PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wroclaw University of Economics and Business

Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu
Research Papers of Wroclaw University of Economics nad Business
Year 2023, Vol. 67, No. 2
ISSN 2392-0041

A System for Filling Store Displays: Pitting a Single Model against a Set of Demand Forecasting Models

Artur Myna

Maria Curie-Skłodowska University in Lublin

e-mail: artur.myna@mail.umcs.pl

ORCID: 0000-0002-2089-6604

Jacek Myna

SAS Institute

e-mail: jacekmyna@gmail.com

Quote as: Myna, A., and Myna, J. (2023). A system for filling store displays: Pitting a single model against a set of demand forecasting models. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, *67*(2).

DOI: 10.15611/pn.2023.2.09
JEL Classification: C53, C52

Abstract: The aim of the paper was to develop the concept of retail display space allocation as a system and to assess the quality of very slow-moving products demand forecasting models (that have not yet been used by retail companies in Poland) as its key subsystem. Forecasts were made using the example of a clothing company. The quality of these models was assessed using the Weighted Mean Absolute Percentage Error. The first step was to build the individual models. Later, the authors built separate models for brick-and-mortar and online stores as well as brands, creating a set of six models. The findings show that the classification approach for very slow movers provides as precise results as the regression approach. No single model or set of models (built with a particular machine learning method) could be identified that made the best demand forecasts for brick-and-mortar stores, as statistical tests generally did not confirm the significance of the differences between the median forecasts.

Keywords: Extreme Gradient Boosting, logistic regression, random forest.

1. Introduction

Retail networks include stores that are diverse in terms of location, with which comes also spatial variation in demand and supply. Such chains use statistical methods, e.g. trend analysis and linear regression, as well as the knowledge, experience and intuition of internal and external experts to forecast demand. Forecasting is often based on Excel (Khan et al., 2020) and the average of past demand, a moving average or a weighted moving average (Ren et al., 2017), in which little weight is given to demand from periods in the distant past. Forecasts made with traditional techniques and statistical methods do not satisfy store managers, as they do not match local demand (Kilimci et al.,

2019). Consequently, turnover is lower than it could be if the supply of products matched the demand patterns present in a given location and time (Kharfan, Chan, and Efendigil, 2020), the efficiency of the retail chain decreases, and storage costs and losses increase.

However, the use of forecasting supported by machine learning (ML) models is often in its nascent stages among clothing and other retail chains (Erhard and Bug, 2016). Even if they sell their products online or handle customer cards, they generally do not use machine learning demand forecasting to fill their store displays or hire data analysts for this purpose. In Poland, the first retail companies are introducing machine learning modules in their systems. Most are just getting ready to take this step, while facing challenges related to data collection and processing (data quality issues, the steps that occur before ML implementation). The aim of this study was to develop the concept of retail display space allocation as a system, and to assess the quality of very slow-moving products demand forecasting models (that have not yet been used by retail companies in Poland) as its key subsystem. Forecasts were made using the example of a clothing company (henceforth the Apparel Company).

In the literature, forecasting demand often concerns fast-moving products, and Extreme Gradient Boosting forecasts have low errors. The demand forecasting models for slow-moving products are generally regression models. The empirical studies lack in-depth research to forecast the demand for very slow-movers (Efendigi and Cameron, 2021). Our study used standard ML models to forecast the demand for clothing, very slow-moving products (i.e. real sales of a single item at a particular store once every ten days). It contributes to the literature by showing that the classification approach (for very slow movers) makes as precise results as the regression models. Furthermore, ML models offered by leading vendors are standardised as much as possible, with building generally one model for the entire dataset. The contribution of this study is building different models for differently rotating types of stores and products**.** This type of customisation is not popular due to the fact that open-source products are still dominated by license-based solutions.

The store display systems could not be designed with using traditional information infrastructure and traditional forecasting techniques or methods (Mohana and Saranya, 2020; Seyedan and Mafakheri, 2020). Given the low product turnover – which in turn establishes no clear trend in product sales – and the absence of autocorrelation in the random variable, the Extreme Gradient Boosting regression model was used to forecast demand. Random Forest and Classified Logistic Regression models were also trained, validated, and tested to determine whether a given volume of a given product would sell at a given location. The quality of the forecasting models was evaluated both for the entire sales network (of brick-and-mortar and online stores), and a specific month, and at the 'intersection' of store, product, and product size.

The authors used Weighted Mean Absolute Percentage Error (WMAPE), a measure for which the forecast error weights for individual products are proportional to their contribution to the total forecast value (Chase, 2013; Louhichi, Jacquet, and Butault, 2012), so that errors for less popular products are given less weight than those sold frequently. In contrast, the literature generally uses an unweighted MAPE measure (e.g. Singh, *Booma, Eaganathan,* 2020), often without specifying the level of granularity of the forecast, which makes it difficult to compare the quality of forecasting models.

2. Theoretical background

CART (classification and regression trees) are decision trees that set data partitioning rules, where *n* observations from a training set are located at a tree node, then partitions the data into *n* disjoint subsets and determines the predicted value as the mean in the leaf node. One weakness of the tree structure is the instability of the estimator in response to small changes in the data, whereas random forests combine multiple unstable decision trees into one stable model with highly accurate predictions. The probability of a classification error by a random forest increases as the correlation coefficient between trees increases, and decreases as the classification quality of a single tree

increases (Breiman, 2001). The random forest method therefore entails the construction of trees that are weakly dependent in relation to one another. It uses random sampling of observations following a uniform distribution, with each training pseudo-sample drawing (with returns) *n* observations (or fewer if the model overfits the data – the so-called subsample method). In random forest trees, unlike in the case of boosting algorithms, node splitting is based on *m* attributes randomly selected from *p*; thus, *p* – *m* attributes are not taken into account. The predicted values are an average of the prediction results of all trees.

Gradient Boosting creates a predictive model in the form of a 'committee' of decision trees. This machine learning method iteratively combines multiple low-quality models into a single high-quality predictive model.

Algorithm Gradient Boosting:

$$F_0(x) = arg\,min_\gamma \sum_{i=1}^{N} \Psi(y_i, \gamma). \tag{1}$$

For *m* = 1 to *M* do:

$$\bar{y}_{im} = -\left[\frac{\partial \Psi\, y_i, f(x_i))}{\partial f(x_i)}\right]_{F(x_i)= F_{m-1}(x)}, i = 1, N$$

$$\{R_{lm}\}_1^L = L - \text{terminal node tree } (\{\bar{y}_{im}, x_i\}_1^N)$$

$$\gamma_{lm} = arg\,min_\gamma \sum_{x_{i\,\epsilon\, R_{lm}}} \Psi(y_i, F_{m-1}(x_i) + \gamma)$$

$$F_m(x) = F_{m-1}(x) + v\,\gamma_{lm} 1\,(x \in R_{lm})$$

End for

where: $\Psi(y_i, F(x_i))$ – differentiable loss function, *M* – number of trees, *N* – number of observations, $\bar{y}_{im}$ – pseudo-residuals, $\{R_{lm}\}_1^L$ – set of terminal nodes, *v* – learning rate, γ_{lm} – values predicted by the m^{th} pseudo-residual tree, $F_m(x)$ – values predicted by m^{th} predictive model.

In the algorithm (Gradient_TreeBoost), an important step is to calculate values γ_{lm} predicted by the pseudo-residual tree. For the loss function described here, for leaves containing only one element the value returned by the tree is identical; if several pseudo-residual values have entered the hypercube, then the predicted value is their arithmetic mean. In the last iteration of the loop, the algorithm creates m^{th} predictive model based on the model with number *m* – 1 and values γ_{lm} predicted by the pseudo-residual tree, multiplied by learning constant *v*. Introducing randomness to the data improves the accuracy and reduces the execution time of the algorithm, in which each model is trained on a subsample randomly selected from the training set. The approximation made by the Extreme Gradient Boosting method is more accurate than that of the basic version of the algorithm. In order to find a better path for identifying the minimal value of function Ψ, the second derivative of the loss function is calculated as $\Psi(y_i, F(x_i))$. The second difference lies in the objective function, which adds a regularisation component to the loss calculated on the training or validation set, which allows the model to be more generalisable.

Logistic regression, on the other hand, models the probability that an explanatory variable belongs to individual classes (Hastie et al. 2001). In linear regression modeling, probability values can be negative or greater than 1; conversely, when using function $logit(v) = \ln\left(\frac{v}{1-v}\right)$, they are bounded by interval [0, 1]. The logistic regression model assumes the following form:

$$ln\frac{\hat{p}(1|x)}{1-\hat{p}(1|x)} = \beta_0 + \beta_1^T x, \quad (2)$$

where: $\hat{p}(1|x)$ – estimator of probability that the dependent variable belongs to the class 1, β_0 – intercept from the linear regression equation, $\beta_1^T x$ – regression coefficient multiplied by some value of the predictor.

The following explanatory variables were taken into account: month, weekday, trading day, quarterly total sales in the store, quarterly total sales of the given product in the store.

The expression $\frac{p}{1-p}$, known as *odds*, is the ratio of the probability of success to the probability of failure, and ranges from zero to infinity. The logit function transforms the probability into a logarithm of the odds. By transforming the model, one obtains the following estimators of the probability of observations belonging to classes 1 and 0 in a binary classification.

3. Conceptual assumptions and framework

Identical data structures across three brands of the Apparel Company made it possible to develop a retail display space allocation system for all three (Figure1) in which the forecast module (demand forecast) plays a key role. Stores (showrooms) have limited capacity, so the system must determine what volume of a particular product should be in any given store of the network on any given day in order to minimise the risk of shortages of the product as well as of overstocking the store with it when the product is characterised by low turnover (the rate at which it disappears from store shelves). Therefore, the first assumption was that experts employed by the Apparel Company would determine the minimum *turnover threshold* as a criterion for the selection of products for which the system would make predictions. The choice of whether to fill store displays with products whose turnover does not exceed the specified threshold was left to the discretion of these experts.

The introduction of new products and the withdrawal of some products from the displays was identified as a problem. Thus, a second assumption underlying the study was that forecasting would include products with a sales history of at least six months, while products with a shorter history would fill the displays based on expert recommendation. For stores that had been open for less than three months, it was assumed that demand forecasts would be formulated on the basis of models analogous to a 'similar' store identified by the expert, and for stores that had been open for more than three months but less than a year, it was assumed that store displays would be filled based on the 'similar' store identified by the expert. Another assumption was that stores are restocked and introduce new products on a daily basis.

The Apparel Company formulated the requirement that expert adjustments could be made when it disagreed with the recommendation that the system generated. It was therefore assumed that expert rules would determine which products should be in specific stores regardless of the outcome of the forecast, which products could not be in specific stores (in which case the Apparel Company would determine their availability in a given store of the network), and which were "acceptable but not necessary" for specific stores (in which case the system would determine whether to make a specific product available in a given store). After combining the demand forecast (the first module of a store display filling system) with the expert rules, a ranking of products that should be found in each store (the *product range* as the second module) was created.

The third module of the system was to create an allocation matrix in which there is such a volume of a particular product in the store's stock at the intersection of the row (store identifier) and the column (product identifier × size identifier) that the product will not run out. The schedule of product deliveries from the warehouse, which are made once a day, between one and five days a week, changes once a week. This varies depending on the area (capacity) of the store, thus creating a potential challenge.

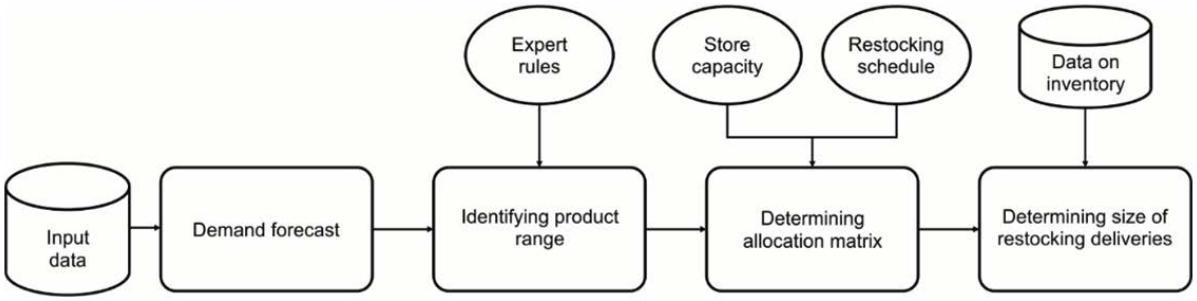

Fig. 1. Conceptual framework for a store display filling system

Source: own work.

This problem is solved by an optimisation algorithm that applies a store capacity constraint filter to the product. After taking into account the next delivery date, the optimisation module calculates how many units of each product should be in a particular store. The last module calculates the difference between the number of units of each product determined by the system and the number of units that are visible in store displays. One complication here is the differentiation of each product by year of production. In accordance with the preferences of the Apparel Company, the system was designed so that filling the display runs in the order from the 'longest-lingering' product that needs restocking to the shortest.

The optimiser objective function is to maximise stock for each product with constraints such as delivery schedule, product availability at the warehouse and expert rules. The decision-making process is based on the optimiser that solves a system of many linear equations.

Demand forecasting and evaluations of the product range were run once a month, the allocation matrix was revised once a week, and delivery size estimations were made daily. The configuration of the system allows each module to be run separately.

4. Methodology for learning and validating predictive models

Month was used to establish the basic unit of time in dividing the data into training, validation, and testing data. The models were tested using two months with low sales and two with high sales. A validation set was created from the data for December 2019, which was the last month of high sales volume. Two separate test sets were built for January and February 2020, so that the months from the two years' worth of sales prior to December 2019 could serve as the training set. As the research design entailed that the models were to be used for two high sales months, August 2019 was chosen as the second month on which the models would be tested. The models could not be taught when based on data for the same period as the one for which they were tested, so August 2019 was removed from the training set and August 2017 data were added in its place. Each month occurred in the training set for two different years, so to avoid over-training the models, data from September, October, and November 2017 were not included.

The training set contained 32 million observations, while the test and validation sets ranged from nearly 240,000 (February 2020) to over 485,000 observations (August 2019). Given the identical data structure for the three Apparel Company brands, an assumption was formulated to build a single supergroup prediction model.

The first models were built using the Extreme Gradient Boosting method implemented in the Python language using the *xgboost* library. The regression models were built using *XGBRegressor*. The following parameters were modified in the course of tuning the model: *tree method*, *max depth*, *base score, subsample, objective function*, *learning rate,* and number of trees (*n estimators*). The process of training the model took less than an hour, while the prediction less than 30 seconds.

The parameter of the *early stopping* method was set to ten iterations. The following objective functions were tested (Hastie et al., 2001):

- Root Mean Square Error (*RMSE*),
- Mean Absolute Percentage Error (*MAPE*),
- Weighted Mean Absolute Percentage Error (*WMAPE*)

$$WMAPE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{\sum_{i=1}^{n} y_i} \cdot 100, \tag{3}$$

- and Forecast/Actual ratio, a custom measure defined as the proportion (quotient) of the sum of the forecast values and the sum of the empirical values:

$$\frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} y_i}, \tag{4}$$

where *n* is number of observations, $\hat{y}_i$ is the forecast value for i^{th} observation, and y_i the empirical value for i^{th} observation.

The methodological approach of building a single Extreme Gradient Boosting model (and then a set of two and six models) was used in building the random forest and logistic regression models. The random forest models were built in classification form because in the test sets, less than 1% of the rows contained values of the explanatory variable greater than 1. As a result, the classification was reduced to a binary form indicating whether at least one item of a particular product of a particular size would sell at a particular store on a particular day. The study used a random forest implementation from the *scikit-learn* library. The parameters *max depth*, *complexity parameter (ccp alpha)*, *min samples leaf* were set in such a way that the models did not overfit the training data. The models were tuned using the parameters *class weight*, *n trees*, *max depth*, *complexity parameter*, and *min samples leaf.* As the random forests were being built, the authors used the configuration *class_weight='balanced,'* in which each observation is given a weight that is inversely proportional to the class to which it belongs.

In light of the inadequate performance of the random forest models, logistic regression models were then built, which often achieve good results for large datasets. Once again, an implementation from the *scikit-learn* library was used. *Liblinear*, a linear classification algorithm for large datasets, was chosen as the *solver, the* key optimisation parameter. The classes were weighted in the same way as in the random forest model. During the learning process, in order to avoid overfitting the model to the training data, the regularisation factor L_2 was applied. Regularisation L_2, also called the Ridge regression, is the method of reducing overfitting the model to the training data.

Parameter $C = \frac{1}{\lambda}$ was used to set regularisation parameter λ, the Ridge regression penalty for overfitting the model.

The quality of the forecasting models was tested on a set of 355 stores and 48 products (shirt types), selected so that each intersection was present in each test set. The forecasted and empirical values were added up to compare demand forecasts with actual sales. At low levels of granularity (STORE x MONTH and PRODUCT x MONTH), the authors analysed WMAPE, along with the difference between predicted and empirical values $(\hat{y}_i - y_i)$, which was called *forecast-actual*. At high levels of granularity (STORE × PRODUCT × SIZE × WEEK, STORE × PRODUCT × WEEK, STORE × PRODUCT × SIZE × MONTH, and STORE × PRODUCT × MONTH), the study used only the *WMAPE* measure.

A non-parametric median test (Mood's test) was conducted to compare *c* samples, where $c \geq 2$ (Aczel and Sounderpandian, 2008). The authors formulated two hypotheses:

H_0: the medians of all *c* samples are equal,
H_1: not all medians of *c* of trials are equal, where *c* is the number of trials in the test.

The results of the test (which was chosen because the samples were not normally distributed), would determine whether the null hypothesis could be rejected.

The analytical part of the paper was developed mainly in Python due to its rich set of libraries. The Python libraries *XGBoost*, *Scikit-learn*, *SciPy*, and *Matplotlib* were used to build, validate, assess the quality of, and visualise the learning results.

5. Results

At high levels of granularity, the lowest forecast error for a single Extreme Gradient Boosting model was recorded for December 2019 for the validation set, while the highest was registered for two months with low sales totals: January and February 2020. To improve the quality of the forecasts, a set of two Extreme Gradient Boosting models was created: one for online stores and one for the stationary stores. For the ABC and GHI online stores, the demand forecasts in the validation set exceeded actual sales (by almost 300 shirts), while for the DEF online store, they underpredicted it by almost 250 shirts (in the context of a low share of DEF records in the training set structure). Mood's test for December 2019 (at 0.05 significance level) showed no significant difference between the median forecasts of a single model and the ensemble of two Extreme Gradient Boosting models.

Given the undertuning of models for online stores and the lack of improvement in forecast quality for brick-and-mortar stores, a set of six Extreme Gradient Boosting models was built for each brand and store type (online and brick-and-mortar), which simultaneously addressed the problem of unbalanced brands in the training set. Mood's test did not show statistically significant differences between the medians of the sum predictions from the set of six models and the single Extreme Gradient Boosting model (at 0.05 significance level). However, creating the set of six models had a positive effect on the tuning of demand forecasts for the online stores, as reflected by the marked decrease in the *forecast-actual* difference for each of the three online stores examined.

The higher the level of granularity in the analysis of the results, the greater the difference in model quality in favour of Extreme Gradient Boosting. For January and February, at the STORE × PRODUCT × SIZE × MONTH level, the *WMAPE* for the random forest model was almost 70%, and almost 50% at the STORE x PRODUCT x MONTH level (Table 1). Thus, a set of six random forest models was built (splitting the training set into six parts allowed models with more trees to be trained with limited RAM) and the prediction results were compared with those of a single Extreme Gradient Boosting model. At the STORE x MONTH level, for December 2019, the differences between the two forecasts were found to be positive for half of the stores and negative for the other half. The results of Mood's test showed that, at 0.05 significance level, there was no basis to reject the null hypothesis that the medians of the forecasts were identical. At the STORE x MONTH level, for the set of six random forest models, *WMAPE* surpassed 20% for more than 120 stores (and for only 50 in the case of the Extreme Gradient Boosting model).

Table 1. WMAPE for Extreme Gradient Boosting, random forest, and logistic regression models at high levels of granularity

Level of	Month	Extreme Gradient		Boosting	Random	forest	Logistic regression	
Granularity		single	a set of	a set of	single	a set of	single	a set of
forecasts		model	2 models	6 models	model	6 models	model	6 models
		WMAPE (in %)						
STORE × PRODUCT × MONTH								
	August 2019	17.31	17.63	17.36	28.25	27.73	19.90	19.23
	December 2019	16.20	17.42	16.25	30.00	24.36	16.90	16.05
	January 2020	21.91	22.49	22.19	46.73	43.95	24.81	25.58
	February 2020	23.05	22.44	22.15	46.78	44.62	25.83	27.15

STORE x PRODUCT × SIZE × MONTH								
	August 2019	31.28	32.28	31.96	43.10	41.36	32.29	32.90
	December 2019	29.34	30.61	29.95	42.80	36.85	27.92	29.13
	January 2020	35.48	36.30	36.28	69.69	64.78	39.61	41.16
	February 2020	34.40	34.01	34.06	67.71	62.70	38.46	40.31

Source: own calculations.

For August 2019, the sum of forecasts made by the set of six random forest models totalled 103.4% of actual shirt sales, marking a substantial difference vis-à-vis the demand forecast made by the single Extreme Gradient Boosting model, which was higher by 9 p.p. For January and February 2020, the forecast was lower than actual shirt sales (by 8.0% and 5.6%, respectively). Mood's test for January and February confirmed the significance of the difference between the median forecasts of the set of six random forest models and the single Extreme Gradient Boosting model (at 0.05 significance level), while for August the test was not statistically significant. The set of six models, as well as the single random forest model, recorded the largest errors in forecasts made for stores with high sales and products with the shortest sales history. For online stores, the predictions of the set of six random forest models had higher errors than the set of the six Extreme Gradient Boosting models, although the difference was not as pronounced as for brick-and-mortar stores. Extracting separate random forest models for the online stores, however, resulted in an improved model fit to the validation set. Overall, compared to a single random forest model, the decrease in *WMAPE* was most pronounced at high levels of granularity.

With 32 million observations in the training set, the logistic regression models fit the validation set well. At the STORE × MONTH and PRODUCT × MONTH levels, for December 2019, the forecast results of the single logistic regression model and Extreme Gradient Boosting were not significantly different in terms of *forecast-actual* and *WMAPE*, although at the PRODUCT × MONTH level, the logistic regression predicted demand with slightly greater accuracy. At the STORE × MONTH level, for August 2019, the logistic regression model overestimated demand by 15.2% relative to actual sales – 3 p.p. more than Extreme Gradient Boosting. As the logistic regression model underestimated demand for online stores on the validation set to a greater degree than the single Extreme Gradient Boosting model, the errors for the three stores were significantly lower for the test sets. At the STORE × MONTH and PRODUCT × MONTH levels the online store for the DEF brand in particular stood out. For January 2020, the model underestimated demand relative to actual sales by 70 shirts for this store (an outcome shared by most DEF stores). The training set contained the most observations for the GHI brand (almost three times as many as the DEF brand, which was the least abundant), and the ABC brand had the most observations in the test set. The models thus learned to predict ABC and GHI product sales better than DEF. The effect of training joint models for the three brands was to underestimate demand forecasts for DEF brand stores and products.

Given the underestimation of demand for the DEF brand, a set of six logistic regression models was built. Analysing the *forecast-actual* differences for the entire network, it was found that the set of six models run on the validation set erred by 25 shirts for a single store at most, and *WMAPE* generally did not exceed 20% at the PRODUCT × MONTH level. At a high level of granularity for the STORE × PRODUCT × MONTH level, for December 2019, the forecast performed by the set of six logistic regression models had the lowest *WMAPE* of all models (Table 1). However, with a good fit to the validation data, the forecast results for brick-and-mortar stores were not significantly different from the single logistic regression model. In addition to the clear underestimation of demand for product #48 (which was also observed for the single logistic regression model), there was a clear overestimation of sales of products that were being withdrawn (by over 160%).

6. Discussion

Evaluating the quality of the forecasts made for the Apparel Company using the Extreme Gradient Boosting and logistic regression methods, the authors found it to be consistent with the evaluation of the demand forecasts for products offered by 78 different companies in (Chaman 2003). The average forecast error for the 78 companies' products was 12% – i.e. in the middle of the range of forecast errors made in the analysis conducted here for the Apparel Company (10-14%). At low levels of granularity (CATEGORY and its counterpart), the errors for the forecasts made for the Apparel Company were no different from the average error of the five 'best' forecasts reported by Chaman (2003). At high levels of granularity (here: January 2020), the average forecast error did not exceed 24% at the STORE × PRODUCT × MONTH level and 37% at the STORE × PRODUCT × SIZE × MONTH level. In other words, the average forecast error reported here was only 1 p.p. to 3 p.p. higher than the forecast error for a set of products offered by 78 companies (at the 'SKU' level, which corresponds to STORE × PRODUCT × SIZE × MONTH). At the same time, for sales forecasts made using the logistic regression method by Bajracharya (2010), *WMAPE* was recorded in an equally wide range (from 8% to 52%, depending on product type) as for forecasts made using this method for the Apparel Company (at the PRODUCT × MONTH level).

Fischer et al. (2018), as well as Swami et al. (2020), found that Extreme Gradient Boosting model forecasts had the lowest errors, and logistic regression provides an alternative to it in forecasting phenomena that lack nonlinear dependencies. Extreme Gradient Boosting also proved to be the best among the tested methods in forecasting GDP growth rate (Premraj, 2019). The ranking of the methods from most (Extreme Gradient Boosting) to least accurate (Random Forest) in forecasting demand for the Apparel Company was generally consistent with their rankings in the literature, and the higher the level of granularity, the more pronounced the difference in forecast quality between the two models in favour of Extreme Gradient Boosting. The study's findings show that the classification approach for very slow movers provides as precise results as the regression approach.

A significance analysis of the variables showed that the Extreme Gradient Boosting models mostly used the quarterly aggregate of sales of a specific product of a specific size and in a specific store to perform the forecast. The random forest used this key variable less frequently because node splitting in the random forest tree was performed using a random set of m out of p variables, which resulted in greater *WMAPE* errors for products with short sales histories. Additional explanatory variables increased the prediction quality of the random forest (in line with Fischer et al. 2018), with the high prediction errors of the random forest reported in our study resulting not from the number of explanatory variables, but rather to their unbalanced predictive power (one variable was much stronger than the others). The random forest can therefore be used in demand forecasting provided that the training set contains several equally 'strong' variables and several (or slightly more) other less important variables.

7. Conclusions

The reasons for the prediction errors made by the models differed between model types (single model versus a set of models). Building separate demand forecasting models (Extreme Gradient Boosting, random forest, logistic regression) for online and brick-and-mortar stores did not improve the fit of online store sales to the validation set. Demand forecasts for the DEF brand products, which were the least numerous in the training set, were found to underpredict demand, whereas for the ABC and GHI online stores the models overestimated sales, with higher forecast errors accompanying higher levels of granularity.

Thus, the authors chose to build separate models for both store types (brick-and-mortar and online) and brands, creating a set of six models. However, only online stores showed improved model fit to the validation set. In view of the variation in the quality of demand forecasts made for different brands,

the study analysed the structure of the training and test sets. It was found that, despite the differences in sales volumes between the three brands, a larger number of observations in the training set facilitated the detection of unusual phenomena (such as sales of products that were being withdrawn), regardless of the forecasting method.

It is not possible to indicate one model or set of models built with a particular machine learning method that made the best demand forecasts for brick-and-mortar stores, as statistical tests generally did not confirm the significance of the differences between the median forecasts. The sales forecasting module for the Apparel Company should consist of: a single Extreme Gradient Boosting model for brick-and-mortar stores, which to minimise errors should be learned on the full training data set, an Extreme Gradient Boosting model for the ABC online store, and two separate logistic regression models for the DEF online store and the GHI online store. In future research, the authors propose using deep neural networks as well as least squares support vector machines (LS-SVMs) in demand forecasting. These methods solve binary classification problems, thus finding a use in demand forecasting where regression is employed (e.g. LS-SVM).

References

Aczel, A., and Sounderpandian, J. (2008). *Complete business statistics.* Boston: McGraw-Hill.

Bajracharya, D. (2010). *Econometric modeling vs artificial neural networks: A sales forecasting comparison.* Borås*:* University of Borås*.*

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Chaman, J. (2003). Forecasting errors in the consumer products industry. *Journal of Business Forecasting Methods & Systems*, *22*(2), 2-4.

Chase, C. (2013). *Demand-driven forecasting: A structured approach to forecasting*. Hoboken: Wiley.

Efendigi, E., and Cameron, K. (2021). *Inventory management for slow moving and high volatility items* (Submitted to the Program in supply chain management in partial fulfillment of the requirements for the degree of master of applied science in supply chain management). Cambridge: MIT.

Erhard, J., and Bug, P. (2016). *Application of predictive analytics to sales forecasting in fashion business*. Reutlingen: Reutlingen University.

Fischer, T., Krauss, Ch., and Treichel, A. (2018). *Machine learning for time series forecasting – a simulation study*. FAU Discussion Papers in Economics, (02).

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, (38), 367-378.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Khan, M., Saqib, S., Alyas, T., Ur Rehman, A., Saeed, Y., Zeb, A., Zaree, M., and Mohamed, E. (2020). Effective demand forecasting model using business intelligence empowered with machine learning. *IEEE Access*, *8*(1), 116013-116023.

Kharfan, M., Chan, V., and Efendigil, T. F. (2020). A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. *Annals of Operations Research*, (6), 1-16.

Kilimci, Z.H., Okay, A., Akyokus, S. et al. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Complexity*, (7), 1-15.

Louhichi, K., Jacquet, F., and Butault, J. (2012). Estimating input allocation from heterogeneous data sources: A comparison of alternative estimation approaches. *Agricultural Economics Review*, *13*(2), 83-102.

Mohana, S., and Saranya, S. (2020). Sales prediction using machine learning algorithm, *International Journal of Advanced Science & Technology*, *29*(3), 1049-1055.

Premraj, P. (2019). *Forecasting GDP growth: a comprehensive comparison of employing machine learning algorithms and time series regression models*. Bergen: Norwegian School of Economics.

Ren, S., Chan, H.L., and Ram, P. (2017). A Comparative study on fashion demand forecasting models with multiple sources of uncertainty. *Annals of Operations Research*, *257*(1), 335-355.

Seyedan, M., and Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *Journal* of Big Data, (7), 1-22.

Singh, K., Booma, P. M., Eaganathan, U. (2020). E-Commerce system for sale prediction using machine learning technique. *Journal of Physics: Conference Series* 1712, 1-8.

Swami, D., Shah, A. D., and Ray, S. (2020). *Predicting future sales of retail products using machine learning*. arXiv:2008.07779.

Wong, W. K., and Guo, Z. X. (2010). A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. *International Journal of Production Economics*, *128*(2), 614-624.

System zapełnienia ekspozycji sklepowych: pojedynczy model a zespół modeli prognozowania popytu

Streszczenie: Celem artykułu jest opracowanie koncepcji zapełnienia ekspozycji sklepowych jako systemu oraz ocena jakości modeli prognozowania popytu (które w Polsce nie są jeszcze wykorzystywane przez sieci handlowe) bardzo wolno rotujących produktów jako jego kluczowego podsystemu. Jakość modeli oceniono za pomocą miary *Weighted Mean Absolute Percentage Error* na różnych poziomach szczegółowości: dla całej sieci sprzedaży i określonego miesiąca oraz na „na przecięciu" sklepu, produktu i rozmiaru produktu. Najpierw zbudowano pojedyncze modele, następnie zaś odrębne modele dla sklepów stacjonarnych i internetowych, jak również marek, tworząc zespół sześciu modeli. Poprawę dopasowania modeli osiągnięto tylko dla sklepów internetowych. Wyniki pracy wskazują, że podejście klasyfikacyjne dla bardzo wolno rotujących produktów charakteryzują równie precyzyjne wyniki prognoz jak podejście regresyjne. Nie można wskazać jednego modelu lub zespołu modeli (zbudowanego określoną metodą uczenia maszynowego), który wykonał najlepsze prognozy popytu dla sklepów stacjonarnych, gdyż istotności różnic median prognoz na ogół nie potwierdzono testami statystycznymi.

Słowa kluczowe: *Extreme Gradient Boosting*, regresja logistyczna, las losowy.