MSiA 421, Data Mining
Assignment 4: CLV and Churn Modeling

Due: Tuesday, February 17 by 11:59pm

1. The data set `np.csv` is space delimited with a header line and the value "." indicates missing. In R you will want to set `na.strings="."`. It has been set up to run discrete time survival models with one record for each customer decision. You have a sample of digital-only subscribers without any left censoring. `SubscriptionId` uniquely identifies a subscriber and `t` is the month number in the customer's life. You have the following variables

    - `churn`: indicator if customer churned this month
    - Overall reader **engagement** variables
        - `regularity`: number of reading days this month
        - `intensity`: number of page views (PVs) per reading day this month
    - **Payment** variables `trial`, `currprice`: indicates if the reader is paying a trial rate and the price paid this period.
    - **Content** variables `sports1`–`opinion1`: number of PVs in each section this month
    - **Location** variables `Loc1`–`Loc4`: number sessions in four different locations this month. Remaining PVs are from other locations.
    - **Source** variables `SrcGoogle`–`SrcLegacy`: number of sessions from different referring sources this month
    - **Device** variables `mobile`, `tablet`, `desktop`: number of sessions on different devices this month

    The purpose of this exercise is to do an exploratory analysis to understand what factors are associated with churn/retention. Insights from this analysis will be used to allocate resources to improving aspects of the product. I think of regularity and intensity as measures of reader engagement, the content variables are about the product, device variables tell us about distribution and the user experience, source variables tell us about promotion and acquisition, and location might help us in targeting acquisition efforts and deciding where to assign reporters.

    (a) For all parts use logistic regression. To avoid issues of simultaneity, predict churn next month from reading behaviors this month. Create a variable `nextchurn` indicating churn next month by customer. Hint: see here for help using the `dplyr` commands `lead` and `group_by`. Also create a lead version of `currprice` and call it `nextprice`. Submit your R code.

    (b) Make `t` a factor variable so that you don't have to use `factor` in every model below. Submit a `table`.

(c) Run the following models:

```
nextchurn ~ t+trial+nextprice+regularity+intensity
nextchurn ~ t+trial+nextprice+regularity
nextchurn ~ t+trial+nextprice+intensity
```

What do you conclude about the effects of trial, price, regularity and intensity. Note that it's always a good idea to examine diagnostics like correlations and VIFs. What is the trial effect telling you, given that (1) most trial offers are 1 month, (2) many customers did not have trial offers, and (3) you already have a dummy for month 1 in the model with the `t` variable?

(d) Fit the following model to study content:

```
nextchurn~t+trial+nextprice+sports1+news1+crime1+life1+obits1+business1+opinion1
```

We need to be careful about multicollinearity. Do your conclusions change if you include regularity in the model?

(e) What can you conclude about the effect of location on churn? Fit these models:

```
nextchurn~t+trial+nextprice+loc1+loc2+loc3+loc4
nextchurn~t+trial+nextprice+regularity+loc1+loc2+loc3+loc4
```

(f) What can you conclude about the effect of source on churn?

(g) What can you conclude about the effect of device on churn?

(h) Do your conclusions change if you fit a model with payment, content, location, source and device variables all in at the same time? What if you use lasso with cross validation rather than statistical significance?

(i) Considering all of your analyses, put the variables into the following categories:

- No association with churn
- Strong drivers of churn (do less of these things)
- Strong drivers of retention (do more of these things)
- Questionable drivers of churn
- Questionable drivers of retention

2. Consider a migration model with k states and transition matrix $\mathbf{P}(k \times k)$. Suppose that at some initial point in time there are n_{0i} customers in state $i = 1, \ldots, k$. Let $\mathbf{n}_t = (n_{t1}, \ldots, n_{tk})^\mathsf{T}$ be the number of customers in each of the k states at time t. Suppose that during each period, $\mathbf{a}(k \times 1)$ new customers are acquired, e.g., if state 1 is for new customers then $\mathbf{a} = (a_1, 0, \ldots, 0)^\mathsf{T}$. Thus, the number of customers in each state at time $t+1$ is

$$\mathbf{n}_{t+1}^\mathsf{T} = \mathbf{n}_t^\mathsf{T}\mathbf{P} + \mathbf{a}^\mathsf{T}, \qquad t = 0, 1, 2, \ldots \tag{1}$$

Show that the expected number of customers at time t equals

$$\mathbf{n}_t^\intercal = \mathbf{n}_0^\intercal \mathbf{P}^t + \mathbf{a}^\intercal (\mathbf{I} - \mathbf{P}^t)(\mathbf{I} - \mathbf{P})^{-1}. \tag{2}$$

Answer: Write out the first few terms to see the pattern. For $t = 1$ we have

$$\mathbf{n}_1^\intercal = \mathbf{n}_0^\intercal \mathbf{P} + \mathbf{a}^\intercal$$

For $t = 2$ we have

$$\mathbf{n}_2^\intercal = \mathbf{n}_1^\intercal \mathbf{P} + \mathbf{a}^\intercal = (\mathbf{n}_0^\intercal \mathbf{P} + \mathbf{a}^\intercal)\mathbf{P} + \mathbf{a}^\intercal = \mathbf{n}_0^\intercal \mathbf{P}^2 + \mathbf{a}^\intercal \mathbf{P} + \mathbf{a}^\intercal$$

For $t = 3$ we get

$$\mathbf{n}_3^\intercal = \mathbf{n}_2^\intercal \mathbf{P} + \mathbf{a}^\intercal = \mathbf{n}_0^\intercal \mathbf{P}^3 + \mathbf{n}_0^\intercal \mathbf{P}^2 + \mathbf{a}^\intercal \mathbf{P} + \mathbf{a}^\intercal = \mathbf{n}_0^\intercal \mathbf{P}^3 + \mathbf{a}^\intercal \sum_{i=0}^{2} \mathbf{P}^i$$

We now see the patters. The last term is a geometric series of matrices and we can substitute the formula from class.

$$\mathbf{n}_t^\intercal = \mathbf{n}_0^\intercal \mathbf{P}^3 + \mathbf{a}^\intercal \sum_{i=0}^{t-1} \mathbf{P}^i = \mathbf{n}_0^\intercal \mathbf{P}^3 + \mathbf{a}^\intercal (\mathbf{I} - \mathbf{P}^t)(\mathbf{I} - \mathbf{P})^{-1}$$

3. A news site has eight segments (states). There are four life stages: registered user (prospect), trial subscriber, full-price subscriber, and churned. The life stages are crossed with two levels of the regularity of reading (number of days per month), low and high. For all parts, assume a monthly discount rate of $d = 1\%$. The transition matrix, average number of page views per month, and the starting customer counts are as follows (available in csv file on Canvas):

Period t	Period $t+1$								
Lifestage	Prospect		Trial		Full		Churn		Page
Regularity	Low	High	Low	High	Low	High	Low	High	views
Pros L	603	76	83	88	50	39	8	1	7.7
Pros H	146	534	15	132	7	41	1	3	380.3
Trial L	0	0	45	17	309	147	26	7	13.4
Trial H	0	0	9	14	134	691	12	32	278.5
Full L	0	0	2	0	4310	614	223	19	4.2
Full H	0	0	0	7	955	4150	52	118	250.7
Churn L	0	0	9	3	18	4	1296	81	2.1
Churn H	0	0	4	6	1	6	154	424	163.6
Starting	5,000	5,000	1,000	1,000	3,000	3,000	4,000	4,000	

(a) Compute the transition probability matrix $\mathbf{P}$.

(b) The value vector has two components, subscription fees and advertising. Assume that the trial rate is \$1/month, full-rate \$10/month, and that there is no subscription revenue from prospects or churns. Suppose that the paper makes \$0.002 for each page view (PV). Find the value vector.

(c) Find CE, letting t go to infinity.

(d) If you cut the number of advertisements in half, so that the ad revenue is \$0.001/PV, by how much does CE change?

(e) For the remaining parts assume that the news organization is only interested in projecting cash flows for the next 36 months (rather than to infinity). Find CE assuming ad revenue of 0.001/PV. Hint: use Equation (1) to find $\mathbf{n}_{t+1}$ from $\mathbf{n}_t$ and $\mathbf{a}$, multiply $\mathbf{n}_t$ by the value vector, and add them up. You could do this with a loop in R or Python, but it might be helpful to try it in Excel the first time using the `mmult` function.

(f) Now suppose that you acquire 200 new prospects each month, 100 with low regularity and 100 with high regularity. What is 36-month CE and how many total subscribers (trial plus full) do you expect to have? By how much did CE increase from the previous part?

(g) Reducing the the ads should increase retention rates. Perhaps you have also started a newsletter to stimulate regularity, which will also increase retention rates. You would have to do a test to know exactly how much the probabilities change, but for this exercises reduce the following transition probabilities by 1%: $p_{1:2,1:2}$ and $p_{3:6,7:8}$. Also increase the following by 1%: $p_{1:2,3:4}$ and $p_{3:6,5:6}$, What is 36-month CE and how many total subscribers (trial plus full) do you expect to have? By how much did CE increase from the previous part?

(h) For you to think about but not turn in: what would the news site have to do to get to 30,000 subscribers and RE=\$15 million? Where is there the most sensitivity? Should they invest in reactivating churned customers?