Using Survival Models to Estimate User Engagement in Online Experiments

Praveen Chandar*, Brian St. Thomas*, Lucas Maystre, Vijay Pappu, Roberto Sanchis-Ojeda,
Tiffany Wu, Ben Carterette, Mounia Lalmas & Tony Jebara
vijay.pappu@onepeloton.com
Peloton
praveenr,brianstt,lucasm,robertoo,tiffanywu,benjaminc,mounia,tonyj@spotify.com
Spotify

ABSTRACT

Online controlled experiments, in which different variants of a product are compared based on an Overall Evaluation Criterion (OEC), have emerged as a gold standard for decision making in online services. It is vital that the OEC is aligned with the overall goal of stakeholders for effective decision making. However, this is a challenge when the overall goal is not immediately observable. For instance, we might want to understand the effect of deploying a feature on long-term retention, where the outcome (retention) is not observable at the end of an A/B test.

In this work, we examine long-term user engagement outcomes as a time-to-event problem and demonstrate the use of survival models for estimating long-term effects. We then discuss the practical challenges in using time-to-event metrics for decision making in online experiments. We propose a simple churn-based *time-to-inactivity* metric and describe a framework for developing & validating modeled metrics using survival models for predicting long-term retention. Then, we present a case study and provide practical guidelines on developing and evaluating a time-to-churn metric on a large scale real-world dataset of online experiments. Finally, we compare the proposed approach to existing alternatives in terms of sensitivity and directionality.

CCS CONCEPTS

• **General and reference → Experimentation**.

KEYWORDS

Experimentation; Long-term metrics; Surrogacy

ACM Reference Format:
Praveen Chandar*, Brian St. Thomas*, Lucas Maystre, Vijay Pappu, Roberto Sanchis-Ojeda,, Tiffany Wu, Ben Carterette, Mounia Lalmas & Tony Jebara. 2022. Using Survival Models to Estimate User Engagement in Online Experiments. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512038

1 INTRODUCTION

Ideas in online services are generated at a rapid pace, ranging from minor changes in the user interface to complicated under-the-hood algorithmic changes. Measuring the value of these ideas efficiently and reliably is essential to the success of these services. Online controlled experiments, or A/B tests, are a reliable way to evaluate ideas and ensure that the changes are aligned with the overall goal of the company. Typically, an *Overall Evaluation Criterion* (OEC) is used as a primary metric for evaluating the success of an A/B test. Therefore, choosing an OEC that captures the overall goal of the company is crucial.

Developing an OEC for experimentation is a challenging problem [34], as it needs to be tightly coupled with the company's goals. Online service companies with subscription-based revenue models need to continuously improve their products and provide quality services to retain customers in the long run.

This translates to keeping users continually engaged and satisfied on the platform, in both the short and long-term. Several industry studies have pointed out the importance of focusing on the long-term impact rather than selecting an OEC that emphasizes short-sighted gains [16, 28, 32].

A common approach is to identify short-term proxy metrics, or surrogates, that align with long-term outcomes. Recently, Athey et al. [3] proposed the surrogacy index to combine multiple surrogates, providing theoretical bounds necessary for proving statistical surrogacy – an important property required for causal validity. The practical challenges in using surrogates in online experiments have been less explored.

In this work, we present an approach to developing long-term metrics using survival analysis and discuss practical challenges in using them for decision making in A/B tests. We approach this problem through metric development for recommendation systems within Spotify, a large audio streaming platform, and provide an empirical study of online experiments. Traditionally, recommendation systems applications have relied on OECs that focus on short-term metrics such as consumption or click/stream-through rate. However, such metrics often provide a myopic view on success, ignoring both user's and company's long-term goals.

Based on past research on user engagement, we hypothesize that a satisfied user will retain and remain engaged with the platform. We, therefore, propose a churn-based metric that captures the time it takes for the user to become inactive for an entire week. A major

[*] Equal contribution authors.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9096-5/22/04...$15.00
https://doi.org/10.1145/3485447.3512038

challenge with this metric is that it takes a long time for users to churn, which delays A/B test decisions. We propose to model the time it takes a user to become inactive by casting it as a time-to-event problem. We adopt the survival analysis framework to model time-to-inactivity as it naturally accounts for censorship bias. We then use the model to combine short-term behavioral metrics such as consumption as a proxy to predict our long-term metric.

We use the survival estimates, i.e., the time-to-inactivity metric, to measure treatment effects in online experiments. The framework allows us to make A/B test decisions aligned with long-term outcomes without delays. While the use of survival analysis for modeling user engagement is not new, its use in A/B for decision making is less explored in the literature. We emphasize that our goal is not to find the best possible survival model but to open a discussion of their utility.

Therefore, we use the Cox Proportional Hazard model and present a set of validations and checks that help use predicted time-to-inactivity metrics in a reliable and trustworthy manner for experimentation. Further, we present a case study of online experiments to demonstrate the practical value of the time-to-inactivity metrics. We use Spotify A/B tests to empericially verify the directionality and statistical surrogacy of the metric. We also show that the metric is more sensitive than naive retention metrics, which enables faster, more accurate decision making.

The major contributions and findings of this work include:

- a simple time-to-inactivity metric for use in A/B tests and provide guidelines for its development using survival models.
- a set of validation checks for using predicted time-to-event for decision making in online experiments.
- a case-study of Spotify‘s online experiments to show that the time-to-inactivity improved sensitivity over retention.

2 RELATED WORK

2.1 User Engagement & Churn

There is a rich body of literature focusing on the measurement of user engagement in online platforms. Ranging from simple count-based metrics such as DAU, WAU, MAU [1] to more complex measures of success, often a combination of behavioral signals such as clickthrough-rate, and dwell time [4]. The use of behavioral signals such as clicks, mouse movements, eye-tracking to measure user satisfaction have been studied extensively in the context of search, recommendation, and advertising [29]. However, interpreting their relationship to long-term user engagement is an open issue. Dupret and Lalmas [18] proposed to address this by measuring user loyalty using absence time – the time between visits– metric. They showed that continuously tracking engagement was an effective way to optimize ranking algorithms and was extended by Kapoor et al. [30] to improve its accuracy. In our work, we also use Cox Proportional hazards to model time-to-inactivity but define our metric with expected survival time instead of hazard rate.

Our work is also related to the literature on churn prediction and customer lifetime value (LTV) [23]. Various models have been explored using parametric survival functions and forecasting future retention [21, 26]. However, they do not focus on decision making in online experiments. The work in [36], and the recommendations presented in [27] about using survival models for long-term metrics is more closely related to our paper. Here, we build on this domain by focusing on the practical aspects of validating the predicted long-term metric.

2.2 Surrogates & Proxies for Long-Term Metrics

Our work is also related to the literature on estimating long-term causal effect [35]. The use of short-term metrics as surrogates for modeling long-term causal effects is a common strategy. Early works have relied on strong surrogacy assumptions that require the short-term surrogates to fully mediate the long-term effect. Recently, Athey et al. [3] showed that when multiple surrogates are used, even if none of the individual metrics satisfy the statistical surrogacy criteria by itself, they may collectively satisfy the statistical surrogacy criteria. Recent works have also represented surrogates using sequential models, and used surrogates for policy optimization, e.g. [39]. Surrogate metrics for online experiments are further developed in Duan et al. [17], showing that surrogate function based metrics underestimate the variance of the target metric, leading to inflated type 1 error rates. We adopt theoretical findings from these works and provide a case study to highlight the practical challenges in long-term effect estimation in online experiments. We note that in survival analysis there is no parametrization of the residual variance, so the impact on type 1 error rates is non-trivial to estimate and plan to tackle in the future.

2.3 Online Experimentation

Online experimentation has evolved significantly over the past decades. Work in this area has focused on topics ranging between experimental design, hypothesis tests, and organizational culture around experimentation. Several papers have focused on common pitfalls, lessons learned, and recommendations for reliable online experiments [8, 12, 32–34, 37]. Recently, efforts have gone into automatic detection of biased experiments [20] and automatic ramp up to gradually expose treatments to users in order to mitigate risk [38]. The importance of measuring long-term outcomes [28] and the challenges in running long-term experiments have been studied in the context of search and advertising [11]. The primary focus of this paper is practical guidance for developing and validating metrics that are aligned with long-term outcomes, enabling reliable and fast decision making.

Measuring and improving metric properties has also become a key part of developing metrics for online experimentation. Directionality establishes how well a metric aligns with long-term business outcomes [8, 12, 32]. For instance, Dmitriev and Wu [13] discuss having a range of subject matter experts construct ground-truth labels of directionality for a set of experiments. Hohnhold et al. [28] construct directionally aligned metrics using a corpus of long-term experiments. In this work, we construct our metric as a surrogate and check statistical surrogacy assumptions to ensure directional alignment with long-term outcomes, which can be done without curated ground-truth labels or long-term experiments.

The second metric property, sensitivity, measures how likely a metric will detect a treatment effect in an experiment and is key in enabling product decisions [16, 33]. Several works have shown that

[1] DAU, WAU, & MAU indicate Daily, Weekly, and Monthly Active Usage respectively.

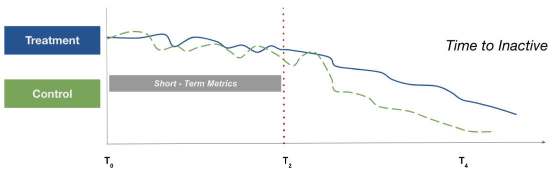

Figure 1: Illustrative example of the WAU retention and survival curves used to estimate time-to-inactivity.

historical data can be used to improve sensitivity through variance reduction [9, 14], or optimizing for sensitivity directly [31]. Others have shown sensitivity could be improved through adapting the method of treatment effect estimation [10, 15]. To measure the sensitivity of the surrogate survival metric presented in this paper, we use the objective Empirical Bayes prior proposed by Deng [7] for two-sample hypothesis tests. We note our approach could be used in combination with the above works on variance reduction and treatment effect estimation.

3 PROBLEM DEFINITION & CONTEXT

We describe a typical data-driven decision-making scenario faced by practitioners in online platforms. We highlight the need for developing long-term engagement metrics, formally define the problem, and enumerate the assumptions required to use our proposed approach in practice.

As discussed above, online service companies rely on A/B tests using engagement metrics such as consumption, clickthrough rate, or a combination of similar metrics for making data-driven decisions, but these short-term metrics do not always align with the long-term goals of the company [28]. Longer-term retention and engagement metrics are easier to align with the company's north-star goals, but there are non-trivial challenges in designing long-term experiments to observe the long-term engagement outcomes [11]. In our setting, a typical long-term outcome we observe is Weekly Active User (WAU) retention, i.e., if the user had streamed in the last 7 days. The main idea behind WAU retention at time t (e.g., WAU Week 4) is to capture the user's long-term satisfaction under the assumption that satisfied users will return to the platform.

We define *time-to-inactivity* as the amount of time that elapses before a user becomes inactive for an entire week. By constructing time-to-inactivity as a statistical surrogate of actual long-term WAU outcomes, and checking statistical surrogacy assumptions, experimenters can trust directional alignment with long-term outcomes.

In the context of online experiments, the time-to-inactivity metric measures the ability of a treatment (feature changes on the platform) to keep the users continually engaged. We use Figure 1 to illustrate the computation of the time-to-inactivity and WAU retention at time t metrics. Here, WAU Week 2 retention can be computed by counting the total number of active users. [2] To compute the time-to-inactivity, we estimate the expectation of the survival curves for *control* and *treatment* represented by green dotted and blue solid lines respectively, in the figure. The survival curve provides the ratio of users who have stayed active up until time t, and we estimate the expected time it takes for a user to become inactive using short-term engagement metrics (e.g., first two weeks). In our example, the time-to-inactivity would be higher for treatment users since they remain engaged for a longer time as indicated by the blue line compared to control (green line).

3.1 Time-to-inactivity Metric

Here we construct the time-to-inactivity metric to estimate the long-term treatment effect on WAU retention using short-term engagement metrics. We are specifically interested in using time-to-inactivity in online experiments for decision making. We will estimate the treatment effect, given an experimental dataset (D_E) in which users are exposed to different treatments over a short period of time (e.g., two weeks), with the following information

$$D_E = (X_i, S_i, W_i) \tag{1}$$

where X_i represents characteristics about the user, S_i is the observed short-term or immediate metrics such as consumption, and $W_i \in \{0, 1\}$ is the treatment assignment status of each user. We indicate the set of treatment or control users as $W^c := \{i|W_i = c\}$ and the number of users in a set as $N_{W^c} := |W^c|$. Our goal is to estimate the average treatment effect (ATE), denoted δ_E, as measured by true long-term engagement, denoted Y_i:

$$\begin{aligned}\delta_E =&\mathbb{E}_{D_E}[Y_i(1) - Y_i(0)]\\ =&\frac{1}{N_{W^1}}\sum_{i\in W^1} Y_i(1) - \frac{1}{N_{W^0}}\sum_{i\in W^0} Y_i(0)\end{aligned} \tag{2}$$

In practice, true long-term engagement Y_i (such as time-to-inactivity or long-term retention) is hard to observe. However, we do observe true long-term engagement in historical data $D_H = (X_i, S_i, Y_i)$, which is collected prior to the experiment. We define estimated long term engagement $\hat{Y}_i$ as:

$$\hat{Y}_i = \mathbb{E}_{D_H}[Y_i|X_i, S_i] \tag{3}$$

To estimate this conditional expectation, we learn a function f_H which can be trained on the historical data D_H, to generate predictions of $\hat{Y}_i$ as

$$\hat{Y}_i = f_H(X_i, S_i) \tag{4}$$

In Section 4, we will describe the use of survival modeling framework for estimation of f_H using historical data D_H, with time-to-inactivity as our long term outcome. We represent the time-to-inactivity outcome as a sequence of survival outcomes, with $\hat{Y}_i^t$ indicating the probability of survival at time t. The average treatment effect (ATE) is then computed using the expectation of $\hat{Y}^t$, i.e., $(\mathbb{E}[\hat{Y}_i^t(1)] - \mathbb{E}[\hat{Y}_i^t(0)])$.

Note that the time-to-inactivity metric relies on $\hat{Y}$ for ATE estimation, and $\hat{Y}$ is obtained from f_H fit on D_H, not D_E. Therefore, it is important to show the directional alignment of the predicted metric with long-term outcomes. For this to hold, i.e., for the ATE on $\hat{Y}$ to be an unbiased estimator of the true ATE δ_E, we require that the assumptions in Section 3.2 are satisfied. Further, we also require that the empirical function $\hat{f}_H(X_i, S_i)$ is recovered using an unbiased estimator from samples such that it converges to the true function f hypothesized to govern the relationship $Y = f(X, S)$ and that the function class H contains f.

[2] If the user had streamed in the last 7 days they are considered as weekly active user.

3.2 Assumptions

We rely on the theoretical foundations laid out by Athey et al. [3] that are required to show that the estimated average treatment effect (ATE) on $\hat{Y}$ is an unbiased estimate of the true ATE on long-term outcome Y. This ensures decisions made with the predicted metric are consistent with decisions had the long-term outcome been observed and ensures that confounders are accounted for. The three main assumptions are:

Positivity and Ignorability Assumptions. The treatment assignment does not depend on the long-term outcome Y and that users have a non-zero probability of assignment to the treatments. These assumptions are satisfied for our problem since the experimental data is collected from a randomized controlled experiment.

Surrogacy Assumption. The treatment affects the long-term outcomes only via short-term metrics. In other words, the set of surrogate or short-term metrics completely mediates the causal effect between the treatment and long-term outcome [35]. In practice, this assumption is difficult to directly test. In Section 5.2, we present the use of the likelihood ratio test as a sanity check to verify this assumption in the presence of the Comparability Assumption.

Comparability Assumption. The conditional distribution of the long-term outcome given the user characteristics & short-term metrics remains the same in both experimental and historical datasets. One implication is the function f should be the same when estimated using historical data D_H or experimental data D_E. In practice, D_H and D_E are sampled from different time periods, raising the issue of temporal model drift. As a sanity check, we use out-of-time model validation and measure the performance of f on data sampled from the time period that matches D_E.

4 SURVIVAL ANALYSIS FOR LONG-TERM EFFECT ESTIMATION

In this section, we describe the use of the survival analysis framework for estimation of f_H, i.e., the time-to-inactivity long-term metric. Our metric intends to estimate the time it takes for an event to occur, where the event is a user becoming inactive.

There are challenges in using traditional ML models for time-to-event problems. Consider, for instance, some users might continue to be active for long periods of time. Therefore, the event of interest (inactivity) might not occur at the time we want to make inferences; these users are referred to as *right-censored* individuals. One might be tempted to ignore them and use traditional ML models such as binary classifiers to fit f_H. However, this would introduce a censoring bias in our estimation. To illustrate this, consider two cohorts of users U_1 and U_2, where the true average time-to-inactivity is 2 and 36 weeks, respectively. If we wish to estimate the time-to-inactivity at week $t = 10$, then, ignoring the right-censored users would result in severe underestimation of the true time-to-inactivity metric.

Therefore, we choose survival analysis to model the relationship $\hat{Y}_i = E[Y_i|X_i, S_i]$, which is a natural framework to study time-to-event problem as it accounts for the censoring bias [6]. Next, we give a brief overview of a commonly used survival regression model as well as common goodness of fit criteria. In survival analysis we want to learn the survival function F, which gives the cumulative probability of an event occurring after time t, so that $F(t) = P(z > t)$ where z is the time of the event. The hazard function gives the probability that the event occurs at time t,

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq z + \delta \leq t + \delta | z \geq t)}{\delta} \tag{5}$$

and the hazard function $h(t)$ and survival function $F(t)$ are related through

$$F(t) = \exp\left(-\int_0^t h(z)dz\right) \tag{6}$$

so recovering the hazard function is equivalent to recovering the survival function.

We use the Cox Proportional Hazards model, noting there may be other survival models that would recover better estimates of the true survival function [1]. However the framework we propose in this paper for validating the resulting predicted metric would also apply to other choices of survival model.

4.1 Cox Proportional Hazards Model

The Cox Proportional Hazards model [5, 6], or the Cox model, assumes that the hazard function has the following form for a user i with covariates X_i:

$$h(t|X_i) = \lambda_0(t)\exp(\beta X_i) \tag{7}$$

In cases where the time to event exceeds the observation period, this data is right-censored. In our problem, right censorship occurs when a user is active every week, in which case their corresponding time-to-inactivity is recorded as the length of the observation window. The Cox model assumes that censoring is uninformative, meaning that the user's time-to-inactivity Y_i is statistically independent of the censoring mechanism, which is satisfied by having a fixed censoring time.

Notice that the baseline hazard function λ_0 is the only part of the hazard that varies with time, and so the ratio of hazards $h(t|X_i)/h(t|X_j)$ between users i and j will always be constant for every t. The likelihood for an event occurring at time t is given by

$$\begin{aligned} L_i(\beta) &= \frac{\lambda_0(Y_i)\exp(\beta X_i)}{\sum_{j:Y_j \geq Y_i} \lambda_0(Y_i)\exp(\beta X_j)} \\ &= \frac{\exp(\beta X_i)}{\sum_{j:Y_j \geq Y_i} \exp(\beta X_j)} \end{aligned} \tag{8}$$

Assuming statistically independent observations, the joint likelihood can be given by $L(\beta) = \prod_i L_i(\beta)$, which can be maximized using a Newton-Rhapson algorithm [19].[3] The corresponding survival model $\hat{F}_H(t|X_i)$ is then used to generate the *time-to-inactivity* metric at time t as

$$\hat{Y}_i = \hat{F}_h(t|X_i) \tag{9}$$

4.2 Model Validation

An essential part of developing the time-to-inactivity, a predicted long-term metric, is to ensure that the estimated metric $\hat{Y}_i$ is close to the true metric value Y_i. In other words, we need to verify that the empirical function $\hat{f}_H(X_i, S_i)$ fit using the Cox PH model, converges

[3] The authors used the implementation available through Lifelines https://pypi.org/project/lifelines/

to the true function f that recovers the true long-term outcome Y. Traditional metrics such as precision, recall, or root mean squared error is not suitable for this case since they cannot account for the censoring bias. In this section, we present three evaluation metrics for validating survival models.

4.2.1 Concordance Index. The concordance index (C-Index) is a measure of the model's ability to correctly order the individuals based on their risk score [2, 22]. It is computed as the ratio of concordant pairs to comparable pairs, where a pair of users $< i, j >$ is considered comparable if the $y_j > y_i$ and y_j is uncensored. And, a pair is concordant if the estimated risk score $\hat{y}$ aligns with the observed survival time y, $(\hat{y}_j > \hat{y}_i \quad y_j > y_i)$. For right-censored data, which is the case for our problem, pairs in which both individuals are censored are ignored.

4.2.2 Integrated Brier Score. The Brier score for survival problems with censored information was introduced by Graf et al. [24]. The metric is computed by accounting for censoring information in the dataset and re-weighting the individuals.

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} w_i(t) [\hat{y_i}(t) - y_i(t)]^2 \tag{10}$$

Here, y_i is the observed time-to-inactivity, $\hat{y}$ is the estimated time-to-inactivity for a given time t. And, w_i is $(1-C_i)/G(y_i)$ if $y_i <= t$ and $1/G(y_i)$ if $y_i > t$. where, C_i indicates if the individual is censored and G is the censoring distribution, typically estimated using the Kaplan-Meier estimator [1]. We report the integral of the scores at different time points computed using the trapezoidal rule.

4.2.3 AUC-ROC. It is often desired to use the estimated risk scores or survival probabilities to identify high-risk users. Here, we use the ROC curve to compare the false positive and true positive rates, and the area under the ROC curve ($AUC - ROC$) plotted for each time-point to measure effectiveness.

5 METRIC VALIDATION & SENSITIVITY

In general, when developing new metrics to be used for decision making in online experiments, two key qualities must be established: directionality and sensitivity [8, 32]. Since the proposed survival metric is learned using an ML model that relies on surrogates (i.e., short-term metrics), we validate directionality through checking statistical surrogacy assumptions [35]. In this section, we give practical guidance to verify the directionality of the metric, validate the surrogacy assumptions, and to measure metric sensitivity.

5.1 Directionality

Directionality is a metric quality indicating if the sign of the detected effect on the metric agrees with the true impact on user experience [8, 32]. To verify directional alignment, we can simply compare against the observed metric values. However, in the case of time-to-event metrics, it is challenging to make such comparisons since this requires following users for long periods of time (which requires a corpus of long-running A/B tests). Instead, we compare the predicted survival probabilities from the model against observed retention at time t. Since the predicted survival at the intermediate time interval directly influences the expected time of survival (i.e., time-to-inactivity), we argue that this is a reasonable check for directionality. In this work, using a corpus of historical A/B tests, we compare the predicted week 4 retention against observed WAU retention @ Week 4, since running month-long A/B tests is reasonable.

5.2 Surrogacy Assumption Test

Directionality alone is not sufficient to conclude the predicted metric is a statistically valid proxy for true long-term outcomes. We need to demonstrate that the predicted metric satisfies the surrogacy criteria by showing the effect of treatment on long-term outcomes is completely mediated through surrogates or short-term metrics [3, 35]. While we can't directly test this assumption, we suggest the following sanity check.

To check the surrogacy assumption for survival-based surrogates, we propose to use the likelihood ratio test (LRT). We need the difference between the log-likelihood statistic of two functions: $f(X, S, W)$ and $f(X, S)$; where, W is the treatment status of the user and X & S are the user characteristics and short-term metrics respectively. The LRT test statistic is computed as follows:

$$LRT = 2ln(\mathcal{L}) - 2ln(\mathcal{L}_{\mathcal{W}}) \sim \chi^2_\rho \tag{11}$$

Here, $\mathcal{L}_{\mathcal{W}}$ denotes the log-likelihood of the $f(X, S, W)$ model and $\mathcal{L}$ is the log-likelihood $f(X, S)$ model. ρ is the number of predictor variables being assessed. We note that this is equivalent to using conditional independence tests such as Hilbert-Schmidt Independence Criterion (HSIC) [25]. The key issue is that conditioning on S opens a collider path, creating spurious relationships between Y and W, so this check may produce significant results when the assumptions are not violated. Still, significant results from this test can flag instances where more thought is needed whether on whether surrogacy assumptions are reasonable.

5.3 Metric Sensitivity

We discuss "metric sensitivity", which measures the ability of a metric to detect differences between treatment groups. A high metric sensitivity allows experimenters to detect small changes with fewer users, thereby increasing experimentation efficiency. We measure sensitivity using the decomposition in [7].

For each group $g \in \{0, 1\}$ in an experiment, we observe the metric average $\bar{Y}_g$, the variance σ_g^2 from i.i.d. experimental units of sample size N_g. We perform a Wald test for two-sample difference in means with the test statistic

$$Z := \frac{\Delta}{\sqrt{\sigma^2/N_E}} \tag{12}$$

with difference in means $\Delta = \bar{Y}_0 - \bar{Y}_1$, effective sample size $N_E = 1/(1/N_0 + 1/N_1)$, and pooled variance σ^2 satisfying $\sigma^2/N_E = \sigma_0^2/N_0 + \sigma_1^2/N_1$. We define $\mu := E(\delta) = E(\Delta)/\sigma$ as the scaleless average treatment effect. Hypotheses of the Wald test correspond to $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$, and we reject H_0 if $Z > z_\alpha$, where z_α corresponds to a type 1 error rate α of the experimenters choosing.

Using a Bayesian interpretation, the probability that a treatment has an effect and we can detect it is

$$\underbrace{P(H_1)}_{\textit{discriminative power}} \quad \times \quad \underbrace{P(|Z| > z_\alpha | H_1)}_{\textit{statistical power}} \tag{13}$$

To estimate $p := P(H_1)$, we use an Empirical Bayesian EM procedure introduced in [7]. For H_0, we place a unit information prior on the difference in means, $\mu \sim N(0, 1)$, while for the alternative hypothesis we give a prior of $\mu \sim N(0, V^2)$. Note that as V increases, the scaleless treatment effect is likely to take on larger values so the metric is more likely to return statistically significant result for a given sample size and α. The prior is also symmetric about zero, so that no directional bias is included.

The two key parameters driving metric sensitivity are then p, which measures the discriminative power, and V, which is proportional to the statistical power. Given an experiment corpus, we estimate these parameters as $\hat{p}, \hat{V}$ using the EM procedure given in Algorithm 1 in [7]. A critical feature of the EM procedure is that, unlike the Naive estimate which is biased by the choice of α, the EM estimates depend only on the posterior odds from each experiment and are independent of α.

To avoid high sensitivity resulting from a high false positive rate, we also simulate A/A tests to ensure the distribution of observed p-values is close to uniform. We simulate A/A tests by re-randomizing the treatment assignments on the observed exposure logs from our experiment corpus. This ensures that the targeting and power of the simulated experiments matches the experiments we run in production environments. We note that this does not return the false positive rate of $\hat{Y}$ using Y to label true positives, since Y is unobserved. To validate that notion of false positive would require adjusting the metric variance for model errors in $\hat{Y}$, however since there is no closed-form error term in the Cox model, we suggest this as future work.

6 EXPERIMENTS & RESULTS

In this section, we use Spotify's recommendation applications as a use case to demonstrate the development and validation of our proposed time-to-inactivity metric. The goal for teams working on search and recommendation applications at Spotify is to provide a satisfying experience for users accessing and listening to music and podcasts on the platform. In addition, the product changes and algorithmic improvements are made with the goal of engaging users not only in the short-term but to keep them engaged in the long-term. Traditionally, engagement metrics such as consumption or stream-rate were used for optimization with week t retention reserved for long-running tests. A major challenge with using week t retention is the time it takes to make a deploy/no-deploy decision. We propose to use the estimated time-to-inactivity metric that accounts for long-term goals without compromising for the delay in decision making. We use historical experiments run within Spotify as a case study to illustrate the development of the time-to-inactivity metric. We empirically compare the predicted survival metric to observed retention and other commonly used engagement metrics.

6.1 Experiment Corpus

To validate the time-to-inactivity metric, we sampled a set of 51 A/B tests run on Spotify's recommendation and search products between March and December 2020 (each test involved millions of users). We restricted our sample to those experiments that were run for at least 28 days following a 7 day intake period. The experiments reflect a diverse range of interventions from UI changes to algorithmic improvements in recommendation systems. While we do not segment on the type of product changes, we note that the type of product change can be an explanatory variable in the relationship between two A/B testing metrics and plan to investigate this in the future [28].

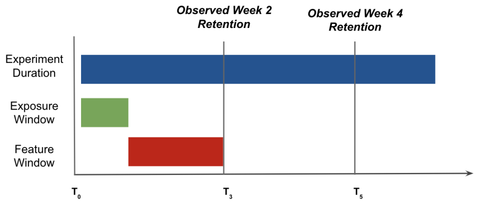

Figure 2: Experiment dataset illustration.

The structure of the experiments in our corpus is illustrated in Figure 2. For each experiment, there is an initial intake period of 7 days and only users exposed during this intake period are considered for analysis. The time window after exposure is used to compute the metric of interest. For instance, we record the observed WAU[4] retention at Week 2 using the data from 14 days since exposure. Similarly, engagement metrics such as consumption, stream-rate are calculated using the data from the duration of the experiment excluding the intake period.

This dataset allows us to compare our predicted time-to-inactivity metric against observed metrics. We expect a good predicted time-to-inactivity metric to directionally align with observed retention and at the same time be sensitive to changes. We discuss in more detail, the accuracy and sensitivity results of time-to-inactivity below.

6.2 Predicted Time-to-Inactivity Metric

We describe our setup to develop the predicted time-to-inactivity metric, a long-term engagement metric, for Spotify's use case. Recall that our goal here is to estimate the average treatment effect of A/B tests without having to wait until the user churns (i.e., become inactive); therefore, we rely on the predicted time-to-inactivity metric for decision-making.

We leveraged various user characteristics, denoted X, such as their historic usage on the platform, device type used, etc., and short-term engagement metrics, denoted S, such as consumption. Short-term metrics observed over the first 14 days since exposure were used to generate the predicted weekly retention metrics over the next 24 weeks. We use Cox Proportional Hazards described in Section 4 to fit a $f_H(X, S)$ on the historical dataset D_H for each A/B test in the experiment corpus. The trained models are then used to obtain the estimated long-term metric $\hat{Y}$ for each user in the experiment, which is used to compute the ATE.

6.3 Model Validation Results

The first step towards building a reliable time-to-inactivity metric is to check the goodness of fit of $f_H(X, S)$. We report the out-of-time and out-of-sample model performance measured using the metrics

[4] WAU is defined by a user streaming any content during the last 7 days.

Metric	Out-of-Time	Out-of-Sample
CI	0.7844 (0.0045)	0.7898 (0.0044)
Integrated Brier Score	0.1591 (0.0047)	0.1633 (0.0010)

Table 1: Concordance index (CI) and integrated brier scores computed on out of time and sample sets averaged over different models trained for each experiment in the corpus.

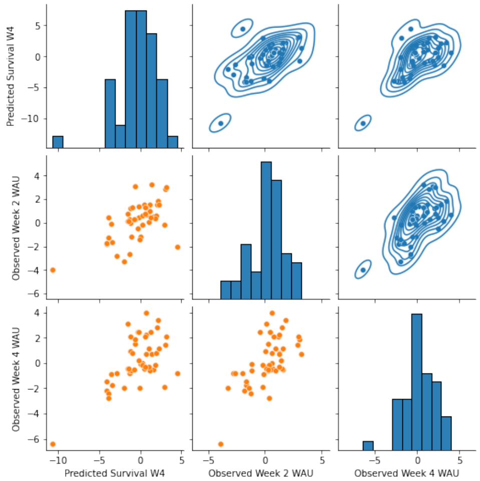

Figure 3: Comparison of expected survival vs observed retention z-scores.

introduced in Section 4.2. We observed high AUC scores of 0.9 and above for earlier weeks and competitive AUCs of 0.83 even for a very longer horizon of 24 weeks, demonstrating the ability of the model to recover the true retention probabilities (see Appendix A.1 for AUC plots). Table 1 again reports high numbers for concordant index and integrated Brier Score defined in Section 4.2, suggesting that the empirical function $\hat{f}_H(X_i, S_i)$ is converging to the true function f to reliably estimates the true long-term engagement.

To test the robustness of the model we measure the out-of-time performance and report CI and Brier scores. Table 1 compares the out-of-time and sample performance of the model and it can be seen that the scores are almost effectively equivalent, suggesting that the f_H is not impacted by time of data collection.

6.4 Metric Validation: Directionality

To understand metric directionality and consistency, we investigate its relationship to other metrics that are more directly interpretable. In Figure 3, we plot the distribution and pairwise relationships of the time-to-inactivity and the observed WAU retention metrics.

Given the approximate normality of each of the three metrics, we apply a two-tailed Pearson non-correlation test between the predicted Week 4 survival and the observed Week 2 WAU metric ($\hat{\rho} = 0.60$, $p = 4.6 \cdot 10^{-6}$) and Week 4 WAU metric ($\hat{\rho} = 0.61$, $p = 2.9 \cdot 10^{-6}$)[5] and reject the null hypothesis that the metrics are uncorrelated. For comparison, the correlation between the two observed retention metrics is observed as ($\hat{\rho} = 0.64, p = 6.4 \cdot 10^{-7}$).

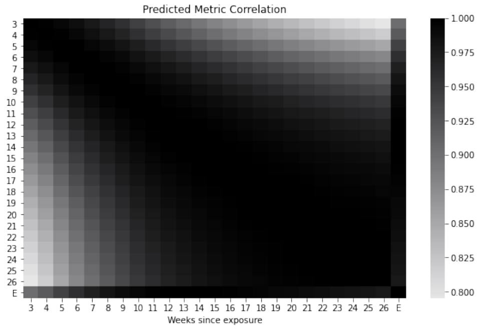

Figure 4: Correlation between predicted survival metrics at given week. E indicates the expected time-to-inactivity.

Metric	Correlation to W4 WAU (95% CI)	p-val
$\hat{Y}_i^4$	0.61 (0.40, 0.76)	$2.858 \cdot 10^{-6}$
W2 Consumption	0.63 (0.43, 0.77)	$8.353 \cdot 10^{-7}$
W2 Rec. consumption	0.42 (0.16, 0.62)	0.002

Table 2: Pearson Correlation to Week 4 WAU retention of observed consumption and predicted survival.

We also check the ATE correlation structure within the predicted metrics to ensure that metric movements are internally consistent. In Figure 4 we show the correlation structure of the 24 weekly $\hat{Y}^t$ metric ATEs, and the expected time-to-inactivity ATE. We see the expected structure, that metrics closer together in their predicted time horizon have a high correlation, while metrics that are far apart in time are less correlated. Overall, the correlation is high with no pairwise correlation less than 0.8. We also note that the expected time-to-event correlation is highest with the longer-term metrics.

A third check is between the predicted metric, observed retention, and engagement metrics together. In Table 2 we show ATE correlations of observed Week 4 WAU retention to $\hat{Y}_i^4$, consumption in the second week since exposure, and recommended content consumption in the second week since exposure. This shows that there can be directional alignment between the underlying engagement metrics the surrogate is constructed with, and also that there is directional alignment of the predicted outcomes.

6.5 Metric Validation: Surrogacy Check

Next, we check if the surrogacy assumption is reasonable for the predicted metric. To do this, we apply the loglikelihood ratio test (LRT) check described in Section 5.2.

We fit two functions $f(X, S, W)$ and $f(X, S)$ on each of the experiments in the corpus described in Section 6.1. We then compare the two functions using LRT to demonstrate the conditional independence on treatment W. Each treatment arm adds one degree of freedom, so in our setting $d.f. = 1$. We observed an average

[5] We do not apply a multiple testing correction to the p-values, but note that both tests would yield significance under a Bonferroni correction at $\alpha = 0.05$.

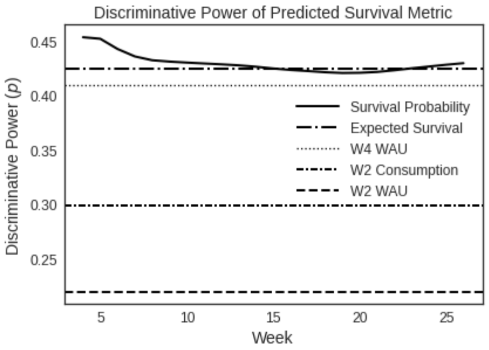

Figure 5: Discriminative power $\hat{p}$ of predicted survival compared to observed metrics engagement and WAU retention metrics, estimated over 51 online experiments

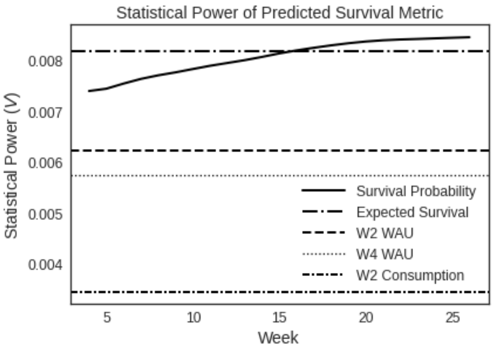

Figure 6: Statistical power $\hat{V}$ of predicted survival compared to observed metrics engagement and WAU retention metrics, estimated over 51 online experiments

test statistic of -6.085 ($p > 0.05$). This suggests there may not be chronic violations of the surrogacy assumption.

6.6 Metric Sensitivity

Finally, we use the sensitivity decomposition described in Section 5.3 to show that the predicted time-to-inactivity enables decisions on A/B tests. To investigate this, we compute the prior probability of movement $\hat{p}$ (discriminative power) and the prior variance of the scaleless treatment effect under the alternative hypothesis $\hat{V}$ on the predicted time-to-inactivity metrics, the observed week 2 WAU, and observed week 4 WAU. The results are shown in Figures 5 and 6, respectively.

We see that the discriminative power for Week 4 WAU is about twice that of Week 2 WAU, and that the predicted time-to-inactivity metrics have comparable discriminative power to the Week 4 WAU metric. We observe the weekly predicted metrics have higher discriminative power using predictions with short time horizons, while longer-term predictions approach that of the expected time-to-inactivity metric. The statistical power parameter estimate $\hat{V}$ also shows that there is similar statistical power between Week 4 WAU and Week 2 WAU. We also observe that the predicted metrics have higher statistical power than either observed metric.

We note that $\hat{V}$ tends to increase with the prediction horizon, while $\hat{p}$ tends to decrease. We investigated both the distribution of the relative differences δ and pooled variances from each metric, and observed that the relative differences in each metric tend to increase with the prediction horizon, which corresponds to a higher $\hat{V}$. Similarly, there is an increase in the pooled variances σ over the prediction horizons, which depresses the discriminative power estimate $\hat{p}$. One consideration is that the Cox model generates diverging survival curves up to a time horizon specific to the data set, so the increase in metric variance could be a result of the choice of survival model and time horizons associated with this analysis. The upward trend in statistical power would not continue indefinitely, as we would see a decrease in statistical power at long time horizons as the predicted survival probabilities approach zero.

Given that both $\hat{p}$ and $\hat{V}$ are measured higher for the predicted metrics than for the observed metrics, this suggests that the set of predicted metrics are more sensitive than the observed metrics. The A/A test results suggest this result is trustworthy since there is not an increased false positive rate.

To ensure that the metric is not exhibiting artificially high sensitivity through a high false-positive rate, we conducted A/A tests by re-randomizing the treatment assignments on the observed exposure logs from our experiment corpus. We looked at the distribution of 432 p-values for each of the predicted metrics resulting from A/A test simulations and as expected, found the distributions to be uniform, suggesting the high sensitivity of time-to-inactivity reported earlier is valid (see Appendix A.2 for details). This is only a validation of the type 1 error rate of the surrogate metric and not the target metric [17].

7 CONCLUSION AND FUTURE WORK

In this work, we introduced a long-term engagement metric – time-to-inactivity – and described the use of survival modeling to estimate the metric using short-term metrics. Then, we presented a set of validation checks necessary to reliably use the metric for experimentation. We first discussed a set of survival metrics to validate the model and then presented a set of checks to verify the directionality and sensitivity of the metrics. In addition, we presented checks to verify the statistical surrogacy assumption when the predicted long-term metric is modeled using a survival function. Finally, we demonstrated the development of the time-to-inactivity metric for a real-world application using experiments at Spotify. We used historical online experiments and showed that the validity checks hold for our proposed metric. The results show the survival metrics improved sensitivity over baseline while preserving directionality.

In summary, using a large-scale empirical study, we presented practical guidelines and validations needed to use the previously proposed survival analysis based engagement models as metrics in A/B tests. By leveraging immediate short-term engagement to predict longer-term churn, our work will allow online service companies to make more rapid decisions that are more aligned with north star company goals. We plan to investigate the impact of different survival models on metric directionality and sensitivity in the future.

REFERENCES

[1] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang. 2020. A Survey on Churn Analysis in Various Business Domains. *IEEE Access* 8 (2020), 220816–220839.

[2] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. 2005. A time-dependent discrimination index for survival data. *Statistics in Medicine* 24, 24 (2005), 3927–3944.

[3] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. *The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely.* Technical Report 26463. National Bureau of Economic Research.

[4] Albert C. Chen and Xin Fu. 2017. Data + Intuition: A Hybrid Approach to Developing Product North Star Metrics. In *Proc. of WWW.* 617–625.

[5] D. R. Cox. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 2 (1972), 187–202.

[6] D. R. Cox. 1975. Partial likelihood. *Biometrika* 62, 2 (1975), 269–276.

[7] Alex Deng. 2015. Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments. In *Proc. of WWW.* 923–928.

[8] Alex Deng and Xiaolin Shi. 2016. Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proc. of KDD.* 77–86.

[9] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proc. of WSDM.* 123–132.

[10] Drew Dimmery, Eytan Bakshy, and Jasjeet Sekhon. 2019. Shrinkage estimators in online experiments. In *Proc. of KDD.* 2914–2922.

[11] Pavel Dmitriev, Brian Frasca, Somit Gupta, Ron Kohavi, and Garnet Vaz. 2016. Pitfalls of long-term online controlled experiments. In *2016 IEEE international conference on big data (big data).* 1367–1376.

[12] Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 2017. A dirty dozen: twelve common metric interpretation pitfalls in online controlled experiments. In *Proc, of KDD.* 1427–1436.

[13] Pavel Dmitriev and Xian Wu. 2016. Measuring metrics. In *Proc. of CIKM.* 429–437.

[14] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proc. of WWW.* 256–266.

[15] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2017. Using the Delay in a Treatment Effect to Improve Sensitivity and Preserve Directionality of Engagement Metrics in A/B Experiments. In *Proc. of WWW.* 1301–1310.

[16] Alexey Drutsa, Anna Ufliand, and Gleb Gusev. 2015. Practical aspects of sensitivity in online experimentation with user engagement metrics. In *Proc. of CIKM.* 763–772.

[17] Weitao Duan, Shan Ba, and Chunzhe Zhang. 2021. Online Experimentation with Surrogate Metrics: Guidelines and a Case Study *(WSDM '21).* 193–201.

[18] Georges Dupret and Mounia Lalmas. 2013. Absence time and user engagement: evaluating ranking functions. In *Proceedings of the sixth ACM international conference on Web search and data mining.* 173–182.

[19] Cameron Davidson-Pilon et al. 2021. *CamDavidsonPilon/lifelines: v0.25.8.*

[20] Aleksander Fabijan, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. 2019. Diagnosing sample ratio mismatch in online controlled experiments: a taxonomy and rules of thumb for practitioners. In *Proc. of KDD.* 2156–2164.

[21] Peter S Fader and Bruce GS Hardie. 2007. How to project customer retention. *Journal of Interactive Marketing* 21, 1 (2007), 76–90.

[22] T. Gerds, M. Kattan, M. Schumacher, and C. Yu. 2013. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine* 32 13 (2013), 2173–84.

[23] Nicolas Glady, Bart Baesens, and Christophe Croux. 2009. Modeling churn using customer lifetime value. *European Journal of Operational Research* 197, 1 (2009), 402–411.

[24] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* 18, 17-18 (1999), 2529–2545.

[25] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. [n.d.]. A kernel statistical test of independence.

[26] Sunil Gupta, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V Kumar, Nathaniel Lin, Nalini Ravishanker, and S Sriram. 2006. Modeling customer lifetime value. *Journal of service research* 9, 2 (2006), 139–155.

[27] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35.

[28] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. 2015. Focusing on the long-term: It's good for users and business. In *Proc. of KDD.* 1849–1858.

[29] Liangjie Hong and Mounia Lalmas. 2020. Tutorial on Online User Engagement: Metrics and Optimization. In *Proc. of KDD (KDD '20).* 3551–3552. https://doi.org/10.1145/3394486.3406472

[30] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. 2014. A hazard based approach to user return time prediction. In *Proc. of KDD.* 1719–1728.

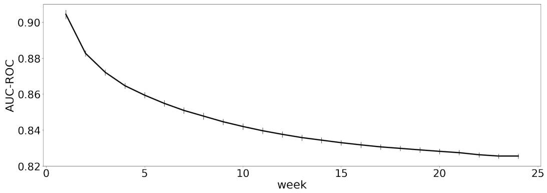

Figure 7: Out-of-Sample AUC plotted averaged across 38 different models and plotted across 24 weeks.

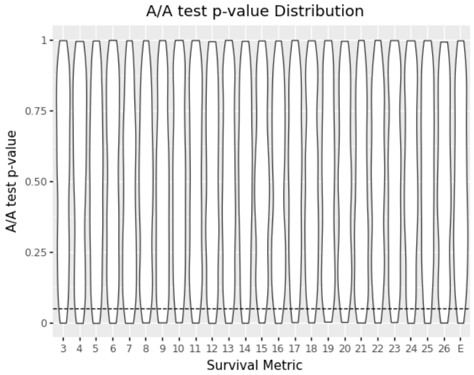

Figure 8: Violin distribution plots of p-values observed for each survival metric across 432 A/A test simulations.

[31] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning sensitive combinations of A/B test metrics. In *Proc. of WSDM.* 651–659.

[32] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proc. of KDD.* 786–794.

[33] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven rules of thumb for web site experimenters. In *Proc. of KDD.* 1857–1866.

[34] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled Experiments on the Web: Survey and Practical Guide. *Data Min. Knowl. Discov.* 18, 1 (2009), 140–181.

[35] Ross L Prentice. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine* 8, 4 (1989), 431–440.

[36] Markus Viljanen, Antti Airola, J. Heikkonen, and T. Pahikkala. 2017. A/B-Test of Retention and Monetization Using the Cox Model. In *AIIDE.*

[37] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proc. of KDD.* 2227–2236.

[38] Ya Xu, Weitao Duan, and Shaochen Huang. 2018. SQR: balancing speed, quality and risk in online experiments. In *Proc. of KDD.* 895–904.

[39] Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. 2020. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835* (2020).

A ADDITIONAL RESULTS

A.1 Model Performance: AUC Scores

To validate the performance of the survival model for our given task, the problem can be viewed as a binary classification task where the goal is to identify high-risk users. Here, given the estimated probability of the inactivity at time t indicates the probability to churn. We use the ROC curve that compares the false positive rate against the true positive rate, and the area under the ROC curve ($AUC - ROC$) is plotted for each time point to validate the model performance.Figure 7 shows the out-of-sample AUC-ROC computed for 24 weeks. The high AUC scores for earlier weeks demonstrate the ability of the model to recover the true retention probabilities, with competitive AUCs of 0.83 even for a long horizon of 24 weeks.

A.2 Sensitivty Analysis: A/A Tests

The goal with A/A tests is to show that the distribution of p-values is uniformly distributed over the interval $[0, 1]$. This would imply that at any value of α, the type-1 error rate is α. As pointed out in [17], surrogate metrics have an inflated type-1 error rate, using movements to the target metric Y as the ground truth labels. Because we are using censored survival outcomes, we can not validate against movements in Y as the ground truth. However we can validate that the surrogate metric itself does not have have other sources inflating the type 1 error rate, such as violating the independence assumption between observations. We validate that this distribution is roughly uniform in Figure 8. We note that simulating additional A/A tests is relatively easy, and that the type-1 error rates at specific α values can also be estimated from the proportion of positive tests at that α from this type of data.