Lab: Machine Learning

1. estimating NFL in-game win probability

we are given a dataset of 1st down and 10 plays

- i = index of play
- $y_i = 1$ if the team with possession wins the game, else 0
- game-state X
  - score differential
  - game seconds remaining
  - pre-game point spread relative to the team with possession
  - yardline
  - down
  - yards to go
  - timeouts for each team

- Fit a win probability model

$$\widehat{WP}(x) = \hat{\mathbb{P}}(\text{win} \mid \text{game-state } x)$$

using either a Random Forest or XGBoost, Pick your poison.

* RF: use a random forest of Regression trees. Either tune the parameters yourself using validation logloss, OR use 500 trees, mtry = 2, nodesize = 200 as in Lock and Nettleton (2014)

http://homepage.divms.uiowa.edu/~dzimmer/sports-statistics/nettletonandlock.pdf

* XGB: use a tree-based XGBoost with logloss loss function. Either tune the parameters yourself using validation logloss, OR use Baldwin's parameters. A tuning tutorial and his params can be found in his blogpost "NFL win probability from scratch using XGBoost in R"

https://opensourcefootball.com/posts/2021-04-13-creating-a-model-from-scratch-using-xgboost-in-r/

- Visualize win probability using partial dependence plots:
  - * Line plot of
    $\widehat{WP}$ (y axis) vs. yardline (x axis) for various values of score differential (color) and time remaining (facet) holding other covariates fixed
  - * Heatmap of $\widehat{WP}$ (color) as a function of score differential (x axis) and time remaining (y axis) holding other covariates fixed
  - * Line plot of
    $\widehat{WP}$ (y axis) vs. yardline (x axis) for various values of point spread (color) and time remaining (facet) holding other covariates fixed

- Quantify uncertainty in win probability point estimates by bootstrapping:

fit $B=100$ bootstrapped WP models $\{\widehat{WP}^{(b)}\}_{b=1}^{B}$ (or if this takes too long, go as low as $B=25$).

Because of a dependence structure in our dataset of football plays, we can't simply use the standard bootstrap, which assumes iid rows.

The outcome variable y_i is the same exact draw of win/loss for all plays in the same game.

y is independent across games but not within games.

The bootstrap works by mimicking the data generating mechanism, so in bootstrapping WP you need to sample games with replacement rather than sample rows with replacement.

* Create 95% WP Confidence intervals for each play in the dataset.
  What is the distribution of CI widths? How does this vary by time Remaining and down? What does this tell you about WP estimates?

* Create a Line plot of
  $\widehat{WP}$ (y axis) vs. time Remaining (x axis) for various values of score differential (color) and field position (facet)
  holding other covariates fixed.
  Create shaded colored Region for the WP CI (geom_ribbon in R). How fat are those CIs?