HW 11: Code Competition Austin Funcheon & Viraj Rane

* Looking at the dataset, it varies from 0 to 1, for most of the predictors. As such, there is no scaling to be applied to the data.
* With 710 variables, number of observations are fairly close to balanced data, which poses a challenge for many methods. Trimming proved mildly effective, but overall didn’t improve AUC score all that much, with a modest 5% of predictors trimmed.
* Also, reviewing the data, it is highly unbalanced.
  + 0:452 1:3548
  + As such, dataset for training set will need to be balanced to generate an appropriate model, in this case, for some models such as KNN and RF, data was under sampled to avoid extreme bias. For other models, such as GBM and ANN, oversampling was used, with only limited bias to the final test outcome.

Models evaluated:

* + - Naïve Bayes
    - Random Forest
    - ANN
    - SVM linear
    - GBM
    - RF Trimmed.

Reviewing methods, GBM seemed to take an early lead with the methods, with a 63.61% AUC. Running with the final validation set, this was slightly reduced.

For Final Validation set, GBM model resulted in an AUC Score of 60.72%

Summary:

|  |  |
| --- | --- |
| Method | AUC Result |
| NB | 52.4% |
| RF | 57.32% |
| SVM linear | 55.77% |
| ANN | 51.4% |
| RF Trimmed | 57.76% |
| GBM | 63.61% |
| Val GBM | 60.72% |