Predicting Online Video Engagement Using Clickstreams

Everaldo Aguiar*, Saurabh Nagrecha†, and Nitesh V. Chawla‡
Dept. of Computer Science and Engineering,
University of Notre Dame. Notre Dame, IN 46556, USA
Email: *eaguiar@nd.edu, †snagrech@nd.edu, ‡nchawla@nd.edu

***Abstract*—As access to broadband continues to grow along with the now almost ubiquitous availability of mobile phones, the landscape of the e-content delivery space has never been so dynamic. To establish their position in the market, businesses are beginning to realize that understanding each of their customers' likes and dislikes is perhaps as important as the offered content itself. Further, a number of companies are also delivering content, product previews, advertisements, etc. via video on their sites. The question remains – how effective are video engagement channels on sites? Can that user engagement be quantified? Clickstream data can furnish important insight into those questions using videos as a communication or messaging medium. To that end, focusing on a large set of web portals owned and managed by a private media company, we propose methods using these sites' clickstream data that can be used to provide a deeper understanding of their visitors, as well as their interests and preferences. We further expand the use of this data to show that it can be effectively used to predict user engagement to video streams, quantifying that metric by means of a survival analysis assessment.**

I. Introduction

The constant growth in volume, speed, availability, and functionality of the Web brings with it not only a variety of challenges and risks, but also a number of opportunities. While there have been a series of major advances in the field over time, one that has been given a considerable amount of attention in more recent years is that of *personalization*.

Data about users' online activity is continuously captured and analyzed. Advanced recommendation systems are now able to tell us what products we might be interested in buying [1], the books we will enjoy reading [2], what movies we should watch next [3], and even which diseases we are at risk of contracting [4]. From a business perspective, the benefits of being able to understand customers in this level of detail are unquestionable.

Methods for capturing user data on the Web are also becoming increasingly efficient. As described in [5], the browsing behavior of individual users can be recorded at the granularity of mouse clicks with little to no work needed to be done. A number of services, both free and proprietary, offer user tracking solutions that can be implemented and deployed within minutes. However, the feedback that one usually gets from these tools is often in the form of simplistic aggregate statistics that do not offer a deeper understanding of user behavior.

With the above in mind, we set to analyze the application of some of these ideas to a specific context, while having as our major goal the understanding of each user as an individual unit. For this study, we analyzed a large dataset that describes user clicks generated within a two-month span and across a number of websites managed by a large media communications company.

We describe the process of analyzing and drawing inferences from the user-generated clickstream data. The objective here is to quantify user engagement in viewing video content and the development of a model to predict early exits in viewership. We begin by showing, from a more general perspective, how this type of data can be used to identify particularly interesting trends in user interest, and to further illustrate the usefulness of this information, we describe how we applied methods to predict user engagement to video streams and discuss their effectiveness.

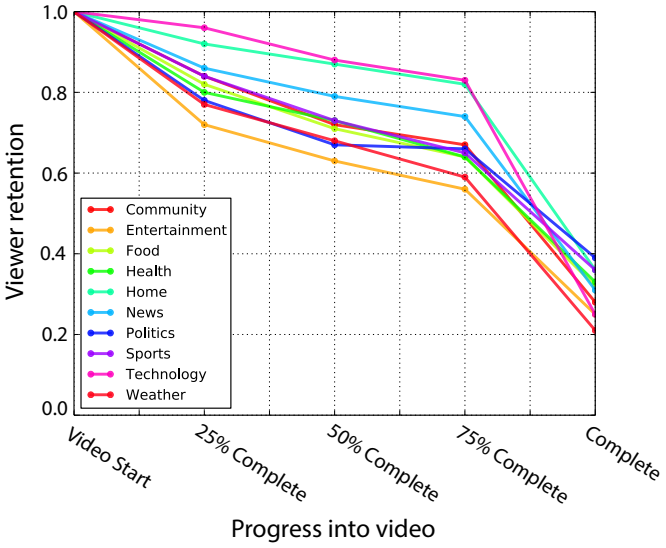

Fig. 1: Video viewership drop-off by category of content

Distributing content that encourages user engagement and captures large audiences is the ultimate goal of all web media

providers. Measuring and forecasting these variables, however, is not an easy task. As Figure 1 depicts, as time goes by, the amount of users that remain tuned to video streams, as illustrated by our dataset, dramatically decreases. For certain categories, the percentage of users that actually watch videos to completion can be as low as 20%.

To address this adverse outcome, we propose the development of clickstream-based models that can learn the individual preferences and characteristics of each user, and utilize this information to predict how engaged they will be to a particular video stream. Content providers can estimate the expected engagement of a particular video based on how past videos have fared with a given demographic. This helps them evaluate content production, while identifying particularly engaged audiences. Being able to know, in advance, if a user is likely to exit a video prematurely allows them some leeway to implement personalized intervention strategies aimed at maximizing viewership retention.

The remainder of the paper is organized as follows: The next section gives an overview of the most recent related literature. That is followed by a detailed coverage of our clickstream data representation and a description of our particular dataset. We then elaborate on the methods applied in this study, the results obtained and their importance in the subsequent sections. Finally, the last section draws conclusions about this exercise and argues for the latent potential that resides in user-generated clickstream data.

II. Related Work

Interest in analyzing the online activities of users is as old as providing consumer content itself. This problem has piqued the interest of multiple fields: marketing, psychology and computer science to name a few.

Mulvenna and Büchner et al. explored the need for data driven marketing intelligence in e-commerce [6], [7], [8]. Their work motivated how data could be leveraged to attract, retain, prevent the departure of customers, and promote cross sales. One of the earliest popular works in the area [9] primarily delineates the process of harvesting data at various levels, through a process of feature extraction from unstructured web data for pattern recognition, which is then used to create business intelligence. It also acknowledges the desire to incorporate privacy into such an analysis.

Since user activity provides an immense amount of measurable secondary data, various models to predict multiple aspects of their behavior have been proposed. User interaction has been studied at various levels — from gaze tracking [10] to broader patterns of path traversal within a website [11], [12]. Simple duration and dwell-time [13] can be used to predict when a user exits the site. User classification [14] can be used to identify what the user is specifically looking for and even morph the website [15] according to the custom tastes of that particular user profile. Personalized content based on click history has been implemented and widely adopted by commercial content providers [16], [17].

With the distribution of online video content becoming mainstream, the way we study user engagement has been greatly enriched. Studies like [18] have measured the role of video content quality in influencing user engagement, but did not utilize clickstreams to contextualize the video views. Online video engagement for Massive Open Online Courses (MOOCs) [19] has shown that the lessons learned from analyzing video views can be used to improve video authoring, editing, and interface design. It also emphasizes the value of *video dropout* as a metric for engagement. Though the MOOC work lacks the contextual history of the users, in this paper we leverage that and many other clickstream aspects to predict video engagement.

III. Clickstream Data Representation

Clickstream data consists of a "virtual trail" that users leave behind as they interact with a given system, website or application. More specifically, data that describes the state of a user's current session is recorded each time a click is performed, and the aggregation of that produces a *clickstream*, which can be used to reconstruct all actions taken by the user while he or she utilized that product.

While applicable to a variety of scenarios, the collection and analysis of clickstreams has become most notably popular in the context of Web-based tools and websites. As highlighted by Srivastava et. al. [9], the analysis of such information has potential applications in a number of areas such as website personalization and modification, system improvement, business intelligence, and usage characterization. Our contribution aims to cater to all of these facets of web based content delivery.

A. Data

The data we utilized for this study originated from a large U.S.-based media communications company that operates in the radio, TV, newspaper and online media domains. This company also manages a few dozen media websites, all of which are equipped to capture clickstreams of their visitors.

As part of the data collection process, user activity is continuously captured by numerous servers across the country, and is then concatenated at the end of the day in the form of daily "dumps". We utilized 59 of these files that covered the period ranging from December 4, 2012 to January 31, 2013. Altogether, they contain an upwards of **65 million click instances**.

Each click instance recorded is characterized by a large number of *features* (161 in this case). Table I lists a small subset of the most relevant features, and a brief description of each. With that information we are able to determine (1) how users reached the website, (2) what attracted them there, (3) what actions they performed while on the site, and (4) how they eventually exited.

Note that while there is no feature that captures the event of a user leaving the website, we followed the industry standard and assumed that when users are inactive for a period longer than 30 minutes (i.e., no click events originate from them during that time), they have likely exited the site [20].

Feature Type	Feature Name	Description
Nominal		
	Browser	The browser that was used
	Channel	The site that the page view belongs to
	City	The city the user accessed the page from
	Cookies	Whether the user had cookies turned on or not
	Country	The country the user accessed the page from
	Domain	Domain of the user's ISP
	Exclude hit	Identifies web crawlers
	First hit page	URL the user first landed on the website
	Frequency of visits	Denotes hourly, daily, weekly, monthly or yearly visit
	IP	Refers to the IP address of the user
	New visit	Determines whether the user is new to the site, based on cookies
	Referrer	Lists the URL of the website that referred this user
	Region	Refers to the state or region the user was in
	Search Keywords	The search string which led to the particular page
	Section	The section of the website where the click took place
	Subsection	Subsection of the website where the click took place
Numeric		
	First hit time	Timestamp of when the user first landed on the website
	Last click	Time stamp of when the last click was made by the user
	Last visit	Refers to when the user visited the site last
	Time & Date	Timestamp of when the click instance happened
	Visit number	Refers to the number of times the user has visited the site

TABLE I: Dataset features described. The above table is a description of the most important subset of features in the data.

This assumption allows us to group these click events from the original data into user sessions, which illustrate the path a user takes while browsing the website, and can be used to identify areas that attract more (or less) traffic. Figure 2 illustrates one individual session chosen at random from our data. We can see that the user in this case was referred to the domain through a link that they found on a social network website, and that their visit consisted of several hops, most of which happened in the *news* section.

Aggregating these sessions allows us to visualize which areas of the website are more popular, as well as which links connecting different sections are traversed the most. Take for instance the example illustrated in Figure 3. To generate this particular figure, we isolated the sessions corresponding to a certain newspaper's website, its 12 most popular sections, and the traffic between them. Among other observations, we noticed that the readers of this particular newspaper website were often prone to navigating to the *sports* section and reading multiple articles there.

Lastly, we note that based on information retrieved from specific features of our data, it is also possible to determine if a user is simply browsing text-based articles, viewing image galleries, or streaming online video. In the following sections, this property of our dataset is used to aid in the development of predictive models for video viewership engagement.

IV. Methods

In the context of the clickstream data representation as per Section III, we revisit the problem statement in terms of identification of class boundaries (video exit points), data parameters, and eliminate redundant/non-important features. We provide the relevant details in the following subsections. We then quantify user engagement using survival analysis and create a predictive framework to answer the question of *when* a given user is likely to exit given video content and *whether* they would exit early. Since the models must result in actionable insights and be explanatory, we also used simple models such as Naive Bayes and Decision Tables, in addition to decision trees and ensemble methods.

A. Identification of Video Exit Instances

When a user clicks a link that redirects them to a video, a flag is activated and from then on, while the user remains connected to that video, a separate log entry is made every time he or she finishes watching a certain percentage of that stream. Further, to ensure that this streaming activity can be uniquely identified, a *player ID* value remains constant. Concretely, this makes the corresponding clickstream log reflect a cumulative history of the viewer's progress within that video, with new entries being added to it at the 25, 50, 75 and 100% video completion marks.

Table II illustrates how video streaming events are recorded as part of the overall user clickstream. In that example, a user identified by his IP address arrives to a video after clicking on a link to it likely found in a news article. The user exited that video before reaching the 50% mark and immediately began streaming a different video, which was eventually watched to completion.

IP Address	Media Type	Player ID	Percentage Complete
123.1.2.0	news article	-	-
123.1.2.0	video	100	0
123.1.2.0	video	100	25
123.1.2.0	video	101	0
123.1.2.0	video	101	25
123.1.2.0	video	101	50
123.1.2.0	video	101	75
123.1.2.0	video	101	100

TABLE II: Simplified clickstream representation of a video streaming event.

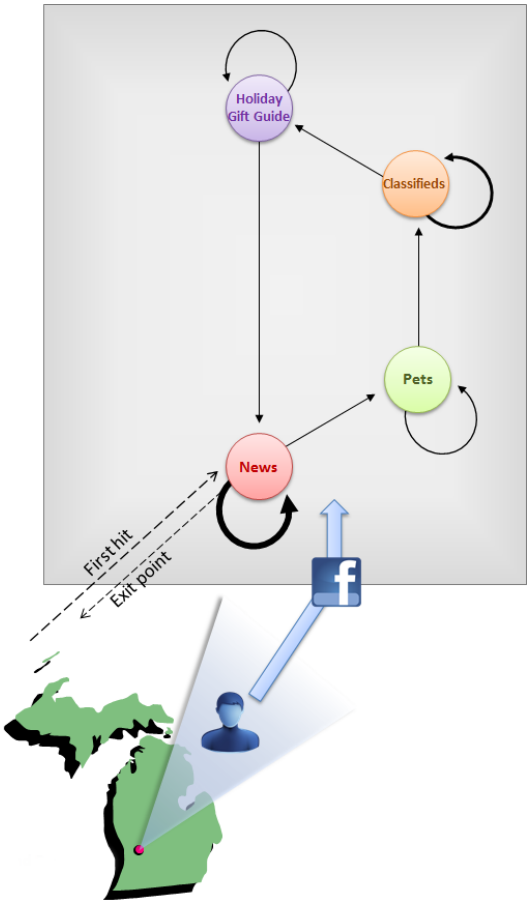

Fig. 2: A simple illustration of the clickstream of a typical user

The above narrative can be obtained at scale for the entire dataset's video viewers as follows. We obtain an individual user's browsing history on the monitored websites using their *unique visitor ID* provided by the data collection tool. The *unique visitor ID* is a persistent cookie which takes into account factors like IP address and user agent. Their video viewing history is marked by the corresponding entry in "media type". Individual video player instances for a given user's browsing session are assigned a *player ID*, which can be used to distinguish between multiple videos being played in the same session. The last entry in each video view for a user is considered to be their exit point for that video.

Due to nature of data collection of clickstreams in this dataset, we get a coarse-grained estimate of when the user reached a certain percent of the video. Entries are made in the clickstream upon starting the video, reaching 25%, 50%, 75% and watching the video to completion. If the last entry shows that the user reached the 50% marker, it can be inferred that the user exited at any point between 50% and 74% of the video.

Since one of our goals is to identify users who are likely to exit a video stream early, we assume that users who exit the video at the beginning, or having only reached the 25% marker, to have exited "early". As described above, this would correspond to users who have viewed 0 to 49% of the video (since the next marker only starts counting from 50% onward).

It should be noted that we can leverage two types of unique IDs towards different objectives— the *unique visitor ID* can be used in order to index users for segmentation and engagement statistics for better personalization, whereas the *video ID* can be used in order to index video content for feedback on how that particular content was received. Thus, our framework offers a pivotable user-centric and content-centric approach.

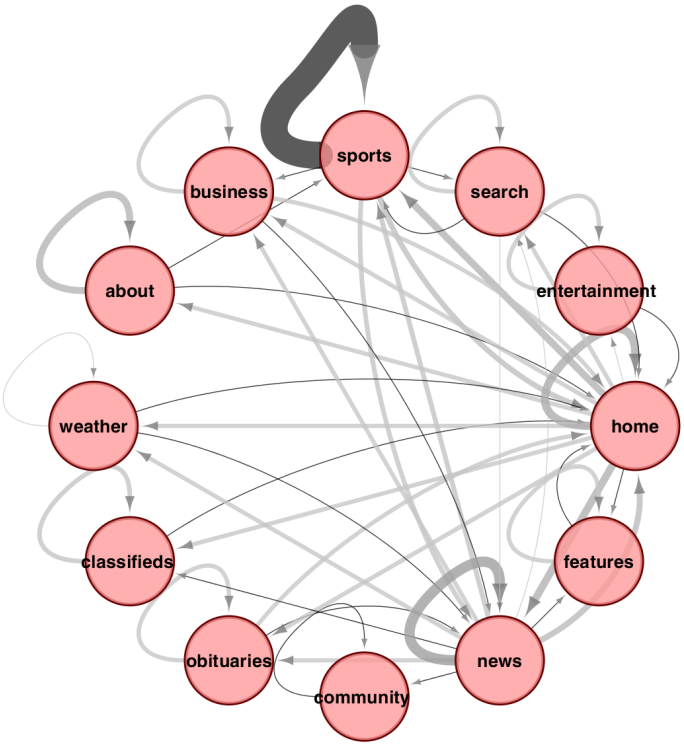

Fig. 3: Clickstream network for a news-media website. The various nodes displayed here represent different sections. The direction of the arrows illustrates user traffic flowing between these sections and the thickness is indicative of the volume of said traffic.

B. Feature Selection

Using various feature selection methods, we reduced the size of our dataset from the original 161 features to the 12 overall best descriptors. Among these were features like IP, location, content annotations, and referrer information. Out of the 161 features in a typical video exit instance, we also found that 40 were mutually redundant, and 32 were constant in value. This motivated the need to find a self-contained set of features that best described the target outcome variable (in this case, the percent of video the user watches before exiting).

Each of the applied methods aim to remove redundant or irrelevant features using different statistical means, all of which having distinct strengths. Though a popular choice in machine learning, correlation based feature selection (CFS)

was not considered due to the sparse nature of the data [21]. For a more detailed coverage of specific feature selection approaches, we refer the reader to [22], [23].

Below we give a brief description of each method utilized in this study:

1) Chi Squared: The chi squared (χ^2) method measures how much deviation is seen in the observed data when compared to the case where the class and feature values are independent of each other. It evaluates whether the feature and class co-occurrences are statistically significantly related.

2) Information Gain: Information gain [24] measures the amount of information about the class being predicted, when only a single feature and the corresponding class distribution are available. More specifically, it quantifies the expected reduction in entropy (uncertainty associated with a random feature).

3) Gain Ratio: A known drawback of information gain is that it favors features that are take a large number of values in the data. To address that issue, gain ratio [25] performs a similar measurement while attempting to reduce that bias by taking into account the number and size of the subsets that would be created when using each feature to partition the dataset.

4) One R: One R formulates a set of simple relationships between each of the available features and the corresponding classes. It then ranks these features based on the error rate produced by their associated rules.

5) Symmetric Uncertainty: Similar to gain ratio, symmetric uncertainty [26], [27] aims to compensate for the inherent bias associated with information gain measurements. In this specific case, that is achieved by dividing the information gain value by the sum of the feature and class individual entropies.

A summary of our feature selection exercise is presented in Table III. The features in the table are the ones which appear in the top 10% most frequently. These features can be said to most consistently indicate the likelihood that a user will exit a video at a given point.

The time of viewing influences at what point people are prone to exit the video. IP address, in conjunction with location, and ISP indicate who is watching the video and thus offer a personalized facet to the prediction. The data collection tool we used also provides its own cookie-based *unique visitor ID*. The number of pages viewed by a person and frequency of visits can be perceived to be reflective of the person's interest in the site. The referrer which brought the viewer to the site can influence the viewer's engagement; a viewer coming from a social network link interacts differently than one who had the site bookmarked on their browser. The entry point is the first page the viewer saw in their current viewing session; this determines their interest in consuming further content. The actual title of the story includes the section which the video is under. As we had observed in Figure 1, users viewing "Weather" related videos were empirically less likely to exit than those viewing "News" related videos.

Features	Chi	IG	GR	oneR	Symm
Time	1	1	7	-	2
IP	2	2	9	-	3
First hit referrer	3	3	5	2	5
First hit page	4	5	10	-	7
Story title	5	4	2	1	1
Search engine	6	7	3	3	8
City	7	6	-	-	9
ISP	8	8	-	-	10
Referrer type	9	10	1	-	4
# Pages viewed	10	9	8	-	6
Search page num	-	-	4	4	-
Frequency of visits	-	-	6	5	-

TABLE III: Feature Selection Rankings.

C. Survival Analysis

In order to statistically quantify what type of content is the most engaging for users, and what type of users respond positively to such content, we use a survival analysis model. This serves the dual purpose of segmenting content in terms of viability, as well as users in terms of engagement. Under the assumption that a user has been inactive for more than 30 minutes, we have an uncensored list of video exit events.

We isolate each video exit event and create a list of such "last known" video view clickstream logs. We would like to investigate the role of session duration on early exit behavior in users; so, we use it to measure when the user stops watching the video. A "birth" event would be registered when a user viewing video content makes their first click in the session. It is to be noted that this may or may not be a click pertaining to video content. A "death" event corresponds to an early exit, as defined in Subsection IV-A above.

Each of these events is characterized using the feature-set obtained in Subsection IV-B. This provides customized groupings according to each feature value in the data. Since we are observing the full extent of the user's interaction with the site's content from individual "birth" to "death" events, there is no left or right censoring at play.

We fit a Kaplan Meier model [28] to estimate the survival function for each of these events in the data. The survival function is a probabilistic estimate of whether the user will stop watching the video prematurely. The survival function estimate used in this paper is as per [29].

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

Here, d_i is the number of death events at time t and n_i is the number of subjects at risk of death just prior to time t. From this survival function, we get a global estimate, as well as feature-wise groupings of segments of users and corresponding content. This approach is implemented in the Python package "lifelines" [30], which was used in this paper.

The survival curve, as shown in Figure 4, is a graphical representation of the probabilistic model learned from the data. The "timeline" represents the session length (in seconds) up to the point of viewing the video. The Y-axis represents the probability of a user's "survival", which, in this case is the

event where they continue to watch more than 50% of the video content. Since the survival function is based on the idea that the number of "deaths" is a non-decreasing function, the survival curve itself is a non-increasing function of time.

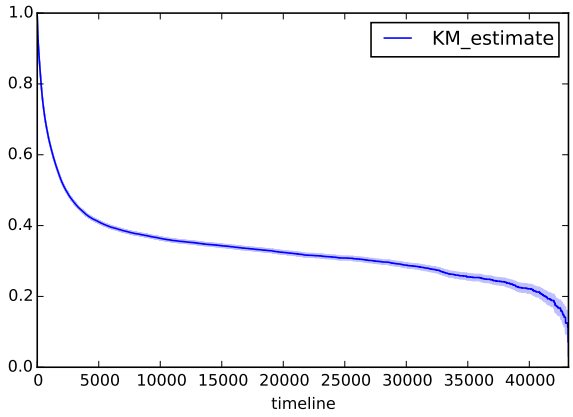

Fig. 4: The overall survival curve for video viewership in the clickstream data. The vertical axis represents probability of the user watching at least half of the video content.

D. Classification

Our aim in this specific scenario is to predict how much of the video a user will likely watch before exiting. In our dataset, we find that this is represented by 5 distinct markers (as previously mentioned and illustrated in table II), which correspond to the percentage of the video the user watched before exiting. Using these markers as our classes, we formulate two classification tasks – to predict what percentage of the video is watched, and whether the user exits the video "early" (before reaching the 50% mark).

The latter can essentially be seen as the binary prediction of these "early exits". The classes would then be a merger of the previously mentioned 5 classes, with the first two combined to form that of "early exits", and the latter 3 being those who did not exit early. This simplification is depicted in the representative expected confusion matrices for both of the classification tasks in Figure 5.

To generate predictions for each of these scenarios, we utilized a variety of classification methods. When selecting these classifiers, we took several factors into consideration, keeping in mind that while the ultimate goals was to produce predictions that are highly accurate in nature, we also had to ensure that a good degree of interpretability existed so as to allow businesses to quickly derive actionable insights from the models, as well as explain them. For that reason, the range of methods we evaluated covered models from simplistic to more complex. We present the results yielded by five of these, and a brief description of each follows.

1) Naive Bayes: Among the simplest and most primitive classification algorithms, this probabilistic method is based on the Bayes Theorem [31] and strong underlying independence

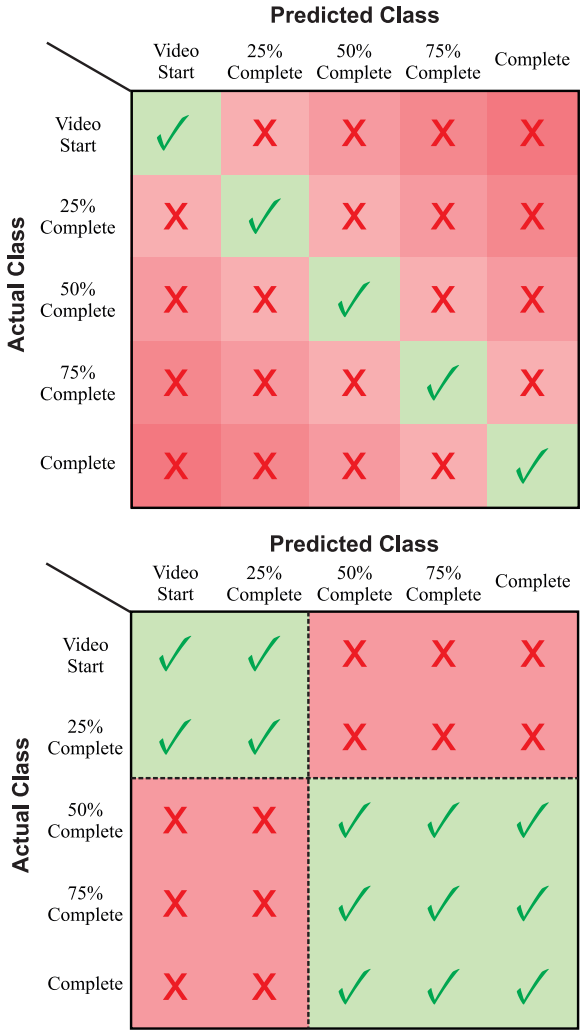

Fig. 5: Converting the Percentages Classification to Early Exit Classification: The 5 class problem (top) is reduced to a binary classification problem by merging classes (bottom).

assumptions. That is, each feature is assumed to contribute independently to the class outcome.

2) C4.5 Decision trees: C4.5 Decision Trees [25] work by building a tree structure where split operations are performed on each node based on information gain values for each feature of the dataset and the respective class. At each level, the attribute with highest information gain is chosen as the basis for the split criterion.

3) Repeated Incremental Pruning to Produce Error Reduction: RIPPER [32] is a rule based classification tree learner. It is algorithmically faster than C4.5, having a complexity of $O(n(log(n))^2)$ as opposed to C4.5's complexity of the order $O(n^3)$. RIPPER constructs an initial set of rules and then iteratively optimizes it according to a tunable parameter.

4) Decision Tables: Decision Table classifiers [33] are built by concatenating a series of rules derived from the feature set to corresponding class outcomes. This method has as its major advantage the fact that it is easy to interpret and notably

Dataset	Classifier	Acc	AUROC
Multiclass	NB	0.416	**0.718**
	C4.5	0.547	0.699
	RIP	0.547	0.629
	DT	0.543	0.717
	ST	**0.569**	0.652
Binary	NB	0.772	0.753
	C4.5	0.809	0.794
	RIP	0.826	0.697
	DT	0.806	**0.805**
	ST	**0.846**	0.739

TABLE IV: Summary of results obtained for each classifier and dataset. The classifiers used are NB: Naive Bayes, C4.5: C4.5 decision tree, RIPPER: Repeated Incremental Pruning to Produce Error Reduction, DT: Decision Table, ST: Stacking using random subspaces of decision trees

efficient.

5) Stacking: Stacking [34] is a meta-classification scheme which employs an ensemble of classifiers and performs the learning task on two levels. First, the classifiers in the ensemble are trained on the data, then the meta-classifier learns from *their* predictions and the training labels of the data.

E. Key Performance Metrics Utilized

Our key performance indices are the accuracy of prediction of when the user will drop-off in the video, and the area under the ROC curve for the classifier. To generate our predictions, we perform 10-fold cross-validation on the available data using various classification methods. In 10-fold cross validation, the data is randomly partitioned into 10 equal-sized subsets, 9 of which are used for training, with the remaining portion serving as the test set. The process is repeated multiple times until all 10 subsets have been used for testing. These predictions are then aggregated to provide the overall performance of the classifier.

The baseline for the performance metrics corresponds to a perfectly random predictor. For a binary classification problem, this would be 50% accuracy and 0.5 area under the ROC curve. For a 5 class problem, the baseline accuracy would be 20% and the baseline area under ROC curve can be calculated as per [35]. Any classifier which delivers statistically greater accuracy than these respective baselines, is considered to be better than a random predictor.

A system tuned to increase accuracy does not necessarily make it a good predictor. Relying on accuracy alone does not provide insights into the nature of misclassified instances, and can be deceiving when the provided class is imbalanced. ROC curves are a way to quickly compare multiple classifiers.

V. Experimental Results

The methods described in Section IV aim to quantify engagement and create predictive models for the clickstream data. In this section, we present and discuss the results of each of these overarching tasks first by elaborating on the outcomes observed from our survival analysis, followed by our classification task results.

A. Survival Analysis

Though the Kaplan Meier survival function as per Figure 4 gives us a *global* estimate of engagement in terms of session time, the true strength of this analysis lies in segmenting these aggregate statistics into actionable feature-wise insights. We segment users based on their feature values, and highlight trends and usable insights from their survival curves. In general, between two survival curves, the more "engaged" one is the curve which is significantly higher than the other in terms of survival function value. Each of the results displayed also contains the global estimate labeled **KM_estimate**. We display some of the salient results from the survival analysis below:

1) **Website:**
   Since the media communications company in our data manages multiple websites, it becomes an interesting study to compare engagement across them. We can use the survival analysis results as a metric to compare and rank websites delivering similar video content in terms of viewer engagement. In Figure 6, we can see that Website #1 is clearly more engaging than the rest, whereas Website #4 is less engaging than the average for all users.
2) **Content Type:**
   Viewer acceptance of contents of various types serves as valuable feedback to content providers, as well as content production. In our data, each video instance has been tagged with a singular content category by the content provider. As per Figure 7, we can see that users differ in their amount of engagement when it comes to type of video content. One can see that live content starts off as being significantly more engaging than any other content type, and progressively becomes less engaging as time elapses. News content is the least engaging throughout user sessions. Not only does this analysis help evaluate content production, but it also helps segment out the user-base in its taste.
3) **Referrer Type:**
   Whether a user was directed to the content via some referrer or directly searched for such content influences their level of engagement at the very outset. Figure 8 clearly shows that users who arrive at a video via links from the same site tend to stay significantly more engaged than average. On the other hand, users who land on content using bookmarked links tend to exit much earlier than average. This could be due the them being prone to more targeted sessions. Search engine referrals can be used as a metric of how accessible the news site's results are for a given search engine. Social network referrals can reflect on the marketing efficiency of certain news items.

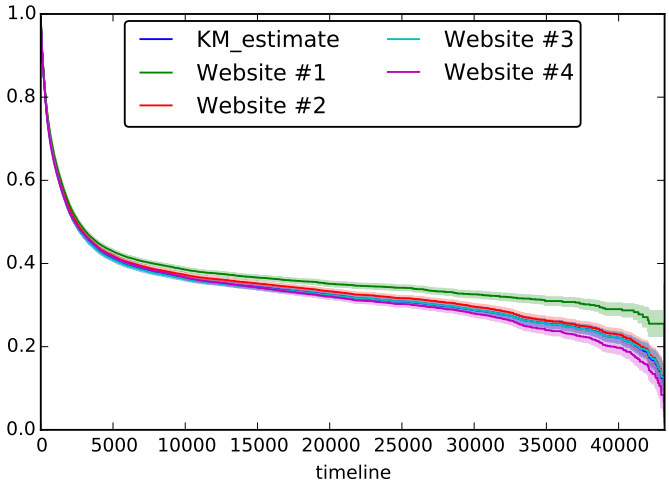

Fig. 6: Survival Curves for various websites. These offer a visual comparison of website engagement. Website #1 is more engaging than the rest of the websites depicted here.

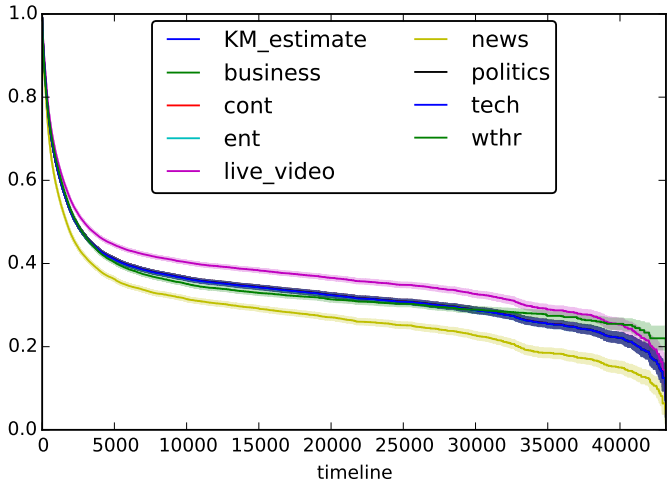

Fig. 7: Survival Curves segmented by content type. This serves as feedback for content providers, as well as to segment the user-base.

4) **Time of Last Visit:**
One would expect that the time of the user's last visit would be indicative of their future level of engagement, but as the data suggests in Figure 9, the survival curves are significantly similar. Someone who hasn't visited the site in more than 30 days has the same level of engagement as someone who visited recently within the past 24 hours. This can also be attributed to the fact that this feature is dependent on cookie data, which upon deletion by user, resets the "last-visited" date to NULL. Since it is prone to a lot of user-intervention and noise, it proves to be a bad indicator of engagement. Non-obvious insights like this make the use of data-driven tools invaluable. One would thus expect the importance of this feature to be very low as per the discussion in Subsection IV-B and Table III.

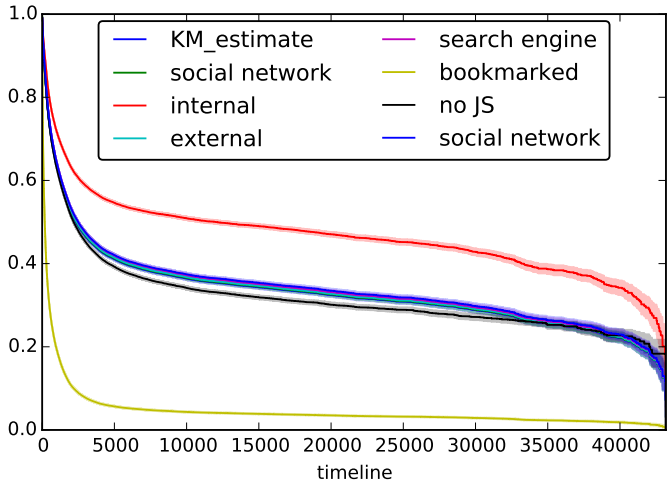

Fig. 8: Survival Curves for various referrer types. Users who arrive using bookmarks tend to have more targeted browsing sessions.

B. Classification

We evaluated the performance of each of the classifiers used, with 10-fold cross validation for both the multiclass and the binary classification predictions. Table IV summarizes the results of all experiments.

Both tasks have slightly different objectives– the multiclass problem aims to find out *when* the user is likely to stop viewing the video, whereas the binary format attempts to predict *whether* the user will stop viewing the video at an early stage or not. We use multiclass prediction in this data since the objective class (% of video content viewed) is recorded as a discretized ordinal variable. In the event where one has a continuous numeric value for percentage video viewed, it can be discretized and treated as a multiclass problem.

1) Multiclass Prediction: We see that in terms of overall accuracy, the stacked classifiers performed slightly better than the other methods, achieving an accuracy of 56.9%. With regards to AUROC, however, it is seen that Naive Bayes performs much better, closely followed by Decision Tables. These simple classifiers might not have the best accuracy, but they outperform the others based on more informative metrics. Further, because of their simplicity and efficiency, such models are, in fact, better suited for deployment in a real production scenario.

2) Binary Class Prediction: In this second scenario, we associate a semantic meaning to the drop-off percentage point and predict if the user will exit early or not. This refinement of the problem statement gives us a much better performance across the board. The stacked classifiers, for instance, achieve a remarkable accuracy of 84.6% when predicting which users exited their video streams prematurely. As it was the case with the multiclass problem, we again saw that Decision Tables

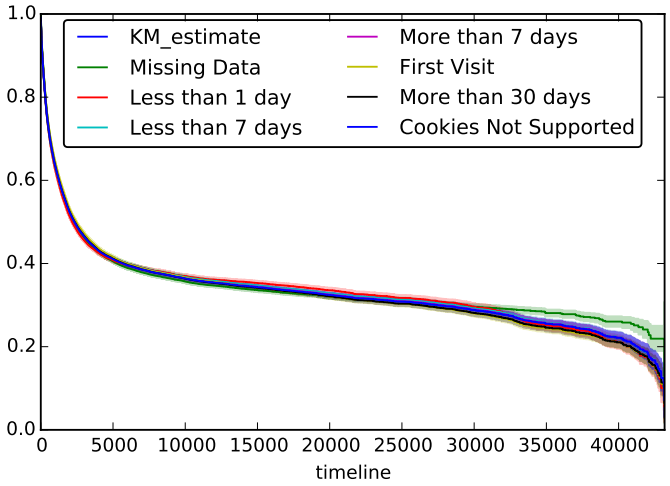

Fig. 9: Survival Curves for various user visit frequencies. Users who have markedly different visiting patterns show similar survival curves.

and Naive Bayes surpassed the other classifiers in terms of AUROC values.

Though stacked classifiers give greater accuracy, they are not as good as Decision Tables or Naive Bayes in predicting early drop-off. This is still reflective of the general trends observed in the multiclass problem as though we have merely merged classes, the underlying data remains the same.

In both, the multiclass and binary class prediction, it is observed that simpler rule based learners outperform complicated meta-classifiers. This is documented in [36], showing that stacking does not always outperform the best classifier.

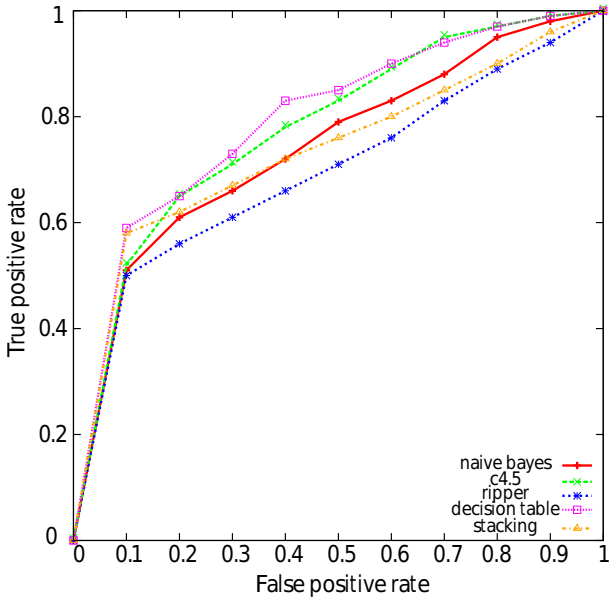

Fig. 10: ROC curves for the binary class problem. A comparison of various classifiers to predict early exit behavior.

We see that simple classification algorithms can be used to achieve comparable, or even better performance than the more complex classifiers, including ensemble methods. Besides the performance superiority, it might be more desirable to use simpler classifiers on grounds of computational complexity as well as comprehensibility.

VI. Conclusions

We demonstrated how clickstream data can be utilized to quantify and predict user engagement to online videos. For that task, we explored a large number of features and identified a subset of 12 that were deemed most important by a number of different feature selection algorithms. We used survival analysis to quantify the "engagement level" of content with the user. Finally, we used classification algorithms to predict what might be the exit point of a user in a video stream. Our work provides a foundation to understand and quantify video engagement using click-stream level behavioral data, which can then be used by media companies to not only proactively personalize video distribution, but also to aid in video advertisement placement, and on deciding the optimal video length for any given content category.

The scope of engagement can be expanded beyond the realm of video content, making it a viable solution for any web-based content delivery solution. With the capacity to model this particular behavior, media companies can take a proactive and personalized approach to how they produce and deliver new content, precisely adapting that strategy so as to optimize viewership retention. Businesses can make use of this paper as a prescriptive blueprint for engagement analysis of their online content. This is facilitated by all of the methods being off-the-shelf techniques. Our analysis suggests that non-obvious features such as *time of the day* and *referrer type*, can be strong predictors of whether or not a given user will watch a video to its completion. This solution serves the dual purpose of user segmentation and the provision of key performance indices to quantify content and news channel effectiveness.

References

[1] J. B. Schafer, J. Konstan, and J. Riedl, "Recommender systems in e-commerce," in *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, 1999, pp. 158–166.

[2] Z. Huang, W. Chung, T.-H. Ong, and H. Chen, "A graph-based recommender system for digital library," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2002, pp. 65–73.

[3] J. Bennett and S. Lanning, "The netflix prize," in *Proceedings of KDD cup and workshop*, vol. 2007, 2007, p. 35.

[4] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, "Time to CARE: a collaborative engine for practical disease prediction," *Data Mining and Knowledge Discovery*, vol. 20, no. 3, pp. 388–415, 2010.

[5] B. Clifton, *Advanced web metrics with Google Analytics*. John Wiley & Sons, 2012.

[6] M. Mulvenna, A. Büchner, M. Norwood, and C. Grant, "The soft-push: mining internet data for marketing intelligence," in *Working Conference: Electronic Commerce in the Framework of Mediterranean Countries Development, Ioannina, Greece*, 1997, pp. 333–349.

[7] M. Mulvenna, M. Norwood, and A. Büchner, "Data-driven marketing," *Electronic Markets*, vol. 8, no. 3, pp. 32–35, 1998.

[8] A. G. Büchner and M. D. Mulvenna, "Discovering internet marketing intelligence through online analytical web usage mining," *ACM Sigmod Record*, vol. 27, no. 4, pp. 54–61, 1998.

[9] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 2, pp. 12–23, 2000.

[10] X. Dreze and F.-X. Hussherr, "Internet advertising: Is anybody watching?" *Journal of interactive marketing*, vol. 17, no. 4, pp. 8–23, 2003.

[11] A. Banerjee and J. Ghosh, "Clickstream clustering using weighted longest common subsequences," in *Proceedings of the web mining workshop at the 1st SIAM conference on data mining*, vol. 143. Citeseer, 2001, p. 144.

[12] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty, "Modeling online browsing and path analysis using clickstream data," *Marketing Science*, vol. 23, no. 4, pp. 579–595, 2004.

[13] R. E. Bucklin and C. Sismeiro, "A model of web site browsing behavior estimated on clickstream data," *Journal of Marketing Research*, pp. 249–267, 2003.

[14] W. W. Moe, "Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream," *Journal of Consumer Psychology*, vol. 13, no. 1, pp. 29–39, 2003.

[15] J. R. Hauser, G. L. Urban, G. Liberali, and M. Braun, "Website morphing," *Marketing Science*, vol. 28, no. 2, pp. 202–223, 2009.

[16] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 271–280.

[17] J. Liu, P. Dolan, and E. R. Pedersen, "Personalized news recommendation based on click behavior," in *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 2010, pp. 31–40.

[18] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. A. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," *SIGCOMM-Computer Communication Review*, vol. 41, no. 4, p. 362, 2011.

[19] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," in *Proceedings of the first ACM conference on Learning scale conference*. ACM, 2014, pp. 31–40.

[20] Omniture, "White paper: SiteCatalyst Metrics- Visits and Unique Visitors," 550 East Timpanogos Circle, Orem, Utah 84097, April 2008. [Online]. Available: http://www.webmetric.org/white_paper/Unique_Visitors.pdf

[21] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[22] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol. 97, 1997, pp. 412–420.

[23] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[24] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[25] Quinlan, John Ross, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993, vol. 1.

[26] Witten, Ian H and Frank, Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[27] J.-H. Eom and B.-T. Zhang, "Machine learning-based text mining for biomedical information analysis," *Genomics & Informatics*, vol. 2, no. 2, pp. 99–106, 2004.

[28] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[29] J. P. Costella, "A simple alternative to kaplan-meier for survival curves," *Peter MacCallum Cancer Centre Working Paper (No*, 2010.

[30] C., Davidson-Pilon, "Lifelines," https://github.com/camdavidsonpilon/lifelines, commit = 63fc4f0ada0c61248957a478f3544efb4eeb2ccf, 2015.

[31] M. Bayes and M. Price, "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M.A. and F.R.S." *Philosophical Transactions (1683-1775)*, pp. 370–418, 1763.

[32] W. W. Cohen, "Fast effective rule induction," in *ICML*, vol. 95, 1995, pp. 115–123.

[33] R. Kohavi, "The power of decision tables," in *Machine Learning: ECML-95*. Springer, 1995, pp. 174–189.

[34] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[35] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[36] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine learning*, vol. 54, no. 3, pp. 255–273, 2004.