**Grupo Bimbo Inventory Demand**

Kaggle Competition: Grupo bimbo Inventory Demand. Site: <https://www.kaggle.com/c/grupo-bimbo-inventory-demand>

**Business understanding**

From the competition site description:

In this competition, Grupo Bimbo invites Kagglers to develop a model to accurately forecast inventory demand based on historical sales data. Doing so will make sure consumers of its over 100 bakery products aren’t staring at empty shelves, while also reducing the amount spent on refunds to store owners with surplus product unfit for sale.

In this competition, you will forecast the demand of a product for a given week, at a particular store. The dataset you are given consists of 9 weeks of sales transactions in Mexico. Every week, there are delivery trucks that deliver products to the vendors. Each transaction consists of sales and returns. Returns are the products that are unsold and expired. The demand for a product in a certain week is defined as the sales this week subtracted by the return next week.

**Data**

There were some files to download and join them. The train files has more than 74.000.000 rows, in the case of test files almost 7.000.000 rows. The other files were related to products, clients an towns.

**Data Exploration and transformation**

It had been decided to separate the files by week in order to facilitate the processing, it

*Client* file has some duplicates that had been removed. The file *product* has the NombreProducto feature with some useful information, that is why it had been processed using Regular Expressions, and it had been obtained features like *brands* (esp. Marcas), *measurements* (esp. medida), *weights* (esp. pesos), and some others.

Then the features obtained had been joined with train and test to get train by week. That process gave us files with around 10.000.000 rows each, then the Data Science techniques have to deal with this file sizes. This week based processing got rid of some useless features, some of them had been converted to factors.

The feature engineering that had been made getting information from product description had been very useful and important to get some insights about the data behavior.

The dataset had features related to the past product demands that obviously doesn’t have the test dataset. With those features it had been created some new ones like the average demand per client and product in order to have a better approximation to the projected demand required, of course trying to avoid as much as possible overfitting. The last step in the dataset processing was to transform some features using One Hot Encoding.

**Model and evaluation**

The target feature in this dataset was numeric i.e. the sales amount. The model used was xgboost, one of the reasons to use this model was that it could process a 10 million file and more or less 200 features in about half an hour approximately in a Macbook Pro 8Gb of RAM.

One of the main problems was the files size in the exploration and transformation process. In the case of Random forest it took almost forever and it was not possible to get a higher rank using the competition evaluation metric that was Root Mean Squared Logarithmic Error.

**Conclusion**

It had been a very good experience because of the big files sizes an the possibilities my computer has processing large amount of data, at least for me. The decision of processing the files a week at a time was not a good one, as far as I have seen some results published in the competition. It was great experience an fun doing Data Science.