BUS501: AI in Digital Marketing

Topic 2: Anomaly detection

Question 1. Today is your first day as a data scientist intern at a 3rd-world credit card company, 3RDCard, that customers are frequently exposed to counterfeit POS (Point of Sales). Since the onset of the service, 3RDCard has been fighting for the numerous types of fraud transactions, thus have hired a pack of experts working 24/7. The company now has a long list of potential fraud types that 24/7 workforce uses as a reference point, but the mafias on the street are getting better and better at cheating. The fraud detection team has become the largest team in the company, or simply put the major cost center.
Due to the pressure for cost-cut, 3RDCard now wants to test if machine learning can be any good use. The company hired two data scientist interns, one from software engineering department and the other from SIAI's MBA AI/BigData. One of you will end up with a job offer, depending on the model performance.
Since both of you are interns, the company's management hesitated to share a full set data. What you have is a data, which seems like a PCA-ed version. You protested that it forces you to work like a blind man, but the 3RDCard boss does not know the importance of data labeling, and your contender does not seem disturbed. Given a lot of constraints, you have below modeling strategy.

1 You try to use the computational efficiency as an indicator, but wonder if MSE works as expected. You wonder what happens if you replace MSE by F-1 score or accuracy ratio. Provide your logic.

2 Given the scarcity of fraud case (label 1), you doubt if simple logit, probit, or SVM can work. Provide your reasoning. You also wonder if sub-sampling can be of any help. Does the 24/7 workforce's long list can be of any help? Provide detailed strategy of sub-sampling with your logic for the list's value.

3 By the fact that there are a number of variables in the data set, you hypothesized that PCA-ed data may actually work better because it may have already circumvented curse of dimensionality. Your opinion? By any chance, can you recover the original data set from PCA-ed data?

4 By the time-stamp, you have figured that it is actually two full day data, which may have some sort of time series properties. You have two contending models between AR(I)MA and Early stopping. Provide your argument what option do you choose. Both, one of them, none.

5 Does partitioning the data can be of any help? How much homogeneity within a partition and heterogeneity between partitions can you find?

6 You now wonder if you can construct the Grubbs's test. Execute the test, and describe why or why not it works.

7 What about LOF (Local Outlier Factor)? Is there any possibility of LOF (or any other kind of grouping) for the fraud detection?

8 Let's assume that you have some compromise with data's distribution. You can rely on your results from 2). Construct an MLE, and argue whether on-line method can be used for this case.

9 Can you find any advantage by re-setting threshold in terms of computational efficiency? The answer may change depending on your answer on 1).

10 Let's assume that your contender, the software engineering guy, just used a simple RNN algorithm with days of hyper parameter tuning. He wonders why you waste your time for modeling, and in fact, he did not even show up to the office for days, until his computer finishes the "learning". Since the algorithm is straightforward, let's assume you could build the same model. Compare yours to the simple RNN.

Bonus Assume that you kicked out the other intern, and you are now a proud full-time employee. You are about to ask your boss un-PCA the data. Given your estimation strategies, what other data sets might you need for further improvement?