Rationale of the algorithm

The challenge is a binary classification, aiming at detecting and reporting suspected transactions. Based on the given data set, we acquire that the ratio of fraud to non-fraud is 3:7, the variables contain mixed formats, ranging from discrete, continuous to timestamp, and the train data set and test data set is 14000, 6000 respectively.

Certainly, traditional statistical method does not work effectively, here innovated method must be introduced to solve the problem. As we all know, artificial intelligent (AI) has already surpassed human in many fields, like image classification and text parser. Combining the given data and AI, especially concerning the mixed variables, we figure out tree model and some well-designed neural network, like DeepFM, are well suited potential solutions. When talking about tree model, we choose ensemble tree model, random forest (bagging technique) and XGBoost (boosting technique), because ensemble model is much more robust and strength than single one. For the neural network models, as far as I know, DeepFM is specially designed for hybrid variables and excels in recommendation area. Based on the above concern, we finally choose Random Forest, XGBoost and DeepFM as candidates.

Every model has its own advantage and limits. Random forest is robust to outliers, has fast compute speed with parallel calculation and low risk of overfitting. While XGBoost (as with other boosting techniques) is more likely to overfit than bagging does (i, e. random forest) but with a robust enough data set and conservative hyper parameters, higher accuracy is the reward. However, as for neural network, it is usually “the more the merrier”. when dealing with massive data set, neural network can converge with the same number of parameters to lower generalization error. But for smaller data set XBGoost typically converges faster and with smaller error.

The experiment results show that XGBoost got the highest f1-score and DeepFM performed the worst. Concerned the limited train data set amount, only 14,000, the performance of DeepFM is not exceptional. However, in real situation with tones of data, there is high possibility DeepFM will surpass XGBoost.