COMPUTING LAB- DATA WAREHOUSE PROJECT PROPOSAL

Sílvia Ariza, Stephen Carmody, María Fernández, Joan Verdú

DS-BGSE, December 2014

Object definition

- Predict actual stock needs for all the products belonging to the top10 sales products according to the whole record history. Do do that, we compare actual stock with one-month prediction, for each product.

The main outcome is a table comparing stock with monthly sales predicition. If we have overstock according to prediction (upper bound at 95% confidence interval), it looks for the customer 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. If we have understock, report the name of the supplier we have to call.

Data extraction, transformation and loading

Since commands are quite sparse, we decided to predict in a month horizon, which we think is feasible for a stock optimization.

Amount of data needed (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 month of every product.

Explanatory variables(x): For each product of the top10 sales and each day, we do a prediction of monthly 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. For the rest of the products in its category (all added):
   - **X5:** Sales last day (standardized)
   - **X6:** Sales last week (standardized)
   - **X7:** Sales last month (standardized)
   - **X8:** Sales last trimester (standardized)
   - **X9:** Sales last semester (standardized)
3. For the rest of the products in other categories (all added):

- **X10:** Sales last day (standardized)
- **X11:** Sales last week (standardized)
- **X12:** Sales last month (standardized)
- **X13:** Sales last trimester (standardized)
- **X14:** Sales last semester (standardized)

Analysis

Since we are dealing with sparse integer data (quantity of product in a seldom command), we decided to apply Poisson with GLM.

In order to compare between models, we also included a simple linear regression, and two models who penalize the number of effective parameters: Poisson GLM Lasso and Poisson GLM Ridge.

For the penalizing models, we found an R function that searches and chooses lambda value such that minimizes the error.

We chose between these four models according to a validation process, done with ¼ of data devoted to this purpose.

We show in a table the summary of models chosen and its RMSE in percent. In general, Poisson GLM and linear regression were the selected methods. As a methodological conclusion, penalizing models don't improve validation error compared with standard Poisson GLM. Since in this case we have 14 features, it seems reasonable to keep all of them and improve prediction, which is anyway not easy and typically involves RMSE of around 50%.

Anyway, after seeing data, results and prediction errors, we think other models should be explored. In particular, since mean value of monthly sales is clearly lower than its variance, due to sparsity of commands, negative binomial GLM models could be a good option.

As an example, we tried to show in a graph the observed and predicted future monthly sales from the range of dates where we had enough information.