Data Mining and Text Mining
Project Report

10529039 Andrea Cappelletti, 10532096 Sandro Maglione

12th June 2020

1 Context

The company provided us a data set containing data regarding the sales of different products. More in detail we know the sku (unique identifier), the pack type (single or multi pack), the size in grams, the brand (there are five different brands), the price, the position exposed the previous week, the % of volume in promo of the previous week and the sales of the previous week of each product.

The total number of distinct products is 43, but the target prediction will be designated only to 12 products (in the scope).

The data set is divided between training and testing.

The data are collected with a weekly frequency starting from the 10th of December 2016 to the 22nd of June 2019 for the training set and 29th of June 2019 to the 14th of December 2019 for the test set.

The task assigned is to forecast the target sales of the 12 products in the scope for the next 25 weeks, from 29th June 2019 to 14th December 2019.

The performances of each model in our analysis are **evaluated by means of the MAPE (Mean Absolute Percentage Error) metric** on the test set.

2 Preprocessing

Data preprocessing is a fundamental step to take in order to obtain consistent information and train the model. The aim of this phase is to obtain a data set with correct and useful data.

- Check for missing week/s: The frequency of the data should be weekly, but some week/s may be missing. We checked for it and the result is that no week is missing in the 3 year time window.

- Check for missing values: The first week of sales was missing for all the product. We filled them using the next week sales. The rationality behind this choice is that there is a very low variance between consecutive weeks.
- Statistics: We analysed the standard indicators of statistical analysis and plotted the result to better understand the distribution of the data.
- Cleaning: We removed the term 'WE' for our analysis and given the name 'date' to the attribute pointing to the week date.

3 Data analysis and Transformation

We analysed the data with a visual inspection of the material provided. We plotted different scenario with different features of our data set in order to understand its nature. We underline some relevant aspects:

- The products in the scope belong to BRAND2 and BRAND4 only
- The product with SKU = 1608 is an outlier.
  Its average sales are a lot higher than any other product in the dataset. We analyzed its characteristics to find any useful information to explain this high amount of sales.
  However, we did not find any meaningful information. Therefore, we assume that this specific product is just very popular.

We observed that the trend of the sales looks similar among different products, so we decided to perform clustering with the data set in order to exploit some meaningful relationships.

We normalized the sales of each product time series and decided to perform a hierarchical clustering because the accuracy is higher with respect to K-Means and the computational time will result very fast with the relatively small amount of data we have.

The result is that the 43 products are very uncorrelated, but we notice that the products in the scope have a very similar sales history.

Actually we can divide them in two clusters. This lead to the decision to **use only the products in the scope to train our model** in order to have a better prediction.

4 Model

In order to build our model we notice that the attributes pack, size and brand do not change for each product, so they don't add any meaningful information and we don't consider them to train the model.

The problem we are facing requires to take into account how sales, prices and the other variables change over time.

We build a first Random Forest model in order to visualize and understand the importance of each feature in the training set. As a result we obtained that the most relevant feature is the price of the current week to predict.
Moreover, the analysis reports that the volume on promotion and the number of stores in which the product was put on evidence do not convey any meaningful information for our prediction.

We also made an Autocorrelation analysis on the sales. We found that the most relevant number of steps in the past to predict the current sales target is 2.

Based on this analysis, **we construct the final dataset using the price of the current week, the sales of the current week, and the sales of the current week divided by the price of the current week**. The training set contains one entry for each week in the original data set. Each entry has the 3 features highlighted above. Therefore, we constructed a new data set containing 133 rows for each of the 12 products in the scope.

4.1 10-Fold Cross Validation

In order to select the best model to predict the target based on the past data at our disposal, we choose to utilize the 10-Fold Cross Validation procedure.
We believe cross validation is ideal in our situation for two main reasons:

- Since we are going to build 12 different models for each SKU in the scope, we only have 133 weeks in our training set for each model.
  Cross validation allows us to utilize all the training data at our disposal by splitting it in 10 folds and using 9 of them for training and 1 for testing. We are therefore able to exploit all the 133 weeks.
- Splitting the set in training and testing would have introduced some significant problems in our analysis. We cannot be sure that the data in a fixed time period in the past can effectively yield a sound result on another fixed time period of the test set. On the other hand, by using cross validation we can mitigate this problem by testing the results on different testing sets.

4.1.1 Random Forest

We choose Random Forest as model to predict the sales in the future. We can train the model on our training set, which consists of 2.5 years of weekly sales data. Then we use the model to predict the following 6 months.

4.1.2 AdaBoost

We tried the same cross validation procedure with the AdaBoost algorithm. Since AdaBoost is able to improve its prediction by taking into consideration the error of the previous model, we reasoned that the model might improve the performance of the overall procedure.

However, we found that the results using AdaBoost were worst with respect to Random Forest in the end.

4.1.3 Lasso

Another option we considered is Linear Regression with Lasso Regularization. We reasoned that linear regression may yield a better prediction for future data compared to the other two models.

Unfortunately, the results of Lasso are worst with respect to both AdaBoost and Random Forest in the cross validation procedure.

5 Conclusion

The result obtained with our model presents a certain level of accuracy characterized by a **MAPE of 11.26** for the products in the scope.
This result has been obtained with a 10-Fold cross validation procedure using Random Forest with 1000 estimators.
Random Forest provides sufficiently precise predictions.

Other alternatives may yield similar or better results, some models are available such as ARIMA or Facebook Prophet.

6 Notebooks

We produced five different notebooks. The notebooks are pipelined: Notebook1 exports a new dataset which is used by Notebook2. Notebook2 also exports a new dataset which is then used by Notebook3, Notebook4, and Notebook5.

- Step1-MissingValues
- Step2-Clustering
- Step3-RandomForest
- Step4-TrendSeasonality
- Step5-TimeSeriesVisualization