# Preliminary scan

# Clean

Zero variance

Similar columns

# Add features

# of N/A’s

# of N/A’s as a %

# Feature importance

# Grid Search-XGB (All)

# Cross-validation – nrounds

# Grid Search-XGB (Kaggle best)

# Ensemble –h20

# Deep Learning

# Results

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sub# | Add features | Drop features | Data | Model | Parameters  (method) | nrounds | AUC  Cv | Accuracy | AUC  Test @  20% | AUC  Kaggle |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | scipy optimize-minimize - ensemble- python |  |  |  |  |  |  |
|  |  |  |  | Ridge |  |  |  |  |  |  |
|  |  |  |  | Lasso |  |  |  |  |  |  |
|  |  |  |  | H20 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| **best** |  |  |  | **XGB** | **booster = "gbtree", nthread=2** |  |  |  |  |  |
| #12 |  |  | All data |  |  | 13 |  |  |  | 0.828475 |
| #16/21 |  |  | Sd=113, 0.13, T sp=0.8 |  | Eta=0.125  md=7 | 30 |  | Up 154 |  | 0.833306 |
|  |  |  |  |  |  |  |  |  |  |  |
| **Kaggle** |  |  |  | **XGB** | **booster = "gbtree", watchlist = watchlist, maximize = FALSE, ss=0.6815, csb = 0.701** |  |  |  |  |  |
| #24 |  |  | Sd=113  0.13T  0.8=sp | xgb | Eta=0.0202048  md=5 | 150 |  | Up 504 | 0.8403631 | 0.38741 |
| #29 |  |  | All | xgb | Eta=0.0202048  md=7 | 515 |  | Up 40 |  | 0.83892 |
|  |  |  |  | xgb | md=6 | 540 |  | Up 376 |  | 0.839914 |
|  |  |  |  | xgb | md=6 | 500 |  | Up 232 |  | 0.840257 |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | **Super-learner** |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | Sup-learn. | Knn,glm, rF | Warn fit 0/1  Not conv. |  |  |  | 0.882732 |
| #17 |  |  | 0.13 | Sup-learn. | Glm,rF |  |  |  |  | .745064 |
| #18 |  |  | 0.13 | Sup-learn. | Nnet, rF |  |  |  | .882 | .779387 |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | **Subsem-ble** |  |  |  |  |  |  |
|  |  |  | F1k,s= 0.5 | Subsemb. | glm, rF  sub=2 |  |  |  | .778 |  |
|  |  |  |  | Subsemb. | Knn, glm, rF  Sub=5 |  |  |  | .753 | 50 + warn  0/1 rank def.  no conv. |
|  |  |  | 0.13, sd=113  spl=0.8 | Subsemb. | Nnls, glm, rF  Sub=1 |  |  | A= 0.969697  ll=0.10098 | .867 |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Standard |  |  | Split=0.8  Seed=  113 |  |  |  |  | Threshold =0.5 | AUC @  S=0.8 |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Standard |  |  | Split=0.7  Seed=  15071 |  |  |  |  | Threshold =0.5 |  |  |
|  |  |  |  | Glm best subset | Really.big=T | 5 vars |  | 0.92333 @0.09 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  | N/A | N/A | F10k | Glm |  |  |  | 0.952333  @0.2 |  |  |
|  | All vars |  |  | Glm |  |  |  | 0.94@.5 |  |  |
|  | 6 vars (2) |  |  |  |  |  |  | 0.939@  0.15 |  |  |
|  |  |  | F10k | CART | Minibucket  =200 |  |  | 0.94333 @0.9 |  |  |
|  |  |  | F10k | RF | ntree=200, nodesize=25 |  |  | 0.962333  @0.1 |  | 0.503 |
| #14 |  |  | F10k | Nn | size = 2, rang = 0.7,decay = 5e-4, maxit = 500) |  |  | 0.955333 @0.1 |  | 0.733892 |
|  |  |  |  | Naïve Bayes |  |  |  |  |  | N/R |
|  |  |  |  | SVM |  |  |  |  |  | N/R |
|  |  |  | F1k | Mboost |  |  |  | 0.93333  @0.1 |  |  |
|  |  |  |  | gamboost |  |  |  | 0.92333 @0.09 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

1. var15+ind\_var25\_cte+ind\_var30+num\_var26\_0+num\_var22\_ult1+num\_var43\_emit\_ult1
2. var15+imp\_op\_var40\_comer\_ult1+ind\_var17\_0+ind\_var31\_0+num\_var22\_ult1+num\_var43\_emit\_ult1