**Corporación Favorita Grocery Sales Forecasting**

shiying Chen, Jianjun Du, Bo Huang, Yanyan Zhu

**Introduction**

This is a regression problem predicting unit sales of more than 5000 products in 54 stores of the company across Ecuador. The training data includes the record of over 120 million transactions. The training dataset, train.csv, includes date, units sold, item ids, and store ids. There are also several other datasets giving information about location of stores, daily oil prices, and type of stores, and other information. All these extra datasets can be used to construct features for model.

**Preprocessing and Feature Engineering**

Because there are 120 million records in the train.csv, and the size of it is more than 4 G. Training such a big data is impossible.

The steps for preprocessing:

1. Convert the type of data in the different columns. For example, a int64 can be changed to int32, and other categorical features can be changed to int8. This step shrinks the data size to 2 G.
2. There are 54 stores. The assumption is the sale of every stores are independent. As a result, the dataset can be separated to 54 sub datasets.

The steps for feature engineering:

1. Extract information from the date column. This generates year, month, day of week, and the days from beginning. It is shown from the attached Python codes, all of these factors are impacting the sale.
2. It is said in the competition’s website that oil price has some impact on the economy of the country. Therefore, daily oil price is also added as feature.
3. Change the month, and day of week to categorical features, and create dummy variables.

Preliminary models:

Ran three models: Extreme gradient boost, Random forest, and SVM.

It is shown that Extreme gradient boost has the best performance.

Attachment: two files, and you need to use Jupyter to open them.