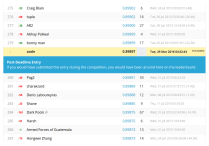
Assignment 3

Ankita De, Daqi Xu

March 29, 2016

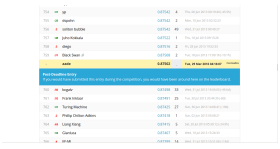
Exercise 1

Private Score : 0.89897
After some trys of default logistic regression model and model tuning (mainly use for-loop to find a better regularization strength), we adopt strategy from Miroslaw's code for feature engineering to get a better predict result with the logistic regression model.
Strategy :
1. One hot encoding
2. Feature Engineering
The dataset includes 9 variables, to find the inter-connection between each variables, we group several features together and increase the dimension of variables, then find the good features among those variables until the greedy search can not give back a better prediction. Group of 3 and 4 variables have been tested. Save the selected fea
3. Tuning regularization strength



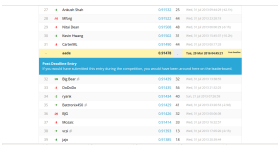
Exercise 2

Private Score : 0.87502
We did not use the rollup1 and role code features. We then tuned the parameters in the order : max_depth and min_child_weight, gamma, subsample and colsample_bytree and reg_alpha using GridSearchCV from scikit learn.



Exercise 3

Private Score : 0.91478
Strategy : We used an ensemble of the following :
1. Xgboost with tuned parameters : average of 10 different predictions, each produced using a random seed. Also dropped the roll_rollup1 and role_code features.
2. Our logistic regression : using feature selection and one hot encoding
3. The logistic regression of the BSMan part of the winning solution.
We used weighted average of their predictions in the corresponding weight ratio of 1:27:27.



References

1. https://www.kaggle.com/c/amazon-employee-access-challenge/forums/t/5283/winning-solution-code-and-methodology
2. http://mlwave.com/kaggle-ensembling-guide
3. https://github.com/pyduan/amazonaccess