**COMPUTING LAB- DATA WAREHOUSE PROJECT PROPOSAL**

Sílvia Ariza, Stephen Carmody, María Fernández, Joan Verdú

DS-BGSE, December 2014

**Objectives:**

* Predict actual stock needs for all the products belonging to the most valuable category according to last trimester value of sales.
* If we have overstock at 95% confidence interval, choose whose customer will be willing to buy it according to the ranking of week mean consumption related to last trimester week consumption. Also report which employee should be selling it, according to the ranking of sales for each particular product (or the affinity to this particular customer). If we have understock, report the name of the supplier we have to call. This objective 2 can be done all in SQL.

We make the assumption that stocks are planned in a weekly basis, that’s why we predict weekly consumption. That helps us to average weekly and smooth the data.

**Amount of data (rows):** One row per day, but since we need backwards data up to six months, we will not be able to do any prediction of the oldest six months!

**Dependent variable (y):** For each day, sales of the next week of every product.

**Explanatory variables(x):** For each product of the chosen category and each day, we have to do a prediction of weekly sales, which involves data of past sales with this added explanatory variables (all sales refer to quantity):

1. For this product:
   * **X1:** Sales last week (standardized by sales per week) ([value-mean]/st.dev.)
   * **X2:** Sales last month (standardized)
   * **X3:** Sales last trimester (standardized)
   * **X4:** Sales last semester (standardized)
2. Include also data for other products that are correlated (standardized). This might need a previous step before starting with predictions.
3. **X5:** For all the products in this category (summation of quantities): same as before (standardized)
4. **X6:** For each of all remaining categories (standardized)
5. Week of the year
6. Month of the year
7. Trimester of the year

**Analytics:**

For every product, apply GLM but using a linear regression algorithm that shrinks some values to 0 (that is LASSO). Choose lambda value such that minimizes MSE. (as we did in our practice).

Once we have beta estimates and its confidence intervals, do Montecarlo simulations to estimate confidence intervals of prediction, and report the mean prediction and the 95% confidence interval high prediction (for actual data). If possible, validate mean and max prediction with past data.

**Graphs/table needed:**

Table of parameter estimates, errors, t values, etc (the standard one).

Graph showing time series of prediction vs actual value at validation (if available). Maybe we can show all products in one graph (one line per product).

Graph showing mean prediction, max prediction and actual stock for every product of the category. Bar graph?

Table showing : per product (actual stock- max weekly predicted sales). If positive (overstock), show which client should be most interested, and the employee that could try to sell it. If negative (stock should be immediately rised by this amount), show the name of the supplier for this stock. So a table with five columns: productname, overstock, BestCustomer, BestEmployee,Supplier. BestCustomer and BestEmployee filled only if overstock. Supplier only filled if understock.