

# Towards Context Aware Sales Prediction

Indrė Žliobaitė, Jorn Bakker, Mykola Pechenizkiy  
Eindhoven University of Technology  
P.O. Box 513, NL-5600 MB, Eindhoven, the Netherlands  
zliobaite@gmail.com, {j.bakker, m.pechenizkiy}@tue.nl

## Abstract

*Sales prediction is a complex task because of large number of factors affecting the demand. We present a context aware sales prediction approach, which selects the base predictor depending on the structural properties of the historical sales. First of all we learn how to categorize the sales time series offline into four categories (“flat”, “frequent”, “occasional” and “seasonal”) based on structural features. Next for each product classified to a particular category, we online apply a prespecified base predictor. In the experimental part we show that there exist product subsets on which, using this strategy, it is possible to outperform naive methods. We also show the dependencies between product categorization accuracies and sales prediction accuracies. A case study of a food wholesaler indicates that moving average prediction can be outperformed by intelligent methods, if proper categorization is in place, which appears to be a difficult task.*

## 1 Introduction

Demand prediction is essential part of business planning. An accurate and timely sales prediction is essential for stock management. That is a crucial part for food wholesales and retail profitability. The stock includes large assortment of goods, some of them require special storage conditions, some are quickly perishable.

There are general and product specific causes of the demand fluctuation. The variations in consumer demand may be influenced by price (change), promotions, changing (rapid or gradual, global or local) consumer preferences or weather changes [15]. Furthermore, a large share of the products sold in that market is sensitive to some form of a seasonal change. Seasonal changes occur due to different cultural habits, religious holidays, fasting. All these factors imply that some types of products have high sales during a limited period of time.

Although seasonal patterns are expected, the predictive

features that define these seasons are not always directly observed. Therefore, fluctuations in sales which are accommodated by the changing seasons are often difficult to predict. Besides, the historical data is often highly imbalanced. For example, occasion specific seasonal products would have only a few weeks of the sales peaks per year.

Several success studies have been reported in the literature. However, some surveys initiated by the several large companies in the food industry indicate that the majority of food companies still suffer from poor sales predictions, which lead to losses for style goods and perishable items, lost opportunities, and decrease of the service level [1].

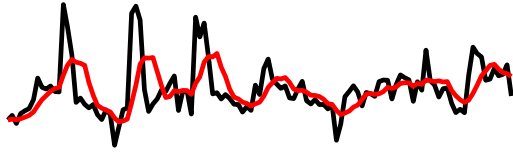
We present a context aware approach for sales prediction. For each product we classify it to one of the predefined categories based on structural properties of the sales time series. Then we apply a category specific predictor to each product.

We make a distinction between context in time and context in space. In this work, when referring to context we talk about context in space, i.e. types of time series. We leave context in time as the future work. That is the case when time series could migrate to different categories over time.

The rest of the paper is organized as follows. In Section 2 we highlight the challenges related to food sales prediction. In Section 3 we present context aware sales prediction approach. A case study of food wholesaler Sligro Food Group N.V. is carried out in Section 4. Section 5 reviews related work. Finally, Section 6 discusses future prospects and concludes.

## 2 Challenges of sales prediction

In this study we address food wholesales prediction. Still the observations and methods discussed here can be generalized to other sales prediction. Sales time series generally cannot decrease below zero and they are also bounded from the top. On the contrary to e.g. trade in financial securities, sales of fast moving goods are limited by physical capacities (outlet size, stock, working hours).



**Figure 1. Predicting with moving average as a baseline. Black - the original series, red - 6 weeks moving average.**

Variation in sales figures can be classified into short term fluctuations (e.g. a buyer went to a pub today, thus will buy groceries tomorrow), medium term seasonal patterns (e.g. June vacations) and long term trends (e.g. changing economic situation, developing new eating habits). In this study we particularly focus on spotting medium term seasonal patterns.

Different horizons of the food sales predictions are required for performing business operations (next day, next week, next month predictions). The basic approaches are often used: a moving average of different lag or simple regression models. In this setting baseline predictions are often overridden by a human. Managers use their domain expertise (knowledge about how products are doing during this time of the year, good or bad weather, school holiday period, promotion and advertisements possibilities, etc) to make the adjustment for the baseline predictions.

Weekly predictions are essential for wholesaling of food and food-related products and it is considered to be the most challenging, and accuracy is more crucial for business decisions (primarily for stock management). Thus we focus on the weekly predictions.

Predictions based on moving averages may work reasonably well when demand is close to level. When demand follows trend or seasonal patterns these methods do not perform well due to the slow reaction to the changes. A typical case of such behavior can be seen from Figure 1. Each time the sales start to rise the prediction rises a couple of weeks later and vice versa when demand decreases. The most important points in the pattern are the increase at the beginning of the season and the decrease at the end of the season. In this case the stock should be prepared for the beginning of the season and decreased at the end of the season.

These typical drawbacks of the moving average method are the main reasons that predicting is mostly verified by human expertise. Managers often try to improve the performance in seasonal peak periods by increasing the number of safety days which results in a higher than necessary stock.

However, this method only works in case the managers are experienced enough and even then it is still difficult to predict at which point in time this should be done. Another typical approach is to have a number of triggers and reminders that should hint about the coming (school, national or religious) holidays, (warm or cold, sunny or rainy) weather and other factors that influence demand of particular food products. However, human factors like frustration, overload of information, lacking expertise (especially with a new personnel or for a new set of products), and simply forgetfulness may result in mistaken predictions and poor decision making.

All these effects can be related to the issue of seasonality. Seasonality does not necessarily imply strict periodicity. We define seasonality as external time dependent factors, implying deviations of the sales patterns.

### 3 Context aware sales prediction approach

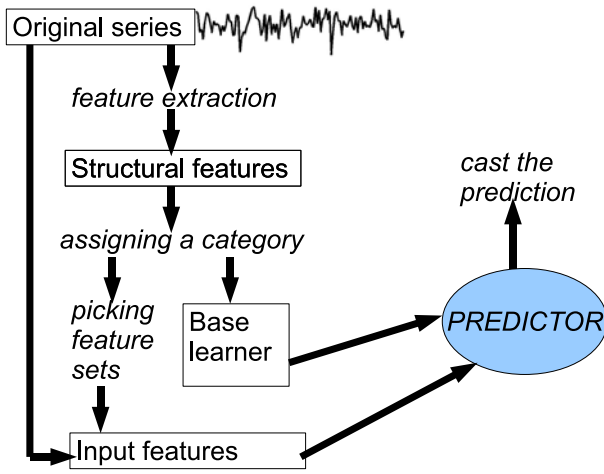
In this section we present context aware sales prediction approach (CAPA). By context we mean space dimension, i.e. categorization of individual time series. First we give the general overview and then explain individual parts of the approach.

#### 3.1 Motivation

The main idea of the context aware approach is to select the predictor based on the structural properties of time series. Different products have different sales behavior and different dependence on calendar events (seasonality). If we can identify and extract distinct categories of products, specific input data construction procedures and specific predictors could be employed for each category.

One could argue, that an ensemble approach does that automatically. All possible input features can be collected and then apply rigorous feature selection and predictor selection from an ensemble. This approach has limitations with respect to a given food sales prediction problem. First of all, the data is noisy and relatively short. For example, a particular food wholesaler company keeps only two years record in their transactional database. If a product is seasonal and peaks only once per year for a particular event, we would have only one or two positive examples in the historical data. By defining the context, we filter out a part of noise.

Secondly, some series share common patterns. For example, New Year peaks are common for a large subset of products. By categorizing the time series based on their structural properties, we narrow down the job for the particular predictor, allowing to focus on the peculiarities of a particular series.



**Figure 2. Online operation of context aware prediction approach (CAPA).**

Another related approach would be a multitask learning ([2]). Yet instead of taking a “black box” approach we use explainable judgement to filter out relevant part of the task.

### 3.2 Decision support with CAPA

In this section we present a snapshot (one time step) of the decision support approach we developed. We discuss the ingredients needed to produce one week ahead sales prediction for a single product.

CAPA consists of two blocks - training (offline) and operational (online). First of all we present operational online part and then describe, how the model is trained and parameterized offline. For now let us assume that the model has already been trained offline:

1. the categories have been fixed;
2. mapping of time series to the categories is established;
3. “local” expertise of each predictor is known.

Figure 2 presents online operation of CAPA.

Let us take a particular product we are interested to predict online. First of all we extract structural features from the original sales time series of the product. Then we assign the product to one of the categories <sup>1</sup>. We pick a base predictor specific to a particular category and select input features, relevant to a particular category. That is the context aware part of the approach. The contexts are specified by predefined categories. Having the original series, the base

<sup>1</sup>Generally we can assign a product to more than one category and use voting, but to keep the focus we delimit the mapping to one category

learner and the input features we can cast the prediction. Note, that the predictor does not use the structural features as were used for assigning the product to a category.

The prediction output now can be used for a decision making in a business process.

### 3.3 Training CAPA

The core part of context aware prediction approach is to match the product categories and the base predictors.

**Offline part.** A limited set of base predictors  $(G_1, G_2, \dots, G_m)$  needs to be preselected based on domain knowledge and expectations in order to delimit full state space search. Then for each product  $m$  parallel experiments are carried out using each of the predictors. Next, the products are grouped into  $m$  categories based on the best performing predictor. There might be empty or small categories, then we discard the corresponding predictors. Each obtained category serves as a basis for constructing categorization rules.

**Online part.** The goal of the training process is to learn to assign a product to one of these defined categories online, having only a fragment of the series. When we have the categorization rules, an unseen product can be processed as described in Figure 2. First of all the category of the product is determined, say  $C_j$ . Then the corresponding predictor  $G_j$  is used to output the prediction.

### 3.4 Structural features

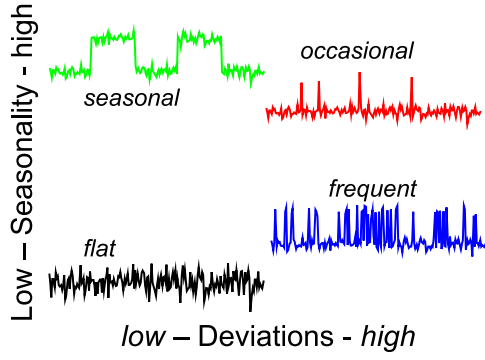
Structural features will be extracted from the input products online and they need to be length independent. In addition, we want the structural features to be related to observable seasonal patterns.

Using external knowledge and visual observations, we can categorize the series using two dimensions: seasonality and deviations (see Figure 3).

For example, bread sales can behave like “flat” series. “Frequent” series represent the products, which are bought in large quantities, following no particular seasonality. “Occasional” product sales increase sharply in relation to particular occasions, like eggs for Easter. Icecream is a “seasonal” products with respect to weather. The series in Figure 3 are artificial, for illustration purposes.

We normalize the values of the series to be in a range  $(0, 1)$  before defining structural features. We extract the following structural features:

- $F_{s1-s3}$  |mean - median|, standard deviation, shift;
- $F_{s4-s8}$  threshold  $h$  crossing ratios;
- $F_{s9-s10}$  normalized power of the frequency  $p$  in the frequency spectrum;



**Figure 3. Structural categories of product sales**

- $F_{s11-s12}$  local variation features: interquartile range and unequal neighbors.

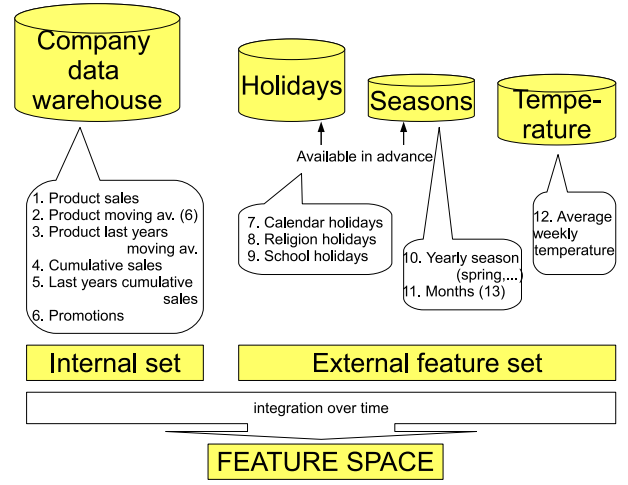
The intuition behind these structural features is related to the categories of the series we defined. “Flat” and “seasonal” series are expected to have similar mean and median, while “occasional” and “frequent” series are expected to have upward deviations and thus larger difference between the mean and the median. “Seasonal” and “occasional” series are expected to have fewer crossings of the upper threshold than the other two.

The features  $F_{s1-s3}$  capture global characteristics of the series  $y$ . Shift is the mean number of points for which  $y_t < \mu$  minus the median of the number of points for which  $y_t < \mu$ , where  $\mu$  is the mean of the time series

In order to capture the behavior of the time series, we define a number of threshold values and note the number of times the signal crosses these thresholds (features  $F_{s4-s8}$ ). This is done for the threshold values  $h = 0.5 \pm 0.1, 0.7, 0.3, 0.8, 0.2$ . This number is then scaled to the total length of the signal to obtain the ratio. This ratio is an indicator of the structural nature of the signal.

Seasonal patterns could manifest themselves as, for instance, yearly or bi-yearly changes. This information should appear in the frequency spectrum of the time series as a relative high power in the frequencies  $p = 1/52$  and  $2/52$ . In order to obtain these features we apply Fast Fourier Transformation (Cooley-Tukey implementation [3]) and extract the corresponding frequencies (features  $F_{s9-s10}$ ).

We aim to capture local variation using features  $F_{s11-s12}$ . Unequal neighbors is the mean number of times  $y_t \neq y_{t-1}$ . The interquartile range of  $y_t - y_{t-1}$  is a robust measure for the spread of variation in the signal. Any outliers that fall in the upper or lower 25% of the difference distribution will not affect it.



**Figure 4. Formation of the Feature Space**

### 3.5 Input Space Construction

In previous section we discussed extraction of structural features, which are used for time series categorization (see Figure 2). In this section we discuss the input space. These features are used by the prediction models (see Figure 2).

The feature space is formed using internal and external data. Internal data comes from a food wholesales company database, where historical sales are stored. External data (holidays, temperature, seasons) is formed using information from the ministry of culture, meteorological institute and common knowledge. The feature set is specified in Figure 4.

The internal features are interrelated. Moving average ( $F_{p2}$ ) is calculated using ( $F_{p1}$ ). Last years moving average ( $F_{p3}$ ) is formed taking the values ( $F_{p1}$ ) and using one year lag. Cumulative sales ( $F_{p4}$  and  $F_{p5}$ ) include the sales of all the products in quantity terms. We checked corresponding series in monetary sales terms to verify that there is no significant distortion due to different quantity measures. The company organizes promotions ( $F_{p6}$ ) for selected products, this can be known in advance.

The external features ( $F_{p7-p11}$ ) are available in advance. Average weekly temperature ( $F_{p12}$ ) can be predicted sufficiently accurately one-two weeks in advance. The external features add on much to dimensionality. Holidays ( $F_{p7-p9}$ ) are described in 16 binary features. Seasons ( $F_{p10}$ ) are described in 4 binary features, and months ( $F_{p11}$ ) are described in 13 binary features. Temperature ( $F_{p12}$ ) is described in a single numerical dimension.

### 3.6 Learning categorization rules

We would like to learn categorization rules, which would assign a given product to one of the defined categories based on its structural features.

We take two approaches, which we call *bottom up* and *top down*. In the first approach we use the training accuracies to label the training products and using these labels try to learn categorization in a supervised manner. In the second case we visually pick a set of representations from the four categories (“flat”, “frequent”, “occasional” and “seasonal”) defined earlier and use them as prototypes to learn the categorization.

#### 3.6.1 Bottom up categorization

In order to test whether the category assignments are learnable, we train a classifier on the “true” labels of the categories generated using the training dataset. The labels are obtained by running all the classifiers from the pool for all the products and then ranking the accuracies for each product. A product gets the label, corresponding to the best performing classifier.

If the categorization is learnable we should be able to assign classes based on the structural features, such that the intelligent predictor methods perform better than the baseline.

**Training.** In principle, we want to be able to interpret the decision boundaries, used for product categorization. For this purpose a decision tree classifier is used to make the mapping from structural features to classes. The training data is the same as in the predictor evaluation. We use the 12 structural features described in Section 3.4. Since it is unknown a-priori which features will result in the best classification, we perform a brute-force search in the structural feature space. That means that for every possible subset of features a classifier is trained and evaluated. For each of the classifiers the mean score of the intelligent predictors is used as an evaluation measure. This score is the MASE with respect to a baseline predictor (next week sales are predicted to be equal this week sales), averaged over all instances in the class assignment made by the classifier.

**Validation.** The best classifiers should be able to generalize the classification. We use cross validation on the training set to select the categorization rule. From each of the instances in the validation set we compute the features (selected in the training step) on the first 57 time stamps in the series. Then the mean scores of the classifiers are computed.

#### 3.6.2 Top down categorization

We visually pick a number of products to represent a category. Then we extract structural features (defined in Section

3.4) from each of the picked series. We average the structural features within each of the four categories and the averages serve as the four prototypes. Finally, we cluster the products to the four categories, using prototypes as fixed cluster centers.

## 4 SLIGRO case study

In this section we study a case of Sligro Food Group N.V. (SLIGRO) sales prediction. The company is engaged into food wholesales. SLIGRO works with corporate clients, mainly food retail and food service companies (restaurants), although there are some direct consumers as well. SLIGRO has around 40 outlets in the Netherlands. The group pursues a multi-channel strategy, covering various forms of sales and distribution (cash-and-carry and delivery service) and using several different distribution channels (retail and wholesale). SLIGRO trades about 60000 products.

### 4.1 Prerequisites

We aim to test if the base predictor, which is selected based on time series category, consistently outperforms alternative models in terms of prediction accuracy.

Our experimental field consists of 538 product sales quantities over two years period (from July 2006 to October 2008). The sales are aggregated on weekly basis, thus each series is of 119 weeks length. Each series represent the sales of one product aggregated over all outlets.

We use a regression as the base classifier, equipped with feature selection mechanism. Although we partially preselect the input features, still further selection needs to be carried out. For instance, we have 15 calendar events but only a few of them might be relevant for a particular product. We run Principal Component Analysis and keep the new features, which explain at least 70% of the data variance.

We use discretized labels to achieve comparability between different products. We discretize the outputs to 8 classes from very low sales (1) to very high (8), using Symbolic Aggregate Approximation (SAX) [8].

For model evaluation we use *Mean Absolute Scaled Error* (MASE) [7]:

$$MASE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{MAE(Baseline)} \right|, \quad (1)$$

where  $e_t$  is the prediction error at time  $t$ ,  $MAE(Baseline)$  is the mean absolute error of the baseline method. We use naive one step ahead prediction as the baseline (the prediction for the next week is equal to the factual sales this week). This is a special case of moving average, when lag is equal to one.

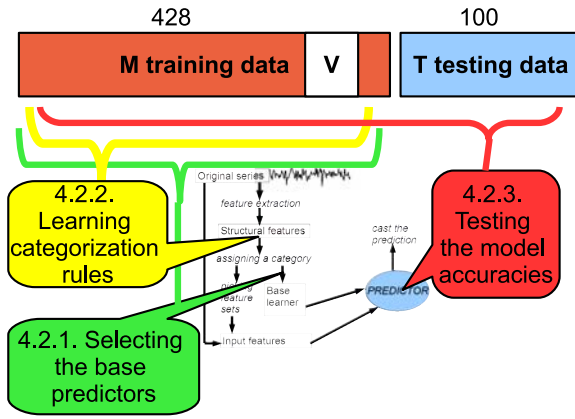


Figure 5. Experimental scenario

Having in mind that we use MASE as the accuracy measurement, we implicitly compare the intelligent classifiers to the naive method.

## 4.2 Experimental set up

Experimental scenario consists of three parts: selection the base predictors, learning categorization rules and testing final model accuracies. In the first part we select the base predictors and obtain “true” labels. We show that there exist product subsets on which it is possible to outperform baseline predictor. In the second part we aim to learn the dependencies between product categorization accuracies and sales prediction accuracies applying two approaches *bottom up* and *top down*. In the third part we test the final model accuracies. The experimental scenario and its relation to CAPA (Figure 2) is sketched in Figure 5.

To quantify the results, we perform controlled experiments using 538 product sales history. We take out random 100 series from the dataset and reserve them for final *model testing*. We call this set **T**. We develop the model using the remaining 438 series. We call the remaining set **M**.

We test the prediction accuracy sequentially. That means at time  $t$  we have all the historical sales up to this point available and want to predict sales level at  $t + 1$ . After the prediction is casted, the value from time  $t + 1$  is included into historical set and then we proceed with predicting the value for time  $t + 2$ . We use discretized outputs, but real valued inputs, normalized to  $(0, 1)$  before regressing them. We rerun discretization procedure after including each new value into historical set.

### 4.2.1 Selection of the base predictors

Having observed four different categories of the series (Figure 3) we narrow down selection of the base predictors to

a regression with different sets of input features. The input features are listed in Table 1. The index in superscript means the time from which the features are taken, assuming we are now at time  $t$  and predict the sales for time  $t + 1$ . For example  $F_{p9}^{(t+1)}$  means school holidays next week and  $F_{p9}^{(t)}$  means this week’s school holidays.

Table 1. Base predictor selection

	Base	Input features
$G_1$	MA(1)	$F_{p1}^{(t)}$
$G_2$	MA(3)	$F_{p1}^{(t-5\dots t)}$
$G_3$	MA(6)	$F_{p1}^{(t-5\dots t)}$
$G_4$	reg.	$F_{p1}^{(t-5\dots t)}, F_{p6}^{(t+1)}$
$G_5$	reg.	$F_{p1}^{(t-5\dots t)}, F_{p2\dots p5}^{(t)}, F_{p6, p11}^{(t+1)}$
$G_6$	reg.	$F_{p1}^{(t-5\dots t)}, F_{p2}^{(t)}, F_{p6, p10}^{(t+1)}, F_{p7\dots p9, p12}^{(t, t+1, t+2)}$
$G_7$	reg.	$F_{p1}^{(t-5\dots t)}, F_{p2\dots p5}^{(t)}, F_{p6, p11, p12}^{(t+1)}, F_{p9}^{(t, t+1, t+2)}$

- $G_1, G_2, G_3$ : moving average with lags of 1, 3 and 6 weeks. These are expected to work well on “flat” type of products with no particular relation to seasonality.
- $G_4$ : a linear regression trained on subsequences of 6 weeks. It is expected to work well on “frequent” or “flat” series, where there is no particular relation to calendar events or seasonality.
- $G_5$ : a linear regression trained on subsequences of 6 weeks, present and last years moving averages, present and last years cumulative sales, months. This classifier is designed for periodic products.
- $G_6$ : a linear regression trained on subsequences of 6 weeks and a range of calendar, temperature and seasonality related features. This method is expected to pick up specific events information for “occasional” products.
- $G_7$ : a linear regression trained on subsequences of 6 weeks and a range of annual patterns related features. This method includes aggregated sales, month indications, calendar events, expecting that there is somewhat annual periodicity in the “seasonal” series.

We run all the predictors on 438 series. We split the series into two parts 57 weeks used as a “warm up” and then the remaining 61 weeks are used for sequential testing. We obtain  $438 \times 7$  matrix of scaled accuracies (MASEs). Then we group the products based on the top ranking predictor. This way we get the “true categories” for given settings. In

**Table 2. Average MASEs for the “true categories”.**

	Size	$G_1$	$G_2$	$G_3$	$G_4$	$G_5$	$G_6$	$G_7$
$C_1$	200	<b>1.00</b>	1.31	1.63	1.63	1.89	1.91	1.93
$C_2$	68	1.00	<b>0.89</b>	1.02	1.25	1.56	1.53	1.62
$C_3$	36	1.00	0.92	<b>0.85</b>	1.04	1.21	1.21	1.22
$C_4$	35	1.00	1.04	1.09	<b>0.85</b>	1.00	1.01	1.03
$C_5$	19	1.00	1.06	1.17	0.96	<b>0.86</b>	0.99	0.93
$C_6$	43	1.00	1.02	1.11	0.97	0.98	<b>0.85</b>	0.94
$C_7$	37	1.05	1.08	1.18	1.02	0.98	0.98	<b>0.90</b>

Table 2 the average MASEs for each of the “true categories” are provided.

The best results appear on the diagonal in bold. The results below 1 mean that *it is possible to outperform moving average* if we do the correct categorization online. Furthermore, there are more than a single case per line, which outperform the baseline predictor. It means that there are alternative methods which can outperform the baseline predictor within the defined products categories.

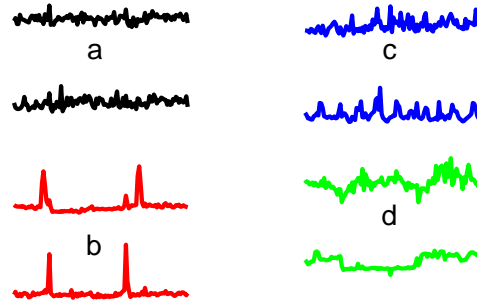
SLIGRO is using 6 weeks moving average, which is  $G_3$ .  $G_3$  is on average worse than the “intelligent” methods  $G_4$ ,  $G_5$ ,  $G_6$  and  $G_7$  in the last four categories which represent the “intelligent” methods.

Let us look at the pool of the predictors from another angle. For each predictor we count the number of products which gave MASE below 1. That is, we count the number of wins against baseline predictor. If there is no method on a given product to outperform the baseline predictor,  $G_1$  gets the score. The results are in Table 3.

**Table 3. Performance counts for the “true categories”.**

	$G_1$	$G_2$	$G_3$	$G_4$	$G_5$	$G_6$	$G_7$
$C_1$	<b>200</b>	0	0	0	0	0	0
$C_2$	3	<b>65</b>	28	16	13	9	13
$C_3$	1	28	<b>35</b>	20	13	17	13
$C_4$	0	12	16	<b>35</b>	21	17	18
$C_5$	0	8	6	12	<b>19</b>	9	14
$C_6$	2	17	17	24	26	<b>41</b>	32
$C_7$	2	15	11	21	28	25	<b>35</b>

The first three elements of the diagonal sum to 2/3 of the total training set and these are the moving averages. That means 1/3 of the products in this set could be better predicted using intelligent methods with additional features.



**Figure 6. Examples of the prototype products (a) “flat”, (b) “frequent”, (c) “occasional”, (d) “seasonal”.**

#### 4.2.2 Learning the categorization

We test the two categorization approaches *bottom up* and *top down* described in Section 3.6. We delimit the task to four categories, which we expect to correspond to the product categories presented in Figure 3. Thus we group the three moving averages ( $G_1, G_2, G_3$ ) into a single category  $C_1$ . A basic regression ( $G_4$ ) forms the second category  $C_2$ . The third category  $C_3$  is the calendar based regression ( $G_6$ ). And the fourth includes both annual patterns related regressions ( $G_5$  and  $G_7$ ).

In both cases we use the full length of the series (119 weeks), normalized to fit the range (0, 1).

In *bottom up* approach we use the training set  $\mathbf{M}$  for obtaining the “true” labels and learning the categorization tree.

For *bottom up* approach we visually pick four products for each category (see Figure 6 for examples), sixteen products all in all.

Figure 7 shows the averages of the structural features for the resulting categories and the prototypes. there is a clear distinction between the categories in prototypes (b), however looking at all the products (a) the distinction between the categories is not so clearly expressed. This suggests that there are mixed series within the product pool, or categories are changing in time.

We present the results of the categorization procedure in the next section together with the final accuracies.

#### 4.2.3 Prediction accuracies

We develop and validate CAPA model using the training set  $\mathbf{M}$ . Then we test the model assuming online settings on the testing set  $\mathbf{T}$ . We are interested in MASEs within each product category, assigned using a categorization rule.

The testing protocol is as follows. We split each series

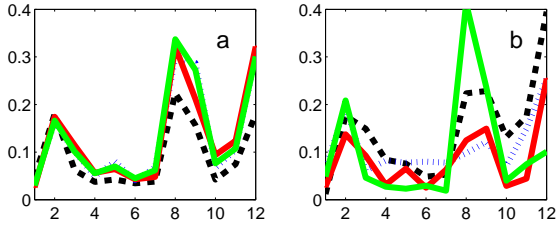


Figure 7. Structural features of (a) the obtained categories, (b) the prototypes.

into “warm up” (57 weeks) and testing (61 weeks). For the dataset **M** we use the categorization, which was obtained during the training phase. We categorize the products from the dataset **T** using only “warm up” part and the categorization rules obtained in the training phase.

We run sequential testing of the four models  $G_1$ ,  $G_4$ ,  $G_6$  and  $G_7$ , which we assume to correspond to each of the four product categories. We run the four models on each of the series from both datasets **M** and **T**. We compare the four accuracies for each product series. We aim to minimize the MASE for each pair of model-category.

First, we present the accuracies of the training data **M**, which corresponds to offline settings. In Table 4 average MASEs for the obtained categories are listed. It can be seen that in training *bottom up* approach (a) the selected base predictors  $G_4$ ,  $G_6$  and  $G_7$  outperform the baseline predictor in the corresponding categories  $C_2$ ,  $C_3$  and  $C_4$ .

Finding a categorization rule that results in a classification with clear dominance of a particular method is not hard. Out of 4096 total classifiers (i.e. feature combinations) 3541 manage to find a classification that meets this requirement. We do 10-fold cross validation on the training data to pick a single classifier. The features selected to for a decision tree were: 2, 5, 8, 10, 12 (see Section 3.4). We get 81% categorization accuracy on the training set.

To validate the results, we do the following experiment. We assign each product to a random category and then calculate average MASEs correspondingly. We present the results in the same Table 4 (c). It can be seen from the accuracies above 1 that in random categorization case the baseline predictor prevails. The *top down* categorization method (b) does not outperform the baseline predictor, however it gives better than random results, leaving prototyping approach as promising future direction. The poor performance of the method can be attributed to categorization accuracy, which is only 43% on the training data. However, random categorization would be only 25%. Thus we are better than random, but not enough to beat the final accuracies of the baseline predictor.

Let us check, how accurate should be the categorization,

Table 4. Average accuracies for the training data **M** (a) trained categorization *bottom up*, (b) trained categorization *top down*, (c) random categorization.

		Size	$G_1$	$G_4$	$G_6$	$G_7$
(a)	$C_1$	328	<b>1.01</b>	1.43	1.68	1.71
	$C_2$	34	1.00	<b>0.89</b>	0.97	1.00
	$C_3$	31	1.00	1.00	<b>0.94</b>	0.97
	$C_4$	45	1.00	0.98	0.95	<b>0.91</b>
(b)	$C_1$	158	<b>1.01</b>	1.69	2.13	2.20
	$C_2$	32	1.00	<b>1.34</b>	1.40	1.40
	$C_3$	7	1.00	0.95	<b>0.94</b>	0.92
	$C_4$	241	1.00	1.07	1.12	<b>1.12</b>
(c)	$C_1$	100	<b>1.01</b>	1.36	1.59	1.54
	$C_2$	108	1.00	<b>1.35</b>	1.66	1.59
	$C_3$	117	1.01	1.28	<b>1.60</b>	1.51
	$C_4$	113	1.00	1.35	1.63	<b>1.56</b>

so that the baseline predictor is still outperformed. In Figure 8 we depict MASE of each of the four categories as a function of categorization accuracy. 100% corresponds to the “true categories”, 0% corresponds to random categorization. We fill in the figure by increasing share of random categorization. For example, 80% means that we randomly select 20% of the products and assign random categories, while the remaining part is assigned the “true categories”.

The area within the ellipse in the Figure 8 is the region of interest. This is the area where the three categories ( $C_2$ ,  $C_3$ ,  $C_4$ ) outperform the baseline predictor in terms of MASE. This shows, that application of CAPA makes sense if we are able to assign the products to prespecified categories with the accuracy higher that 85%.

Now let us have a look how the categorization works on the testing data for online settings. Note, that we use only “warm up” part of the data here for categorization.

Table 5 presents average MASEs of the obtained categories on the dataset **T** for *bottom up* (a) and *top down* (b) categorization approaches. Along we present the results of random categorization (c). The results do not show MASE below 1 for the target categories. This is due to not sufficient categorization accuracy, which is 47% for (a) and 43% for (b). However, random categorization would give only 25% accuracy. Thus we managed to learn some categorization and these are promising results.

In Table 6 the performance counts for the test data **T** are listed. For example, the cell ( $C_1$ ,  $G_4$ ) means that in the category 1 there were 18 cases when the predictor  $G_4$  outperformed the baseline predictor.

Finally, let us have a look to a few prediction examples.



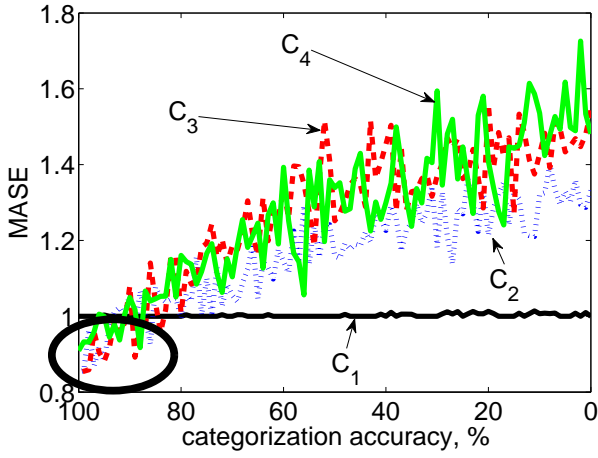


Figure 8. MASE as a function of categorization accuracy

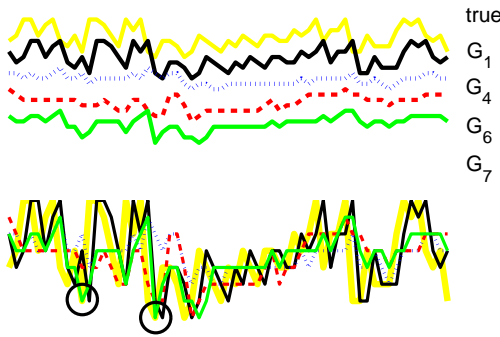


Figure 9. Example of "seasonal" predictions

In Figures 9 and 10 discretized prediction of two ("occasional" and "seasonal") series using  $G_1$ ,  $G_4$ ,  $G_6$  and  $G_7$  series are depicted. The bottom of the figures show true signal and the predictions, while in the top part the shapes of the predictions are laid separately.

In Figure 9 ("seasonal")  $G_7$  is the prevailing classifier. The two circled areas show  $G_7$  predicting on time, while the other classifiers mostly the moving average. In Figure 10 ("occasional")  $G_6$  predictor is the most accurate. The circled area indicates, how it spots the decline, while the peer predictors remain flat.

### 4.3 Implications for business decision making

We presented a case study of SLIGRO. In their business process the predictions are made using moving average but human corrections are made on top of the obtained predic-

Table 5. Average accuracies for the testing data T (a) trained categorization *bottom up*, (b) trained categorization *top down*, (c) random categorization.

		Size	$G_1$	$G_4$	$G_6$	$G_7$
(a)	$C_1$	69	<b>1.00</b>	1.37	1.57	1.65
	$C_2$	3	1.00	<b>1.41</b>	1.63	1.60
	$C_3$	20	1.00	1.62	<b>1.84</b>	1.85
	$C_4$	8	1.00	1.17	1.26	<b>1.28</b>
(b)	$C_1$	39	<b>1.00</b>	1.29	1.37	1.43
	$C_2$	4	1.00	<b>1.22</b>	1.39	1.42
	$C_3$	1	1.00	1.21	<b>1.23</b>	1.36
	$C_4$	56	1.02	1.32	1.67	<b>1.70</b>
(c)	$C_1$	24	<b>1.00</b>	1.30	1.50	1.59
	$C_2$	33	1.00	<b>1.37</b>	1.57	1.63
	$C_3$	15	1.00	1.72	<b>1.92</b>	1.93
	$C_4$	28	1.00	1.39	1.58	<b>1.63</b>

Table 6. Performance counts for T.

(a)	$G_1$	$G_4$	$G_6$	$G_7$
$C_1$	45	18	17	19
$C_2$	2	1	1	1
$C_3$	15	4	2	2
$C_4$	4	3	4	3

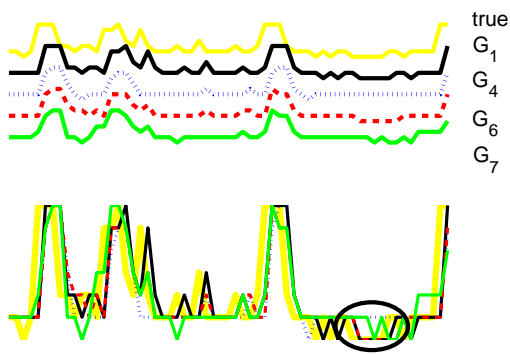
tions. Human factor is a volatile part of the decision making. Thus we aimed to summarize the seasonality related factors to build intelligent models with integrated context knowledge.

We presented a set of controlled experiments to justify our claims regarding context aware prediction. The data represent a static snapshot of the database. In reality the company would have a moving window of the sales history, i.e. all two years data could be used to cast the prediction for the next week. After the true sales values for the predicted weeks are obtained, it will be possible to include it into the training data. This way it would be possible to reassign product category online. Next cycle would be performed on an updated database.

CAPA is generic in a sense that we can reassign the product to another category at each time step based on recent development of sales. If a concept drift is happening domain expert could interfere using the feedback loop.

## 5 Discussion of related work

In many real-world domains the situations seen in the past might partially repeat, which is referred as reoccur-



**Figure 10. Example of "occasional" predictions**

ring contexts [19]. Seasonality comes very close to the concept of reoccurring contexts, with an emphasis that it is not known with certainty, when the contexts will reoccur.

The methods designed to handle reoccurring contexts can store concept descriptions [19], employ instance [4, 5] or batch [6, 9] selection to look for case bases, or maintain a diverse ensemble of learners [13, 12, 11, 17, 14].

Ensemble learning maintains a set of concept descriptions, predictions of which are combined using a form of voting, or the most relevant description is selected online. An alternative approach is referred as meta learning [16], where the relevant learners are selected based on offline predefined criteria. A two level learning model is presented by Widmer [18] perceived context changes are used to focus the learner specifically on the information relevant to the current context. Klinkenberg [10] developed an approach, where at each time step not only the training window but also the type of base learner and its parametrization is selected from a fixed set of learners, using cross validation.

In this study we introduced a generic sales prediction approach with context awareness. We incorporate external information and behavioral observations to categorize the target series based on their structural properties and then add relevant external features for prediction. In contrast with the discussed meta learning approaches [18, 10], we incorporate domain expertise and observations in categorization and base predictor selection process.

## 6 Conclusion

Sales prediction is a complicated task. There are different seasonality patterns across the product assortment. We developed context aware sales prediction approach, via introducing background knowledge and visual observations into predictor selection process.

In SLIGRO case study we showed that distinct categories exist, where the intelligent learners can outperform naive predictors if online categorization is accurate enough (in SLIGRO case 85%).

We showed that it is possible to learn the "true", we obtained 47% accuracy on the testing set, while random categorization gives only 25% accuracy. However, we did not reach 85% accuracy which is necessary to have an advantage of intelligent methods at the final prediction. One of the reasons is that for online categorization only one year sales history is available, which does not generally allow for reoccurring contexts to appear.

Further improvement of the obtained categorization accuracy could be achieved by finding more representative structural features, introducing multiple category assignments, adding more domain knowledge to the selection of the base predictors.

Furthermore, product sales are not independent. It would be interesting to analyze the semantic relations between the products and product groups in a context of sales prediction. Hierarchical prediction would give sales estimates on aggregated level, product groups and individual products simultaneously.

In this study we primary handled space context issues, fixing an assumption that the category of a given product is static over time. Next practical step would be to employ the approach in a dynamic setting, where the structural type might change over time as well. For example, a new marketing strategy comes into play promoting a particular beer to be the beer of rainy days.

## References

- [1] D. Adebajo and R. Mann. Identifying problems in forecasting consumer demand in the fast moving consumer goods sector. *Benchmarking: An Int. Journal*, 7(3):223–230, 2000.
- [2] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
- [3] J. W. Cooley and J. W. Tukey. An algorithm for the machine computation of the complex fourier series. *Mathematics of Computation*, 19:297–301, 1965.
- [4] S. J. Delany, P. Cunningham, A. Tsymbal, and L. Coyle. A case-based technique for tracking concept drift in spam filtering. In *The 24th SGAI Int. Conf. on Innovative Techniques and Applications of Artificial Intelligence*, pages 3–16. Springer, 2004.
- [5] W. Fan. Systematic data selection to mine concept-drifting data streams. In *KDD '04: Proc. of the 10th ACM SIGKDD int. conf. on Knowledge discovery and data mining*, pages 128–137. ACM, 2004.
- [6] M. B. Harries, C. Sammut, and K. Horn. Extracting hidden context. *Machine Learning*, 32(2):101–126, 1998.
- [7] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *Int. J. of Forecast.*, 22(4):679–688, 2006.

- [8] J.Lin, E. J. Keogh, L. Wei, and S. Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Min. Knowl. Discov.*, 15(2):107–144, 2007.
- [9] I. Katakis, G. Tsoumakos, and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: an application to email filtering. *Knowledge and Information Systems*, 2009.
- [10] R. Klinkenberg. Meta-learning, model selection and example selection in machine learning domains with concept drift. In *Ann. Workshop on Machine Learning, Knowledge Discovery, and Data Mining (FGML-2005) Learning - Knowledge Discovery - Adaptivity (LWA-2005)*, pages 164–171, 2005.
- [11] J. Z. Kolter and M. A. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *J. Mach. Learn. Res.*, 8:2755–2790, 2007.
- [12] K. Stanley. Learning concept drift with a committee of decision trees. CS Dept., University of Texas-Austin, 2001.
- [13] N. W. Street and Y. S. Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *KDD '01: Proc. of the 7th ACM SIGKDD int. conf. on knowledge discovery and data mining*, pages 377–382. ACM, 2001.
- [14] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Dynamic integration of classifiers for handling concept drift. *Information Fusion*, 9(1):56–68, 2008.
- [15] J. van der Vorst, A. Beulens, W. de Wit, and P. van Beek. Supply chain management in food chains: improving performance by reducing uncertainty. *Int. Transactions in Operational Research*, 5(6):487–499, 1998.
- [16] R. Vilalta and Y. Drissi. A perspective view and survey of meta-learning. *Artificial Intell. Review*, 18:77–95, 2002.
- [17] H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *KDD '03: Proceedings of the ninth ACM SIGKDD int. conf. on Knowledge discovery and data mining*, pages 226–235, New York, NY, USA, 2003. ACM.
- [18] G. Widmer. Tracking context changes through meta-learning. *Machine Learning*, 27(3):259–286, 1997.
- [19] G. Widmer and M. Kubat. Effective learning in dynamic environments by explicit context tracking. In *ECML '93: Proc. of the European Conf. on Machine Learning*, pages 227–243. Springer-Verlag, 1993.