**Competition 04: Google Analytics Customer Revenue Prediction**

Team: C4-11 Sakshi Jawarani, Bowei Sun, Jingyi Luo

*Who might care about this problem and why?*

The results will help the company better prepare marketing budgets that may attract more customers and boost the overall revenue. The model may also be used to guide future decision making regarding the products to attract customers produce most of the revenue. Identifying customers who are not so interested in the products would also help the company improve its products for these customers. Profiling a user can help a company in customer retention and future promotions for that particular user. Target marketing can be performed once user profiles are identified.

*Why was this problem challenging?*

The data has many missing values. It is unclear how to handle missing values without proper domain knowledge. JSON columns need additional parsing. The large size of the data makes processing difficult and time consuming.

*What other problems resemble this problem?*

Predicting revenue for online shopping/ordering websites resemble this problem. For these websites, it will be useful to how which group of customers buy more products so that the companies could customized the product for these customers. As for the not-so-interested customers, the companies may want to identify them as well. Different types of products could be offered to these less-interested customers.

*What might account for the differing performance levels of the mandatory models?*

First, the linear model and nonlinear model performance differently. As the relationship between the revenue and the variables used is non-linear, it explain the difference in performance between the linear and non-linear models. For example, Random Forest works well for multiclass problem while SVM are usually used for two classes.