Predicting Session Length in Media Streaming

Theodore Vasiloudis*
RISE SICS
Stockholm, Sweden
tvas@sics.se

Hossein Vahabi
Pandora Media Inc.
Oakland, USA
puya@pandora.com

Ross Kravitz
Pandora Media Inc.
Oakland, USA
rkravitz@pandora.com

Valery Rashkov
Pandora Media Inc.
Oakland, USA
vrashkov@pandora.com

ABSTRACT

Session length is a very important aspect in determining a user's satisfaction with a media streaming service. Being able to predict how long a session will last can be of great use for various downstream tasks, such as recommendations and ad scheduling. Most of the related literature on user interaction duration has focused on dwell time for websites, usually in the context of approximating post-click satisfaction either in search results, or display ads.

In this work we present the first analysis of session length in a mobile-focused online service, using a real world data-set from a major music streaming service. We use survival analysis techniques to show that the characteristics of the length distributions can differ significantly between users, and use gradient boosted trees with appropriate objectives to predict the length of a session using only information available at its beginning. Our evaluation on real world data illustrates that our proposed technique outperforms the considered baseline.

KEYWORDS

User Behavior; Survival Analysis; Dwell Time; Session Length

ACM Reference format:
Theodore Vasiloudis, Hossein Vahabi, Ross Kravitz, and Valery Rashkov. 2017. Predicting Session Length in Media Streaming. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 07-11, 2017,* 4 pages.
https://doi.org/10.1145/3077136.3080695

1 INTRODUCTION

More and more people these days use online services to consume media. Whether they are consuming videos or music, users start an interaction with a service, consume a number of items, and after some time end their interaction. We refer to this complete interaction as a *user session*, and the time it takes from start to finish as the *length* of the user session.

Being able to predict the length of a session is important, because it allows the service provider to optimize the user experience along with its business goals. In terms of user experience, session length can be a useful signal for a recommendation engine. For a music streaming service, where users typically consume multiple items in a single session, the recommendation system can be tuned to be more exploratory or exploitative, based on the expected length of the session. On the business side, services often need to present ads to users to generate revenue, but there is a limit on how many ads a user will tolerate within one session [6]. The length of a session provides a vital data point for the trade-off between user satisfaction and generated revenue. By having an estimate of the length of a session early on, ads can be rescheduled so that the revenue target (i.e. number of ads presented) is maintained while minimizing the annoyance to users.

The length of user sessions can be hard to predict, because of two main factors: First, there exist a number of extraneous parameters that can affect session length, that are difficult to model based only on data that are available to an online service. Sessions can start and end for any number of reasons: users entering/exiting the subway, arriving at home or work etc. Second, user interactions commonly exhibit long-tail distributions; see for example dwell time studies [7, 9], phone call duration [10] and Section 3 of this study. This, in combination with the lack of predictive features makes it harder to correctly place the probability mass of predictive models. In this work we mitigate this issue by using an appropriate objective function for our model.

Most of the related research has focused on website visits, modeling the time spent on a clicked result (*dwell time*) after a search [2, 7] or an ad click [1, 8]. This type of interaction is very different from the way users consume items on a media streaming service. In a web search or ad click scenario, users enter a query or click on an ad, check the result and leave the website, in a "screen-and-glean" behavior [9]. In a media streaming service, users may interact very little with the platform, but have very long sessions ("lean-back" behavior), or have exploratory sessions, constantly revising their selection until settling down or abandoning the session. We show in Section 3.3 that 44% of the users exhibit "negative-aging" length distributions, i.e. sessions that become less likely to end as they grow longer.

*Part of this work was performed during an internship at Pandora Media Inc.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

ACM ISBN 978-1-4503-5022-8/17/08…$15.00
https://doi.org/10.1145/3077136.3080695

To summarize, the key contributions of this study are the following:

- We provide an analysis of user session length in an online media streaming service, using the Weibull distribution in Section 3.
- We develop a predictive model for session length using contextual and user-based features with appropriate objective functions in Section 4 and present experimental results in Section 5.

2 RELATED WORK

Survival analysis and prediction of dwell time has come into focus recently, with many studies using it as a proxy of user satisfaction in search and ad click scenarios. One of the first studies using the Weibull distribution analyzed the dwell time of users visiting web sites after performing a search [9]. The study indicated a strong "negative aging" effect for websites visited after a search , i.e. the probability of abandoning a page decreased with increasing dwell time (see Sec. 3.1).

The distribution of dwell time can also be used to measure and improve the satisfaction of users with search results [7]. In this study, the authors segment the results according to attributes like "readability level" and model the distribution of dwell time for each segment. Their findings include that the dwell time distributions for satisfied and dissatisfied clicks differ, and that it is possible to use characteristics of these distributions to better predict satisfied clicks.

Finally, predicted dwell time has also been used to improve the ranking of ads [1, 8]. The predicted dwell time is incorporated into an "ad-quality" model, which may include other aspects of an ad and a user, summarized as the probability of a user clicking on an ad. The authors develop predictive models and use features about the user and ad landing page to estimate the dwell time and bounce rate. They perform an evaluation on historical data and online experiments to measure the effect that using the ad quality model has on user engagement.

3 WEIBULL ANALYSIS OF SESSION LENGTH

In this section we perform an analysis of the session length distribution for users in our sample. We provide a brief introduction into Weibull analysis then move on to the results and discussion.

3.1 Weibull Distribution Review

The Weibull distribution is attractive for survival analysis because it allows us to model different kinds of failure rates, when the probability of failure changes over time. The probability density function (PDF) of the distribution is:

$$f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k}, t \geq 0 \quad (1)$$

The distribution has two parameters, k and λ, which correspond to the *shape* and *scale* of the distribution. The shape, k, determines how the elapsed time affects the rate of failure. The scale, λ, affects the spread of the distribution: the larger it is, the more spread out the distribution becomes.

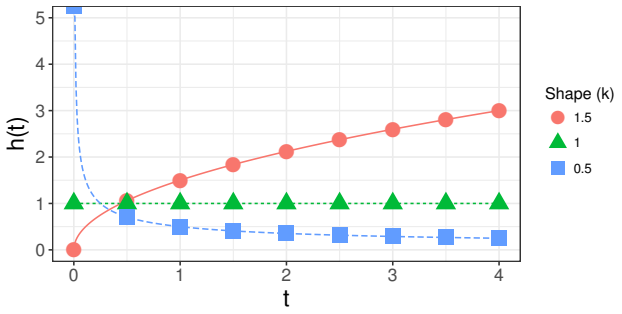

Figure 1: The failure rate of the Weibull distribution for different values of the shape parameter, k. We set $\lambda = 1$.

The effect of k can be better illustrated by the *hazard rate* (or hazard function) which gives us the failure rate of an item that has survived up to time t. For the Weibull distribution it is given by:

$$h(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{(k-1)} \quad (2)$$

The effect of k is illustrated in Figure 1. For values $0 < k < 1$ the hazard rate decreases as time increases. This behavior is often described as "negative aging" or "infant mortality" failures, where defective units might fail early on, but as time goes on and defective units get weeded out, the probability of a unit failing decreases. For $k > 1$ the probability of failure increases with time. This type of failures are called "wear-out" failures, when units become more likely to fail with time. For $k = 1$ the failure rate is constant and the distribution is equivalent to the exponential distribution.

3.2 Data

The dataset we use comes from user interaction data from a major ad supported music streaming service. We define a user session as a period of continuous listening, demarcated by breaks or pauses of 30 minutes or longer, i.e. a new session is started if a user stops or pauses the music for 30 minutes or more. We gathered data from a random subset of users for a period of 3 months (February-April 2016), resulting in 4,030,755 sessions.

In Figure 2 we can see a histogram for the session length data. For confidentiality reasons the x-axis has been normalized to 1000 bins. The distribution is highly skewed to the right, with a very small number of sessions going all the way up to the cutoff.

3.3 Analysis of user session length distribution

For our analysis, we fit a Weibull distribution on the data of each user using Maximum Likelihood Estimation with the `fitdistrplus` R package [4].

In Figure 3 we can see the empirical cumulative distribution for the shape parameter. We observe that the users in our sample are split approximately down the middle, with 44% of the users exhibiting Weibull distributions with $k <= 1$ and the rest $k > 1$. Although not directly comparable, we note that for the dwell time on web sites after a search, 98.5% of the web sites visited have dwell time distributions with $k < 1$, exhibiting almost exclusively the "negative aging" effect [9].

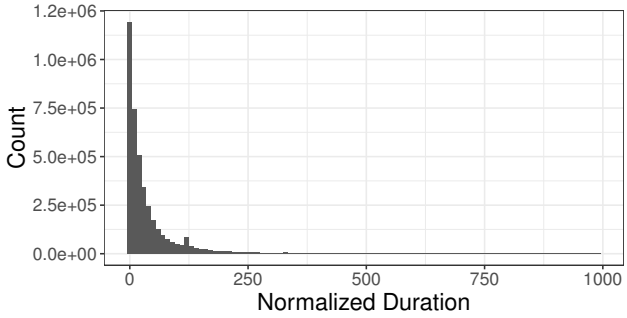

Figure 2: Histogram plot of session length. The x-axis has been normalized to the 1-1000 range.

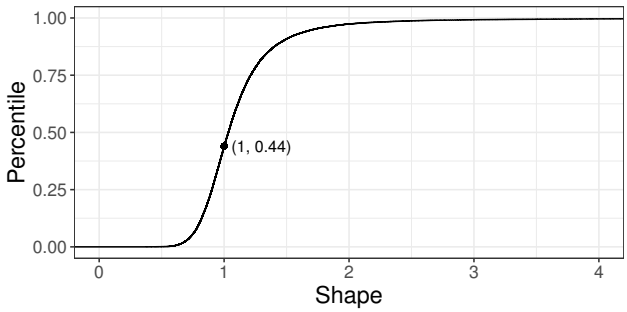

Figure 3: The empirical cumulative distribution for the shape parameter per user. The x axis has been truncated at $x = 4$ for readability (99.5 % of data points shown).

One consideration we should note here is that the variability in k could also be caused by sampling variability between users. We aim to investigate this through hypothesis testing in an extended version of this work.

4 PREDICTION

Apart from investigating the distribution of session length, ultimately we would like to be able to predict the length of a session. To that end we gathered features about the users and sessions, and treated the problem of predicting the length of a session as a regression problem.

4.1 Features

For each of the users and sessions available in our sample we collected a number of features. Some features, which we call "user-based" are features that we assume do not change between sessions, for example the gender of a user. Other features which we call "contextual" can change every time a user starts a new session, for example the type of network or device that a user was using when they started the session, or the length of their last session. We provide a summary of some of these features in Table 1, separated into user-based and contextual features.

Table 1: Example user-based and contextual features used in the models.

Feature	Description
Gender	The gender of the user
Age	The age of the user
Subscription Status	Whether the account is ad-supported
Device	The device used for the session
Network	The type of network used for the session
Previous duration	The duration of the user's last session
Absence time	Time elapsed since the last session

4.2 Model

We selected gradient boosted trees (GBTs) [5] as our model for a number of reasons: First, because our dataset contains missing data, the algorithm we chose had to be able to handle them explicitly, which decision trees do.

Second, the method should allow for proper modeling of non-negative data. For such data it is possible to log-transform the dependent and use a squared error objective, but using an objective function that is better fitted to the distribution of the dependent is often desirable, which is a common use-case for Generalized Linear Models. GBTs provide a flexible optimization framework that allowed us to do just that. To test both approaches, we first log-transformed the dependent and used a root mean squared error objective, then ran the same experiments again, this time selecting the log-likelihood objective of a Gamma distribution with a log link function, which allows for explicit modeling of right-skewed, non-negative data. In Section 5 we refer to these models as *linear* and *Gamma* respectively.

Finally, we tested two versions of each model. One aggregated where a single model was created using the data of all the users, and one per-user, where we separately trained one model per user, using only the data originating from that user for the training and testing. This meant that only contextual and not user-level features could be used to train the per-user models. This way we tested the trade-off between the statistical power that a large dataset provides versus having personalized models.

5 EXPERIMENTS

The baseline model we tested against is the per-user mean session length; that is, we calculated the mean session length in the training set for each user, and used that value to make all predictions for each session that user had in the test set. This gave us a baseline that is simple, but personalized to account for the differences in listening habits between users.

Because we are focusing on direct length prediction rather than survival probability [1], thresholded duration classification [8], or distribution parameter estimation [7, 9], most of the related work models used in search and ad click scenarios are not directly applicable. Therefore we don't include them in our comparison.

We used 10-fold cross validation, and stratified our sample per user to ensure that every user had data points both in the train and test set of each split.

Table 2: Performance metrics for length prediction task. We report the mean value across the 10 CV folds, and the standard deviation in parentheses.

Method (*Objective*)	Normalized MAE	nRMSE
Baseline	1 *(0.001)*	1.16 *(0.005)*
Aggregated (*Linear*)	**0.71** *(0.008)*	1.23 *(0.008)*
Aggregated (*Gamma*)	0.93 *(0.007)*	**1.10** *(0.005)*
Per-user (*Linear*)	0.83 *(0.002)*	1.29 *(0.004)*
Per-user (*Gamma*)	0.86 *(0.001)*	1.31 *(0.003)*

To ensure that we have enough data points per user, we only retained users that had at least 20 sessions recorded. The resulting dataset had 3,563,544 sessions. Due to the size of the dataset we chose to use the *xgboost* [3] variant of GBTs which is implemented with scalability in mind, utilizing parallel, cache-aware, and out-of-core computation to handle massive data sets. The parameters for xgboost were selected through cross-validation on a separate validation set.

5.1 Metrics

We chose two evaluation metrics to measure the performance of our algorithms. The first was the Root Mean Square Error (RMSE), which is a common choice for regression problems. In particular we used the normalized variant of the measure (nRMSE), which is simply the RMSE scaled by the mean value of the dependent, $\bar{y}$.

Large errors can be observed more often when the distribution of the dependent variable is highly skewed as it is in our case (see Figure 2). Therefore, we include the Median Absolute Error (MAE) in our analysis due to its robustness to outliers. This way a few very large errors will not affect the metric disproportionately, compared to taking the mean. For confidentiality we normalize all the MAE measurements by the baseline so that it has has an error of 1, and lower measurements are better.

5.2 Results

We report the performance of the various approaches in Table 2. As mentioned before, we refer to the models using the RMSE objective as *linear* to avoid confusion with the nRMSE metric. The linear aggregated model outperforms all models in terms of Median Absolute Error, but cannot beat the baseline on nRMSE. The aggregated model using Gamma regression has the best performance in terms of nRMSE, but has worse MAE than its linear counterpart. We note that it's the only model that beats the baseline in nRMSE.

The per-user linear models outperform the baseline for MAE but not for nRMSE, similarly to the aggregated linear model. The per-user models using Gamma regression perform similarly to the linear per-user models, indicating that the change in objective function becomes less important in small data domains.

What these results indicate is that the aggregated model using the Gamma regression objective is able to place the mean of the distribution more accurately because it places more probability mass in the right tail of the distribution. The linear models place more of their probability mass closer to the origin, allowing them to better capture the shorter sessions that are over-represented in the data, but as a result miss many of the longer (outlier) sessions. This causes their mean-based metrics to suffer, while median metrics benefit.

We also see that the per-user models mostly perform worse than their aggregated counterparts. This can be explained by the fact that per-user models are mostly trained on few data points, and for a flexible model like GBTs, they are likely to overfit. In this case the trade-off between having a single model trained with all the data versus having personalized models trained on each user's data favors the aggregated model.

6 CONCLUSIONS

In this paper we perform the first, to the best of our knowledge, analysis of session length in a mobile-focused online service. Our analysis showed that the probability of a session to end evolves differently for different users, with some exhibiting "negative aging" and others "positive aging". We also used a state of the art prediction algorithm to predict the length of a session based on user-level and contextual features, and demonstrated the differences between using a root mean squared error objective function and a Gamma regression objective that is more suited to right-skewed, non-negative data such as session length.

In the future we aim to demonstrate the utility of session length prediction, by using the predictions made by our model as input to a recommender system, or an ad scheduling algorithm. We also plan to improve our predictive models by incorporating features whose values evolve during a session (like interactions with the application), and keep a running estimate for the duration of a session as it evolves.

REFERENCES

[1] N. Barbieri, F. Silvestri, and M. Lalmas. Improving post-click user engagement on native ads via survival analysis. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 761–770, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[2] A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. A context-aware time model for web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 205–214, New York, NY, USA, 2016. ACM.

[3] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[4] M. L. Delignette-Muller and C. Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015.

[5] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[6] D. G. Goldstein, R. P. McAfee, and S. Suri. The cost of annoying ads. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 459–470, New York, NY, USA, 2013. ACM.

[7] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 193–202, New York, NY, USA, 2014. ACM.

[8] M. Lalmas, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. Promoting positive post-click experience for in-stream Yahoo Gemini users. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1929–1938, New York, NY, USA, 2015. ACM.

[9] C. Liu, R. W. White, and S. Dumais. Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 379–386, New York, NY, USA, 2010. ACM.

[10] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec. Mobile call graphs: Beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 596–604, New York, NY, USA, 2008. ACM.