Assignment 4

Brian Sohn

1. Fitting AFT Models to determine rate of events

We start by fitting three AFT models that describe customers' behavior for deauthorization, purchasing subscription, and cancelling subscription. All three models will be based on a (censored) Weibull distribution, with group, engagement, and interaction (=group*engagement) terms as covariates that determine difference in rate of event for each person.

The deauthorization model will find out the rate at which customers will delete their account after registration. We will use a Weibull distribution to fit this behavior, where the rates for each person are accelerated by the effects of their own covariates. However if a customer decides to purchase subscription before they delete their account, we will not be able to observe the moment that they would actually delete their account (Here, we are assuming deleting/purchasing are mutually exclusive events. There is no person in the data that had both events occur). For a customer that neither deleted nor purchased, their deletion time would also be unobserved. Thus we can say that for those who did not delete their account, their deletion time was censored by the purchasing time/current time, both of which are recorded as T in the dataset.

We model the purchase model in the same manner. As a person will not be able to purchase subscription after deleting their account, we cannot observe purchase time for those who deleted their account before purchasing. We cannot observe purchase time for those that neither deleted nor purchased as well. Thus we can say that for people who did not purchase the subscription, purchasing time was censored by deletion time/current time, both of which are recorded as T in the dataset.

Finally, for those who purchased subscription, we can model the subscription length. This will be modelled in the same way as the previous two models using an AFT model with a Weibull distribution, where their subscription length is censored for those who did not cancel within our timeframe of study.

The following table summarizes how censoring occurs for each customer segment in each model.

	Model_deauth	Model_purchase	Model_cancel
No Action	Censored	Censored	NA
Delete	Uncensored	Censored	NA
Purchase, not cancel	Censored	Uncensored	Censored
Purchase, cancel	Censored	Uncensored	Uncensored

2. Calculating Expected Lifetime Value of Customers

Once we have a fitted model for all three events, we can use them to estimate the expected lifetime value for each customer. This can be calculated as probability of purchase * subscription length * price.

For a user that has neither deleted nor purchased, we should estimate the probability of purchase p using our parameter values. For each parameter value (of model_deauth and model_purchase) in the sample of the posterior, we can get a posterior predictive (of uncensored Weibull distributions) for deauthorization time and purchase time, for each person. For a particular user, p would be estimated as the proportion of posterior predictives where purchase time < deauthorization time. Finally, subscription length can be estimated as the expected subscription length from the cancellation model (one way to get this would be using sample mean from the posterior predictive of an uncensored Weibull distribution[1] with same parameters).

A user that has deleted an account already will have a probability of purchase and therefore would have lifetime value of 0.

For a user that has purchased but did not cancel a subscription, we know that the probability of purchase is 1 already. To estimate the true subscription length, we can calculate the expected subscription length for this user conditional on the condition that subscription length is bigger than the observed subscription length[2] (time between purchase and now).

A user that has purchased and cancelled a subscription will have a probability of purchase of 1, and the subscription length is already observed.

	Probability of purchase	Subscription Length
No Action	p	E(sub_length)
Delete	0	NA
Purchase, not cancel	1	E(sub_length | sub_length>observed sub_length)
Purchase, cancel	1	Observed sub_length

Once we have calculated probability of purchase and subscription length for each person, we can multiply the two and then multiply by the price they were assigned to calculate their expected lifetime value.

Finally we would compute the average LTV for each price group and compare which is bigger to determine which price to impose to all customers.

[1] I sampled a posterior predictive in my notebook, but it is the posterior predictive of a censored distribution, right? So I would create a new variable that follows a Weibull distribution with same parameters but uncensored for each of the three models. The we could use a posterior predictive distribution of that variable. Not implemented due to long sampling time.

[2] Is this a good way to do it? I didn't want to use E(sub_length) since it could be smaller than observed sub_length.