IMC463: Machine Learning Project

- Purpose: build models to predict churn of a media subscription service. Churn means that the subscriber stops paying for the service, which is a dichotomous outcome.
- Market: Each team will have data from one of eight markets. Each market will be studied by two teams independently to see if they come to the same conclusions. These were assigned randomly:

Data set	Team 1	Team 2
Market 1	We love Ed	Data minding
Market 2	Authentic hard worker	XSWL
Market 3	Insighter	Intelligent robots
Market 4	LML	Human learning
Market 5	Homoscedasticity	ABCC Association
Market 6	YYDS	Future billionaires
Market 7	People learn machine learning	Data 4 You
Market 8	AI Geeks	CJPP

- Data: one record for each month of a customer's life until either the customer churns or we reach the end of time (censored).
  - `subscriptionid`: identifies a subscriber
  - `nextchurn`: your dependent variable, which equals 1 if a customer churns next month and 0 otherwise.
  - `train`: equals 1 for training set and 0 for test set.
  - `thismonth`: calendar time
  - `t`: customer time (month number of the customer's life)
  - Measures of **habit**: `regularity`, `sessions`—the number of days and sessions, respectively, in the current month. `NHomePage` is the number of visits to the homepage.
  - `PVs` total number page views
  - Session counts for **devices**: `DevMobile`, `DevTablet`, `DevDesktop`, `DevApp` (these may not sum to the session variable because there may be some additional sessions from unknown devices)

- Session counts by referring **source**: `SrcSearch` Google, Bing, Yahoo, duckduckgo; `SrcSocial` Twitter, Facebook
- Session counts by **market**: `MktOutside`
- Page view counts by content **topic**:
  - `TopicBreakNews`: breaking news and trending stories
  - `TopicLocalCom`: local communities
  - `TopicNatWorld`: nation world
  - `TopicLocGov`: local government
  - `TopicStateGov`: state government
  - `TopicNatGov`: national government
  - `TopicHealth`: health
  - `TopicCrime`: crime, courts
  - `TopicElect`: elections
  - `TopicColSport`: college sports football, basketball, other
  - `TopicProSport`: pro sports football, basketball, baseball, other
  - `TopicHSsport`: high school sports
  - `TopicFireAccident`: disasters, fires, accidents
  - `TopicImmigration`: immigration and racism
  - `TopicEduc`: education
  - `TopicOpenClose`: business openings and closings
  - `TopicJobEcon`: jobs, economy, finance
  - `TopicLocBus`: corporations, local business news
  - `TopicRealEstate`: real estate development
  - `TopicWeather`: weather
  - `TopicEnviron`: environment
  - `TopicRestDine`: restaurants and dining
  - `TopicEvent`: events
  - `TopicCelebrity`: celebrities
  - `TopicEntertain`: entertainment
  - `TopicTourism`: tourism
  - `TopicTraffic`: transportation, commuting, traffic
- Newsletter subscriptions (`NewsLetSubs`) and unsubscriptions (`NewsLetUnsubs`)

- Deliverables
  - **Regularity model**: Build a model using only regularity. What model and transformations did you use? Visualize the model with a spline or partial dependence plot.
  - **Best churn model**: Build your best model and tell what worked, i.e., model, variables used, key transformations.
  - **Parsimonious model**: Drop less important variables from your best model. The model should have roughly the same performance as the best model, but with a few variables as possible.
  - Compare the models on the test set with an ROC plot, AUC, and gains table.
  - Suggest "save" strategies that are financially justified (assume subscription generates $15/month). Discuss the implications of your ROC/gains analysis, e.g., which customers should recive a "save" contact point or points?

  The point of this project is not exploratory insights. Instead, we seek a parsimonious churn model. How well can we do with regularity alone? How much does adding other variables help?

- Suggestions
  1. Examine descriptives statistics and a correlation matrix.
  2. Factor analysis could group variables, but this may not help. It might give you insights on multicollinearity.
  3. Try different predictive models. At the minimum I suggest trying logistic regression (with transformations, stepwise, lasso and ridge, etc.); GAM; and RF/GBM (with variable importance and partial dependence or ALE plots).
  4. Is there anything you can say about whether temporal patterns of regulairty matter? For example, is the most recent value of regularity the only predictor necessary versus including previous values of regularity (e.g., `lag(regularity)`)? A more advanced and elegant approach would be to use *exponential smoothing*, which I'll discuss in class.

- Timeline
  - Teams will present preliminary results to the class on a Thursday
  - June 6 or 7: Schedule a 15-minute time to present your work to me over Zoom or in person.
- Report. Powerpoint deck and 10-minute presentation, 5 minute of questions. You should format the deck for a presentation, but you might want to have appendix slides with additional analyses.