Retail Promotion Forecasting: A Comparison of Modern Approaches

Casper Solheim Bojer, Iskra Dukovska-Popovska, Flemming Christensen, Kenn Steger-Jensen

Retail Promotion Forecasting: A Comparison of Modern Approaches

Casper Solheim Bojer[1], Iskra Dukovska-Popovska[1], Flemming Max Møller Christensen[1] and Kenn Steger-Jensen[1,2]

[1] Centre for Logistics (CELOG), Materials and Production, Aalborg University, Denmark
[2] Faculty for Technology and Maritime, Department of Maritime Technology, Operations and Innovation, University college of Southeast Norway, Norway
cbojer14@student.aau.dk

Abstract. Promotions at retailers are an effective marketing instrument, driving customers to stores, but their demand is particularly challenging to forecast due to limited historical data. Previous studies have proposed and evaluated different promotion forecasting methods at product level, such as linear regression methods and random trees. However, there is a lack of unified overview of the performance of the different methods due to differences in modeling choices and evaluation conditions across the literature. This paper adds to the methods the class of emerging techniques, based on ensembles of decision trees, and provides a comprehensive comparison of different methods on data from a Danish discount grocery chain for forecasting chain-level daily product demand during promotions with a four-week horizon. The evaluation shows that ensembles of decision trees are more accurate than methods such as penalized linear regression and regression trees, and that the ensembles of decision trees benefit from pooling and feature engineering.

Keywords: grocery retail, promotion forecasting, machine learning.

1 Introduction

Grocery retail supply chains are facing pressure on margins because of the tougher competition, more demanding customers, and increasing focus on reduction of food waste. Managing the supply chain efficiently is important, particularly for perishables, as a mismatch of supply and demand leads to lost profit due to lost sales, markdowns or waste.

Forecasts at the product level are crucial for the alignment of supply and demand in the retail supply chain to ensure a smooth and timely flow of products. They are needed by manufacturers for planning of capacity and materials, and by retailers as input to the replenishment process at both stores and distribution centers. However, forecasting at the product level is challenging due to several characteristics of the retail environment, such as stockouts, intermittency and promotions [1]. Promotions' demand is particularly challenging to forecast due to the often-limited history of similar promotions. Research

has shown that stockouts are more frequent during periods of promotion [2-3], cementing that they pose a challenge for the supply chain.

Several researchers [4-6] have dealt with the challenge of forecasting promotions at product level. One of the widely used methods in retail, especially as a benchmark in the literature to justify the use of more complex methods, is the base-time-lift approach. This approach uses an estimate of the regular demand and adjusts for promotions using a multiplicative factor, the "lift" [1]. Different methods that achieve greater predictive performance than the base-time-lift have been proposed, ranging from linear to non-linear methods. However, the evaluation conditions for the proposed methods differ in terms of aggregation level, forecast horizon and available data, which are context specific. The methods also differ in terms of: (1) model scope, i.e. whether the model deals with forecasting of both regular and promotion demand, or promotion demand only, (2) the level of pooling, i.e. whether multiple SKUs at a given aggregation level are included in the same model, and (3) the variables constructed given the available data, also known as feature engineering. Most of the previous research only compares the proposed methods to the simple baseline method and thus it is unclear how the methods compare to each other under different conditions and modelling choices. This paper adds to the previously evaluated methods in literature a class of emerging techniques, based on ensembles of decision trees, and aims to provide a comprehensive comparison of different promotion forecasting methods, under different levels of pooling and feature engineering.

2 Background

The previous work in the area of product level promotion forecasting differs widely in their approaches and evaluation conditions.

A number of papers have evaluated ordinary least squares (OLS) linear regression models. Foekens et al. [7] examined forecasting accuracy for weekly product demand at different aggregation and pooling levels using the log-linear SCAN*PRO model. The best performing model at both chain and market level in terms of mean absolute percentage error (MAPE) and median relative absolute error was the chain-specific store-level SCAN*PRO model without weekly seasonality indicators. However, they did not include any other models in the comparison. Cooper et al. [4] presented a linear regression based model formulation for forecasting promotional product demand at store level. Information on price, advertising, display conditions, major events and historical performance of promotions were included. Additionally, products were categorized into slow movers or fast movers, while promotion events were categorized based on their duration with one model created for each combination of product and duration category. Thus, pooling was conducted with regards to stores and items. The authors find that the model is superior to using historical averages of matching display and advertising conditions in terms of forecast error measured in cases. Van Donselaar et al. [8] presented a linear-regression model for forecasting the lift factor at chain and product level including information on price, competitive information, advertising, display, baseline sales, weight and shelf life. They compare pooling at category level and

one model for all categories. They find that the best pooling level differs by category and that the regression model outperforms a moving average of historical lift factors in terms of MAPE and root mean square error (RMSE). Huang et al. [9] examined forecasting weekly product demand at chain level considering competitive information using an autoregressive distributed lag (ADL) model, and not using pooling. They found that the ADL model incorporating competitive information is more accurate than the base-time-lift method using a variety of error measures, including mean absolute error (MAE) and MAPE.

A number of recent papers have dealt with more modern statistical or machine learning approaches, including penalized linear regression, decision tree models, support vector machines and neural networks. Ma et al. [6] considered the use of competitive information using a multi-stage LASSO model with an ADL formulation for forecasting weekly product demand at store level without the use of pooling. They found that the model improves upon base-time-lift and that including competitive information outside the product category only helps marginally in terms of a variety of error measures.

Gür Ali et al. [5] compared a variety of methods for forecasting weekly product demand at store level, including linear regression, support vector machines and regression trees, using MAE. Different pooling schemes were considered: one model for all observations, pooling by store and pooling by subgroup. Compared to the previously mentioned papers, they use a more data-driven approach with extensive feature engineering as is often seen in the machine learning community. They found that regression trees outperform the other models considerably, particularly during promotions. The best pooling scheme for the regression tree was one model for all observations, and the feature engineering improved forecast accuracy. A later study by the same lead author compared the aforementioned models as well as penalized linear regression and neural networks for forecasting both weekly and daily product demand at store level using MAE and MASE [10]. Pooling was conducted at subcategory level, as well as more extensive feature engineering. They found that regression trees and penalized linear regression show similar performance for weekly forecasting, whereas penalized linear regression is best at daily forecasting.

In addition to the published research, a retail forecasting competition was held by Ecuadorian grocery retail chain Corporacion Favorita. The challenge given was to forecast weekly product sales at store level given transaction data including historical sales and promotion indicators, but not price. The top five performers of the competitions used gradient boosted decision trees, neural networks or a combination of both.

To sum up, most of the studies conducted only evaluate their proposed models against simple baselines. In addition, it is difficult to get a unified overview due to differences in modeling choices and evaluation conditions. It is therefore unclear how these proposed models stack up against each other. Two exceptions are the studies conducted by Gür Ali et al. [5] and Gür Ali [10], which compare several different techniques of varying complexity. However, it remains unclear how large the performance gap is between regression trees and penalized linear regression, and under which conditions one outperforms the other. In addition, the studies do not include recent advances in the field of machine learning such as random forest [11] and gradient boosted decision trees (e.g. [12]), which have shown great promise in forecasting competitions and are

widely used by machine learning practitioners. In this study, a comparison of the different methods presented in literature will be conducted, including recent developments within the area of machine learning. The aim is to provide further evidence as to which models are most accurate for product level retail forecasting, and thereby also contribute to the question of whether non-linear models, specifically decision tree-based models, are superior to linear models and warrant the added complexity.

3 Method

The purpose of the paper is to compare and evaluate the main methods used for forecasting demand during promotion events at a daily product level. More specifically, the regression-based methods - ordinary least squares linear regression, penalized linear regression using LASSO and regression trees, are compared to modern machine learning methods based on ensembles of decision trees - random forest [11] and XGBoost [12], which are non-linear. In addition, the historical average is used as benchmark. The historical average method simply forecasts the historical average under matching conditions of price and advertising, with price as the dominant condition in case of no full match. The fallback forecast in case of no match is simply a naïve forecast of the last promotion. Base-time-lift is not included as a benchmark, as it is not possible to use it for items not sold outside promotion periods.

The method comparison is conducted on promotional sales of fresh meat and fish from a large Danish discount grocery chain, as these items present a major challenge due to their perishable nature. A chain level forecasts four-weeks prior to a promotion are sent to the suppliers for creation of a shared plan for meeting promotion demand. Promotions primarily have a duration of one week, and a forecast is required at the daily level for the promotional period. The case company uses a relatively simple promotion strategy based on price discounts advertised in a weekly flyer and occasionally on TV & radio. Most products are promoted at two or three price points. The data available consists of aggregated POS data, product master data and promotion master data, including price and advertising information, for the period of January 2015 to November 2018. Information on display conditions were not available in the case company databases. Only data from promotional periods are used for fitting the models. The data is split into training, validation and test sets, where five weeks are used as the validation set to tune hyperparameters, and twenty weeks are used as the test set. A total of 152 SKUs are present in the test dataset, with 48 of the SKUs having less than five promotions in the training and validation period. These items have therefore not been possible to forecast using item level models, although the use of pooling allows for forecasting new products with few or no observations. We evaluate the forecasting accuracy on these SKUs separately to illuminate which method performs best for products with short promotion history and what accuracy can be achieved.

The methods are compared using the basic data set, and with feature engineering. The features constructed include competitive intensity information, historical averages of sales by product, category, promotion conditions etc., similar to the work of Gür Ali [5, 10]. The feature engineering is not included for simple OLS regression, as it does

not have any built-in variable selection method and hence will overfit. Instead, a model formulation similar to that of the SCAN*PRO model is used. In addition, we consider the models at various levels of pooling:

- Full pooling, i.e. a single model used for forecasting all products
- Category pooling i.e. one model per category
- Subcategory pooling, i.e. one model per subcategory
- No pooling, i.e. one model per product.

The simple OLS model is considered with no pooling and subcategory pooling as any higher pooling level is likely to produce bias. The decision tree-based models are evaluated with full pooling and with category pooling, since these methods generally need a larger sample size to be effective. This is in line with Gür Ali et al. [5] that found that including all observations for the regression tree improved performance significantly. Table 1 summarizes the models considered in the evaluation. The methods are evaluated using time series cross-validation also known as rolling origin evaluation [13]. At each time step of one week the model is fitted and a four-week ahead forecast is created for the promotion event. For the hyperparameter tuning, the model is not refitted due to the large computational demands. The forecast accuracy is evaluated in terms of forecast error magnitude and bias using both scale-dependent and scale-independent measures. The volume weighted MAPE (WMAPE) is chosen as it is scale-independent and stable with zero values, while the RMSE is chosen as it is widely used. The mean error (ME) and a mean-scaled version are used to evaluate forecast error bias. The evaluation is carried out in the statistical computing language R [14].

Table 1. Models considered in the evaluation including dataset used and pooling strategy. N - evaluated using basic dataset, B - evaluated using both basic dataset and with feature engineering.

	No pooling	Subcategory pooling	Category pooling	Full pooling
Simple OLS	N	N		
LASSO	N	B	B	B
Regression tree			B	B
Random Forest			B	B
XGBoost			B	B

4 Results

The forecasting accuracy of the best combination of dataset preparation and pooling for each of the evaluated methods for SKUs with at least five historical promotions can be seen in Table 2, whereas the forecast accuracy for newly introduced SKUs are presented in Table 3. From Table 2, it is clear that the best method for SKUs with historical information in terms of all measures of error magnitude is XGBoost with category-level pooling and feature engineering. The results are in general dominated by XGBoost and random forest, but XGBoost is slightly biased, whereas the random forest has lower bias at a very small decrease in accuracy. The regression trees and OLS models perform

significantly worse, not managing to beat out the historical average method. In the middle of the performance spectrum lies LASSO, which is best without feature engineering and at the subcategory level.

From Table 3, it is clear that the LASSO is the best performer for new SKUs, with no feature engineering and category level pooling coming out on top. The WMAPE and bias is much higher for the new items in general, with the WMAPE of 33.1% for the best model, compared to 16.2% for the SKUs with longer history of promotions.

Overall, the results show that the modern machine learning approaches outperform the historical averages, OLS, LASSO and regression trees given that there is more than five historical promotions available for the SKUs being forecasted. In addition, feature engineering and either category or full pooling improves the performance of these methods.

Table 2. Best forecast accuracy for each method on SKUs with historical information.

Method	Dataset	Pooling	WMAPE	RMSE	ME	Scaled ME
XGB	With	Category	0.162	1488	363	0.077
RF	With	Full	0.168	1578	174	0.037
LASSO	Without	Subcategory	0.182	1702	160	0.034
Hist. Avg.	Without	None	0.189	1692	164	0.036
RT	With	Category	0.229	2076	90	0.019
LM	Without	None	0.249	2601	395	0.084

Table 3. Best forecast accuracy for each method on SKUs with few historical promotions.

Method	Dataset	Pooling	WMAPE	RMSE	ME	Scaled ME
LASSO	Without	Category	0.331	522	118	0.125
RF	Without	Full	0.399	588	-129	-0.137
XGB	With	Full	0.420	631	-60	-0.063
RT	Without	Category	0.451	716	-112	-0.119

5 Discussion

The results are impacted by both the forecasting conditions and the modeling choices used in the evaluation. We hypothesize that the modern machine learning methods outperform linear models in situations with strong patterns, interaction effects and a large relevant sample. The daily aggregation level has the effect of increasing the sample size, but also presents more noise than at the weekly level, whereas the chain aggregation level has the opposite effect. Our findings indicate that at this particular aggregation level the sample size and pattern strength is large enough to make the machine learning models superior to linear models. It is difficult to compare the findings to the findings of Gür Ali [10], as they look at forecasting store level demand and do not

include ensembles of decision trees. Contrary to previous research, our results show that for this case, linear regression and regression trees are surpassed by a simple benchmark method: historical averages under matching conditions. This benchmark method has to our knowledge only been used by Cooper et. al. [4], but its strong performance suggests that it should be considered when evaluating forecast accuracy for promotions.

Pooling in general proved useful, as it improves performance for all models except for OLS. This underlines that patterns exist across products, and that the models can effectively use these. The linear models seem to benefit less from pooling, and subcategory level seems to be the best trade-off between bias and variance in coefficient estimates, whereas the decision tree-based models favor more pooling, even performing well with one model for all fresh meat and fish products. This is likely due to the nature of decision trees, as they can effectively choose between pooling and no pooling where appropriate. The feature engineering conducted benefitted the decision tree-based models and led to greater forecast accuracy, particularly for the random forest model. This was not the case for the linear models, which could be due to non-linear relationships between the created variables or high correlations. It is therefore plausible that feature engineering aimed specifically at linear models would have improved their performance. However, this would also demand greater effort on feature engineering caused by the more restrictive nature of linear models.

For the forecasting of products with limited demand history, the results indicate that while the models can provide forecasts for these products, they are not very accurate and it is likely that a judgmental forecast by a category manager can provide better or at least similar accuracy.

6 Conclusion

The comparison of methods for forecasting product level promotion demand found that modern machine learning methods in the form of ensembles of decision trees outperform previously proposed methods such as linear regression, penalized linear regression and regression trees for SKUs with more than five historical promotions on the task of forecasting chain-level daily demand. For SKUs with less promotion history, penalized linear regression is the best performer, although all of the methods have relatively high forecast errors. In addition, the comparison found that the ensembles of decision trees benefit from both feature engineering and pooling, either in the form of category level or full pooling, whereas the linear models perform better without feature engineering at subcategory level. The implications of these findings are that ensembles of decision trees should be considered candidates for forecasting product level promotion demand at daily chain-level. An interesting area for further research is whether this also holds at weekly chain level, where the sample size is reduced by a factor of seven, or at weekly store level, where the sample size is increased significantly at the expense of much greater noise. Limitations of the study include that the results are based on one case only with one particular forecast horizon. A shorter horizon could potentially change the results, as lagged sales thus becomes a valuable form of information not currently considered. In addition, we only use promotional data to fit the model, and it

is possible that the linear models particularly can benefit from non-promotional data to obtain better estimates of seasonality, making this a topic worthy of further research.

References

1. Fildes, R., Ma S., Kolassa, S.: Retail forecasting: Research and practice. Working Paper. Lancaster University Management School (2018). doi: 10.13140/RG.2.2.17747.22565
2. Taylor, J., Fawcett, S.: Retail on-shelf performance of advertised items: An assessment of supply chain effectiveness at the point of purchase. Journal of Business Logistics. 22, 73 – 89 (2001). doi: 10.1002/j.2158-1592.2001.tb00160.x.
3. Corsten, D., Gruen, T.: Desperately seeking shelf availability: An examination of the extent, the causes, and the efforts to address retail out-of-stocks. International Journal of Retail & Distribution Management. 31, 12, 605 – 617 (2003). doi: 31. 10.1108/09590550310507731
4. Cooper, L.G., Baron, P., Levy, W., Swisher, M., Gogos, P.: PromoCast™: A new forecasting method for promotion planning. Marketing Science. 18, 3, 301-316 (1999). doi:10.1287/mksc.18.3.301
5. Gür Ali, Ö., Sayin, S., Van Woensel, T., Fransoo, J.: SKU demand forecasting in the presence of promotions. Expert Systems with Applications. 36, 12340-12348 (2009). doi:10.1016/j.eswa.2009.04.052
6. Ma, S., Fildes, R., Huang, T.: Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. European Journal of Operational Research. 249, 245-257 (2016). doi: 10.1016/j.ejor.2015.08.029.
7. Foekens, W., Leeflang, P., Wittink, D.: A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation. International Journal of Forecasting. 10, 245-261 (1994). doi: https://doi.org/10.1016/0169-2070(94)90005-1
8. Van Donselaar, K. H., Peters, J., de Jong, A., Broekmeulen, R. A. C. M.: Analysis and forecasting of demand during promotions for perishable items. International Journal of Production Economics, 172, 65-75 (2016). doi: 10.1016/j.ijpe.2015.10.022
9. Huang, T., Fildes, R., Soopramanien, D.: The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. European Journal of Operational Research, 237, 738-748 (2014). doi:10.1016/j.ejor.2014.02.022
10. Gür Ali, Ö.: Driver moderator method for retail sales prediction. International Journal of Information Technology and Decision Making. 12, 1-26 (2013). doi: 10.1142/S0219622013500363.
11. Breiman, L.: Random forests. Mach. Learn. 45, 1, 5-32 (2001). doi: https://doi.org/10.1023/A:1010933404324
12. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. ACM, New York (2016). doi: 10.1145/2939672.2939785
13. Tashman, L.: Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting. 16, 437-450 (2000). doi: 10.1016/S0169-2070(00)00065-0
14. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2018). https://www.R-project.org/.