1. In the report, they go through the variable selection, model selection and model implementation procedures. They choose four models (Logistics Regression, KNN, Random Forest and Gradient Boosting) to fit the dataset from ‘application_train.csv’ and ‘buearu.csv’. Gradient Boosting achieve the best performance in cross-validation, with the area under curve score of 0.71 and accuracy of 0.91. While the Kaggle public score of Gradient Boosting is only 0.50, and Random Forest become the best with a score of 0.54.
2. 1) Do basic analysis and preprocessing of variables.

   2) Use various models to fit the dataset and carefully analyze the models and results.

   3) Get the Kaggle public scores of each model.

3. 1) The Outlier(“DAYS_EMPLOYED”) are not handled.

   2) For variables with more than 2 unique categories, label encoding may cause some problems. Because it gives the categories a random order, if we perform the same process again, the labels may be reversed or completely different. So one hot encoding will be better.
4. 4
5. 3.5
6. 4
7. 3