Case 3 : Marketing

SKU-Level Sales Forecasting in Grocery Retail

A Comparative Analysis of Regularization and Machine-Learning Methods

Francesco Luigi Gasco

Siddharth Dey

Oscar Delarbre-Jimeno

Marc Bilanin

Abstract

In this paper, we develop, evaluate and compare several forecasting models to predict weekly sales of orange juice brands across ten stores of Dominick's Finer Foods. Models include a random walk benchmark, forward-stepwise linear regression, LASSO, ridge regression, elastic net, and random forest. Forecasting uses a rolling window estimation method, with parameters updated periodically. Predictor sets include own-price, promotions, competitor prices, price-derived variables, interaction terms, lagged dependent values, and seasonal dummies. Forecast accuracy is measured using MSE, Asymmetric MAE, volume-weighted MSE, coefficient of variation of errors, and pairwise Diebold–Mariano tests. Results show that intricate models consistently outperform the benchmark, and in certain settings, elastic net and random forest show superior results. This highlights the value of regularization and nonparametric methods in capturing correlated predictors and nonlinearities. Our findings also underscore the importance of accounting for store and brand heterogeneity when making modeling choices, as no model dominates across all settings.

1 Introduction

With the growth of advanced data collection systems like barcodes and checkout scanners, supermarkets today gather vast quantities of highly detailed sales-related data. This data tracks daily or weekly sales patterns, providing insights into promotions, competition, and price changes, helping anticipate consumer behavior more accurately. In particular, generating precise forecasts is critical for fast-moving consumer goods, such as orange juice, since their quick turnover makes optimal inventory management difficult.

To expand, accurate sales forecasts help supermarkets effectively manage their stock levels. This avoids under-stocking, which causes a loss in revenue, and excess inventory, which generates unnecessary costs. The challenge is selecting variables and models that accurately capture purchasing dynamics. Given the large set of potential predictors, including current and lagged prices, promotions, seasonality, and competition, selecting an optimal subset is essential for building robust, efficient models.

We analyze and use sales data from multiple brands of orange juice across several stores of Dominick's Finer Foods to develop and compare forecasting models. We first build a traditional random walk model as a benchmark. This is accompanied by modern variable selection and modeling techniques, namely Specific-to-General, LASSO and ridge regression, Random Forest, and Elastic Net. We compare these methods to examine how selection strategies and modeling techniques affect forecast performance and practical utility.

Socially, this study derives relevance from its direct impact on consumer satisfaction, sustainability from reduced food waste, and economic efficiency within the retail sector. Academically, this paper contributes to an ongoing discussion in forecasting research by systematically comparing classical benchmarks with advanced machine learning and regularization methods (Chu and Zhang, 2003; Aras et al., 2017). Chu and Zhang (2003), for example, find that neural networks applied on de-seasonalized data outperform traditional linear models in forecasting aggregate retail sales, while trigonometric models prove ineffective. Of course, the insights to be gained from this study surpass beyond orange juice sales by highlighting methodological guidelines useful in broader categories of consumer goods.

The remainder of this paper is structured as follows. Section 2 introduces and describes the data and motivates relevant transformations and other pre-processing steps. Section 3 presents the models used to forecast sales data and outlines the forecast evaluation techniques that we employ. Finally, Section 4 presents the empirical results.

2 Data

2.1 Dataset

This paper's data originates from Dominick's Finer Foods, a supermarket chain based in the greater Chicago area. Our dataset spans 102 weeks, from September 1989 to August 1991, and comprises weekly sales records for refrigerated orange juice across a selection of ten stores within this chain. Specifically, we examine weekly sales data (in ounces) for 11 distinct store-keeping units (SKUs), each defined by a unique combination of brand and package size. These SKUs include major brands such as Tropicana, Minute Maid, and Florida's Natural, as well as Dominick's private label brand, and varying sizes ranging from 64 to 128 ounces.

For later modeling convenience, we divide the original dataset into separate subsets for each SKU-store combination. In these 110 "sub-datasets", each observation is indexed by week and reports several key variables, including the sales in ounces of the SKU, the corresponding price per ounce (measured in cents) for all SKUs, and two binary variables: one for the presence of feature promotional activity and one for in-store coupons in the given week.

2.2 New Variables & Transformations

To capture more of the relevant variation in sales, we construct a range of additional predictors based on the original dataset. Specifically, we make use of the information in the original variables to illustrate more complex market dynamics, such as post-promotional dips in sales or momentum in sales. Moreover, to accurately consider competition amongst SKUs, predictors based on competitor prices and promotional activities are also created. An extensive list of all potential predictors we consider can be found in Section A of the Appendix.

In addition to creating predictors, we also apply a logarithmic transformation to the sales series to make its magnitude similar in scale to that of other variables. This transformation also helps linearize the relationship between sales and the regressors, as well as stabilize variance across observations, reducing issues of heteroskedasticity that are found in sales data. This is also accompanied by a log-transform of prices. In addition to scaling, which we thought necessary as prices are measured in cents per ounce, this transformation implies that estimated price coefficients in our models simply measure sales elasticity, which provides an easy interpretation. Note that certain models, namely LASSO and ridge, also demand the standardization of variables, but this is discussed in

section 3.1.1 as it pertains to model-specific methodology.

Finally, to prevent significant bias in our estimates, which could affect our models' performance, we identify the outliers for each SKU-store combination using the interquartile range (IQR) method and replace those observations with the relevant lower/upper bound (Tukey et al., 1977).

2.3 Time Series Dynamics

From theory, we know that time series data, like sales, often exhibit autocorrelation properties, which means that lagged values of the dependent variable can serve as a helpful predictor. Therefore, we computed the partial autocorrelation function of the log sales series of each SKU-store combination to explore the appropriate order of autoregressive terms in each SKU-store model (the choice of building a model for each SKU-store combination is discussed and motivated further in Section 3).

In doing so, we used the first 72 observations of the series. This choice avoids look-ahead bias and aligns with the 72-week rolling window used during forecasting. Here we make the assumption that the appropriate lag order will remain constant throughout the sample, such that it does not need to be re-estimated in every window of the forecasting procedure, leaving one round of estimation in the first window of 72 observations sufficient. This assumption is motivated by both pragmatic concerns and that in a sample of 102 observations, 72 should be ample to capture the underlying autoregressive dynamics.

For the sake of exposition, in Figure 1 we present the correlogram of SKU 6 in store 21. From the PACF graph on the right, it is clear that we should select up to the second lag of log sales in this SKU-store combination model. A similar procedure is repeated for all other combinations.

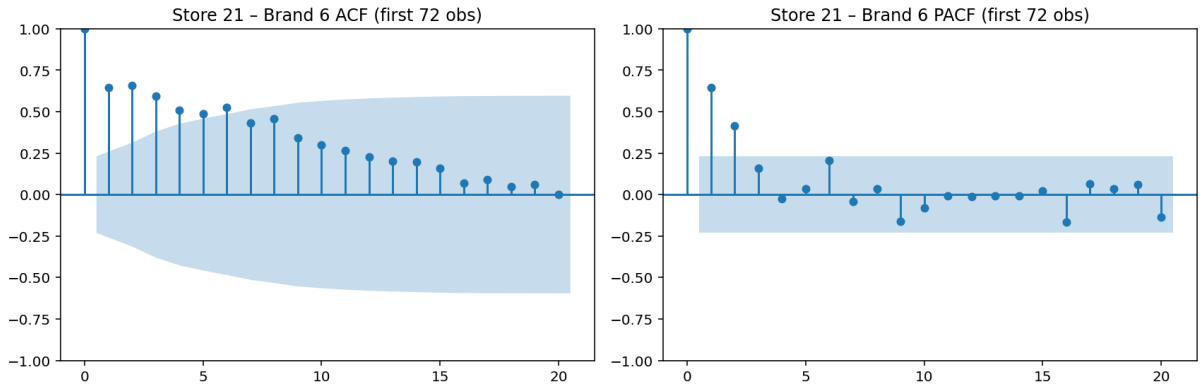

Figure 1. Correlogram of log sales for Store 21 and SKU 6

2.4 Descriptive Statistics of Main Variables

Table 1 presents summary statistics for each of the 11 SKUs, averaged across all ten stores (we consider untransformed variables to calculate these statistics). Several patterns relevant for our modeling choices emerge here. First, there is substantial variation in average sales between SKUs. For example, SKU 10 (Dominick's 64 oz) has the highest average sales, while smaller brands like SKU 8 (Tree Fresh 64 oz) and SKU 3 (Florida's Natural 64 oz) have much lower average sales. This difference implies varying baseline demand levels for each SKU, which suggests that model specifications should incorporate SKU-specific effects or different intercepts.

Price differences across SKUs are relatively small, with average prices ranging from 0.027 to 0.048 cents per ounce. However, when both sales and prices are log-transformed, as in our case, price coefficients represent elasticity, which means that even small price differences, such as the gap between SKU 10 (0.027) and SKU 2 (0.048), can lead to large differences in predicted sales. The promotional variables also show significant variation. Some SKUs have a substantial amount of promotions (for example, SKU 5 and SKU 10, with high averages for both *feature* and *deal*), while others are promoted less often. This suggests that promotional activity will be an important predictor of sales and may interact differently with price across SKUs.

Table 1. Summary statistics by SKU (Averaged Across Stores)

	Sales		**Price**		**Feature**		**Deal**	
SKU	μ	σ	μ	σ	μ	σ	μ	σ
1	14161.7	16265.9	0.044	0.009	0.137	0.344	0.540	0.499
2	8040.3	4313.9	0.048	0.007	0.127	0.334	0.306	0.461
3	3210.3	5304.3	0.044	0.005	0.147	0.354	0.365	0.482
4	16824.5	27742.3	0.035	0.006	0.314	0.464	0.550	0.498
5	18908.5	30532.6	0.034	0.006	0.255	0.436	0.563	0.496
6	5420.1	2453.0	0.040	0.006	0.196	0.397	0.477	0.500
7	6497.9	17855.8	0.035	0.006	0.147	0.354	0.404	0.491
8	3113.2	4996.0	0.033	0.005	0.098	0.298	0.360	0.480
9	3536.0	13257.8	0.032	0.007	0.157	0.364	0.447	0.497
10	24125.1	34840.3	0.027	0.006	0.275	0.446	0.508	0.500
11	9951.4	6948.1	0.029	0.005	0.167	0.373	0.282	0.450

Overall, the variation in both sales levels and promotion frequencies highlights a modeling approach that can capture SKU-specific patterns.

3 Methodology

In this section, we introduce the forecasting models used to predict sales of each SKU in every individual store. Importantly, each of the models described below will be applied to every SKU-store combination (as partly justified by Section 2.4), meaning that different SKU-store combinations will have different parameter estimates. Doing so allows us to properly account for the fundamental differences between stores (e.g., location) and between brands (e.g., premium and budget products).

For notational ease, let us denote the set of all stores in our dataset by $N = \{1, ..., 10\}$, and the set of SKUs in our study by $I = \{1, ..., 11\}$.

3.1 Core models: Specific-to-General, LASSO & ridge regressions

3.1.1 Description of the models

As in most econometric problems, identifying relevant predictors is crucial for the performance of our forecasting model.

A commonly used approach to this problem is the Specific-to-General (or Forward Stepwise selection) method (FS). As described in Section 6.1.2 of James et al. (2023), this model-building approach starts with a model with no predictors and iteratively adds the predictor that results in the biggest improvement in model fit, as defined by the increase in the R^2 of the relative regression. The "best" model is chosen as that which minimizes the Bayesian Information Criteria (BIC), which penalizes more complex specifications.

Recent machine learning literature has introduced more complex regularization techniques to tackle this issue, namely, Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regression. These models are particularly useful in overcoming specific difficulties of other approaches, such as the Specific-to-General method or model selection based on information criteria.

Both LASSO and ridge regressions work in similar ways. Namely, they both estimate parameters by imposing a shrinkage term that penalizes overfitting. Specifically, LASSO

and ridge regressions estimate the parameters $\hat{\beta}_L$ and $\hat{\beta}_R$ as follows:

$$\text{LASSO: } \hat{\beta}_L = \min_{\tilde{\beta}} \left(\sum_{t=1}^{T} \left(\widetilde{\ln S}_{t,(i,n)} - \sum_{j=1}^{K} \tilde{x}_{j,(i,n)} \tilde{\beta}_j \right)^2 + \lambda_L \sum_{j=1}^{K} |\tilde{\beta}_j| \right) \tag{1}$$

$$\text{Ridge: } \hat{\beta}_R = \min_{\tilde{\beta}} \left(\sum_{t=1}^{T} \left(\widetilde{\ln S}_{t,(i,n)} - \sum_{j=1}^{K} \tilde{x}_{j,(i,n)} \tilde{\beta}_j \right)^2 + \lambda_R \sum_{j=1}^{K} \tilde{\beta}_j^2 \right). \tag{2}$$

Here $\widetilde{\ln S}_{t,(i,n)}$ denotes the standardized log-sales of SKU i in store n at time t, and the $\tilde{x}_{j,(i,n)}$'s denote the standardized potential predictors outlined in Section A of the Appendix.[1]

The subtle difference in the last term of each expression has important implications for the applicability of both models. Specifically, while LASSO may set certain coefficient estimates exactly to zero, ridge may only shrink them towards zero, without setting any of them exactly to zero. Put differently, while LASSO may be used to effectively perform variable selection, ridge will only limit the influence of potential predictors. As a result, ridge tends to perform better in settings where many predictors contribute marginally, while LASSO can better handle situations where we only expect a few predictors to be relevant. Given the uncertain nature of the process by which sales are generated, we consider it useful to include both models in our analysis.

Importantly, both LASSO and ridge require all variables to be standardized, as this ensures that the regularization penalty is applied fairly across all variables. Hence, we standardize all predictors within each SKU-store combination, i.e., using the mean and standard deviation of the relevant training window, before estimating these models.

Lastly, these minimization problems are solved numerically using algorithms from the `scikit-learn` package in Python. Specifically, for LASSO, we use coordinate descent, an iterative algorithm well suited for ℓ_1-penalized problems. On the other hand, ridge models are usually estimated via a closed-form solution using singular value decomposition (SVD) when feasible. In case of more complex settings, `scikit-learn` may use iterative algorithms such as stochastic gradient descent.

3.1.2 Selecting optimal λ_L and λ_R values

As is apparent from equations (1) and (2), the coefficient estimates depend significantly on the shrinkage terms λ_L and λ_R. To determine the most suitable values for these parameters, we employ cross-validation.

[1] See penultimate paragraph of this section to understand why we consider standardized variables.

Specifically, we first specify an arbitrary set of potential values for λ_m ($m = L, R$), typically ranging from small values (e.g., 0.001) to large values (e.g., 1000). For each candidate λ_m value, we split the relevant dataset into a prespecified number of folds (in our case, five). We then implement an expanding-window cross-validation approach on these five folds. Namely, we use the first fold to train the model and forecast the second fold, then the first two folds are used to predict the third, and this process continues until all folds have been used. Finally, we select the λ_m value for which the average MSE across all folds is minimized.

3.2 Extensions and benchmark models

3.2.1 Extension models: Elastic Net & Random Forest

To evaluate the forecasting accuracy and efficiency of LASSO and ridge regressions, we include two complementary estimation methods in our analysis, namely, Elastic Net (EN) and Random Forest (RF).

As introduced by Zou and Hastie (2005), EN combines the penalty terms of LASSO and ridge, allowing it to overcome certain difficulties of the LASSO model, specifically in contexts where predictors are highly correlated, while maintaining its model selection capabilities. Contrary to LASSO, which selects sone predictors and eliminates the rest, EN allows correlated coefficients to shrink together, which allows it to better describe the true predictive power of each variable. Note that, similar to LASSO and ridge, EN also requires predictors to be standardized. Our EN models are estimated using coordinate descent in `scikit-learn`'s `ElasticNetCV`. Tuning parameter are still determined using cross validation as described in section 3.1.2.

On the other hand, RF relaxes the assumption of a linear relationship between predictors and the dependent variable. Hence, it's particularly useful to determine whether a linear model is sufficient to describe the relationship between sales and the corresponding list of predictors, or whether considering more complex models could increase the accuracy of our forecasts. Including such a model in our analysis is especially interesting, as the relationship between sales and other variables (such as price or promotions) is often non-linear and may involve complex dependencies. For example, RF could naturally model interactions like: a price drop on SKU A only boosts sales of SKU B if SKU B is not concurrently on promotion. In `scikit-learn`, RF models are estimated by creating an ensemble of decision trees, where each one is trained on a bootstrapped sample of the data. Then, at each split a random subset of predictors is considered, and the prediction is computed by averaging the individual tree predictions.

In the end, these models will help us determine whether LASSO and ridge yield satisfactory sales forecasts over our sample period, or whether more elaborate models are necessary to capture the dynamics of the data.

3.2.2 Benchmark model: Random Walk

To evaluate the performance of the above-mentioned models, we also include a simpler, more commonly used benchmark approach: the random walk (RW) model. The RW model provides a more simplistic approach to the problem, by assuming that $\ln S_{t,(i,n)} = \ln S_{t-1,(i,n)} + \epsilon_t$, where ϵ_t is a white noise process. The optimal one-step ahead forecast, $\widehat{\ln S}_{t+1|t,(i,n)}$, is then given by $\widehat{\ln S}_{t+1|t,(i,n)} = \ln S_{t,(i,n)}$. Including this model in our analysis will help us determine whether the added complexity of our core and extension models leads to significant improvements in our forecasts, in a context where sales may reasonably be approximated by a "random walk".

3.3 Forecasting and comparing predictive accuracy

3.3.1 Out-of-sample forecasting

As mentioned at the beginning of this section, we will carry out our forecasting procedures by estimating 110 SKU-store-specific models (11 SKUs $\times$ 10 stores). To ensure we accurately describe the time-varying relationships in the data, we apply a rolling-window estimation method. Specifically, we selected a window length of 72 weeks as it provides a sufficiently large dataset to accurately estimate the coefficients in each model. Furthermore, this window length allows us to forecast 30 one-step ahead sales values per SKU-store combination (given our dataset of 102 weeks), which we believe is statistically sufficient to meaningfully compare the forecast accuracy between our models.

We re-estimate coefficients (and λ values, if applicable) and smearing factors every two weeks to adapt to changes in predictor relevance while balancing computation.

Importantly, note that all models described in Sections 3.1 and 3.2 provide estimates for log-sales, which therefore have to be converted to interpretable sales per ounce estimates ($S_{t,(i,n)}$). Of course, simply exponentiating the output from our models is an option, but this route fails to correct for the so-called "retransformation bias" (Duan, 1983). Therefore, to obtain meaningful forecasts, we apply a non-parametric smearing transformation to our log-sales estimates, as proposed by Duan (1983). Since this method does not impose any assumptions about the error distribution, it's a more robust and less restrictive choice. Duan (1983) characterizes the smearing estimate as a "low-premium insurance policy" against distributional misspecification. If the errors are normally dis-

tributed, the "premium" is only a slight reduction in efficiency, while this approach offers crucial protection when the error distribution is not normal. Section B of the Appendix details all assumptions required for the smearing estimates to be consistent, along with a discussion of their plausibility in our setting.

3.3.2 Comparing the quality of forecasts

To evaluate the forecast accuracy of each model, we will employ two loss functions: the Asymmetric Mean Absolute Error (A-MAE) and the Mean Squared Error (MSE). The former penalizes under- and over-predictions differently, by employing different penalty terms for each situation. This is particularly useful when forecasting SKUs, as under-forecasting sales is often more costly than over-forecasting, since it leads to a loss in revenue and customer dissatisfaction. Hence, we set the penalty term for under-forecasting to be double that for over-forecasting. On the other hand, MSE penalizes large forecast errors (from both under- and over-forecasting) more severely than smaller ones, which is again particularly interesting where large deviations from the true value of sales are likely to cause significant losses for the store.

While these loss functions can be used to evaluate model accuracy at the SKU-store level, comparing models across 110 SKU-store combinations individually would be pragmatically cumbersome. Thus, to summarize model performance in a meaningful way, we adopt a panel-adjusted variant of the Diebold-Mariano (DM) test as described in section 2.1 of Qu et al. (2024). Specifically, we compare models at the SKU level by averaging forecast errors across the 10 stores for each SKU. This, for example, allows us to statistically test whether RF produces more accurate forecasts than ridge for SKU 1. Hence, rather than testing for predictive accuracy using time-series differences at the individual SKU-store level, this approach aggregates the loss differential series along the cross-sectional dimension, i.e., the stores, for a given SKU.

Precisely speaking, let $L^{(i)}_{m,n,t}$ denote the loss from model $m \in \{1, 2\}$ for store n at time t for a particular SKU i. In line with Qu et al. (2024), we use squared forecast errors as our loss function. Then the loss differential for store n at time t for SKU i is defined as:

$$\Delta L^{(i)}_{n,t} = L^{(i)}_{1,n,t} - L^{(i)}_{2,n,t}.$$

The test then checks whether the average loss differential across all stores and time periods is significantly different from zero. Thus, the null hypothesis is that the two models have

equal predictive accuracy for SKU i, that is:

$$\mathbb{H}_0^{\text{pool}} : \mathbb{E}[\bar{L}_1^{(i)}] = \mathbb{E}[\bar{L}_2^{(i)}],$$

where $\bar{L}_m^{(i)} = \frac{1}{N \times T} \sum_{n=1}^{N} \sum_{t=1}^{T} L_{m,n,t}^{(i)}$ denotes the average loss of model m for SKU i across all stores and time periods. This is tested using the following test statistic:

$$J_{N,T}^{\text{DM},(i)} = \frac{(N \times T)^{-1/2} \sum_{t=1}^{T} \sum_{n=1}^{N} \Delta L_{n,t}^{(i)}}{\widehat{\sigma}(\Delta L_t^{(i)})},$$

where $\Delta L_t^{(i)} = \frac{1}{N} \sum_{n=1}^{N} \Delta L_{n,t}^{(i)}$ is the cross-sectional average of the loss differentials at time t. For $\widehat{\sigma}(\Delta L_t^{(i)})$ we use a Newey-West estimator. Specifically, we use the Bartlett kernel with the truncation lag set to $1.2N^{\frac{1}{2}}$ based on recommendations outlined in Andrews (1991). Under standard regularity conditions, this statistic is asymptotically standard normal

Although this panel-adjusted DM test allows us to compare two models at the SKU level, performing all possible pairwise comparisons across would still be difficult. With 11 SKUs and 6 models, there would totally be $\binom{6}{2} \times 11 = 165$ individual hypothesis tests. Therefore, to keep the analysis manageable while still allowing for meaningful insights, we follow a two-step procedure.

First, we compute aggregated loss statistics for each model at the SKU level in a way that reflects each SKU's relative economic importance across stores. Specifically, let $\widehat{y}_{m,n,t}^{(i)}$ denote the predicted sales for model m at time t, store n, and SKU i, and let $y_{n,t}^{(i)}$ be the corresponding true value. Then the MSE for model m, store n, and SKU i as:

$$L_{m,n}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \left(\widehat{y}_{m,n,t}^{(i)} - y_{n,t}^{(i)} \right)^2.$$

To measure store-level performance in a business-relevant way, we calculate a volume-weighted average of these losses across stores. Let $w_n^{(i)}$ denote the total units sold of SKU i in store n over the forecast horizon:

$$w_n^{(i)} = \sum_{t=73}^{T} S_{t,(i,n)}.$$

Then, the volume-weighted MSE for model m and SKU i is given by:

$$\text{MSE}_m^{(\text{wt},i)} = \frac{\sum_{n=1}^{N} w_n^{(i)} L_{m,n}^{(i)}}{\sum_{n=1}^{N} w_n^{(i)}}.$$

This metric ensures that stores with higher sales volumes for a given SKU add more to the final value, making model selection more economically relevant. Based on these weighted MSE values, we identify the two best-performing models for each SKU. Further, to complement these weighted loss measures, we also compute the coefficient of variation (CV) for each SKU-model combination using the mean squared errors calculated separately for each store, given a SKU and model. For model m and SKU i, we define the MSE in store n as $L_{m,n}^{(i)}$, and compute the mean and standard deviation across stores as:

$$\mu_m^{(i)} = \frac{1}{N}\sum_{n=1}^{N} L_{m,n}^{(i)}, \quad \sigma_j^{(i)} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(L_{m,n}^{(i)} - \mu_m^{(i)}\right)^2}.$$

The coefficient of variation is then given by:

$$\mathrm{CV}_m^{(i)} = \frac{\sigma_m^{(i)}}{\mu_m^{(i)}}.$$

This captures the relative variability of forecast errors across stores. A lower CV implies that the model performs more consistently across the stores, while a higher CV implies greater variance in accuracy, likely due to store-specific dynamics.

In the second step, we apply the panel-adjusted DM test, as described earlier, to compare only these top two models per SKU. This allows us to assess whether the "best" model's performance is statistically significant once forecast errors are aggregated across stores.

Notably, while we compute both MSE and A-MAE to evaluate forecast quality, we use only MSE for statistical testing. This is because the panel-adjusted DM test we use is derived under the assumption of a quadratic loss function. Nonetheless, A-MAE results are reported in section 4.1 to provide a broader perspective on model performance. Finally, the modified DM test is also used to statistically compare each of our core and extension models against RW at the SKU level.

4 Results

4.1 Overview of results

We begin by analyzing the forecasting accuracy of our models when applied to the last 30 weeks of our dataset. Figures 2 and 3 report, respectively, the heatmaps of the A-MAE and MSE values for each model-SKU-store combination.

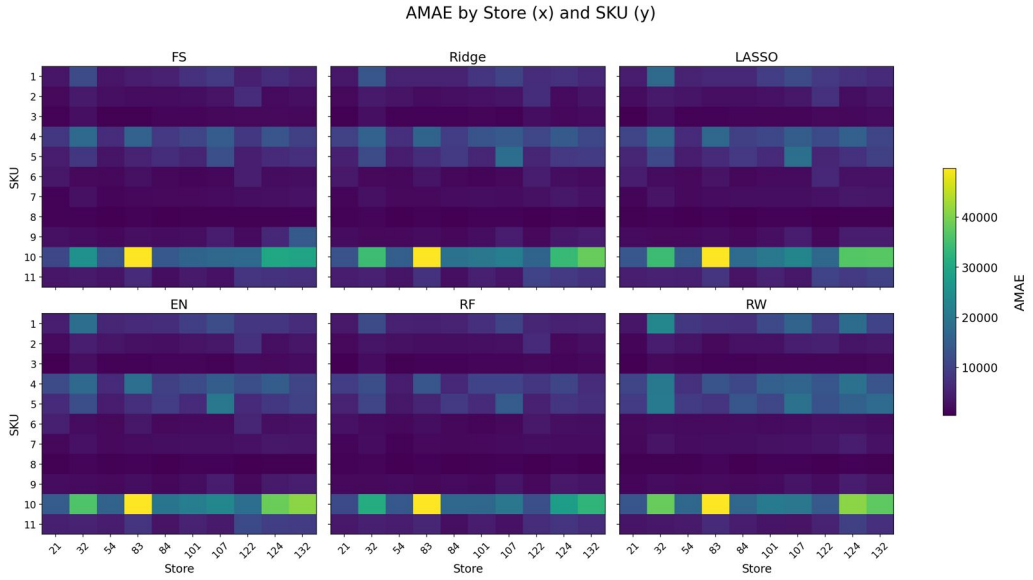

Figure 2. A-MAE values for each model-SKU-store combination

Figure 3. MSE values for each model-SKU-store combination

Firstly, note that the plots for the MSE values appear significantly darker than those of the A-MAE figures. As a darker shade in our case indicates a relatively lower loss function value, this is likely indicating that our models have a tendency to underpredict sales values more often than they over-predict them. As previously discussed in Section 3.3.2, this is not desirable when predicting sales figures. Hence, when applying these models in practice, it may be desirable to first modify them to account for this downside.

Figures 2 and 3 show that all models seem to perform consistently worse for certain SKUs, in particular SKUs 4, 5, and 10. However, this is not particularly surprising, as Table 1 shows that these SKUs exhibit significantly higher sales volatility compared to the others, thus complicating the accurate forecast of sales. One particular SKU-store combination exhibits significantly worse estimates than the rest, namely, SKU 10 in store 83. Although the high volatility of SKU 10's sales likely contributes to this problem, it may be interesting to further investigate this specific combination. Doing so may uncover interesting patterns and relationships in the data that we do not account for in our models.

Interestingly, all models appear to perform similarly across all brand-store combinations. In particular, the Forward-Stepwise method exhibits similar forecast accuracy to its more complex alternatives, suggesting that the added complexity of these models may not translate to better performance. Sections 4.2 and 4.2 will provide formal tests to assess the significance of these differences in forecast accuracy. Note that the loss function values of the RW model are often comparable with those of all other models, reinforcing the idea that sales may follow a near-random process.

4.2 Comparing all models to Random Walk

Let us now examine the relative performance of all core and extension models against that of our chosen benchmark, i.e., the RW model. To do so, we perform a series of DM tests, as explained in Section 3.3.2, of which we report the resulting test statistics and p-values in Table 2.

Firstly, For some SKUs, the complex models perform comparably to the RW benchmark. Specifically, for SKUs 6 and 8, no model significantly outperforms the RW in terms of forecast accuracy, as indicated by the high p-values of the relative DM tests. In other words, during our forecast period, the sales of these SKUs can be accurately predicted by simply projecting last period's sales forward. A possible explanation for this lies in the results of Table 1, which shows that SKUs 6 and 8 exhibit some of the lowest sales volatility across the sample. As a result, it is not surprising that the RW model yields significantly accurate results when considering such stable time series.

Nevertheless, the results in Table 2 indicate that, for most SKUs, our more complex forecasting models (especially ridge, LASSO, EN, and RF) tend to yield significantly better forecasts than the RW model. This is clear from the mostly negative and significant test statistics, often at 1% level. Hence, this shows that orange juice sales are not entirely driven by randomness in market dynamics, but can instead be accurately predicted when correctly selecting the relevant predictors.

Table 2. Panel Diebold-Mariano Test Statistics by Brand

	Model				
SKU	**FS**	**Ridge**	**LASSO**	**EN**	**RF**
1	-3.4181***	-3.4182***	-3.2230***	-3.3188***	-3.1824***
	(0.0006)	(0.0006)	(0.0013)	(0.0009)	(0.0015)
2	-2.7845***	-2.8527***	-2.7841***	-2.7385***	-3.1553***
	(0.0054)	(0.0043)	(0.0054)	(0.0062)	(0.0016)
3	-0.3554	-1.8680*	-2.4812**	-2.2718**	-2.2619**
	(0.7223)	(0.0618)	(0.0131)	(0.0231)	(0.0237)
4	-4.0826***	-4.6727***	-4.3048***	-4.5033***	-4.4736***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
5	-3.4338***	-3.4764***	-3.4782***	-3.5061***	-3.3228***
	(0.0006)	(0.0005)	(0.0005)	(0.0005)	(0.0009)
6	-1.0686	-1.4232	-0.9549	-1.1107	-1.4568
	(0.2852)	(0.1547)	(0.3396)	(0.2667)	(0.1452)
7	-3.3048***	-3.5508***	-3.3469***	-3.4498***	-3.3802***
	(0.0010)	(0.0004)	(0.0008)	(0.0006)	(0.0007)
8	-1.4749	-1.4729	-1.5818	-1.5060	-1.7434*
	(0.1402)	(0.1408)	(0.1137)	(0.1321)	(0.0813)
9	1.6504*	-2.6711***	-1.0084	-2.2378**	-2.8027***
	(0.0989)	(0.0076)	(0.3132)	(0.0252)	(0.0051)
10	-1.7472*	-3.5129***	-3.1452***	-3.2550***	-3.1795***
	(0.0806)	(0.0004)	(0.0017)	(0.0011)	(0.0015)
11	-2.0266**	-2.5593**	-2.1694**	-2.3435**	-2.6122***
	(0.0427)	(0.0105)	(0.0301)	(0.0191)	(0.0090)

Note: The table displays Diebold-Mariano test statistics comparing each model against a random walk (RW) benchmark, evaluated per brand across all stores. P-values are reported in parentheses. Significance levels are denoted by: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Negative statistics indicate the model is better than the RW benchmark, and vice-versa.

Finally, FS often underperforms compared to the other core and extension models. This indicates that the added complexity of ridge, LASSO, EN, and RF models may lead to improved forecasting accuracy, likely due to their more robust variable selection and regularization methods.

Overall, Table 2 shows that ridge and EN tend to display the most statistically significant and lowest test statistics. This may indicate that these models are more effective when forecasting sales across different SKUs.

4.3 Comparing Top Two Models Per SKU

Having seen that most models significantly outperform the RW benchmark, we now focus on a more precise comparison. Specifically, we identify which of the top two models for a given SKU, based on weighted MSE scores, create the most accurate forecasts. Table 3 presents the weighted MSE scores (at the SKU-model level) used to select the top two models for each SKU, and Table 4 presents relevant results of the panel-adjusted DM tests.[2]

For most SKUs, the test statistics suggest that, while one model appears better than the other, the differences are not always statistically significant. This means that, in most cases, retailers may choose between the two strong models without worrying about large drops in accuracy. However, there are a few notable exceptions worth analyzing.

Firstly, for SKU 3, RF significantly outperforms ridge. This may suggest that the nonlinear patterns described by RF are particularly useful for forecasting demand in this SKU. This could be due to complex interactions between promotions and other variables such as price or store characteristics. There is also an interesting result for SKU 5, where FS outperforms ridge. This highlights that a well specified linear model can sometimes beat more complex regularized methods. This is most probable when sales are volatile but mostly explained by a few strong predictors, such as price or promotional activity. In such cases, the FS method is able to single out these important predictors effectively, while ridge may reduce their effect by regularizing a larger set of predictors, including those that add little explanatory power.

For SKU 10, one of the highest selling and most volatile products (see Table 1), EN significantly outperforms LASSO. As EN is more adept with correlated predictors, this outcome may suggest that SKU 10 sees overlapping promotional strategies and shared dynamics across stores, whose correlated nature, plays an important role in explaining sales.

[2]Results for the coefficient of variation for each SKU-model combination can be found in Table C1 of Section C in the Appendix.

Table 3. Weighted Forecast Losses Across Models (MSE-Based)

	Model					
SKU	**FS**	**Ridge**	**LASSO**	**EN**	**RF**	**RW**
1	3.80e7	3.19e7	3.99e7	3.65e7	2.74e7	1.59e8
2	4.94e6	3.92e6	4.78e6	4.51e6	4.43e6	1.02e7
3	1.36e6	1.01e6	1.05e6	1.06e6	8.69e5	1.37e6
4	1.01e8	6.53e7	7.23e7	6.65e7	5.38e7	2.13e8
5	3.45e7	3.99e7	4.08e7	4.12e7	4.31e7	2.10e8
6	3.22e6	2.45e6	3.27e6	2.93e6	2.46e6	4.77e6
7	2.33e6	2.41e6	2.82e6	2.66e6	2.07e6	7.07e6
8	3.98e5	4.07e5	3.86e5	3.95e5	3.34e5	5.69e5
9	1.77e8	3.12e6	4.67e6	3.79e6	2.86e6	5.74e6
10	5.34e8	4.16e8	4.01e8	3.83e8	4.12e8	9.11e8
11	1.66e7	1.43e7	1.64e7	1.52e7	1.36e7	2.71e7

Table 4. Diebold–Mariano Test Statistics Between Best Two Models by Brand

SKU	**Model 1**	**Model 2**	**DM Statistic**
1	RF	Ridge	-0.5719
2	Ridge	RF	-0.5549
3	RF	Ridge	-1.9473*
4	RF	Ridge	-1.2902
5	FS	Ridge	-1.7678*
6	Ridge	RF	0.7798
7	RF	FS	-1.2128
8	RF	LASSO	-1.3872
9	RF	Ridge	-0.7371
10	EN	LASSO	-1.8814*
11	RF	Ridge	-0.5522

Note. The table displays Diebold-Mariano test statistics comparing the two best models against each other, evaluated per brand across all stores. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Negative DM statistic indicates that Model 1 exhibits lower expected forecast loss than Model 2, implying superior predictive accuracy, and vice-versa.

Interestingly, SKUs like 6, 7, and 11 show no significant difference between the top two models. This pattern suggests near equal robustness across modeling approaches for these SKUs, and may also reflect more stable sales patterns across stores. For decision-makers, this means that either of the top two models can be reliably used without significant trade-offs.

Overall, these results highlight that the best forecasting model can vary by product, depending on characteristics such as sales volatility and responsiveness to pricing or promotions. Retailers aiming to use such SKU-level forecasting models should consider this heterogeneity when making decisions.

5 Conclusion

In this study, we developed and compared various models for predicting orange juice sales for a major retail chain based on SKU-store level data. We assessed predictive performance and business relevance by incorporating both traditional econometric methods and more modern machine learning discoveries.

Our conclusion is that no single model consistently outperforms the others for all SKUs. Instead, model performance is product-specific, driven by characteristics such as sales volatility, promotional frequency, and price dynamics. Although all models generally tend to outperform the RW, EN and ridge emerge as the strongest. EN shows particular strength for high-variance SKUs, such as SKU 10, likely due to its robustness to correlated predictors, while RF proves the strongest for SKU 3. FS also holds ground in certain cases, like for SKU 5, where a few key predictors may be dominant.

By carefully analyzing predictive accuracy at the product-SKU level, and applying business-relevant evaluation metrics, we provide evidence of why forecasting strategies can, and should, differ between products. This highlights that retailers should avoid a one-size-fits-all solution for inventory and promotion planning, and adopt an approach that is more tailored to individual products.

However, our study does have some key limitations. Firstly, the results of this paper are based on data from one particular supermarket chain over a short period from 1989 to 1991. This may significantly hinder the generalizability of our results, especially given the age of the dataset, as consumer behavior is likely to change over time. Furthermore, we do not include any contextual information (like weather or consumer characteristics) in our regressors, which may significantly affect sales and therefore improve the accuracy of our models. Future research on the topic of forecasting sales of consumer goods should address these limitations by using more recent and extensive datasets.

References

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858.

Aras, S., İpek Deveci Kocakoç, and and, C. P. (2017). Comparative study on retail sales forecasting between single and combination methods. *Journal of Business Economics and Management*, 18(5):803–832.

Chu, C.-W. and Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics*, 86(3):217–231.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2023). *An Introduction to Statistical Learning: with Applications in R*. Springer, 2 edition.

Qu, R., Timmermann, A., and Zhu, Y. (2024). Comparing forecasting performance with panel data. *International Journal of Forecasting*, 40(3):918–941.

Tukey, J. W. et al. (1977). *Exploratory data analysis*, volume 2. Springer.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Appendix

A Complete List of Potential Predictors

We include the following potential predictors in each SKU-store combination. Let i index SKUs, n index stores, and t index weeks. The design matrix $\mathbf{X}_{i,s}$ comprises the following variables:

1. **Own-brand price and promotion:**
   (a) Natural logarithm of current price: $\ln p_{t,(i,n)}$.
   (b) Promotional dummy indicators: feature $f_{t,(i,n)}$ and deal $d_{t,(i,n)}$.
   (c) Price change: $\Delta p_{t,(i,n)}/p_{t-1,(i,n)}$.
   (d) Discount depth: $\left(\frac{p^{\text{reg}}_{t,(i,n)} - p_{t,(i,n)}}{p^{\text{reg}}_{t,(i,n)}}\right) d_{t,(i,n)}$, where $p^{\text{reg}}_{t,(i,n)}$ is last non–deal price.
2. **Own-brand dynamics:**
   (a) Log-sales lags: $\ln S_{t-k_{(i,n)},(i,n)}$ for $k = 1, \ldots, p$, where $k_{(i,n)}$ is a dynamically determined lag order over the first rolling window, assumed to be constant for the rest of the sample.
   (b) One-week lags of promotions: $f_{t-1,(i,n)}$, $d_{t-1,(i,n)}$.
   (c) Four-week rolling sums of promotion flags: $\sum_{s=1}^{4} f_{t-s,(i,s)}$ and $\sum_{s=1}^{4} d_{t-l,(i,n)}$.
   (d) Four-week rolling mean of log-sales: $\frac{1}{4}\sum_{l=1}^{4} \ln S_{t-l,(i,n)}$.
3. **Competitor aggregates (by week):**
   (a) Mean feature share: $\overline{f}_{t,(-i,n)} = \frac{1}{10}\sum_{j\neq i} f_{t,(j,n)}$.
   (b) Mean deal share: $\overline{d}_{t,(-i,n)} = \frac{1}{10}\sum_{j\neq i} d_{t,(j,n)}$.
   (c) Mean log-sales: $\overline{\ln S}_{t,(-i,n)} = \frac{1}{10}\sum_{j\neq i} \ln S_{t,(j,n)}$.
4. **Competitor price statistics:**
   (a) Average log-price: $\overline{\ln p}_{t,(-i,n)} = \frac{1}{10}\sum_{j\neq i} \ln p_{t,(j,n)}$.
   (b) Minimum log-price: $\min_{j\neq i}\{\ln p_{t,j}\}$.
5. Rolling competitor momentum: $\frac{1}{4}\sum_{l=1}^{4}\sum_{j\neq i} \overline{\ln S}_{t-l,(j,n)}$.

6. **Interaction and relative measures:**

   (a) $f_{t,(i,n)} \times \ln p_{t,(i,n)}$ and $d_{t,(i,n)} \times \ln p_{t,(i,n)}$. Note that these predictors quantify the extent to which features and deals moderate price sensitivity.

   (b) Price gap: $\ln p_{t,(i,n)} - \overline{\ln p}_{t,(-i,n)}$.

   (c) One-week lag of competitor feature share: $\overline{f}_{t-1,(-i,n)}$.

7. Promotion-start indicators:

$$\mathbf{1}\{f_{t,(i,n)} = 1,\ f_{t-1,(i,n)} = 0\}, \quad \mathbf{1}\{d_{t,(i,n)} = 1,\ d_{t-1,(i,n)} = 0\}.$$

8. **Seasonality:** Dummy variables for calendar month.

B List of assumptions required for a consistent smearing estimate

For the "smearing estimate" to be consistent, we require that:

- The retransformation function h is continuously differentiable. This holds in our case as $h(x) = e^x$ is continuously differentiable on $(-\infty,\infty)$, i.e., the range on which the sales of orange juice are defined.
- The regressor matrix X contains an intercept. Again, we ensure this is the case in our regression.
- $\frac{1}{n}X'X \to \Sigma$, where Σ is a positive definite matrix. We assume this to hold.
- $\mathbb{E}\left[\sup_{|t|\leq M}\left(h'(x_0'\beta + \epsilon + t)\right)^2\right] < \infty$ for all $M < 0$. Note that, as mentioned in Duan (1983), the above expression simplifies to $\mathbb{E}(e^{2\epsilon}) < \infty$, when $h(x) = e^x$. Again, we deem this a reasonable assumption for our models and therefore assume it to hold.

C Coefficient of Variation

Table C1. Coefficient of Variation of Forecast Errors by SKU and Model (Raw MSE-Based)

	Model					
SKU	**FS**	**Ridge**	**LASSO**	**EN**	**RF**	**RW**
1	0.990	1.029	1.054	1.052	0.882	0.934
2	1.005	0.898	0.907	0.868	0.945	0.596
3	0.666	0.733	0.704	0.726	0.614	0.524
4	0.592	0.443	0.478	0.464	0.568	0.478
5	0.877	0.943	0.943	0.956	0.869	0.559
6	0.930	0.844	0.810	0.719	0.700	0.405
7	0.479	0.416	0.414	0.385	0.440	0.531
8	0.344	0.390	0.346	0.358	0.365	0.450
9	2.386	0.679	0.917	0.839	0.780	0.776
10	1.289	1.084	0.915	0.871	1.011	0.907
11	0.782	0.762	0.751	0.733	0.724	0.889