MLDS 401/IEMS 404 Project
Due: November 25, noon
Professor Malthouse

Work with your assigned group. All teams should submit the presentation (I suggest using Powerpoint, Google Slides or Beamer) by **Nov 25 noon**. Also, if you have not already done so, schedule a 20-minute slot where you can present your findings to me ideally during Nov 25–27. If we are unable to find an agreeable time during those days then I will consider other days (Nov 24, Dec 1–3). Send me an email with all members copied and some slots that work for you. I will respond with an outlook invite.

Prepare a 12 minute presentation in Powerpoint, and expect five minutes of questions. Do not come in with raw output from R or Python. Answer the question and give **actionable** advice. The presentation should tell a story. Do not tell me about every model you tried; I'm looking for your answer rather than a list of 20 models. You don't have time to present 20 models in 12 minutes. You will be evaluated on the following:

- Quality of the presentation (slides and presentation), including staying within time
- Quality of the conclusions and supporting evidence presented
- Quality of the managerial implications

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="."` and `sep=" "`. It has been set up to run a churn analysis with one record for each customer decision. You have a sample of digital-only subscribers. `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
- **Device** variables `mobile`, `tablet`, `desktop`: number of sessions on different devices this month
- Ignore `Loc1`–`Loc4` and `SrcGoogle`–`SrcLegacy`.

The purpose of this project 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 media product. I think of regularity and intensity as measures of reader engagement, the content variables are about the product, and device variables tell us about distribution and the user experience.

1. For all parts use logistic regression. To avoid issues of around causes happening at the same time as outcomes, 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 a `table`. Run this R code.

```
np = read.table("np.csv", header=T, na.strings=".") %>%
  arrange(SubscriptionId, t) %>%
  group_by(SubscriptionId) %>%
  mutate(nextchurn = lead(churn),
         nextprice=lead(currprice),
         t = t)
```

2. Run the following models where `t` is numerical (not factor):

```
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. Here are a few considerations:

   (a) What is the trial effect telling you, given that (1) most trial offers are 1 month and (2) many customers did not have trial offers.

   (b) What do you conclude about the effects of intensity versus regularity? Which one should an organization develop strategies to encourage?

3. Fit the following model to study content:

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

   Do your conclusions change if you include regularity in the model?

4. What can you conclude about the effect of device on churn?

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

6. Considering all of your analyses, give a final recommendation for which factors retain customers, which factors drive them away, and which factors have no (substantial) effect on churn? In other words, give a summary of your final conclusions. Here are some hints:

    - I suggest making a conceptual framework (DAG) showing the theorized relationship between the variables. Think carefully about pipes, forks and colliders.
    - Think carefully about the managerial implications. For example, suggesting a price reduction should be justified—you earn less revenue with lower prices and the lower revenue must be justified by something else, e.g., better retention. Try to put yourself in the shoes of a media manager here.