Task 4 – Inferring user properties from app usage (40%)

The rise of smartphones to omnipresent means of communication has resulted in a boost of mobile app diffusion and, subsequently, a major increase in the use of telecommunication channels from mobile (3G, 4G, etc.). Providers are eager to learn more about the way their customers interact with the mobile apps and with each other to, e.g., better place product advertisements.

In this final task you will work as a data scientist for a real Chinese provider. Given log files of application usage and information about users and applications you will analyze user behavior and try to predict the gender and the age group of a user.

This time, the data (which you can download here) is – like in most real-world cases – distributed over several sources. Your first task is thus to explore each of the sources and to evaluate which data can be useful for predicting the gender and age group of a user based on EDA. Afterwards, you should continue to merge the relevant parts of the data sources into a single dataset. Note that the resulting data can be quite large, so either make sure you have access to a powerful computer or use sampling techniques to treat subsets of the data.

Again, you will find a training set and a test set. However, this time the test set is not labeled. Therefore, you should evaluate your model by choosing an appropriate resampling method (e.g. cross-validation) on the training set.

Note that in this task there are two major tasks:

1. The first task is to clean, analyse, preprocess and merge the data. You should spend a considerable amount of time with these tasks. In particular, pay attention to missing data, data that is badly formatted and data that cannot be understood immediately. This first task is valued with 65% of the grade of your usual final presentation
2. Based on your merged data, the second task is to build your predictive model. This task is valued with 35% of the grade.
3. As in the previous tasks, present your findings in class.
4. At the end of the semester (September 30), summarize your findings together with those of the other tasks in the final report.