 2], retail CRM [10], the social Web [19, 13, 9] and to a great extent in online and gaming communities [20, 30, 11]. It is a well-known fact that with the advent of social media and news aggregator websites, newspapers large and small are facing a challenge of ever declining print readership, resulting in customer "churn"[25]. Since print has a large margin of profit, reduction in print subscriptions results in drastically reduced sales and advertising revenues. News organizations are therefore looking at ways to rejuvenate print circulation by bundling various digital services along with print subscriptions [25].

In this work, our objective is to explore various structured and unstructured data available within a news organization, from print and on-line properties, to gain insight into various factors affecting user engagement. Further, we want to be able to use these insights, to come up with predictive models for customer churn using features mined from transactional databases or Web-based textual data to determine which factors most impact user engagement and customer churn, using appropriate methods for each. Though entrusted primarily with subscriber transaction data we believe that any available Web data, indicative either of user activity, e.g. search clicklogs [23], or of readership, e.g. Web news [24], could be a rich source of signal for the task of churn prediction, hence these sources are also included into our study to determine factors affecting churn. The intuition behind our use of clicklogs is to find out how digital user behavior impacts print subscription and vice versa, and clicklogs could provide important types of signal related to traffic patterns and preferences of users [23], e.g. searches for items that: (i) cannot be easily or directly found in news or advertisements, (ii) are very related to news or advertised products, or (iii) are unrelated to news. Similarly, Web news offers almost similar content to the subscriber as print. We therefore use Web news as a source of signal to look at the impact of environmental context on consumer behavior [7] hypothesizing that top ranking news items may provide valuable cues into user engagement [18, 22, 26], and hence may be correlated.

Thus our strategy is to mine these Web information sources using an unsupervised learning approach such as LDA-based probabalistic topic modeling [5, 29] well-suited to this task, in order to extract useful features to gauge user engagement [24, 14].

According to *The Search for a New Business Model*—a report released by the Pew Research Center [25], in the embattled newspaper industry of present times, various news organizations conducted business experiments after suffering major economic losses [25], and case studies in the aftermath have shown that improvements are indeed possible by customizing the business model for the community, and by a general push toward improving the editorial product, advertisement, and sales revenue, through various means—be it by offering a full range of on-line marketing services to merchants like The Santa Rosa Press Democrat (California); or a complete overhaul of media properties by building a digital company with a narrowly focused-editorial identity like The Salt Lake City Deseret News; or by a complete sales force reorganization like the Naples Daily News of Florida [25].

The Columbus Dispatch Printing Company, referred to here on as The Dispatch, is a newspaper with a circulation of 1.2 million subscribers. Being the only mainstream newspaper in its home city, it comprises of a news daily and weekly, several websites for local news, sports and events, associated blogs, and a couple of TV stations, all part of the Dispatch Media Group, that has at its disposal large amounts of content, shared by all its outlets and distributed through multiple modes (print, Web and mobile), all delivered by a common server infrastructure. Since The Dispatch operates in a captive market, where its strategy is to offer similar content across all its delivery vehicles, our hypothesis is that traffic across the various modes of delivery, and hence the datasets we consider, are correlated. Our research objective therefore, is to jointly explore these structured and unstructured datasets to gain insight into various factors affecting user engagement.

To our knowledge our work is the first to study patterns of both on-line and offline behavior of customers, by tying together Web and relational databases of user activity, for the task of predicting customer churn, in contrast to previous works in this space e.g. Coussement et al. [6] that only look at features from transactional data of newspaper subscribers. Our experimental results confirm our intuition for using Web features to model subscriber churn and demonstrate the value of extracting signal from the Web.

2. BACKGROUND AND RELATED WORK

Churn is defined in the telecommunications industry, as the number of customers leaving during a period, divided by the average customer base during that period [1, 2]. Some of the problems that the Dispatch is facing with respect to print circulation subscription are: losing existing subscribers, no new or younger subscribers, and a slower turnaround time for changing the business model. Thus the organization wants to use its diverse reserves of electronic and Web data, to come up with better business models that can drive subscriber engagement and retention.

Notable previous works such as by Iwata et al. for extending subscription periods, efficiently extract information from log data and purchase histories using Cox proportional hazards models and survival analysis techniques [17], to find frequent purchase patterns in users with a long subscription period, infer these users' interests and use it to improve recommendations for new users. Coussement et al. apply a SVM classifier to construct a churn model for newspaper subscription using mainly transaction and metadata based features, comparing two parameter-selection techniques [6]. An often-used performance criterion in churn prediction is lift [10], that measures how many times a classification model improves identification of potential churners over random guessing. De Bock et al. investigate use of probability estimation trees (PETS) and alternative fusion rules, to improve lift performance of four well-known ensemble classifiers [10].

Previous work on churn in online communities such as Dror et al. focus on predicting churn in new users, specifically within their first week of activity on a popular CQA website [13], while others such as Karnstedt et al. explore the relation between a user's value within a community in an on-line social network, constituted from various user features, and probability of the user churning [19]. Kawale et al. use influence vector models [20] by taking into account online players' game engagement and social influence from neighbors in influence propagation, in a MMORPG setting, for gamer churn prediction. Dave et al. use a timespent based model [9], speed of discovery and information theoretical analyses to find a subset of informative recommendations that are most indicative of user retention in an on-line personalized content discovery setting.

Maximum entropy is a powerful statistical model, widely applied in information retrieval & text mining [21, 18]. One advantage of such a model is that it enables the unification of information from multiple knowledge sources into one framework where each can be considered as a set of constraints in the model—from the intersection of all these constraints, a probability distribution with the highest entropy can be learned [18]. This study contributes to the existing literature by investigating the effectiveness of the maximum entropy based method [21, 8, 16] in extracting signal from heterogeneous sources of data such as a transactional database (structured) and the Web (environmental) [7], while employing an unsupervised approach such as LDA-based topic modeling to extract the most informative features [5, 29] for this task. Thus, our major contributions include: (i) a unique dataset normalization and modeling approach to carve out a future time-frame from the present data, for prediction, (ii) use of timelines to facilitate joins across heterogeneous data sources to enable studying patterns of both on-line and offline customer behavior in tandem, (iii) temporal analyses looking at short v/s long-term effects, and finally (iv) the use of topic modeling and sentiment analysis to extract signal on the ***nature*** of events driving user activity or readership, i.e. not the topics themselves per se, but what meta information these topics convey, such as the *type* of news that may have influence, e.g. *political* or *financial*, or the *type* of searches that are influential, e.g. *personal* or *general*. We present our unique insights from our set of experiments.

3. DATASET

The Dispatch has at its disposal large amounts of transactional, content and server access log data from their news websites, blogs, print and digital subscriptions. This dataset comprises data from various divisions of the enterprise, viz. news stories, blog content and comments data from thirteen different websites, and newspaper subscriber transactional data with subscription history, viz. current status of a sub-

Dataset	Volume	Time frame
TopChurn$_{TRANS}$	• 3.72 million transactions • 605K unique account histories	Jan 2011–June 2014
TopChurn$_{WEB\text{-}NEWS}$	• 13 websites • 100K news articles	April 2013–Sept 2013
TopChurn$_{WEB\text{-}CLICKLOGS}$	• 3.4TB of Amazon S3 server access logs	July 2013–May 2014

Figure 1: Overview of Dataset

```
ACCOUNT,TRANS_NUM,TRANS_TYPE,START_DATE,END_DATE,MEMO
45720000,75585270,START,01/02/2011,06/25/2011,NYT
16000066,74063439,RESTART,01/02/2011,02/19/2011,sub is
back from vacation
45700000,75705562,COMPLAINT,01/02/2011,01/02/2011,DM
John spoke with George, 2nd Sunday in a row no paper
33720000,75705297,COMPLAINT,01/02/2011,01/02/2011,Note
will be left with IC to follow delivery instructions to
double bag when wet
33450000,1055XYZQP,STOP,01/01/2014,12/31/2999,"did
offer to downgrade sub to Sunday paper only, sub
declined"
67799228,1056XYZQF,STOP,01/01/2014,12/31/2999,"per John
Doe, son called to let us know sub is in hospital, will
call to restart"
```

Figure 2: Customer transactions with memo text

scription, time-stamped transactions for start, stop, changes to or renewal of a subscription and associated memo text. The experiments for this project are performed on the print subscription history portion of the dataset, integrating features derived from activity, stories and sentiment from the Web to create models of churn based on linguistic, temporal and metadata features.

Figure 1 gives an overview of the portion of the dataset accessible to us that was used in our work for both exploratory and predictive analysis: 1) TopChurn`TRANS` — comprising a structured, relational database of transactions containing user complaint text directly tied to customer requests and satisfaction, and 2) TopChurn`WEB` — comprising free textual data from the Web, divided further into: (i) TopChurn`WEB-NEWS` — news stories from 13 websites and (ii) TopChurn`WEB-CLICKLOGS` — server access logs reflecting everyday user activity and website traffic on the 13 websites.

4. METHODOLOGY

We formalize the task of churn prediction as a supervised learning experiment for binary classification. Given the history of a customer account up to date D, we learn one of two class labels, `ACTIVE` or `DROPPED`, i.e. we predict the status of the account–whether the customer will remain subscribed or not, at date $D+K$. We investigate two timescales K: 3 and 6 months.

4.1 Data Preparation and Topic Modeling

The subscriber transactions in TopChurn`TRANS` are as shown in Figure 2 with the actual account ids, transaction ids and names de-identified to protect privacy. There are 5 types of `Status` an account may be in, given by the `TRANS-TYPE` field, viz. `START`, `STOP`, `PRODCHG`, `COMPLAINT` and `RESTART`. The data file of 3.72 million transactions is sorted by the transaction start date so all transactions for an account are in sequential order. We use this transaction data to obtain the sequence of status changes pertaining to an account and generate *account histories* one per subscriber account, corresponding to a *single instance* in our training, and create features from the counts, metadata and content of any complaints occurring for that account. Further we use a portion of the status sequence corresponding to an account history to generate the *gold standard label*, described in detail in *section 4.2*. Other textual features such as sentiment scores & unigram frequencies etc. [4, 12], and metadata features, are also calculated off the ***complaints text*** in the TopChurn`TRANS` dataset, and included as needed into the various models. The text in the TopChurn`WEB` (`NEWS` and `CLICKLOGS`) datasets are processed via unsupervised learning. We use Stanford NLP's TMT v0.4 [15] to create topic models for the TopChurn`WEB` news and search click log data and also calculate sentiment scores for the same [12]. Since we intended to use the Web data to predict churn for subscribers in our database, we had to identify elements that would allow an appropriate join between the two datasets, as the `ACCOUNT` element identifying a customer is present only in TopChurn`TRANS` but not in TopChurn`WEB` (`NEWS` and `CLICKLOGS`). The only element that allows this are the timestamps that are shared across these datasets, so we normalize dates across the board in order to facilitate this join. Thus the TopChurn`WEB` dataset was summarized as follows:

- TopChurn`WEB-NEWS` comprising $>$ 100K articles over 180 days, was summarized by grouping all news articles occurring within a given day and formatting the dates to join correctly with dates in our subscriber transactions database, so as to generate one concatenated document per each day, containing titles and content. We then run topic modeling on this set of documents and get topic distributions for 50 topics for each date for the news.

- We parse 3.4TB of TopChurn`WEB-CLICKLOGS` server log data to produce a unique 4-tuple of the form **{Requestor_IP, TimeStamp, SearchTerms, SearchQuery}** [23]. Our parsed data produced on an average about 1500 Web searches for each day of the available time period, on the 13 websites of the Dispatch. These are grouped similar to news producing a single concatenated document of Web searches on a given day, on which we again ran topic modeling to get distributions over 50 topics for each date for the clicklogs.

Our processed dataset contains 605K unique customer accounts, all of which were active at some time within the last three years. Roughly 41% of them were `ACTIVE` at date D-3 and 59% were at `DROPPED`, and 40% of them were active at date D-6 and 60% were at `DROPPED`. We did not balance the dataset as we wanted to keep our models as close to the real-world as possible. Some basic status change statistics we have are: Of the 246628 `ACTIVE` customers in the dataset 11% changed status from `STOP` to `RESTART` once, 74% never stopped, and 15% changed status from `STOP` to `RESTART` more than once. Of the 359073 `DROPPED` customers in the dataset 6% changed status from `STOP` to `RESTART` once, 79% stopped at some point never to restart, and 15% changed from `STOP` to `RESTART` more than once.

4.2 Label Generation for Predictive Modeling

In order to make our models predictive, and since we don't know what actual state an account will be in, K time steps out in the future, we model this by carving out the time window for prediction K, from the current data itself as follows: 1) For our experiments we consider the unit of time to

Status	DROPPED	OVERALL
TopChurn$_{TRANS}$		
Top-1,2,3	<topic13: 'Incorrect Renewal Processing, Batch Restart'> [1437] <topic05: 'Individual Named Customer complaints'> [1141] <topic02: 'Need Followup To Ensure Delivery, Placement'> [1022]	<topic14: 'Missed Delivery, Today's Paper, Redelivered' > [5044] <topic12: 'Missed Paper, Not delivered, Too early, Late Tomorrow ' > [1933]
TopChurn$_{WEB-NEWS}$		
Top-1,2,3	<topic03: 'Trayvon Martin, George Zimmerman, 2nd Quarter Earnings, Inflation'> [65124] <topic32: 'Local Area School Districts, Community News, Students Council'> [3193] <topic00: 'Memorial Tournament, Tiger Woods, Dublin, Friday'> [1350]	<topic03: 'Trayvon Martin, George Zimmerman, 2nd Quarter Earnings, Inflation'> [108258] <topic32: 'Local Area School Districts, Community News, Students Council'> [4834] <topic00: 'Memorial Tournament, Tiger Woods, Dublin, Friday'> [4391]
TopChurn$_{WEB-CLICKLOGS}$		
Top-1,2,3	<topic05: 'Song, Lyrics, Movie, Download'> [71176] <topic33: 'Police,Missing,Robbery, University, Student, Gunman, Facebook'> [71141] <topic15: 'Today Stories, Boxscore Standings,House Fire, Firefighters, Police'> [70605]	<topic05: 'Song, Lyrics, Movie, Download'> [117779] <topic33: 'Police,Missing,Robbery, University, Student, Gunman, Facebook'> [117688] <topic15: 'Today Stories, Boxscore Standings,House Fire, Firefighters, Police'> [116800]

Figure 3: Top ranking topics by dataset with counts

be *months*, so that prediction windows are set to 3 months, 6 months etc. Since this is set up as a binary classification problem, we use a *partial sequence* of most recent states an account goes through, in chronological order, to generate one of two class labels, ACTIVE or DROPPED. 2) Since not all account histories are of same length, we first normalize the history for a given account by generating a time-step for each month in the account's history until the last recorded transaction date. For each date in the history we then map a status as follows: If the timestamp is an actual transaction date in the data, we use its original status. For every interim date we generate a filler status of either TRUE or FALSE such that if the previous status in the sequence was a STOP or FALSE, then the current filler time-step status is set to FALSE; otherwise it is set to TRUE. 3) Once this normalized history is produced with a sequence of states the account went through, we hold out the statuses for the last K time steps, for label generation. Thus this partial sequence used for label generation is of length equal to the prediction interval K, e.g. it is the 3 last time-steps of account history for K=3. 4) An instance that has in its *partial sequence*, either a STOP|FALSE as the very last status, or a STOP|FALSE followed by nothing other than one or more COMPLAINTs (a subscriber may complain even after they have dropped), gets the label DROPPED. Any other combination of the five possible statuses in any order, gets an ACTIVE label. 5) Once the partial sequence has been used for label generation, it is then *forever discarded* and never used in the final model.

4.3 Exploratory Trend Analysis

Prior to creating models of churn and running experiments on different datasets, we wanted to gain insight into some of the reasons that subscribers either complain most, or express satisfaction, and ask if there might be certain trends within the news or Web searches that might affect readership in print or online. For this we ran LDA-based topic modeling on the datasets [24, 29, 14, 5, 15], annotated the topics, and joined with our processed account histories such that all Web news and all Web searches are mapped to account activity by matching timelines, to obtain a topic-account association. We then ranked the topics that co-occur most frequently with ACTIVE and DROPPED accounts.

Figure 3 shows the highest ranking topics associated with accounts for the TopChurnTRANS, TopChurnWEB-NEWS, and TopChurnWEB-CLICKLOGS datasets. Section 6 further details what insights these topic rankings bring into our experiments. Further, topic analysis done on a smaller subset of the data containing complaint text for only the DROPPED

Topic 3 – "Vacation, Job Loss"	Topic 6 – "Product Offers, Subscription Downgrades"	Topic 11 – "Moving, Will Call to Resume"
Cancel, called, wants, wanted,due,her, cancelled, just, with, then, acct, past, said, job, lost, account, over, vacation, made	offered,declined, rate, lower, offers, mail, online, Sunday, only, offered, downgrade, promoted, line, refused, discount, great, digital	call, when, back, restart, resume, new, going, ready, for, return, address, date,may,moving, settled, once, sure, with, she

Figure 4: Topic models for DROPPED complaints

subscriptions shows finer-grained topics, specific to only accounts that have a STOP status, indicative of reasons why a subscriber may stop their subscription either temporarily or longer term. This analysis reveals several reasons for DROPPED subscriptions such as due to *"Vacation and Job Loss"*, *"Product Offers, Subscription Downgrades"* and *"Moving, Will Call to Resume"*, as shown in Figure 4.

Complaints-based Features	Applicable Models
complaints-YN	baseline+K
• *complaints-3-5-YN* – binary feature for complaints between 3 and 5 • *complaints-gt-5-YN* – binary feature for complaints > 5	complaintstats+K
Temporal Features – temporal features calculated for short and long term	
• *AvgComplaintGap* – Averaged time steps between COMPLAINT statuses in history • *NumComplaints* – Total # of COMPLAINTS in history • *AvgProdChg*- Averaged time steps between PRODCHG statuses in history • *NumProdChg* – Total # of PRODCHG in history • *STOP-RESTART-YNGT1* - # of STOP statuses followed by RESTART in history • *TotalTrans* - Total # of Transactions in history	temporal+short+K temporal+long+K temporal+all+K
Top-100 unigram features – binary features for top 100 unigrams	
missed-YN,todays-YN,redelivered-YN,paper--YN,wsj-YN,auto-YN,account-YN,paper-YN,cut-YN,batch-YN,per-YN,sub-YN,dm-YN,paper-credit-YN,credit-YN,will-YN,pub:1-YN,ppr-YN,ivr-YN,vol-YN,date-YN,ic-YN,on-YN,prior-YN,delivery-YN,nyt-YN,restart-YN,d6-YN,contact-YYN,with-YN,in-YN,ensure-YN,for-YY,new-YN,no-YN,mrs-YN,cc-YN,customer-YN,called-YN,sample-YN,mr-YN,...,complaint-YN,satisfied-YN,that-YN,be-YN,this-YN,at-YN,message-YN,late-YN,upset-YN,offered-YN,vacation-YN,papers-YN,payment-YN,stop-YN,issues-YN,up-YN,door-YN,carrier-YN,ppprs-YN,deliver-YN,her-YN,time-YN,he-YN,been-YN,where-YN,said-YN,have-YN,if-YN,by-YN,back-YN,too-YN,sure-YN,address-YN,posting-YN,get-YN,missing-YN,pls-YN,due-YN,only-YN,coupons-YN,declined-YN,early-YN	topKunigrams+K
Sentiment features – sentiment scores for text in TRANS and WEB	
• *AverageSentiment* – Averaged score of sentiment polarity over associated text • *MaxSentiment* – Maximum score of polarity over associated text	complaintsentiment+K web+news+sentiment+K web+clicklog+sentiment+K
Topic features – topic distribution features for WEB models	
Topic00, Topic01, Topic02, Topic03, Topic04…Topic49	web+news+topics+K web+clicklog+topics+K web+news+all+K web+clicklog+all+K
Service-based Features – keywords related to service in complaints	
is-INCOMPLETE-YN,is-CANCEL-YN,is-REDELIVERY-YN,is-DELIVERY-YN,is-MANAGER-YN,is-UPSET-YN,is-THREATEN-YN,is-PROBLEM-YN,is-CREDIT-YN,is-MISSED-Y,is-NOT-SATISFIED-YN,is-NOT-READING-YN,is-WET-PAPER-YN,is-VERY-UPSET-YN,is-INCOMPLETE-YN,is-CANCEL-YN,is-REDELIVERY-YN,is-DELIVERY-YN,is-MANAGER-YN,is-UPSET-YN,is-THREATEN-YN,is-PROBLEM-YN,is-CREDIT-YN,is-MISSED-Y,is-NOT-SATISFIED-YN,is-NOT-READING-YN,is-WET-PAPER-YN,is-VERY-UPSET-YN	service+K

Figure 5: Feature Categories for Churn models

4.4 Feature generation

We generate features from the text and metadata fields of the TopChurnTRANS and TopChurnWEB datasets and have the following categories of feature sets shown in Figure 5, where K denotes the churn prediction window. The features calculated for the TopChurnTRANS dataset consists of six categories: ***Complaints-Statistics***, ***Service Keywords***, ***Temporal***: Short and Long-term effects, ***Status History***, ***TopK-Unigrams*** (we pick k=100) and ***Sentiment*** [12]. For the long-term temporal model we consider the entire history of the account barring the prediction window, and for the short-term temporal model we consider a portion of the history for a shorter period immediately preceding but again barring the prediction window, set during the feature generation phase, to 6 months. The TopChurnWEB-based feature sets consist of 2 categories: (i) a ***Topic Distribution*** [15] over a set of 50 topics, and (ii)***Sentiment*** scores [12], for processed text in both WEB-NEWS and WEB-CLICKLOGS.

5. EXPERIMENTAL SETUP

We set up churn prediction as a task in supervised learning with the feature sets described in the previous section and train churn models for two prediction windows, viz. 3 months and 6 months into the future, using 10-fold cross validation with a Logistic Regression classifier [8, 16, 27]. We hypothesize that existence of complaints for an account, being tied directly to customer satisfaction, is an important predictor for customer churn [28]. Hence our baseline is a classifier with a single binary `complaints-YN` feature, taking on a Yes or No value depending on whether the account for an instance has any complaints in its history or not. Several experiments were run for the prediction of dropped subscriptions with different sets of features active, as detailed in the results below. Predicting whether the subscriber churns or not is a binary decision, for which we use a logistic regression classifier due to its robustness and many advantages described earlier [21, 8]. We also experimented with classifiers such as C4.5 Decision Trees [16] and Random Forests [16], but since no significant performance gain is observed using these methods, and because the purpose of our experiments is primarily to extract signal, and compare performance of the various models derived from our linguistic, temporal and metadata-based feature sets, hence it is sufficient to apply only maximum entropy based logistic regression.

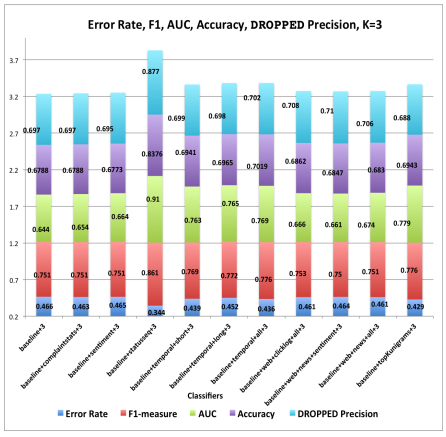

Figure 6: Churn Prediction, K=3

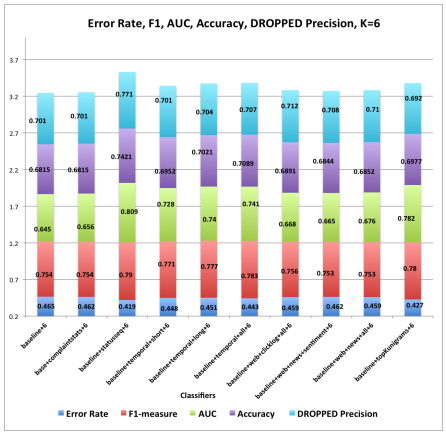

Figure 7: Churn Prediction, K=6

5.1 Results

The results for our experimental models that were run for K=3 and K=6, are detailed in Figures 6 and 7, with performance shown in terms of overall Accuracy, Precision of the +ve (`DROPPED`) class, Error rate, the Area Under ROC (AUC) and the *F1* performance measure of churn prediction. Our simple baseline has an AUC=0.64 and AUC=0.65, for 3 and 6-month intervals. The results show that each of our proposed models adds to the predictive value of the baseline for its timeframe, further validated by performing a two-sided Wilcoxon signed-rank test against the baseline that shows statistical significance with p-value $<<$ 0.01 for each. To gain more insight into which features are most informative for churn prediction, we calculate the Information Gain for each feature for predicting the target variable, for each of our churn models. Figures 8 and 9 show top contributing features sorted in descending order of value of Information Gain with number of features for a model in parentheses.

Feature	Infogain	Rank	Model
complaints-gt-5-YN	0.0421	1	base+complaintstats+3 **(3)**
complaints-3-5-YN	0.0132	2	
AvgSentiment	0.0314	1	base+complaintsentiment+3
MaxSentiment	0.0302	2	**(3)**
StatusSeq	0.6708	1	base+statusseq+3 **(2)**
complaints-YN	0.0706	2	
complaints-YN	0.0706	0	base+temporal+short+3 **(7)**
TotalTrans	0.0626	1	
NumComplaints	0.0391	2	
AvgComplaintGap	0.0272	3	
TotalTrans	0.1311	1	base+temporal+long+3 **(7)**
complaints-YN	0.0706	2	
NumComplaints	0.0685	3	
AvgGapComplaints	0.0604	4	
LongTermTotalTrans	0.1311	1	base+temporal+all+3 **(13)**
complaints-YN	0.0706	2	
NumComplaints	0.0688	3	
ShortTermTotalTrans	0.0626	4	
AvgComplaintGap	0.0604	5	
NumMatches	0.0172	1	base+web+clicklog+all+3
topic05	0.017	2	**(54)**
topic15	0.0167	6	
MaxSentiment	0.0166	9	
AvgSentiment	0.0166	10	
topic33	0.0164	13	
NumMatches	0.0228	1	base+web+news+all+3 **(54)**
topic32	0.0113	8	
topic03	0.0112	11	
topic00	0.0105	22	
...			
AvgSentiment	0.0084	53	
MaxSentiment	0.0083	54	
NumMatches	0.0228	1	base+web+news+
AvgSentiment	0.0084	2	sentiment+3**(4)**
MaxSentiment	0.0083	3	
missed-YN	0.0659	1	base+topKunigrams+3 **(110)**
todays-YN	0.0633	2	
redelivered-YN	0.0452	3	
paper-YN	0.0434	4	

Figure 8: Most informative features, K=3

TotalTrans	0.0831	1	base+temporal+long+6
complaints-YN	0.0721	2	(7)
NumComplaints	0.0659	3	
AvgComplaintGap	0.0595	4	
NumProdChg	0.0586	5	
complaints-YN	0.0720	1	base+temporal+short+6
NumComplaints	0.0343	2	(7)
AvgComplaintGap	0.0247	3	
TotalTrans	0.0197	4	
NumProdChg	0.0177	5	
TotalTrans	0.0831	1	base+temporal+all+6
complaints-YN	0.0720	2	(13)
NumComplaints	0.0661	3	
AvgComplaintGap	0.0595	4	
NumProdChg	0.0588	5	
complaints-YN	0.0722	1	base+sentiment+6
AvgSentiment	0.0317	2	(3)
MaxSentiment	0.0309	3	

Figure 9: Feature comparison with K=6

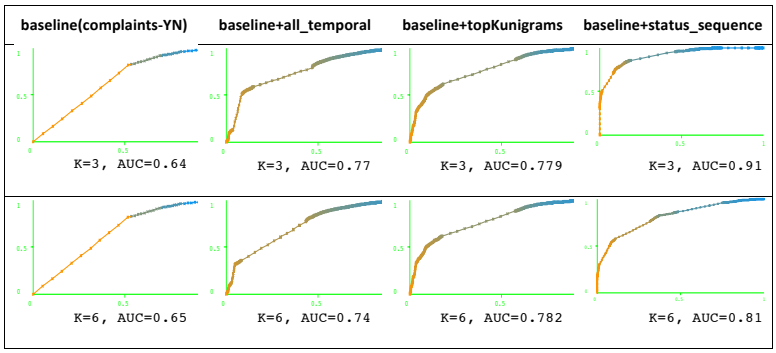

Figure 10: ROC curves for top performing churn models on TopChurnTRANS

6. DISCUSSION & FUTURE WORK

The experimental results for our churn models using the TopChurnTRANS dataset show the temporal, unigram and status history models to be clear winners for all prediction windows as seen in Figure 10, with *status history* outperforming all models with an AUC=0.91 for 3-months out prediction. Predictive power declines across the board for 6-month prediction for TopChurnTRANS, but this is expected given there is greater uncertainty with larger K. However, as seen in Figure 11, for experiments with the TopChurnWEB dataset, comparing purely text-based models of Web news and clicklogs against an augmented baseline having a complaint sentiment feature added to `complaints-YN`, we find the opposite effect, i.e. not only does inclusion of the WEB features increase predictive ability over this new baseline, but predictive power also actually increases slightly for larger interval K, thus the model with all NEWS and CLICKLOGS topics and sentiment features added to complaint sentiment, contributes the most, with AUC=0.69 for K=6—we leave this effect as a topic of future investigation. Another interesting finding from the `complaintstats` model shows that subscribers with >=3 but <=5 complaints tend towards *dropping* subscription, subscribers who *never* complained almost always *drop*, and subscribers with >5 complaints almost always remain ACTIVE—these, we confirm are stable subscribers, but may just be routine callers or vacationers.

For individual feature performance, we find that for the WEB models, sentiment features from CLICKLOGS rank much higher than those for NEWS; where all topics rank higher than sentiment. For temporal models, we find information gain of features flips depending on short or long term effects being studied, so previously less-informative features become more predictive for larger K, e.g. `AvgComplaintGap` and `NumComplaints` start to have more influence in our short-term effects temporal model, for K=6. Intuitively, we do expect the *temporal* models to be more sensitive to such effects, shown in fact by our results to be true. This analysis also reveals that topics found co-occurring highly with our +ve DROPPED class during the exploratory phase, e.g. <topic05: *'Song, Lyrics, Movie, Download'*> and <topic15: *'Today Stories, Boxscore Standings,House Fire, Firefighters, Police'*> from CLICKLOGS are actually very predictive in our experiments, showing that Web searches tend to be both, *quite unrelated* and *highly related* to the news. Also, as seen from Figures 3 and 8, ***local*** news such as given by <topic32: *'Local Area School Districts, Community News, Students Council'*> have more predictive value over ***national*** news,

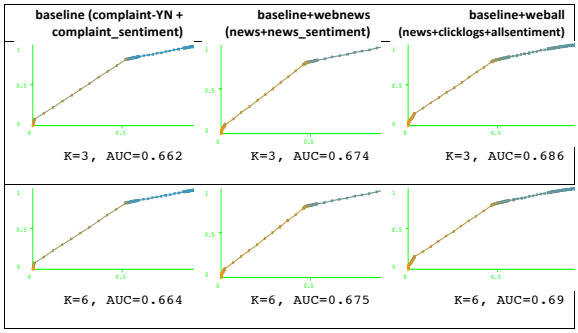

Figure 11: ROC – TopChurnWEB topics+sentiment

such as <topic03: *'Trayvon Martin, George Zimmerman, 2nd Quarter Earnings, Inflation'*> for *this dataset*, both of which also ranked highly during exploratory trend analysis.

7. CONCLUSION

A large media organization typically has at its disposal large amounts of historical transaction data with user complaints, updates to subscriptions, published content on the Web and on-line user activity from search click logs. We build several models of subscriber churn on such data, providing a comparison of the feature sets of these models with respect to their ability to predict subscriber churn. Our models based on temporal, unigram and Web-based topic and sentiment features, show statistical significance, and improved predictability over baseline models that utilize only transaction metadata, showing that features mined from Web news content and on-line user activity do have influence on newspaper subscriber engagement and ultimately, churn. Future research may be to develop and validate the effectiveness of enhanced churn models incorporating semantically enriched content from transactional data and the Web.

8. ACKNOWLEDGMENTS

This work was done as part of the Data Initiative at The Columbus Dispatch Printing Company. We are grateful to Jax Zachariah, Jason Cotter, John Valentine, John Schafer, Nikhil H. and Brian Espin for providing timely access to data, subject matter expertise and useful organizational insights that made this work possible.

9. REFERENCES

[1] Idiro Analytics. Retaining customers.

[2] Idiro Analytics. Rotational churn.

[3] Daniel Archambault, Neil Hurley, and Cuong To Tu. Churnvis: visualizing mobile telecommunications churn on a social network with attributes. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 894–901. IEEE, 2013.

[4] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] Kristof Coussement and Dirk 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.

[7] Mihaly Csikszentmihalyi, Matthias R Mehl, and Tamlin S Conner. *Handbook of research methods for studying daily life*. Guilford Publications, 2013.

[8] Hal Daumé III. Notes on cg and lm-bfgs optimization of logistic regression. *Paper available at http://pub. hal3. name# daume04cg-bfgs, implementation available at http://hal3. name/megam*, 198:282, 2004.

[9] Kushal S Dave, Vishal Vaingankar, Sumanth Kolar, and Vasudeva Varma. Timespent based models for predicting user retention. In *Proceedings of the 22nd international conference on World Wide Web*, pages 331–342. International World Wide Web Conferences Steering Committee, 2013.

[10] Koen W De Bock and Dirk Van den Poel. Ensembles of probability estimation trees for customer churn prediction. In *Trends in Applied Intelligent Systems*, pages 57–66. Springer, 2010.

[11] Thomas Debeauvais, Bonnie Nardi, Diane J Schiano, Nicolas Ducheneaut, and Nicholas Yee. If you build it they might stay: Retention mechanisms in world of warcraft. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, pages 180–187. ACM, 2011.

[12] TextBlob 0.9.0 Documentation. Textblob: Simplified text processing.

[13] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn prediction in new users of yahoo! answers. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 829–834. ACM, 2012.

[14] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[15] The Stanford Natural Language Processing Group. Stanford topic modeling toolbox.

[16] Geoffrey Holmes, Andrew Donkin, and Ian H Witten. Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361. IEEE, 1994.

[17] Tomoharu Iwata, Kazumi Saito, and Takeshi Yamada. Recommendation method for extending subscription periods. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 574–579. ACM, 2006.

[18] Xin Jin, Yanzan Zhou, and Bamshad Mobasher. A maximum entropy web recommendation system: combining collaborative and content features. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 612–617. ACM, 2005.

[19] Marcel Karnstedt, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. The effect of user features on churn in social networks. In *Proceedings of the 3rd International Web Science Conference*, page 23. ACM, 2011.

[20] Jaya Kawale, Aditya Pal, and Jaideep Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 423–428. IEEE, 2009.

[21] Dan Klein and Christopher Manning. Maxent models, conditional estimation, and optimization. *HLT-NAACL 2003 Tutorial*, 2003.

[22] Raymond J Mooney and Loriene Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204. ACM, 2000.

[23] Arnab Nandi and Philip A Bernstein. Hamster: using search clicklogs for schema and taxonomy matching. *Proceedings of the VLDB Endowment*, 2(1):181–192, 2009.

[24] David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *Intelligence and Security Informatics*, pages 93–104. Springer, 2006.

[25] Pew Research Centers Journalism Project. Newspapers turning ideas into dollars, February 2013.

[26] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM, 1994.

[27] Cynthia Rudin, Rebecca J Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. A process for predicting manhole events in manhattan. *Machine Learning*, 80(1):1–31, 2010.

[28] Clay Shirky. Broadcast institutions, community values. http://www.shirky.com/writings/broadcast_and_community.html. Accessed: 2015-03-01.

[29] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[30] Pin-Yun Tarng, Kuan-Ta Chen, and Polly Huang. On prophesying online gamer departure. In *Network and Systems Support for Games (NetGames), 2009 8th Annual Workshop on*, pages 1–2. IEEE, 2009.