B2W Pricing Challenge

André Teixeira dos Santos

30/09/2016

SUMMARY

The main objective of the test was to develop a model to predict the volume of sales for a product in a day, for a given price. The data available was a transactional dataset, which registers sales events and a price monitoring dataset, which registers prices in time from competitors. Several models and techniques were tried, but the successful one gave an error of 16% (mean absolute error / average volume in day), which represents 75% less error than the benchmark (MAE for naive method). Along with the models some analysis were made to gain insights into the data.

ABOUT THIS DOCUMENT

This document is not intended to represent the detailed methodology nor to be a scientific article. The purpose is to provide a guideline for in-loco presentation, to explain the contents of this deliverable and to allow pre-evaluation ,of the results and of the author, to some degree. Any details not covered here can be asked about during presentation.

CONTENTS

In **deliverables directory** you will find the required deliverables for this test.

The **model.r** file contains the final model for product P2 and some plots. The **plots.r** contains plots to help understand the data. All R scripts connects to a local MySQL database. You can restore the dump file contained in **dump.zip** or you can run the queries in the **queries** directory in order**: load.sql; analysis.sql; models.sql**. You should modify the load.sql to point to the original CSVs files with the data. After that, you can modify the R scripts (firsts lines only) to connect to the restored MySQL database. By default the MySQL database is **local**, with user **root** and password **root** (port **3306**).

The remaining of this document is of optional reading. The contents bellow serves has documentation and will be discussed during presentation.

TECHNOLOGIES

The main technologies used in this test were MySQL, R and Metabase. The CSVs with the data were checked manually for structured and loaded into staging area at MySQL. In MySQL, the data were prepared for analysis, i.e., indexing, renaming, horizontal enhancements (e.g., day of week), dimensional aggregations (e.g., summary of a day) and, mostly, creation of specific tables, via query, to be consumed at Metabase or R.

Metabase was used to answer simple questions about the data in a way that is more fast and maintainable (and prettier) the doing in R. R and R Studio were used to answer more complex questions (e.g., histograms) and to develop and test models. Github were used just as code sharing & repository and not as version control due to lack of need.

PROCESS OVERVIEW

Although there was no specific development process nor specific design for the deliverables, the evolution of the test could be divided into the following phases:

- basic data preparation
- basic analytics
- first modeling attempt: linear regression (cross-sectional predictors)
- second modeling attempt: vector autoregression (VAR & SVAR)
- advanced data preparation & analytics
- third modeling attempt: dynamic models (cross-sectional predictors + lagged and differentiated predictors)
- fourth modeling attempt: clustering + dynamic models
- consolidation of results & presentation

PREMISES AND SIMPLIFICATIONS

The following simplifications could be worked out for a real project, but due to the nature of test, which is to evaluate potential, the increase in labor would not be justifiable.

- the nominal/base price for a product, for a given day, was considered as the maximum price for that day
- outliers were not treated
- inner joins were used to combine sales and competitors prices to avoid dealing with missing data, resulting in a decreased dataset
- the models were developed for 1 product (P2) and inter-products effects were not considered (e.g., substitute goods), i.e., it was assumed products sales were not negatively correlated
- neural networks and other non-linear models were not considered due to increase in complexity and labor, and loss in interpretability.
- for all regression models the mean absolute error (MAE) was used to evaluated the model performance and for benchmarking was used the MAE of naive and mean forecasting methods. Other performance metrics like R^2 were not considered.

REMARKS ABOUT THE DATA

- some prices from the competitors monitoring csv looked like were multiplied by 10, so it was divided back in the staging area.
- each product had a different temporal window, being for sale or competitor monitoring.
- the claim that the price was captured twice a day was not precise, i.e., it could be more or less than that and also, it could be at the same time, which has no use.
- price is not a good predictor for volume, as the requirements led to believe.

ANALYTICS

Analysis were made in order to understand better the data and attend to test requirements. Three indicators were developed, for a given product and competitor:

pricing efficiency: how well is the competitor able to keep its prices bellow mine, for each product?

pricing influence: how does the competitors prices impacts my sales, for each product, i.e., when he has lower prices, does my sales volume decreases?

pricing relevancy: what products should I focus on, to give a more intelligent pricing strategy, i.e., what products represents a greater revenue for me and are more sensible to competitive prices?

Some insights were compiled here as a sample from the result of those analysis:

- C5 is the main competitor
- P7 is the main product
- P2, P8 and P7 are the products who mostly require smart pricing
- Despite its prices, it is preferred by consumers to buy P2 at the competitors
- It's probable that consumers looks at C5 for P2 prices first.
- C6 is niche, i.e., consumers buy there no matter the price.

FIRST MODELING ATTEMPT: LINEAR REGRESSION

The first attempt was to use cross-sectional variables to predict the volume of sales, for one product, in a day. From preliminary analysis it was clear that a regression with price as the only predictor was not going to fit well, i.e., price alone is not a predictor for volume, although it influences to some degree. In general, the first regression was tried with a combination of the following predictors: base price, average competitors prices, minimum competitors price, month, day of week, day of month. This approach proved to be fruitless since the MAE for the models were not considerably better than the MAE for naive and mean forecasting methods.

SECOND MODELING ATTEMP: VECTOR AUTOREGRESSION (VAR & SVAR)

Since cross-sectional data alone weren't able to fit a model, time series forecasting techniques were next in line. To fit a VAR model were used the time series of: volume, price and competitors prices with a lag of 5, chosen based on the results of R's VARSelect method. The VAR model do not predict the volume "for a given price", so, although the model were generated (not evaluated), it couldn't be used strictly, given the test requirement. So, to use a autoregressive model to predict for a given price, SVAR should be used, but that would increase considerably the complexity and labor, so this approach was dropped.

THIRD MODELING ATTEMP: DYNAMIC MODELS

The third attempt was to manually combine cross-sectional data with lagged and differentiated values from multiple time series into a linear regression, e.g., prices times series of the most influential competitor, prices times series, volume time series, day of week, etc. It's important to note that the variables were not randomly chosen, i.e., correlations, ACFs plots, CCFs plots and other techniques were used to evaluate possibilities. But again, the mean absolute error was not considerably better the benchmark methods (naive and mean).

THIRD MODELING ATTEMP: DYNAMIC MODELS

Clearly by now there were no obvious predictor for volume, thus it had to be hidden somewhere. And also, considering that "Show methods for clustering" is a differentiating factor has the test requirements says, it was decided to cluster the days of the year, with the motivation that intrinsic characteristic of the market was possibly the best predictor for volume, i.e., days that have a particular effect on the volume behavior, e.g., mom's day, Christmas. So, the normalized sales volume, for all products, were used for clustering, i.e., each product sales volume was column for kmeans clustering. The number of centers was visually chosen to be 6, based on the evolution of the within cluster sum of squares (WSS) for each number of centers, from 2 to 15. A column with the

cluster of each day was added to the model along with base price, competitors minimum price and sales volume of the day before. The mean absolute error for that model was 46.32 (items sold per day), which represents 16% of the volume average of 291.34. The MAE for naive method was 200.94, which represents 70% of the volume average, thus the linear model had an error reduction of 75% when compared with the benchmark error (naive)