Machine Learning Engineer Take Home

Answer:

1. how you validate your model, which, and why you chose such evaluation technique(s)

   -> I perform grid search cross validation (Grid Search CV) with XGBoost. Grid Search CV performs hyperparameters optimization and cross validation simultaneously. In this case Stratified k-fold cross validation is used. Some of advantages using Stratified k-fold cross validation are:

   a. address imbalance data by rearranging data such that each of the folds is a good representative with respect to different classes.
   b. Helps in reducing both bias and variance.

   Here is output from Grid Search CV given a set of hyperparameters :

```
Fitting 3 folds for each of 324 candidates, totalling 972 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 2 concurrent workers.
[Parallel(n_jobs=-1)]: Done  37 tasks      | elapsed:   42.0s
[Parallel(n_jobs=-1)]: Done 158 tasks      | elapsed:  3.7min
[Parallel(n_jobs=-1)]: Done 361 tasks      | elapsed:  8.3min
[Parallel(n_jobs=-1)]: Done 644 tasks      | elapsed: 15.1min
[Parallel(n_jobs=-1)]: Done 972 out of 972 | elapsed: 22.9min finished
model best params: {'eta': 0.05, 'gpu_id': 0, 'lambda': 1, 'max_delta_step': 1, 'max_depth': 3,
'n_estimators': 150, 'nthread': 4, 'objective': 'binary:logistic', 'scale_pos_weight':
13.391408400143591, 'subsample': 0.8, 'tree_method': 'gpu_hist'}
model best score:  0.8554592718744608
```

2. What is AUC? Why do you think AUC was used as the evaluation metric for such a problem? What are other metrics that you think would also be suitable for this competition?

   -> AUC stands for Area Under the Curve, which refers to ROC (Receiver Operating Curve) in machine learning. AUC is good because it is generic to evaluate model performance regardless of score threshold. This can be handy to compare models in some situations, e.g. as sample distribution is skewed like in this challenge.

   Other metrics that can be also used here are F1-score, precision-recall.

3. What insight(s) do you have from your model? What is your preliminary analysis of the given dataset?

-> From the dataset, it's clear that class distribution is very imbalanced.

Based on evaluation the XGBoost model can predict pretty well in general. It didn't show overfitting as well based on training-dev set performance.

4. Can you get into the top 100 of the private leaderboard, or even higher?

   So far the best private score (70% of test data) is 0.86653. This is equivalent to rank 143th in the Leaderboard chart.

   Screenshot from my submission in Kaggle:

cs_pred_nn.csv an hour ago by wandisusanto 5th sub neural net	0.86490	0.85783	☐
cs_pred.csv 2 hours ago by wandisusanto 4th sub	0.86490	0.85783	☐
cs_pred.csv 6 days ago by wandisusanto 3rd sub	0.86650	0.86028	☐
cs_pred.csv 6 days ago by wandisusanto 2nd submission	0.86441	0.85759	☐
cs_pred.csv 6 days ago by wandisusanto 1st submission	0.86653	0.86038	☐