USING RANDOM FORESTS AND COX PROPORTIONAL HAZARDS MODEL TO UNDERSTAND AND PREDICT SUBSCRIBER LONGEVITY IN A SUBSCRIPTION-BASED ONLINE WEB SERVICE

CHRISTOPHER PETERS
Department of Experimental Statistics, Louisiana State University
cpeter9@gmail.com

ABSTRACT

User engagement and is an important goal for subscription-based online webserivce companies. In particular, segmenting subscribers based on early user engagement can help software-as-a-service ("SaaS") companies affect subscriber longevity. The purpose of this study is to assess the degree to which various measures of early user engagement and overall user engagement affect subscriber survival (or longevity). Cox Proportional Hazards and random forest regression models are used to test these relationships in a real dataset of the Treehouse Inc. SaaS company. Finally the results of these approaches are compared using prediction error curves as well as the standard Kaplan-Meier estimator applied to left out data. Results show that early user activity assessed within the first, third, seventh, fourteenth and twenty-first days is not a good preditor for survival. However, models that include other data collected over the full lifetimes of subscribers do have good predictive ability. Finally, it is shown that the relative difference in error rate between Cox Proportional Hazards models and Random Forest models is negligble for early user engagement metrics, but substantial when including a more complete set of covariates.

Key words

Cox Proportional Hazards, Random Forests, User Engagement, SaaS.

1. Introduction

Treehouse (*teamtreehouse.com*) is an online subscription-based web-service company offering programming lessons to users in fields such as web design, web development, and iOS and Android app programming. Treehouse's business model relies on signing up subscribers billed on a monthly basis that in turn receive access to online lessions in the previously mentioned topics. Treehouse is a type of business commonly referred to software-as-a-service ("SaaS"). Other businesses in this category include Netflix, Apple's iCloud, Google Apps, and Salesforce. These

businesses operate with subscription bases and are therefore confronted with the problem of churn. Churn is the monthly fluctuations in subscriptions signups and subscription cancellations. SASS companies attempt to maximize signups and minimize cancellations. Therefore it is very important to SASS companies find ways to lengthen the subscriber lifecylce. User engagement is an important part of this process. Previous studies have shown that early user engagement can have a significant effect on customer logevity. (Cite collected studies).

Treehouse teaches its subscribers to program in various languages for use in fields such as web design, web development and iOS and Android app development. Subscribers generally take a linear path through each course and complete sesquential stages. Each stage contains video tutorials, quizzes, and in-browser code challenges. Upon completion of each stage (each course usually has several stages), the subscriber receives a "badge". Badges are tokens received by subscribers that indicate their completion of a stage; badges are also awarded for completion of courses. This study focuses on badge earning and the viewing of video tutorials as a measure of user engagement. Various early time periods (1, 3, 7, 14, and 21 days, and all-time) are assessed for predictive ability.

The purpose of this study is to first assess if any early user engagement metrics have significant predictive ability, and second to assess if any metrics measured over the whole subscriber lifetime can further improve predictions.

The most common and traditional approaches to survival regression problems are Cox Proportional Hazard models and Scale-Accelerated Failure-Time models. These models both have drawbacks that must be dealt with. Cox Proportional Hazard models require covariates that are not time dependent. Scale-Accelerated Failure-Time models require distributional assumptions about the data. Newer techniques such as Random Survival Forests have recently become popular with the publication of R packages. This study applies Cox Proportional Hazard models and two separate Random Forest techniques to assess the predictive quality of early user engagement metrics as well as those metrics collected over the full subscriber lifetime.[1]

1.1 The Data

Treehouse maintains a database that records many variables associated with each subscriber. This study uses a snapshot of the customer base from May 12, 2012. Since some subscriptions at Treehouse can have multiple users, this study focuses on only those subscriptions with a single user. All subscribers ($n = 12{,}004$) were active subscribers on the site as of May 12, 2012 or have since joined Treehouse.

[1] The performance of Scale-Accelerated Failure-Time models are left as an extention of this research as methods for this type of model do not exist in the "pec" R package, a sophisiticated package to compare model prediction error curves. Instead we focus on assessing the predictive ability of Cox Proportional Hazard models and Random Forests.

By October 18, 2012, the composition of the subscriber base was 9,847 active while 2,157 had cancelled.

1.1.1 Right-censoring: Type 1

Many lifetime data include censored observations. Censored observations in the context of this study are subscribers that had not cancelled their membership as of October 12th. Therefore, no subscriber that cancelled during the time period of May 12th to October 12th is considered censored. Therefore censoring can only occur after 153 days have passed, but uncensored cancellations are randomly distributed from day 1 to day 363, the age of the oldest Treehouse subscriber included in the study.

Type 1 censoring is technically written as,

$$t_i = min(T_i, C_i)$$
$$\delta_i = I(T_i \leq C_i)$$

where t_i represents time-to-cancellation for each subscriber, and C_i represents the time of censoring for each subscriber.

The joint probability distribution of the time-to-cancellation and censoring mechanism here is as follows:

$$f(t_i)^{\delta_i} Pr(T_i > C_i)^{1-\delta_i}$$

and the likelihood function is:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

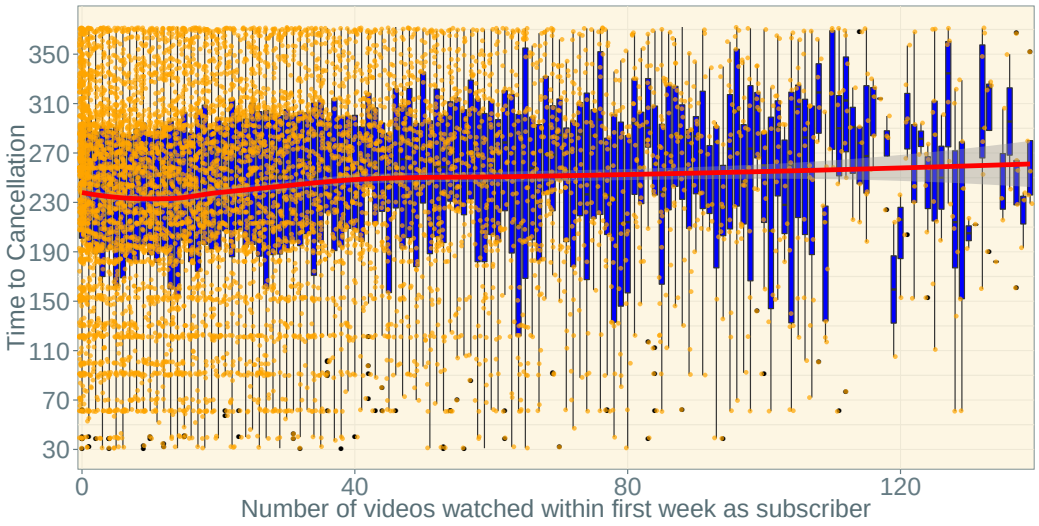
1.1.2 Covariates

A special interest of this study is the effect to which, if there is an effect, that covariates such as the number of badges earned or videos watched by each user. In particular, the hope of this study is to find covariates as measured during the first week of subscribership that influence survival.

The number of badges earned within the first seven days were counted for each user using the Treehouse database. The minimum number of badges earned was one as subscribers earn a badge by simply signing up for the service. Twenty-five percent of users had earned 2 or less badges within their first seven days a member. The median number of badges earned within the first seven days was 5, with the mean being 8.19 indicating that some users earned a large number of badges within the first two weeks. The third-quantile stands at 11 badges earned and the most badges earned stands at 82. #Get total number of badges possible to be earned and insert here.#

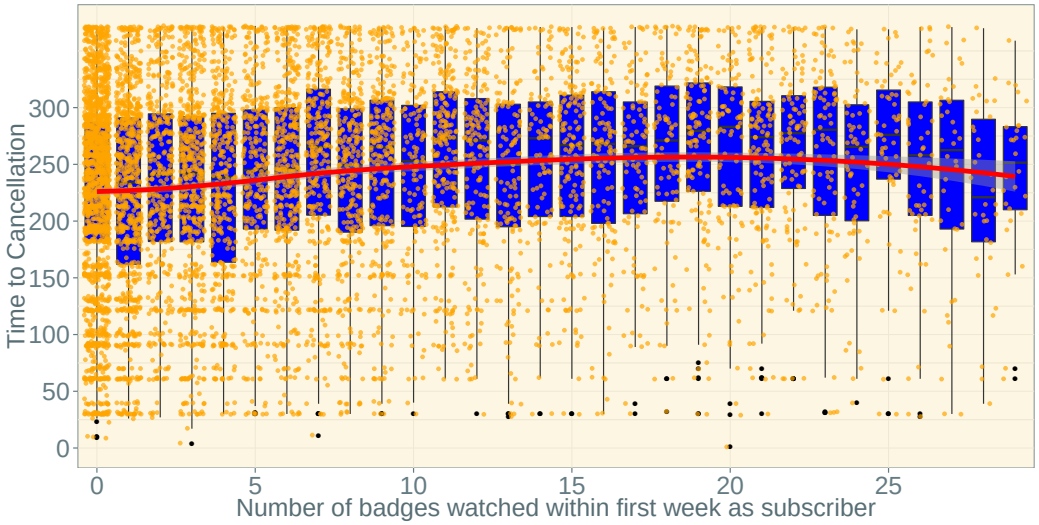
For example, the plots below shows customer longevity (time to cancellation) against the number of badges earned and videos watched by a user during the first seven days, respectively. The blue boxes are traditional boxplots with the lower edge of each blue box represending the 0.25 quantile and the top edge representing the 0.75 quantile. The red line tranversing the boxes is a locally weighted smoother also known as a LOESS (or LOWESS) model. This shows us that the mean of time-to-cancellation dips slightly between 10 to 20 videos watched and increases slightly through 120+ videos watched. The orange dots represent the actual data points themselves.

Figure 1. Customer longevity vs. number of videos watched in the first seven days as a subscriber



A similar figure, but with respect to badges, shows a slight increase in time-to-cancellation from 1 through about 20 badges earned in the first week before declining. It may be possible that the slight decline after 20 badges is because it's possible for heavy-using subscribers to exhaust Treehouse content - or at least the content they are interested in.

Figure 2. Customer longevity vs. number of badges earned in the first seven days as a subscriber



The practical advantage of being able to covariates measured within the first week of subscription to survival is that it would allow for the potential of Treehouse to intervene early on in the customer lifecylce. If a strong relationship is shown, it opens the potential that Treehouse could segment subscribers based on these covariates to attempt to influence the survival curves of subscribers.

Figure 3 on page 6 shows time-to-cancellation against the total number of badges that each user earned. The upward slope is more pronounced. While finding covariates restricted to early user engagement is an important part of this investigation, measuring the total badges a user earned is a useful covariate if it can better explain the survival of subscribers, especially for predictions.

Figure 3. Customer longevity vs. total badges earned in the first seven days as a subscriber

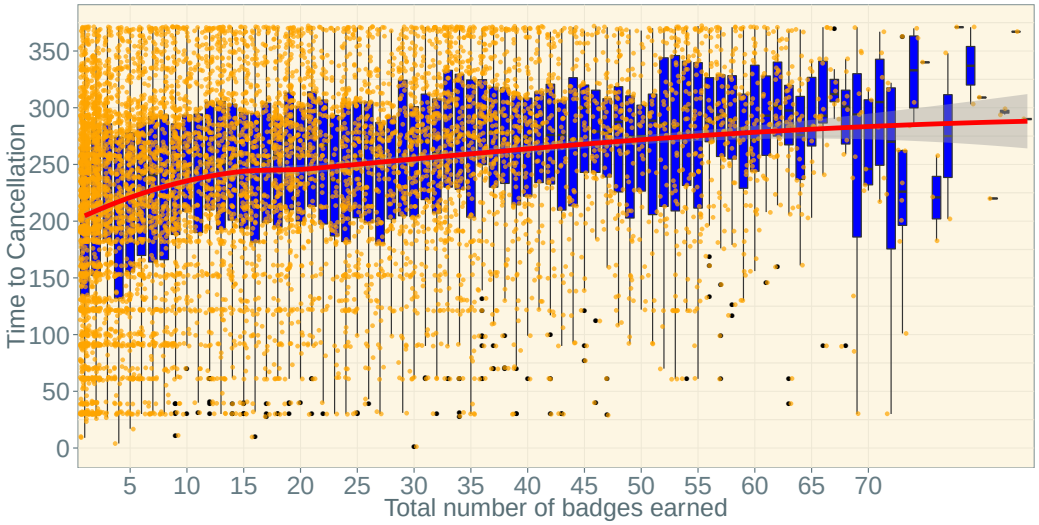
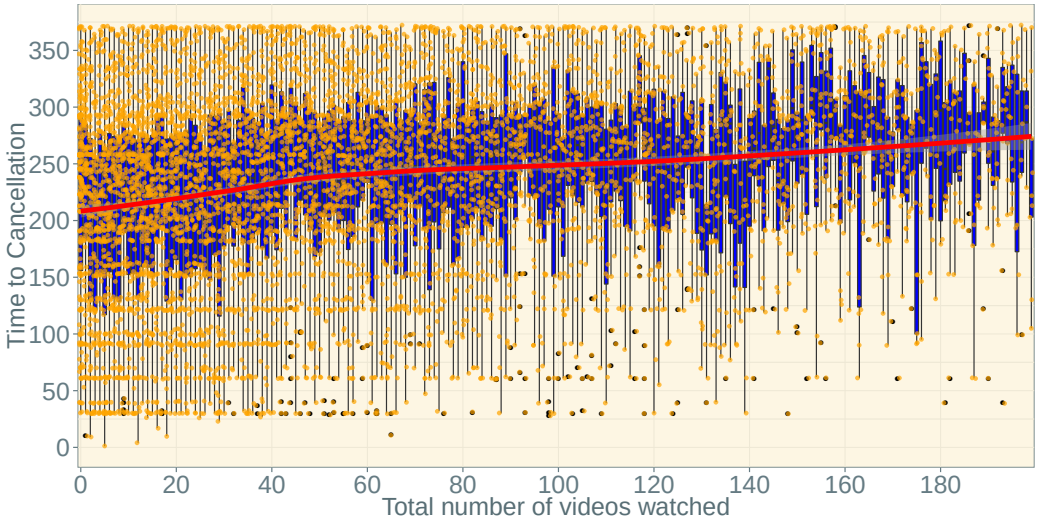


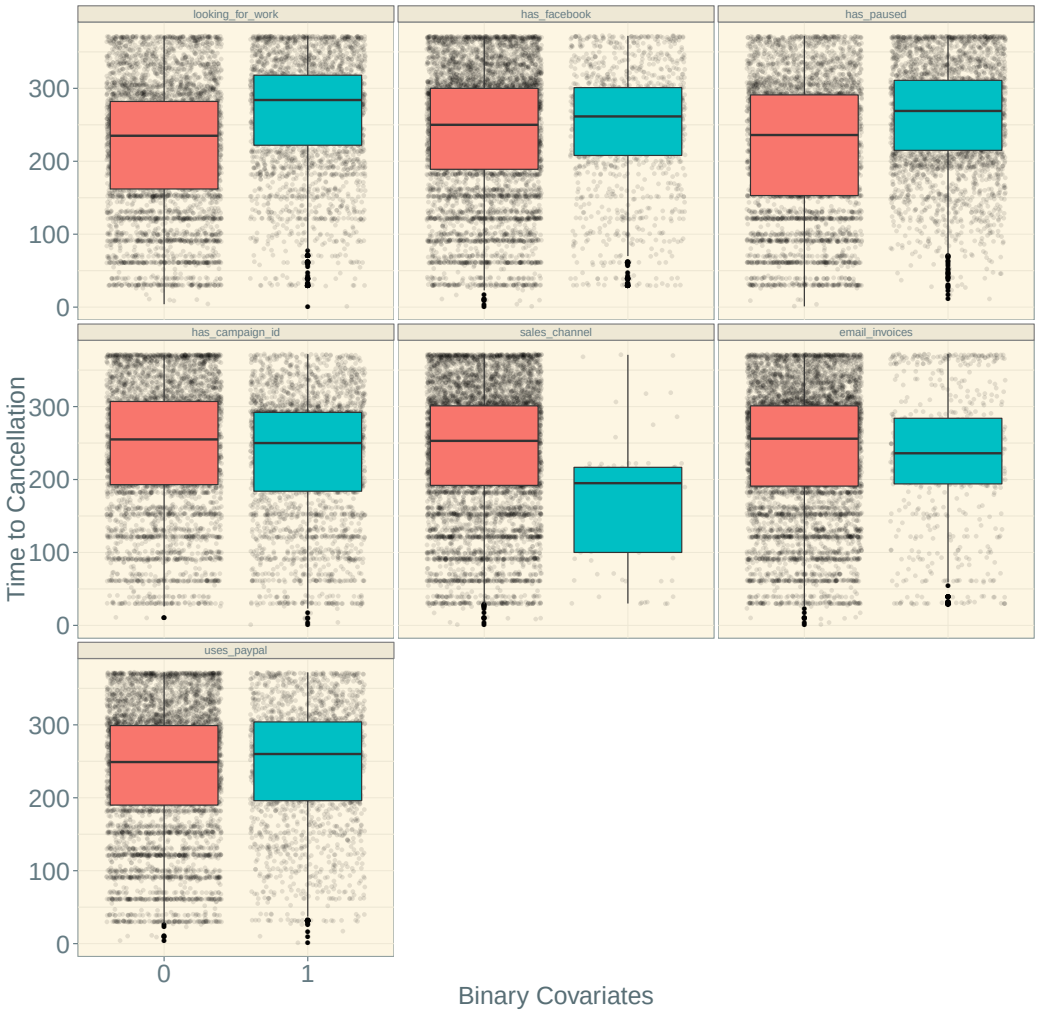
Figure 4 shows a more consistent increase in time-to-cancellation as subscribers watch more video tutorials.

Figure 4. Customer longevity vs. total videos watched in the first seven days as a subscriber



In addition to discrete covariates taking on many values, a series of binary covariates were examined for their usefulness in terms of predicting subscriber longevity. Figure 5 shows this set of covariates that includes whether a subscriber was looking for work, had provided Treehouse with their Facebook ID, had ever paused their account, signed up by way of a marketing channel or sales channel, whether the subscriber receives email invoices and finally whether the subscriber pays by Paypal or some other method (credit card or Braintree).

Figure 5. Customer longevity vs. Binary Covariates



Simple visual comparisons of binary covariates against time-to-cancellation do not yield any major indication of a link betwen time-to-cancellation and said binary covariates. The best apparent relationship is whether a user has selected on the

site whether they are looking for work. Also, a large difference appears for those subscribers that have entered through the direct sales channel, however due to the small amount of data available in this category it is hard to give this result much creedence.

2. Methods

A primary interest in this study is to assess the usefulness of badges earned by user within the first seven days as a subscriber as a predictor for the hazard rate and survival curve of Treehouse subscribers. For this purpose, a set of Cox Proportional Hazards models were used as well as Random Forests.

2.1 Cox Proportional Hazards

The Cox proportional hazard (PH) model is a regression model that relates covariates to the hazard function of lifetime data. In turn this allows a relationship of covariates to the corresponding survival curve to be derived. The basis for this method arrives in the following formulation of the hazard function:

$$h(t|\mathbf{x}) = h_0(t)r(\mathbf{x}) \tag{1}$$

where r($\mathbf{x}$) and $\mathrm{h}_0(t)$ take only positive values.

In this formulation, $\mathrm{h}_0(t)$ is commonly called the baseline hazard function. The term gives the hazard function for each individual its form or shape, while r($\mathbf{x}$) scales the baseline hazard function depending on values of the covariates.

Stepwise regression using the Akaike Information Criterion (AIC) as a criterion for variable selection was used to select the best parsimoneous Cox Proportional Hazard model.

2.2 Random Survival Forest Models

The downside of using a Cox Proportional Hazard model, as previously mentioned, is due to the assumption that estimated hazards must be proportional. That is, they must be related by a multiple at all failure times with respect to the value of covariates. Random forest models, however, do not involve any similar assumption, and therefore allow their users to include variables that would otherwise violate the proportional hazard assumption. Introduced in 2008 in *The Annals of Applied Statistics*, Random Survival Forests are an extention of the popular Random Forest type model created by Leo Breiman (cite).

It is helpful to think of Random Forests as an extention of the popular Decision Tree modeling method (the phrase CART may be familiar). As in decision trees, the dependent variable, in this case survival times, are partitioned into two subsets such that the values of the dependent variable are as different as possible with respect to each explanatory variable. More specifically, the criteria used to split the dependent variable values are commonly Chi-square values, worth, logworth, entropy, and Gini puritiy index values, etc. Once the dataset is partitioned, the same algorithm is applied recursively with all explanatory variables. Decision criteria determine when the process stops, traditionally the minimum number of observations left in a "node", but there are other criteria as well. Random Forests are an extention of this concept by which Decision Trees are "grown" on subsets of a bootstrapped dataset. First, a bootstrap sample of size N is selected from the training data. Next a decision tree is grown until some criteria is met such as a floor on observations in each nodes (these nodes are called terminal nodes). Unlike analysis done with a decision tree, the individual trees of a random forest are not pruned. Finally a prediction for given input data is made by:

$$\hat{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{2}$$

which is simply the average of predictions given by all of the sub-decision trees given a certain set of inputs.

2.2.1 randomSurvivalForest

The first of two random survival forests used for this analysis comes from the R package `randomSurvivalForest`. To the knowledge of this author,the `randomSurvivalForest` package is the only statistical software written to apply Random Forests to survival problem. The default selection, which was used for this analysis, excludes 37% of the data specifically from each bootstrap. Further, nodes are split by variables that maximize the survival difference between daughter nodes using logrank. Finally, each tree terminantes similar to traditional random forests, when each node has no less than three unique deaths. Unlike the Random Forest methodology, the Random Surival Forest method calculates a cumulative hazard function at the end of each terminal node. The authors of the `randomSurvivalForest` package put forth the cumulative hazard function as a new survival data object (to be added to the set of survival, hazard function, probability density function and cumulative density function). Constructing an ensemble cumulative hazard function is the primary way that the `randomSurvivalForest` package creates survival predictions. The CHF is calculated using the Nelson-Aalen estimator:

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \tag{3}$$

where $d_{l,h}$ is the number of deaths at time $t_{l,h}$ and $Y_{l,h}$ is the number of individuals at risk at time $t_{l,h}$. All observations within h have the same CHF.

The ensemble CHF is calculated as follows:

$$H_e(t|x_i) = \frac{1}{B} \sum_{b=1}^{B} H_b(t|x_i) \tag{4}$$

Ultimately predictions of survival are obtained by way of:

$$\hat{S}^{rsf}(t|x) = \exp\left(-\frac{1}{B} \sum_{b=1}^{B} H_b(t|x_i)\right) \tag{5}$$

2.2.2 cforest

The second of two random survival forests, comes from the `party` package in R and is called a conditional inference forest. The specific function is called `cforest`. The difference is that the ensemble methodology is as follows:

$$\hat{S}^{cforest}(t|x) = \prod_{s \leq t} \left(1 - \frac{\sum_{b=1}^{B} d_{l,h}}{\sum_{b=1}^{B} Y_{l,h}}\right) \tag{6}$$

Despite the difference between the two random survival forest methodologies, that which preceeds the ensemble method is the same.

Prediction error calculations referrenced later will use an OOB ensemble CHF estimate in conjunction with the in-sample ensemble CHF to determine model fit.

The OOB ensemble CHF is calculated as follows:

$$H_e(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b(t|x_i)}{\sum_{b=1}^{B} I_{i,b}} \tag{7}$$

where $I_{i,b}$ is an indicator variable that takes the value of 1 for each observation within a respective individual bootstrap that is considered out-of-bag (OOB). OOB are observations that fall within the 37% group of observations excluded from each bootstrap. Otherwise $I_{i,b}$ takes a value of zero. $H_b(t|x_i)$ is the CHF for each respective observation within each bootstrap.

3. Results

3.1 Cox Proportional Hazard Model

As Lawless (2003) states, "The name proportional hazards (PH) comes from the fact that any two individuals have hazard functions that are constant multiples of one another." This introduces an important aspect of modeling with PH models, the assumption that the resulting hazards are proportional. However, random forests require no such assumption. Therefore, this study first investigates the best Cox Proportional Hazard model as determined by backward stepwise selection using the Akaike Information Criterion (AIC) as a decision rule. The residuals of the selected Cox Proportional Hazard model were examined for time-dependence (and hence violation of the proportional hazards assumption).

An important assumption of Cox Proportional Hazard models is that covariates do not affect estimated hazard functions in a way the depends upon time. The table below is a test developed by Grambsch and Therneau (1994) using Schoenfeld residuals and weighted regression to develop a Chi-square test statistic for to test for violations of the proportional hazards assumption.

Stepwise selection via the Cox PH model yields the following model:

A number of the selected variables violate the proportional hazards condition. As a result, two models were tested for predictive performance against the two random forest methods. The first model is the "full" stepwise-selected model despite its violation of the proportional hazards assumption. The purpose will be to assess the effect of neglecting this key assumption on the predictive ability of the model. The second model was obtained by removing variables that fail the #test name#, one-by-one based on smallest p-value. A final model that contains no variables for which the test rejects the null hypothesis of a lack of contributing to non-proportional hazards.

The following table shows test results for the final model:

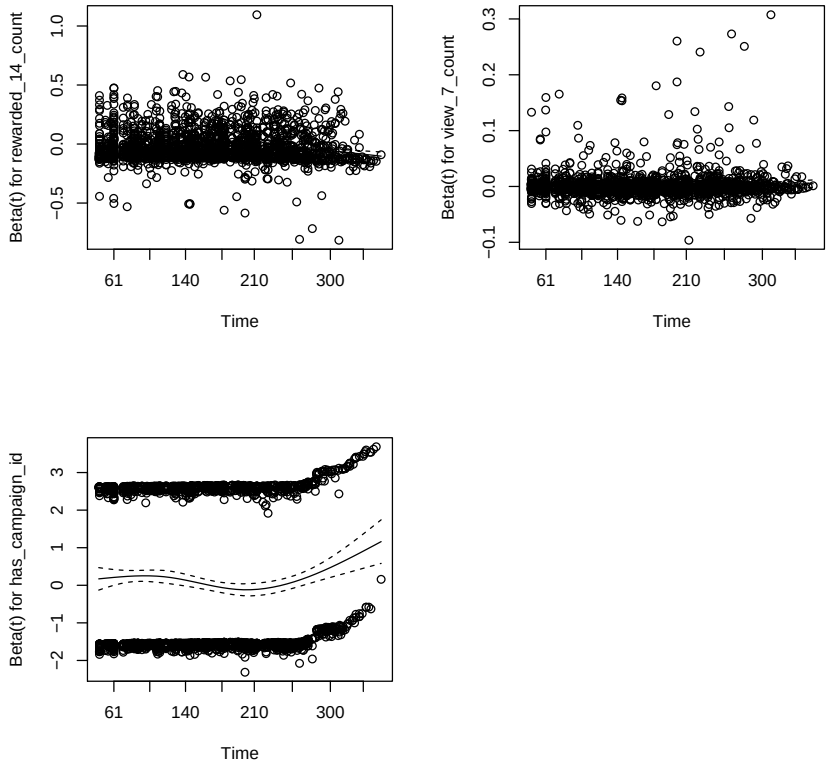
Selected model	rho	chisq	p
rewarded_1_count	0.01	0.35	0.56
rewarded_3_count	0.02	1.04	0.31
rewarded_7_count	-0.01	0.29	0.59
rewarded_14_count	0.01	0.23	0.63
rewarded_21_count	-0.08	10.85	0.00
person_badges_count	0.07	14.65	0.00
view_1_count	-0.02	0.95	0.33
view_3_count	0.02	0.54	0.46
view_7_count	0.01	0.21	0.64
view_14_count	-0.01	0.13	0.71
view_21_count	-0.05	5.10	0.02
total_view_count	0.11	47.97	0.00
looking_for_work1	0.12	26.48	0.00
has_campaign_id	-0.03	1.88	0.17
has_facebook	-0.09	15.44	0.00
has_paused	0.28	160.24	0.00
sales_channel	-0.04	2.88	0.09
payment_methodpaypal	0.05	5.34	0.02
email_invoices1	-0.09	15.53	0.00
GLOBAL		346.03	0.00

Table 1. Assessment of Proportionality Assumption

Final model	rho	chisq	p
rewarded_14_count	-0.02	0.96	0.33
view_7_count	0.04	2.67	0.10
has_campaign_id	-0.00	0.03	0.87
GLOBAL		2.75	0.43

Further, we can visually examine the Schoenfeld residuals (#explain#) for time-dependence. Visually it appears that the stepwise-selected variables counting the number of badges earned within the first 14 days and the number of videos watched within the first 7 days do not exhibit time dependence. In the bottom plot of whether a subscriber arrived via a marketing campaign shows a loess line not apparent in the other two visualizations. It is this last graphic that appears to show some time dependence, especially in older subscribers, though our previous Chi-square test, does not reject the null hypothesis of a lack of time dependence.

Figure 6. A visual examination of Schoenfeld Residuals



The stepwise selected results are as follows:

A Chi-square test of the Schoenfeld residuals for time dependence shows all variables as contributing to the rejection of the null hypothesis that the model is treating the constituent hazard functions as proportional.

After four incremental steps of removing the least significant variable from the model and rerunning, the following model is arrived at:

Table 2. Cox Proportional Hazard Model: stepwise selection

	coef	exp(coef)	se(coef)	z	p
rewarded_14_count	-0.01	0.99	0.00	-4.57	0.00
view_21_count	0.01	1.01	0.00	15.68	0.00
total_view_count	-0.01	0.99	0.00	-14.83	0.00
looking_for_work=1	-0.24	0.78	0.05	-4.78	0.00
has_paused	-0.66	0.52	0.05	-13.43	0.00

	rho	chisq	p
rewarded_14_count	-0.09	16.85	4.05E-05
view_21_count	-0.14	60.01	9.44E-15
total_view_count	0.15	79.11	0.00E+00
looking_for_work1	0.12	26.06	3.30E-07
has_paused	0.28	150.08	0.00E+00
GLOBAL		268.60	0.00E+00

With a final test of the Schoenfeld residuals given as:

	coef	exp(coef)	se(coef)	z	p
rewarded_14_count	-0.03	0.97	0.00	-8.67	0.00
view_21_count	0.00	1.00	0.00	1.86	0.06

	rho	chisq	p
rewarded_14_count	-0.01	0.12	0.72
view_21_count	0.01	0.18	0.67
GLOBAL		0.20	0.90

A plot of the Schoenfeld residuals for each of the two independent variables does not appear to show time-dependence:

Figure 7. A visual examination of Schoenfeld Residuals

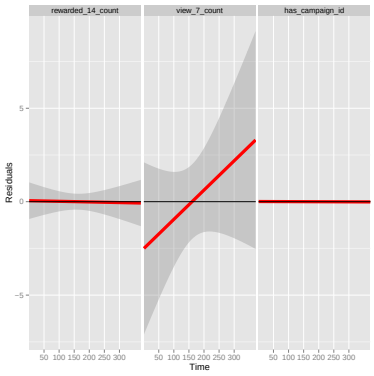
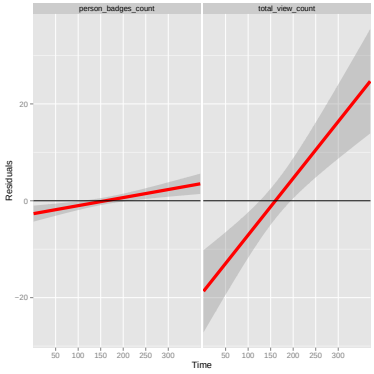


Figure 8. A visual examination of Schoenfeld Residuals



The Cox Proportional Hazards model will first be compared to that of two sets of two random forests. The purpose is to see if while respecting the PH assumption, random forests are better than a Cox Proportional Hazards model for this

application. The second set of models run will include all available explanatory variables, not just those that satisfy the PH assumption. The guiding purpose of this examination is to first, find early user engagement predictors (like badges earned within the first 14 days) that can be useful to Treehouse and allow them to attempt to improve customer longevity. The secondary purpose is to find the best possible model of survival curves for all subscribers over variables that are not measured within a window of time but all time. Such a model could allow Treehouse to make revenue projections, or to find out which marketing channels, or customer behaviors tend to be associated with longer survival times.

3.1.1 Corresponding Random Forests

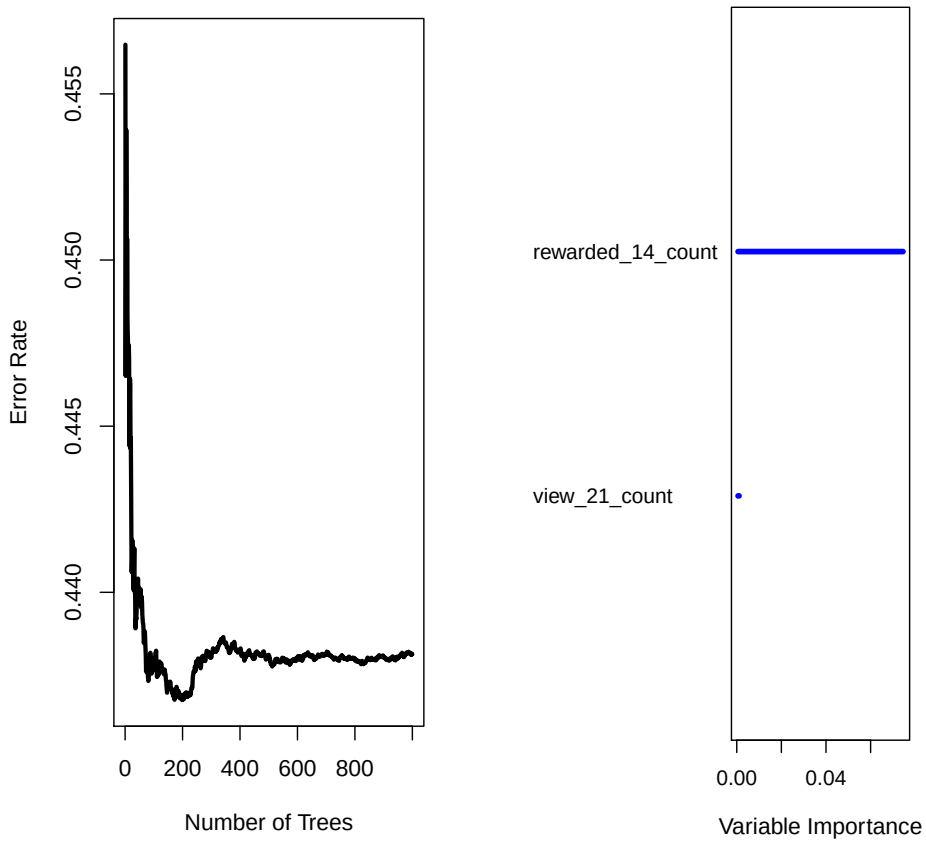
The first random forest presented is the `rsf` call of the `randomSurvivalForest` package:

```
Call: rsf(formula = Surv(ttc, censor) ~ rewarded_14_count +
          view_21_count, data = training.data, ntree = 1000,
          splitrule = "logrank", nsplit = 1)
```

Sample size: 5430
Number of deaths: 1973
Number of trees: 1000
Minimum terminal node size: 3
Average no. of terminal nodes: 112.343
No. of variables tried at each split: 1
Total no. of variables: 2
Splitting rule: logrank *random*
Number of random split points: 1
Estimate of error rate: 43.81

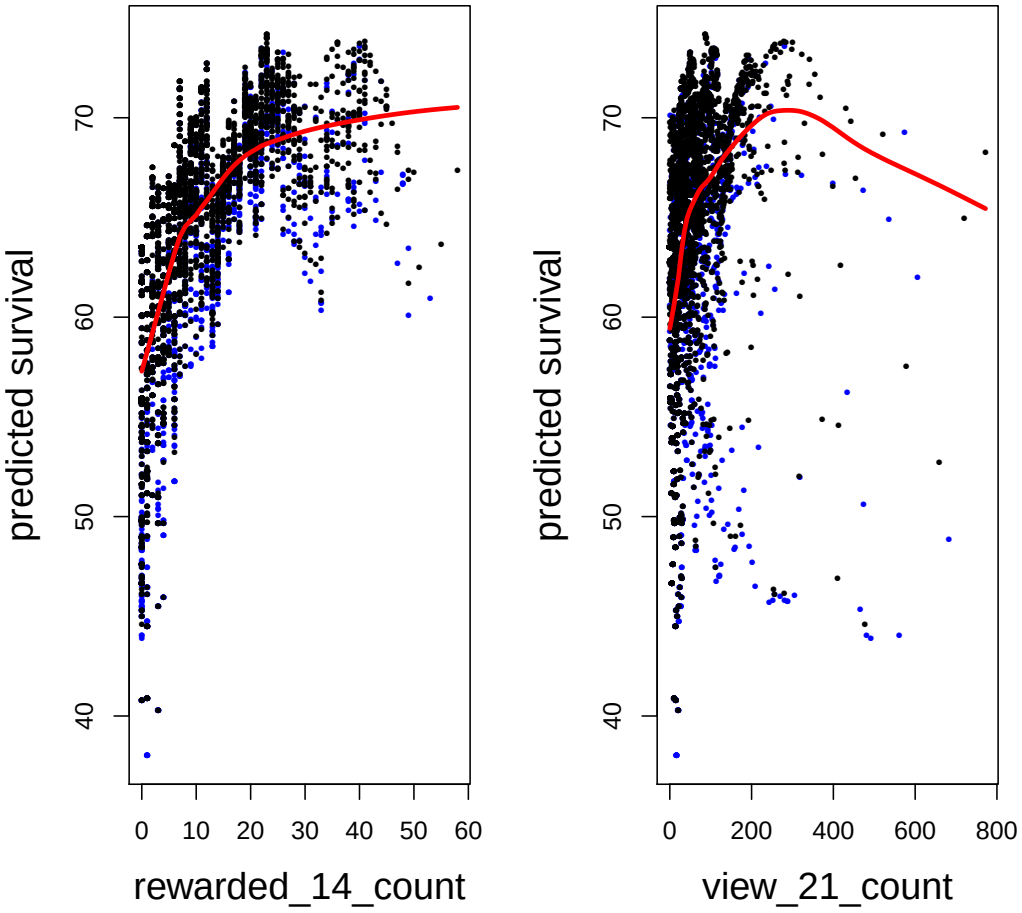
The model above is restricted to those explanatory variables that satisfied the PH assumption for the Cox PH model. The estimated error rate is quite high. The change in error rate as the random forest was grown as well as variable importance is shown below:

Figure 9. Random Forest: rsf (limited to Cox PH vars)



Partial dependence plots show how survival times are predicted to change as explanatory variables change within the random forest. The following figure shows that survival times increase monotonically as the number of badges earned within the first 14 days increase - diminishing returns are seen at about 20 badges. Further, survival times increase as videos watched within the first 21 days increase until about 300 videos, then survival times decrease. This is not surprising as content is finite and it is possible for subscribers to feel that they've exhausted all content that interests them.

Figure 10. Random Forest: partial dependence plots

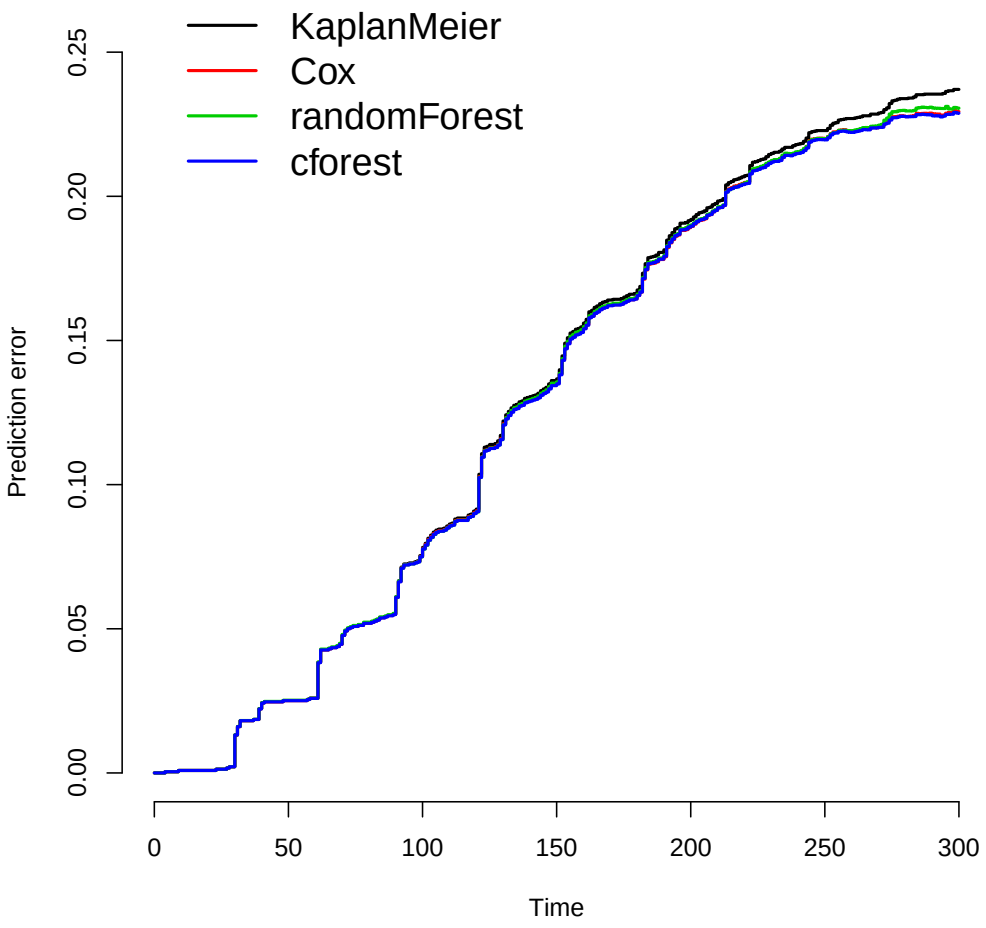


Another version of the random forest found in the `party` package in `R` was run. The specific call is `cforest`. The difference between the `rsf` and `cforest` implementations is as described on page #insert ref#. The details of the `cforest` model are omitted here due to similarity to the results from `rsf` shown above, instead differences in the prediction results are examined.

3.1.2 Prediction Error Curves: Restricted to Cox PH Model Variables

Explain the math behind prediction error curves here#

Figure 11. Prediction Error Curves



	Time	KaplanMeier	Cox	randomForest	cforest
1	30	0.052	0.052	0.053	0.052
2	60	0.091	0.090	0.090	0.090
3	90	0.140	0.138	0.139	0.138
4	120	0.179	0.177	0.177	0.177
5	150	0.205	0.203	0.203	0.203
6	180	0.222	0.220	0.220	0.219
7	210	0.233	0.228	0.229	0.227
8	240	0.239	0.231	0.232	0.231
9	270	0.245	0.237	0.239	0.235
10	300	0.249	0.236	0.230	0.225

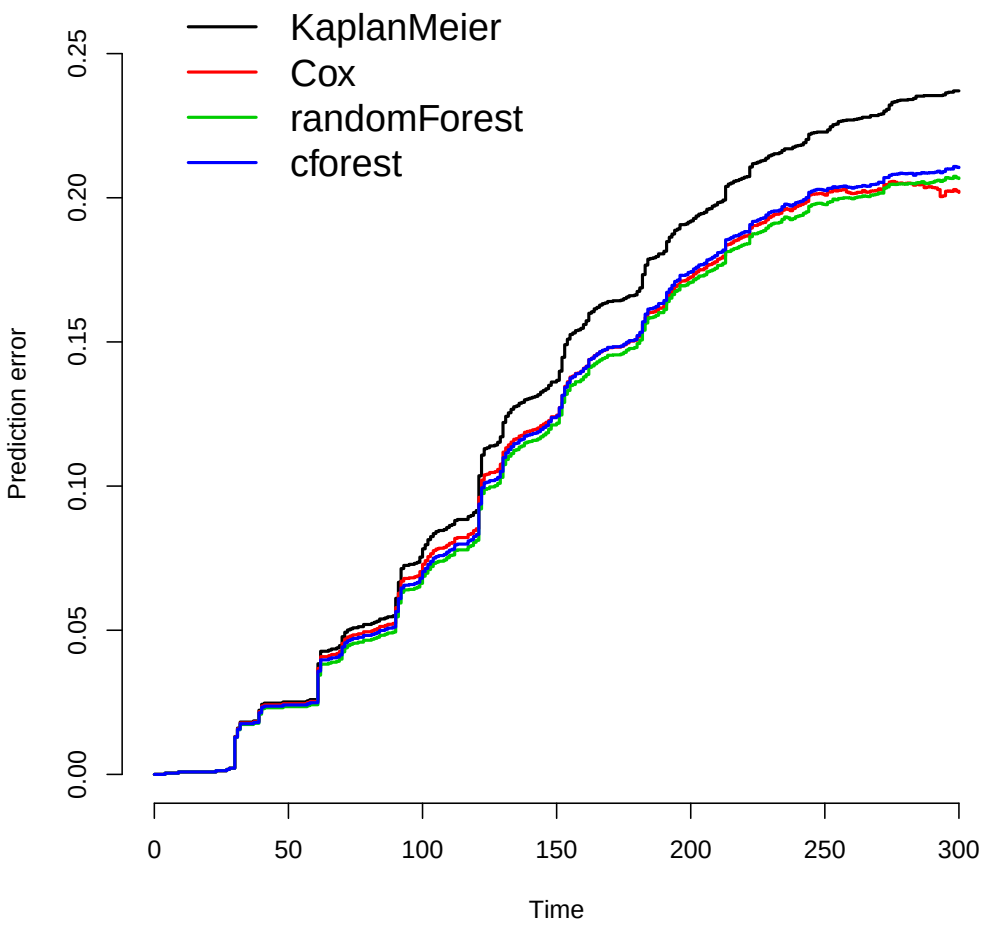
3.2 Unrestricted Random Forests (& Cox PH model)

3.2.1 randomSurvivalForest

```
Call: rsf(formula = Surv(ttc, censor) ~ rewarded_1_count +
      rewarded_3_count + rewarded_7_count + rewarded_14_count +
      rewarded_21_count + person_badges_count + view_1_count +
      view_3_count + view_7_count +      view_14_count +
      view_21_count + total_view_count + looking_for_work +
      has_campaign_id + has_facebook + has_paused +
      sales_channel +      plan_id + payment_method +
      email_invoices.
   data = training.data[training.data$plan_id \%in\% test.data$plan_id, ]
          , ntree = 1000, splitrule = "logrank", nsplit = 1)
```

Sample size: 5423
Number of deaths: 1972
Number of trees: 1000
Minimum terminal node size: 3
Average no. of terminal nodes: 277.779
No. of variables tried at each split: 4
Total no. of variables: 20
Splitting rule: logrank *random*
Number of random split points: 1
Estimate of error rate: 32.02%

Figure 12. Prediction Error Curves



References