Final Project Update, Nov 20, 2017

Team: KKBox - Xiaotong Gui, Minh-Quan Do, Zihao Xu

1. Have you already collected, or do you have access to, all of the data that you will need in order to complete your project? If not, please estimate the percentage of the data that you have, describe any issues that you are having, and what your plan is for getting the rest of the data?

   We have obtained all the data from Kaggle already. But recently there has been some newly released data that needs to be appended to the previous data sets. Or alternatively, we could also conveniently use the new datasets (which are smaller in size than the original datasets) as the test data to evaluate our model performances.

2. What is the single biggest unresolved issue you are having? Please describe it briefly, and what your plan is for resolving this issue.

   a. The biggest problem we are encountering is that the response variable is very imbalanced: the churn to no_churn rate is around 1 : 10000. This creates two problems. First, it it hard for the model to capture the differences in between the groups as there are so few observations in the “positive” class; secondly, we need to find other ways to evaluate our model, like precision and recall or the ROC curve, because accuracy is unreliable (predicting all to be no_churn gives 99.9% accuracy)
   b. We have trained RF and SVM models using the data sets, but both give poor results due to the class imbalance. To address this problem, we might need to oversample the observations from “churn” class, or adopt other more advanced algorithms like neural networks.
   c. One of the datasets, user_log, is still too big to be directly imported into memory. To address this, we will use the online kernel provided by Kaggle, which has 16GB of RAM, pre-process this dataset (groupby user_id and take the mean of all the columns, or simply select the row with most recent timestamp) so that we can load and do computations with this dataset more easily in the future.

3. What are the elements from outside of the course, if any, that you plan to incorporate into your project?

   a. To improve prediction accuracy, we will experiment with other models that are more suitable with imbalanced datasets. We plan to use: Improved Balanced

Random Forest (IBRF) [1], an algorithm that penalizes misclassifications of the group w/ few observations, and autoencoders [2] [3], a form of neural networks that had been proven to work well on imbalanced datasets (fraud detection and etc.).

b. We are also gaining experiences dealing with imbalanced response variables. We need to evaluate the model using methods like the ROC curve, which is a concept untouched in the class
c. Maybe we can use Tableau to present the final interactive visualization?(totally optional if time allows us)

References

[1] Customer churn prediction using improved balanced random forests
http://www.sciencedirect.com/science/article/pii/S0957417408004326

[2] Autoencoders and anomaly detection with machine learning in fraud analytics
https://shiring.github.io/machine_learning/2017/05/01/fraud

[3] Credit Card Fraud Detection using Autoencoders in Keras — TensorFlow for Hackers
https://medium.com/@curiousily/credit-card-fraud-detection-using-autoencoders-in-keras-tensorflow-for-hackers-part-vii-20e0c85301bd