No One-Size-Fits-All: A Contextual Study of Retail Forecasting Methods

Thijs van der Windt, 641286
Xavier Jaeger, 657508
Femke van Peyma, 640975
Leonardo Rossi, 638449
Group 42

20 June 2025, Word count abstract: 147, Word count total: 4877

Abstract

This study evaluates methods to determine forecasting models for one-week-ahead orange juice sales across 11 SKUs and 10 stores in the Chicago area. We compare statistical and machine learning approaches, including Hybrid Subset Selection, Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and Extreme Gradient Boosting (XGBoost), using Mean Percentage Squared Error (MPSE), Quasi-likelihood (QLIKE), and relative RMSE (rRMSE) to assess forecast accuracy. Ridge and Elastic Net consistently outperform others. LASSO underperforms due to aggressive variable selection, while XGBoost underperforms throughout, likely due to overfitting limited data. Store- and SKU-level and seasonal analyses reveal method performance varies across contexts, suggesting targeted model selection improves accuracy. For example, Ridge excels in summer and Elastic Net in other seasons. Volatility analysis shows frequency of forecast underestimates increase in sales variance. This paper's findings highlight the importance of regularisation and feature selection in retail sales forecasting to minimise waste and avoid stockouts.

1 Introduction

Predicting future sales is crucial for store managers aiming to maximise profits, as understocking can frustrate customers and overstocking leads to costly waste. With advancements in technology, store managers now have access to enormous datasets about their operations, capturing everything from customer purchasing patterns to prices and promotional activity. However, not all data is equally relevant and incorporating irrelevant variables into predictive models can result in overfitting, reducing forecast accuracy.

To address this, this paper investigates statistical and machine learning methods that identify the most significant predictors of orange juice sales, evaluating their ability to deliver precise and reliable forecasts. Specifically, we employ Hybrid Subset Selection (hereafter referred to as Hybrid), a statistical approach that systematically evaluates variable combinations to optimise model accuracy (James et al., 2013). Additionally, we utilise Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) regression techniques, which enhance model robustness by shrinking less relevant parameters towards zero, thereby balancing model parsimony and predictive power (Hoerl and Kennard, 1970; Tibshirani, 1996). We also investigate Elastic Net, a combination of the strengths of Ridge and LASSO to capture a broader range of predictor relationships (Zou and Hastie, 2005). Finally, we apply Extreme Gradient Boosting (XGBoost), a relatively new machine learning algorithm introduced by Chen and Guestrin (2016), which uses decision trees.

To evaluate these forecasts, we apply a ranking and statistical scoring system based on three loss functions: Mean Percentage Squared Error (MPSE), Quasi-likelihood (QLIKE), and relative RMSE (rRMSE). After this, we dive into multiple scenarios of factors that might influence method choice, and evaluate method-specific strengths and weaknesses for these different factors.

This study contributes to the growing literature on predictive modelling in high-dimensional time series by addressing the complexity of forecasting retail sales across multiple SKUs with correlated predictors such as prices, promotions, and lagged effects. It provides a comparative evaluation of regularised regression and machine learning approaches, offering practical insights into method performance under multicollinearity and temporal dependence, common challenges in applied forecasting settings. The social relevance lies in improving sales predictions to reduce food waste and prevent stockouts, leading to more efficient retail operations and sustainability.

The remainder of the paper is structured as follows: Section 2 describes the data and preprocessing, such as transformations and the handling of outliers. Section 3 outlines the methodology and evaluation metrics. Section 4 presents and analyses empirical results. Finally, Section 5 concludes by summarising key insights and proposing future research directions.

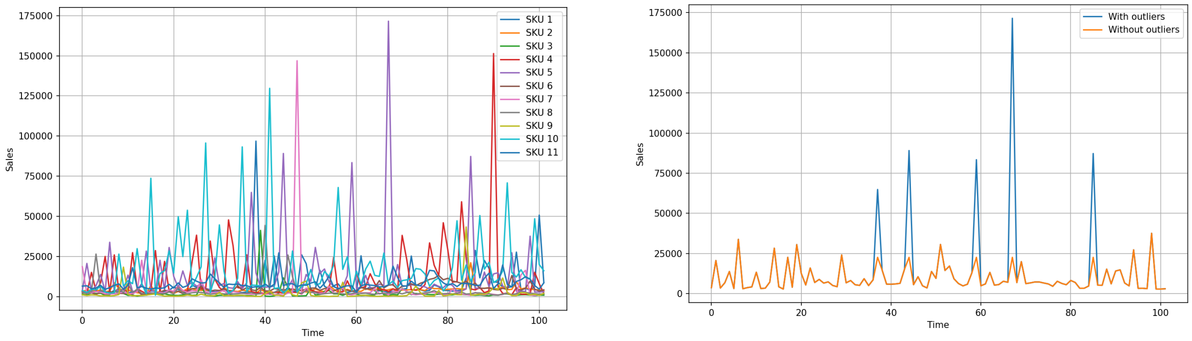
2 Data

The dataset used in this paper is retrieved from the Dominick's Finer Foods chain in the greater Chicago area, as discussed in Wedel and Zhang (2004). The dataset consists of prices, promotional activity, and discounts for 11 brands of refrigerated orange juice (SKUs) for 10 stores. The sample consists of 102 weeks spanning from September 1989 to August 1991.

Figure 1(a) reveals several unusually high observations in the sales data, which may have arisen from entry or measurement errors, or abnormal sales. Including these outliers in our regression models could lead to inaccurate forecasts and skewed evaluation metrics. We define outliers using the Interquartile Range method, following Tukey (1977). We define an outlier as an observation which lies outside of the interval $[Q_1 - 3 * IQR, Q_3 + 3 * IQR]$, where Q_1 and Q_3 are the first (25^{th} percentile) and third (75^{th} percentile) quartiles of the store-and-SKU-specific sales data, and $IQR = Q_3 - Q_1$ is the interquartile range. The multiplier of three is chosen solely to include extreme outliers. Note that here it is assumed it is not necessary to distinguish between promo-deal characteristics, as the data reveals that there is a significant number of observations for each promo-deal combination. This means that the data of lower sales due to no promotion or deals is generally included above the first quartile, and the usual significant increases in sales due to promotions and deals are generally included in the third quartile.

Outliers are replaced with interpolated values derived from the remaining data. To adjust the data, we categorise sales into four groups based on promotional status: (1) no promotions or deals, (2) promotions only, (3) deals only, and (4) both promotions and deals. For each category, we use the median sales value to replace corresponding outliers, as the median is robust to extreme values. We ensure each category within each store-SKU combination contains sufficient data points to support robustness. The analysis proceeds using this adjusted dataset.

Subsequently, the sales data is log-transformed using the natural logarithm. This is useful because it stabilises variance and normalises skewed data. It also allows for modelling percentage changes, which better reflects real-world sales dynamics. Additionally, both sales and price data for all SKUs are standardised to ensure all predictors are on a comparable scale, preventing their original magnitudes from disproportionately affecting the method results.



(a) Orange juice sales for 11 SKUs in store 1, including outliers

(b) Orange juice sales for SKU 5 in store 1, with and without outliers

Figure 1: Orange juice sales in store 1 across 102 weeks (Sep 1989 - Aug 1991), incl. and excl. outliers

Note[1]: Outliers are identified using the Interquartile Range (IQR) method by Tukey (1977), seen as values outside $[Q_1 - 3 * IQR, Q_3 + 3 * IQR]$ for each store-SKU pair. Promo- and deal-related variation is preserved, as promo-deal combinations are well represented in the data. Outliers are replaced using median values within four promo-based categories: (1) no promo/deal, (2) promo only, (3) deal only, and (4) both. Medians are computed per store-SKU-category for robustness.

Table 1: Descriptive statistics for price and sales of orange juice brands after removing outliers

SKU	Brand	Size (oz)	Price/oz	Sales		Features	Deals
				Avg.	St.Dev.		
1	Tropicana Premium	64	0.044	11 615	10 274	14.0	53.9
2	Tropicana Premium	96	0.048	7522	3786	13.0	30.4
3	Florida's Natural	64	0.044	2348	1464	15.0	36.9
4	Tropicana	64	0.035	12 606	14 781	32.0	56.0
5	Minute Maid	64	0.034	13 962	13 151	26.0	56.5
6	Minute Maid	96	0.040	5456	2480	20.0	48.1
7	Citrus Hill	64	0.035	4628	4560	15.0	40.9
8	Tree Fresh	64	0.033	2426	1215	10.0	37.0
9	Florida Gold	64	0.032	1946	1927	16.0	45.9
10	Dominick's	64	0.027	20 715	24 303	28.0	53.0
11	Dominick's	128	0.029	10 409	6541	17.0	29.5

Note[1]: "Price/oz" indicates the average price per fluid ounce of a unit of a certain SKU over time, aggregated over all stores. "Sales Avg." and "Sales St.Dev.", respectively, depict the mean and standard deviation of the sales in fluid ounces per SKU over time, aggregated over all stores. "Features" and "Deals", respectively, represent the average number of features and deals per SKU over time, aggregated over all stores.

Table 1 provides an overview of the average price, sales, deals and features for different orange juice brands, aggregated over all ten stores. This information helps spot products with highly volatile sales as well as those that have a high intensity of marketing tools like features and deals. This breakdown gives useful insights about how customers behave and how brands compete in the orange juice market.

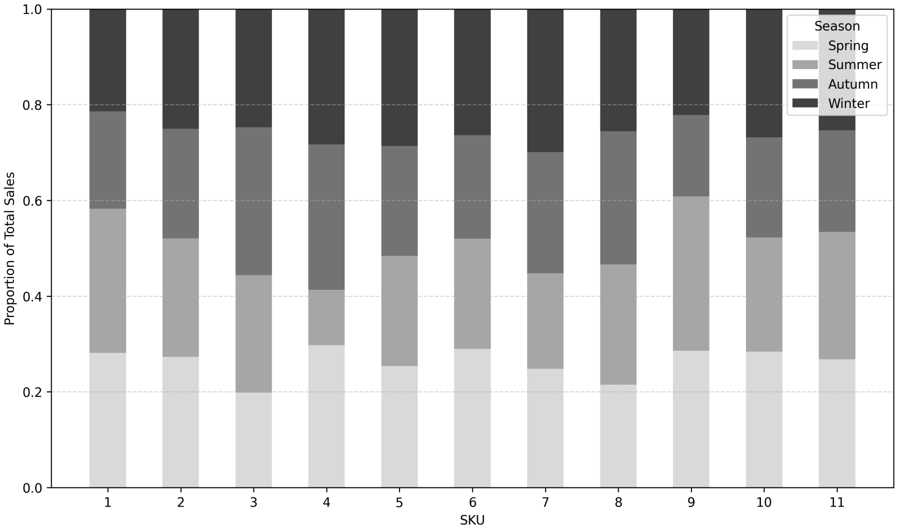


Figure 2: Proportion of total sales for 11 SKUs per season across all 10 stores

Note[1]: Seasons are determined based on typical weather patterns: Spring = March-May, Summer = June-August, Autumn = September-November, Winter = December-February
Note[2]: Week 1 represents the first week of September 1989, week 2 the second, etc., all the way through to week 102.

Figure 2 shows seasonal sales patterns for eleven SKUs, averaged across 10 stores. Ideally, sales would be evenly distributed at 25% per season, but the graph highlights variations. For example, SKU 4 has minimal summer sales, while most brands see a surge during this period. Furthermore, some brands gain popularity in colder months. Promotional patterns do not align with these sales trends, indicating seasonality as the primary driver. Thus, seasonality is a key factor to address in subsequent sections.

3 Methodology

3.1 Potential predictors

The objective of this research is to investigate variable selection methods to determine models that forecast one-week-ahead sales for 11 refrigerated orange juice SKUs using weekly data on sales, prices, and promotional activities. Owing to the high number of potential predictors listed below, the forecasting model must address several econometric challenges: high dimensionality, multicollinearity, and potential overfitting. We start with a linear regression model for the log-transformed fluid ounce sales:

$$\begin{aligned}\ln(\widetilde{S_{i,\ell,t}}) = \alpha_{i,l} + \sum_{i=1}^{11} \beta_{i,l}^{(\mathrm{p})} \, \widetilde{\mathrm{Price}}_{i,\ell,t} + \beta_{i,l}^{(\mathrm{r})} \mathrm{Promo}_{i,\ell,t} + \beta_{i,l}^{(\mathrm{d})} \, \mathrm{Deal}_{i,\ell,t} + \beta_{i,l}^{(\mathrm{int})} \, \mathrm{Promo}_{i,\ell,t} \cdot \mathrm{Deal}_{i,\ell,t} \\ + \sum_{s=1}^{3} \beta_{i,l,s}^{(\mathrm{s})} \mathrm{D}_{i,l,s} + \gamma_{1_{i,l}} \ln(\widetilde{S_{i,\ell,t-1}}) + \gamma_{2_{i,l}} \ln(\widetilde{S_{i,\ell,t-2}}) + \gamma_{3_{i,l}} \, \mathrm{Promo}_{i,\ell,t-1} + \gamma_{4_{i,l}} \, \mathrm{Deal}_{i,\ell,t-1} \\ + \gamma_{5_{i,l}} \, (\mathrm{Promo}_{i,\ell,t-1} \cdot \mathrm{Deal}_{i,\ell,t-1}) + \varepsilon_{i,\ell,t}, \qquad (1)\end{aligned}$$

where $i \in \{1, ..., 11\}, \ell \in \{1, ..., 10\}, t \in \{3, ..., 102\}$, and $s \in \{1, ..., 3\}$. $\widetilde{\ln(S_{i,\ell,t})}$ is the standardised log-transformed fluid ounce sales of SKU i at store ℓ in week t, this index notation holds for all the listed predictors in the full model above. $\text{Price}_{i,\ell,t}$ is the price (own and competitors'), $\text{Promo}_{i,\ell,t}$ the promotion indicator (e.g., feature ad), $\text{Deal}_{i,\ell,t}$ the price reduction or in-store deal indicator, $\text{Promo}_{i,\ell,t} \cdot \text{Deal}_{i,\ell,t}$ the interaction effect between promotion and deal for the focal SKU, $\ln(\widetilde{S_{i,\ell,t-1}}), \ln(\widetilde{S_{i,\ell,t-2}})$ the lagged standardised log sales to capture persistence, $\text{Promo}_{i,\ell,t-1}, \text{Deal}_{i,\ell,t-1}$ the one-week lagged promotion and deal indicators, and $\text{Promo}_{i,\ell,t-1} \cdot \text{Deal}_{i,\ell,t-1}$ the interaction of lagged promotion and deal. We also add seasonal dummies $\text{D}_{i,l,s} = 1$ if time t is in season s.

The economic rationale for including own prices and promotion indicators stems from the fact that lower prices and promotional activities (deals or features) are expected to directly boost sales of the promoted SKU. The inclusion of lagged sales and lagged promotional activity accounts for persistence and post-promotion dips as people over-purchased during promotions. Seasonal dummies are included to capture predictable fluctuations in demand across time, like holidays or weather-driven trends.

In contrast, we exclude promotional indicators for other SKUs to avoid multicollinearity and reduce model complexity, especially in a high-dimensional setting. Moreover, cross-SKU promotional effects tend to be weaker and less consistent, and are often absorbed by the lagged sales terms.

3.2 Statistical/econometric methods

3.2.1 Hybrid Subset Selection (Hybrid) procedure

Hybrid Subset Selection is a statistical variable selection method that combines forward and backward stepwise selection (James et al., 2013). Starting from the empty model, predictors are added based on the greatest improvement in the Akaike Information Criterion (AIC) (Akaike, 1974), which we prefer over the Schwarz Information Criterion (SIC) due to its stronger focus on predictive accuracy over parsimony (Brewer et al., 2016). After each addition, backward selection checks if removing any predictor further improves AIC, repeating until no improvement is possible. Subsequently, we resume the forward selection and repeat the steps above. This process balances the limitations of using forward (backward) selection alone, namely, once a variable is added (removed), it cannot be removed (added). This can lead to suboptimal models. While best subset selection ensures optimal fit, it is computationally infeasible here. The hybrid approach offers a practical and efficient alternative.

3.2.2 Ridge regression

Ridge regression introduces a quadratic penalty on the magnitude of the coefficients, leading to the following optimisation problem, following James et al. (2013):

$$\hat{\beta}_{i,\ell}^{\text{Ridge}} = \arg\min_{\beta_{i,\ell}} \left\{ \sum_{t=1}^{T} \left(\ln(\widetilde{S_{i,\ell,t}}) - \sum_{j=1}^{K} x_{i,\ell,t,j} \beta_{i,\ell,j} \right)^2 + \lambda_R \sum_{j=1}^{K} \beta_{i,\ell,j}^2 \right\}, \tag{2}$$

where K is the number of regressors, $x_{i,\ell,t,j}$ the regressors and $\beta_{i,\ell,j}$ the coefficients in Equation 1. This formulation shrinks the coefficients towards zero, thereby mitigating issues of multicollinearity and overfitting. Although Ridge regression does not set any coefficients to zero, it is particularly effective when many predictors exert small but non-negligible effects on sales. The regularisation parameter $\lambda_R > 0$ controls the extent of shrinkage and the standardisation of the variables ensures the penalty term is applied fairly across all coefficients.

3.2.3 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO regression, by contrast, applies an ℓ_1-norm penalty on the absolute values of the coefficients, following James et al. (2013):

$$\hat{\beta}_{i,\ell}^{\text{LASSO}} = \arg\min_{\beta_{i,\ell}} \left\{ \sum_{t=1}^{T} \left(\widetilde{\ln(S_{i,\ell,t})} - \sum_{j=1}^{K} x_{i,\ell,t,j} \beta_{i,\ell,j} \right)^2 + \lambda_L \sum_{j=1}^{K} |\beta_{i,\ell,j}| \right\} \tag{3}$$

Unlike Ridge, LASSO both shrinks coefficient estimates and sets some to zero. This feature allows it to perform variable selection and estimation simultaneously, which is particularly advantageous for identifying the most influential variables (e.g., brand-specific promotions or competitor pricing) affecting each SKU's sales. The downside is that it tends to underestimate the true effect sizes of variables that are actually important. This bias is the trade-off for reducing variance and achieving variable selection.

3.2.4 Elastic Net

The Elastic Net regression combines the penalties of LASSO and Ridge, leading to the following formulation, based on (Zou and Hastie, 2005):

$$\hat{\beta}_{i,\ell}^{\text{ElasticNet}} = \arg\min_{\beta_{i,\ell}} \left\{ \sum_{t=1}^{T} \left(\widetilde{\ln(S_{i,\ell,t})} - \sum_{j=1}^{K} x_{i,\ell,j,t} \beta_{i,\ell,j} \right)^2 + \lambda_1 \sum_{j=1}^{K} |\beta_{i,\ell,j}| + \lambda_2 \sum_{j=1}^{K} \beta_{i,\ell,j}^2 \right\} \tag{4}$$

This formulation provides a better balance between variable selection and coefficient shrinkage, reducing issues of multicollinearity and overfitting while allowing for the selection of groups of correlated predictors. The penalties simultaneously perform variable selection and shrinkage, minimising the variance against variable selection trade-off as described in Section 3.2.3.

3.2.5 Extreme Gradient Boosting (XGBoost)

XGBoost (Chen and Guestrin, 2016) uses decision trees to make predictions, capturing the non-linear relationships between predictors. These variables often interact in complex ways that linear models may fail to capture accurately. One of XGBoost's strengths is that it can detect patterns even with relatively small datasets, making it a suitable choice for our sample of 102 weeks. In addition, XGBoost includes regularisation to reduce the risk of overfitting and provides insight into which variables are the most important drivers of sales. It works by building many small decision trees, where each tree learns from the errors of the model so far by minimising an objective function that balances prediction accuracy and model simplicity. We set the number of trees to 200, the maximum depth to 8, and the learning rate to 0.1, as described by Chen and Guestrin (2016).

By introducing XGBoost alongside Ridge and LASSO, we aim to evaluate the benefits of non-linear modelling techniques in capturing complex sales dynamics while maintaining a rigorous comparison framework across methods.

3.3 Implementation

3.3.1 Window length

In our forecasting method, we employ a rolling window approach with a window length of 70 observations, which balances method stability with adaptability by capturing structural changes in the data over time.

We first generate 30 forecasts for the most recent data points, namely weeks 73 to 102. These forecasts correspond to the spring and summer weeks. Secondly, to address potential seasonal biases in method performance, we use wraparound forecasting to predict sales for earlier weeks 3 to 24 using later data. These point forecasts correspond to autumn and winter periods. By forecasting across all 52 weeks of the year, we ensure that the method's predictive accuracy is not biased by seasonal patterns, thereby enhancing its robustness and generalisability.

3.3.2 Choosing the optimal λ

We use blocked 3-fold cross-validation (CV) within each expanding forecast window to estimate the optimal regularisation parameter λ for Ridge, LASSO and Elastic Net regression. Unlike standard K-fold CV, which assumes i.i.d. observations and typically involves random shuffling, blocked CV respects the time-series structure by dividing the data into chronologically ordered, contiguous folds (Bergmeir and Benítez, 2012). This is crucial, as observations are serially dependent, and shuffling the data would lead to information leakage from the future into the past, invalidating model and method assessment. While no universally optimal method exists for selecting λ (Arlot and Celisse, 2010), blocked CV balances bias, variance, and computational feasibility, particularly valuable for models like LASSO without closed-form solutions. Given our relatively small expanding window of 70 observations, we set $K = 3$ to retain a sufficient training sample within each fold relative to the number of predictors while enabling robust method evaluation.

For each cross-validation iteration, we divide the 70 observations into three contiguous blocks, preserving the temporal order. We then construct two expanding training sets: the first uses the initial block for training and the second block for validation, the second training set combines the first and second blocks, with the third block used for validation. This blocked approach ensures that each training set only includes past data relative to its validation block, avoiding lookahead bias. Model performance for each λ from a specified range explained below is assessed via mean squared error computed on each validation segment. The λ with the lowest average cross-validated error is selected for final model fitting.

We construct a linearly spaced grid of 100 λ values from 10^8 to 10^{-2}, scaled by the standard deviation of sales, following James et al. (2013). This captures the full range from the empty model (very high λ) to the ordinary least squares fit ($\lambda = 0$). The logarithmic scale ensures a finer resolution for smaller λs, where model coefficients begin to enter. Scaling by std(sales) ensures that the regularisation remains meaningful and comparable as sales levels and volatility vary across time and SKUs.

3.3.3 Transforming log sales

Predictions on the log scale will be retransformed to the sales level using the smearing method by Duan (1983), as taking the exponential without correcting by a factor introduces a bias due to the nonlinearity of the exponential function. This method also accounts for the potentially non-normal distribution of the model residuals. The ML techniques additionally produce a forecast of normalised log sales for SKU i at time T. Before applying the smearing factor, we reverse the normalisation to obtain a prediction for log sales in the original scale, using the standard deviation and the mean of the training data. We obtain the final forecast of sales as:

$$\hat{S}_{i,T} = \exp\left(\widehat{\ln(S_{i,T})}\right) \cdot \text{Smearing Factor}, \quad \text{where Smearing Factor} = \frac{1}{T}\sum_{t=1}^{T}\exp(\hat{\varepsilon}_{i,t}). \tag{5}$$

$\hat{\varepsilon}_{i,t} = \ln(S_{i,t}) - \widehat{\ln(S_{i,t})}$, where $\widehat{\ln(S_{i,t})}$ is obtained from the fitted values of the training data.

3.4 Method performance comparison: Forecast error evaluation

To evaluate and compare our methods' predictive performance, we rely on rankings and statistical scoring of forecast accuracy, measured by three different loss functions. We aggregate loss functions, ranks and scores within varying groups to analyse performance from different perspectives. These groups include stores (aggregating across all weeks and SKUs per store), SKUs, seasons, and a global group (aggregating across all weeks, SKUs, and stores). Data points g within a group G are denoted as: $g = (i, l, t) \in G$, representing SKU i, store l, and time t.

3.4.1 Loss functions

The first loss function is Mean Percentage Squared Error (MPSE), as described in Pindyck and Rubinfeld (1988):

$$\text{MPSE} = \frac{1}{|G|}\sum_{g\in G}\left(\frac{S_g - \hat{S}_g}{S_g}\right)^2 \times 100 \tag{6}$$

MPSE normalises error by actual sales, allowing fair comparison across SKUs with different sales volumes. It expresses errors relatively, penalising large percentage deviations equally, regardless of the scale of sales.

The second loss function is QLIKE, based on Patton (2011):

$$\text{QLIKE} = \frac{1}{|G|}\sum_{g\in G}\left(\log \hat{S}_g + \frac{S_g}{\hat{S}_g}\right). \tag{7}$$

QLIKE uses a logarithmic scoring rule that penalises underprediction more than overprediction. This is valuable in retail, where shortages can lead to dissatisfied customers.

The third loss function is rRMSE, which compares the RMSE of a method's forecasting model to that of a random walk (rw) benchmark, a naïve baseline:

$$rRSME = \frac{RMSE_{model}}{RMSE_{rw}} = \frac{\sqrt{\frac{1}{|G|}\sum_{g\in G}(S_g - \hat{S}_g)^2}}{\sqrt{\frac{1}{|G|}\sum_{g\in G}(S_g - \hat{S}_g^{\text{RW}})^2}} \tag{8}$$

The random walk baseline enables consistent comparisons across SKUs with varying sales patterns. An rRMSE below 1 indicates that the method outperforms the benchmark, while values above 1 signal underperformance.

3.4.2 Ranks and scores

Running the regressions yields 1650 loss values: 5 methods × 110 store-SKU pairs × 3 loss values. To aggregate these, we first assign ranks to each method per loss function within a group G, with rank 1 assigned to the method with the lowest average loss.

Ranks equally reward/punish two methods regardless of whether their performances are statistically significantly different or not. Therefore, we also implement a scoring system. For each pairwise method comparison within G, we assign +1 to the better method and -1 to the worse based on a significance test. We match up all bilateral combinations of methods for each grouping, G, so that each method can score between -4 and +4.

To test for significant differences in predictive accuracy, we apply the Pooled Diebold-Mariano (DM) Panel test for equal predictive accuracy, following Timmermann and Zhu (2019). Instead of the standard squared forecast error loss, our test uses a general loss differential:

$$\Delta L_{g|t} = L_{g|t}^{(m_1)} - L_{g|t}^{(m_2)}, \tag{9}$$

where $L_{g|t}^{(m)}$ denotes the loss values (either MPSE, rRMSE or QLIKE) for method m and combination $g = (i, l, t+1) \in G$, which includes SKU i, store l, and time $t+1$, given that we are only performing a one-step-ahead point forecast. This aligns with Elliott et al. (2005), which shows that taking a more general loss value approach does not violate the required assumption that there is weak serial dependence in the sequence of forecast losses (Timmermann and Zhu, 2019). This has been tested and confirmed with autocorrelation plots for every store-SKU combination: there is always at most one observation that lies outside the 95% confidence interval, implying there is not sufficient evidence to reject the null hypothesis of no autocorrelation. Specifically, considering $e_{g|t}^{(m_i)}$ is the one-point ahead forecast error of index g for method m_i:

$$\text{MSPE:} \quad L_{g|t}^{(m_i)} = \frac{e_{g|t}^{(m_i)}}{S_{g|t}} \tag{10}$$

$$\text{rRMSE:} \quad L_{g|t}^{(m_i)} = \frac{e_{g|t}^{(m_i)}}{e_{g|t}^{\text{RW}}} \tag{11}$$

$$\text{QLIKE:} \quad L_{g|t}^{(m_i)} = \frac{S_g}{\hat{S}_g^{(m_i)}} - \ln\left(\frac{S_g}{\hat{S}_g^{(m_i)}}\right) - 1 \tag{12}$$

We test $H_0^{Gpool} : \mathbb{E}(\overline{L_G^{(m_1)}}) = \mathbb{E}(\overline{L_G^{(m_2)}})$ using the test statistic:

$$J_G^{DM} = \sqrt{|G|} \sum_{g \in G} \frac{\Delta L_{g|t}}{\hat{\sigma}(\Delta L_{g|t})}, \tag{13}$$

where $\hat{\sigma}(\Delta L_{g|t})$ is estimated via Newey and West (1987).

We first analyse methods' forecast performance globally across all store-SKU-time combinations. We then split by season (four time periods), and finally investigate store-specific and SKU-specific results.

4 Results

4.1 Global performance

Table 2 summarises the global performance of the five methods, aggregating across all stores, SKUs, and time periods. The left panel displays the average ranks based on three loss metrics (rRMSE, MPSE, and QLIKE), while the right panel reports Pooled Diebold-Mariano (DM) Panel test-based scores that capture statistically significant pairwise performance differences. Some key insights emerge.

Table 2: Global method performance using average ranks and Pooled Diebold-Mariano Panel test significance scores

	Ranks				**DM Scores**			
Method	rRMSE Rank	MPSE Rank	QLIKE Rank	Rank Average	rRMSE Score	MPSE Score	QLIKE Score	Score Average
Ridge	2.31	2.70	2.35	2.45	0	4	3	2.33
ElasticNet	2.86	2.41	2.46	2.58	0	2	3	1.67
Lasso	3.35	2.89	3.16	3.13	0	0	-1	-0.33
Hybrid	2.86	3.27	2.79	2.97	0	-3	-1	-1.33
XGBoost	3.62	3.73	4.24	3.86	0	-3	-4	-2.33

Note[1]: "Rank Average" is the mean of the three column-rank values. Lower rank indicates lower loss values than competing methods. Higher rank indicates the opposite.
Note[2]: "Score Average" is the mean of the three Pooled Diebold-Mariano (DM) Panel test (Timmermann and Zhu, 2019; Elliott et al., 2005) scores (one per loss function). Positive scores indicate a method significantly outperformed its competitor methods more often than vice versa at the 5% level. Negative scores indicate the opposite.

Ridge regression achieves the lowest average rank (2.45), closely followed by Elastic Net (2.58), with both consistently outperforming others in raw loss values. LASSO and Hybrid perform moderately (ranks 3.13 and 2.97), while XGBoost trails at 3.86. This ordering is broadly intuitive: Ridge and Elastic Net both retain all predictors (to varying degrees), whereas LASSO introduces hard sparsity and XGBoost prioritises flexibility over parsimony. The fact that Elastic Net lies between Ridge and LASSO, both in rank and score, aligns well with its blended penalty structure.

Beyond ranks, score-based comparisons offer a stricter and more statistical view. Ridge scores highest (2.33), followed by Elastic Net (1.67), both significantly outperforming others in pairwise tests at the 5% level. LASSO, Hybrid, and XGBoost perform poorly, with XGBoost clearly underperforming, likely due to overfitting or limited data misaligned with the model's complexity. Elastic Net's gains over Ridge may be marginal and inconsistent across the panel, while Ridge shows stronger and more frequent gains in key subsets. These scores reinforce the ranking while underlining Ridge and Elastic Net's dominance.

Still, not all metrics align. Elastic Net ranks better under MPSE than Ridge (2.41 vs. 2.70), yet Ridge has a higher DM score (4 vs. 2), suggesting more frequent significant wins despite smaller average losses. This highlights the difference between average performance and statistically meaningful gains. The DM test captures these subtleties, offering a complementary perspective to raw averages.

Although rRMSE scores are uniformly zero across all methods, due to volatility in the benchmark error inflating the denominator, this metric does not alter the overall ranking of methods. While useful for general benchmarking, in short and volatile sales series, the random walk often generates large, erratic errors. This inflates the denominator in the rRMSE ratio randomly at different times and obscures real differences between methods. However, since the order remains consistent with that implied by MPSE and QLIKE, we retain rRMSE in our reporting for completeness, acknowledging that while it contributes little to the magnitude of the method scores, it does not affect our core conclusions regarding relative method performance.

While Ridge emerges as the best overall method, the differences are not large enough to conclusively reject Elastic Net as a viable alternative. This motivates the disaggregated analysis that follows, where we examine method performance by season, store, SKU, and volatility regime. The global test averages across heterogeneous contexts. The scores may therefore mask local performance differences, namely, a method might dominate in one setting but lag in another. This contextual variability is central to understanding forecast performance in practice.

4.2 Performance per store

Evaluating method performance at store level helps determine whether forecasting accuracy varies across locations due to differences in scale, pricing dynamics, or promotional sensitivity. Since inventory decisions are made per store, local forecast accuracy is operationally critical and also serves as a robustness check. Methods that perform consistently across stores are more scalable and reliable.

Figure 3 depicts method performance across the 10 stores. Although some variability exists, no extreme outliers or store-specific anomalies dominate the pattern. Most stores show moderate dispersion in score across methods.

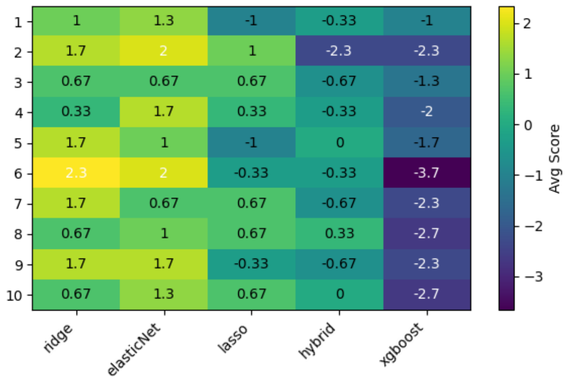


Figure 3: Average Pooled Diebold-Mariano forecast accuracy scores for 10 stores for five variable selection methods

Note[1]: Each cell reflects the method's mean bilateral test score across all SKUs for a given store. The values quantify how often a method statistically outperformed its competitors in one-week-ahead sales forecasting.
Note[2]: A score of +4 means the method significantly outperformed all four other methods for that store-SKU combination. A score of 0 means the method's performance was statistically indistinguishable from others. A score of -4 indicates the method performed significantly worse than all four alternatives.
Note[3]: The scores are aggregated at the store level, across all 11 SKUs and 52 forecast windows, using the Pooled Diebold-Mariano (DM) Panel test for equal predictive accuracy, outlined in Timmermann and Zhu (2019) and Elliott et al. (2005).

Ridge regression and Elastic Net clearly outperform the other methods across all stores, as indicated by their consistently positive scores, highlighting both methods as strong candidates for deployment.

Ridge achieves the highest individual score of 2.3 in store 6, where it outperforms all alternatives. However, a closer comparison reveals some important distinctions that favour Elastic Net. Ridge has a slightly lower average score (1.24) than Elastic Net (1.333). Additionally, the minimum score for Elastic Net is notably higher at 0.67, compared to Ridge's 0.33. This minimum score only occurs twice for Elastic Net, whereas Ridge falls to 0.66 three times. This pattern suggests that Elastic Net not only performs well on average but also does so with greater consistency and fewer extreme under-performances.

Given its stronger minimum performance and comparable top scores, Elastic Net appears to offer more robust and scalable forecasting. This makes it a particularly suitable choice in contexts such as franchise-wide implementation where only a single method can be deployed and consistent reliability across varied locations is essential.

On the contrary, LASSO displays inconsistent results. Although it occasionally performs better than the average (e.g. store 2), it often has negative scores. This means it is beaten significantly more than it beats other methods. This suggests its aggressive variable selection might drop important predictors, which may lead to weaker performance.

XGBoost is the worst-performing method across all stores, with consistently negative scores. This trend reinforces the method's unsuitability to forecast one-day ahead orange juice sales.

4.3 Performance per SKU

Alongside store-level analysis, we also evaluate performance at the SKU level. While inventory decisions are store-based, tailoring methods to individual SKUs can improve accuracy, as SKUs differ in their sales dynamics (some driven by trends, others by promotions or seasonality). If feasible, using different methods per SKU can enhance forecasting and inventory management.

A similar analysis to that in Section 4.2 is conducted at the SKU level, with results presented in Figure 4. The patterns are broadly consistent with those in Figure 3, reinforcing earlier findings: Ridge and Elastic Net emerge as the top-performing methods across most SKUs, while XGBoost consistently underperforms.

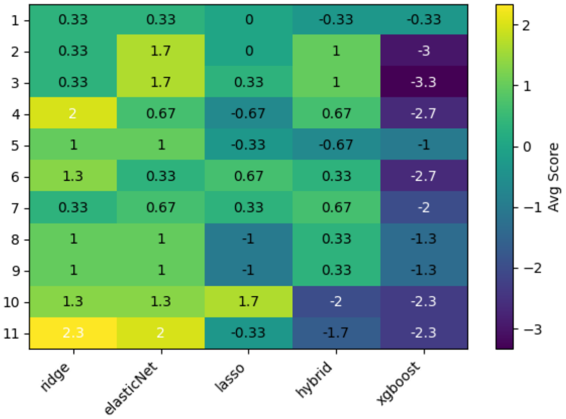


Figure 4: Average Pooled Diebold-Mariano forecast accuracy scores for 11 SKUs for five variable selection methods

Note[1]: Each cell reflects the method's mean bilateral test score across all SKUs for a given store. The values quantify how often a method statistically outperformed its competitors in one-week-ahead sales forecasting.
Note[2]: A score of +4 means the method significantly outperformed all four other methods for that store-SKU combination. A score of 0 means the method's performance was statistically indistinguishable from others. A score of -4 indicates the method performed significantly worse than all four alternatives.
Note[3]: The scores are aggregated at the SKU level, across all 10 stores and 52 forecast windows, using the Pooled Diebold-Mariano (DM) Panel test for equal predictive accuracy, outlined in Timmermann and Zhu (2019) and Elliott et al. (2005).

More specifically, Ridge performs particularly well for all SKUs apart from 2, 3, 10 and 11, while Elastic Net shows strong results for SKUs 2, 3, and 11. Interestingly, LASSO only demonstrates competitive performance for SKU 10, which stands out as an exception. Therefore, a store manager with a larger budget is recommended to deploy said methods accordingly for each of their SKUs.

Comparing these findings with the descriptive statistics in Table 1, SKU 10's high sales volatility suggests that LASSO performs relatively better for SKUs with greater variance. This may be because LASSO's aggressive feature selection focuses on the most influential predictors while discarding weaker ones that, in highly volatile contexts, contribute more noise than signal, ultimately improving forecast accuracy.

4.4 Performance per season

As established in Section 2, orange juice sales exhibit clear seasonal patterns. This suggests that, besides store- and SKU-specific patterns, method performance may also vary across seasons, as different methods might be more (or less) effective at capturing seasonal fluctuations in consumer demand. Figure 5 shows the results of this analysis.

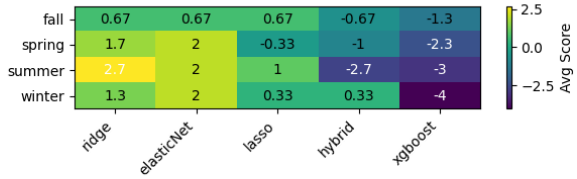


Figure 5: Average Pooled Diebold-Mariano forecast accuracy scores for four seasons for five variable selection methods

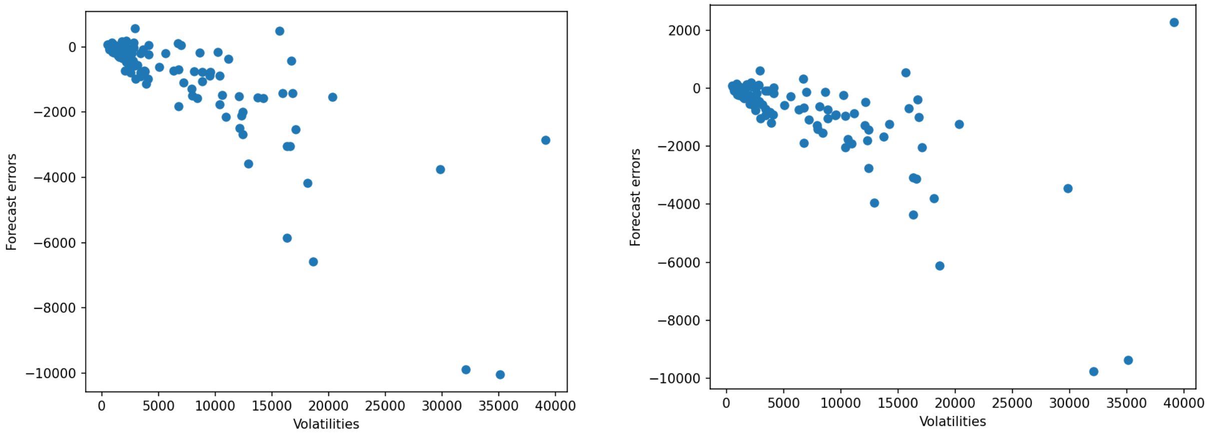
Note[1]: Each cell reflects the method's mean bilateral test score across all stores and SKUs for a given season. The values quantify how often a method statistically outperformed its competitors in one-week-ahead sales forecasting.
Note[2]: A score of +4 means the method significantly outperformed all four other methods for that store-SKU combination. A score of 0 means the method's performance was statistically indistinguishable from others. A score of -4 indicates the method performed significantly worse than all four alternatives.
Note[3]: The scores are aggregated at the season level, across all 11 SKUs, 10 stores and 52 forecast windows, using the Pooled Diebold-Mariano (DM) Panel test for equal predictive accuracy (Timmermann and Zhu, 2019; Elliott et al., 2005).

Figure 5 indicates that Ridge performs best in predicting summer sales, standing out as the clear winner during this season. This aligns with the earlier observation that Ridge performs well for SKUs with a higher proportion of summer sales (SKUs 1 and 9 in Figures 2 and 4). For the other seasons, Elastic Net consistently delivers the best performance. Meanwhile, XGBoost again underperforms relative to the other methods, with particularly poor results in winter.

For stores limited to a single forecasting method, a season-specific strategy is recommended: use Ridge regression in summer and Elastic Net for the rest of the year. This approach balances accuracy and practicality by leveraging each method's strengths across seasons.

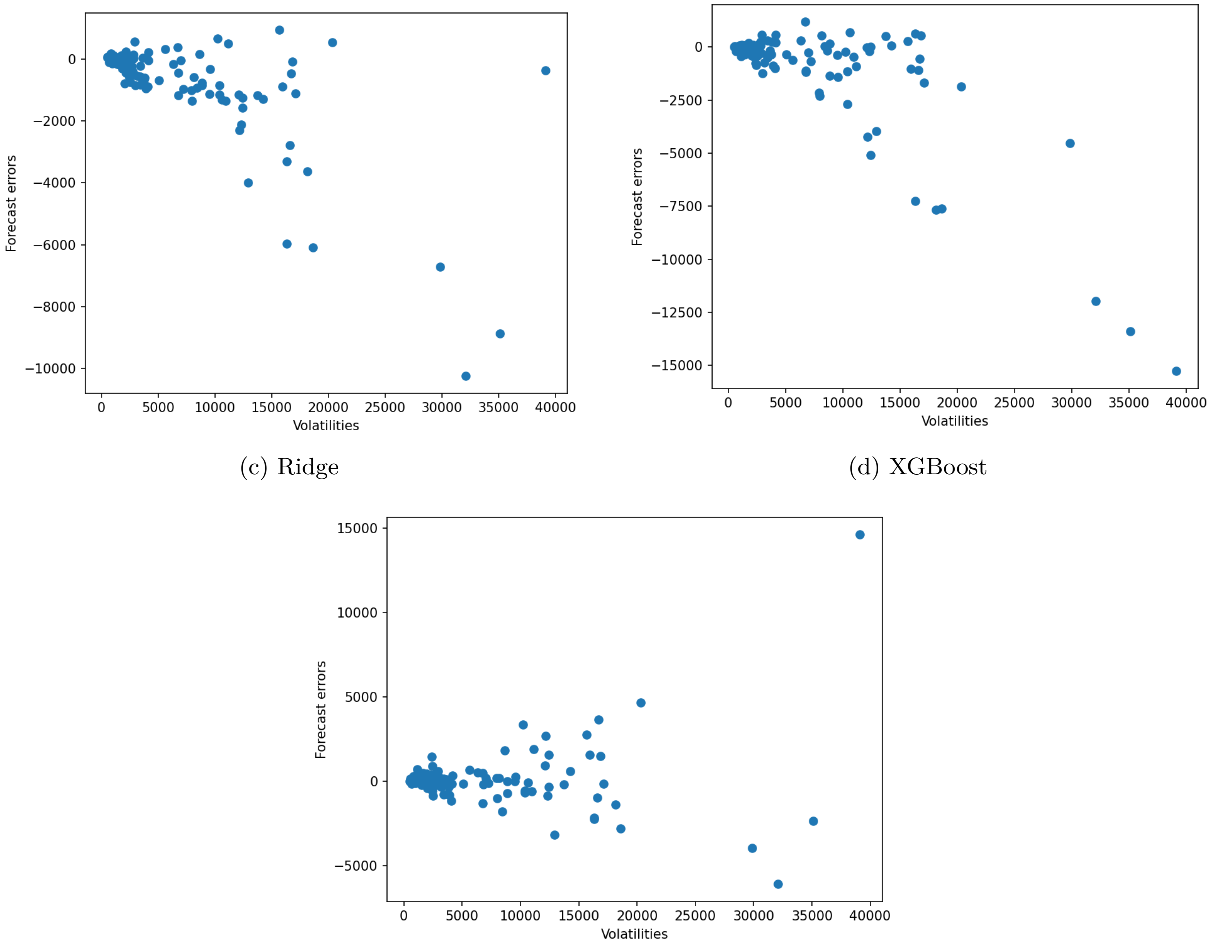
4.5 Performance in various volatility regimes

Besides the factors discussed above, there might be other characteristics affecting sales patterns. Different volatility regimes, driven by macroeconomic indicators, may have an impact on consumer behaviour and hence grocery store sales. The ability to model such effects is of primary importance. For this reason, we examine the relationship between sales volatility and prediction accuracy. Figure 6 presents the relationship between the standard deviation of sales during the forecast horizon and the corresponding average forecast error. The analysis is conducted at the store-SKU level, resulting in 110 data points per method.



(a) Elastic Net

(b) Lasso



(c) Ridge

(d) XGBoost

(e) Hybrid

Figure 6: Scatter plots of volatilities against forecast errors across different variable selection methods

Note[1]: The graphs are obtained by calculating the volatility of the sales across the forecasting period and by calculating the average absolute forecast errors across the forecasting period. This is done for all stores and SKUs, resulting in 110 data points per method.

Most methods show similar patterns, except Hybrid. Forecast errors generally increase with sales volatility, indicating underprediction in volatile environments, particularly in lower-volatility ranges where forecast accuracy consistently declines. This suggests methods struggle to capture the dynamics of frequent or high sales spikes. Hybrid, however, exhibits more symmetric errors around zero, indicating less conservative predictions during high-volatility periods, which helps store managers avoid empty shelves.

A limitation is the limited high-volatility observations, which are underrepresented in the dataset. Nonetheless, the clear trend in low-volatility data confirms that volatility reduces forecast accuracy. This highlights the importance of a store identifying its consumer behaviour and drivers of volatility, such that it does not end up with empty shelves.

5 Conclusion

This study evaluated the forecasting performance of statistical and machine learning methods for one-week-ahead sales predictions of refrigerated orange juice across 11 SKUs in 10 stores, using Dominick's Finer Foods data (1989-1991). We compared Hybrid Subset Selection (Hybrid), Ridge, LASSO, Elastic

Net, and XGBoost methods, assessing their predictive accuracy through a ranking and scoring system based on three loss functions (MPSE, QLIKE, and rRMSE) and the Pooled Diebold-Mariano (DM) Panel test. Out-of-sample forecasts are evaluated across store-level, SKU-level, seasonal, and volatility conditions to provide insights for retail inventory management.

Ridge and Elastic Net outperform other methods both globally (across all store-SKU combinations over time) and across most scenarios. LASSO and Hybrid perform moderately. LASSO occasionally benefits from selecting key predictors but often loses information by zeroing coefficients, unlike Ridge, which retains all predictors. XGBoost consistently performs worst, likely due to its sensitivity to small datasets, which limits its ability to capture complex nonlinear dynamics effectively.

At the SKU level, LASSO performs better for higher variance items due to aggressive feature selection. Seasonally, Ridge excels in summer, while Elastic Net performs better otherwise. We, hence, suggest a season-specific hybrid strategy. Forecast errors increase with sales volatility across most methods, with Hybrid uniquely maintaining symmetric errors, making it less conservative and useful for avoiding stockouts.

For practical applications, we recommend a tailored model selection strategy based on seasonal and store-specific factors. For retailers opening new stores or forecasting for existing ones, deploying Ridge regression during summer periods maximises accuracy. For other seasons, Elastic Net offers the best approach. This dual-method strategy is ideal for store managers with sufficient budgets, allowing them to apply Ridge and Elastic Net selectively for each SKU to capture seasonal dynamics effectively. If a store has a greater budget to differentiate between methods for each SKU, it is recommended to use Ridge, LASSO, and Elastic Net, depending on the strength of the predictors used to forecast the SKU. For franchise-wide implementations where only a single method can be deployed, Elastic Net is recommended for its consistent reliability across diverse locations. This ensures stable performance regardless of store-specific variations.

Despite its contributions, this study has limitations. The 3×IQR multiplier to determine the threshold for outlier removal may discard valid extreme observations, potentially affecting method robustness. Additionally, the 102-week sample limits statistical power, particularly in high-volatility periods, which may weaken DM test significance. A larger dataset would also enable longer rolling windows, potentially allowing XGBoost to exploit its superior ability to capture complexity. The reliance on log-transformation and the smearing estimator relies on estimations, which may distort results.

Future research should explore alternative outlier detection methods, such as robust statistical techniques. Approaches like the Hampel filter (Hampel, 1974) or Tukey's biweight function (Tukey, 1977) reduce the influence of noise and outliers without discarding valid extreme sales events. Expanding the dataset with more recent or longer time series would improve statistical significance and method generalisability. Introducing more volatile periods and SKUs with large variations in sales would allow for further investigation into LASSO's performance in high-volatility contexts. Furthermore, analysing more SKUs could reveal patterns behind forecasting accuracy variation across different SKUs. Additionally, incorporating exogenous variables, such as macroeconomic indicators or consumer sentiment, into XGBoost or Elastic Net may improve performance in volatile settings. Finally, a combined approach combining Ridge for short-term, high-demand periods and Elastic Net for stable ones could further optimise forecasting accuracy.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Brewer, M. J., Butler, A., and Cooksley, S. L. (2016). The relative performance of aic, aicc and bic in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7(6):679–692.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, CA, USA. ACM.

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

Elliott, G., Timmermann, A., and Komunjer, I. (2005). Estimation and testing of forecast rationality under flexible loss. *The Review of Economic Studies*, 72(4):1107–1125.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer, New York.

Newey, W. K. and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 28(3):777–787.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Pindyck, R. S. and Rubinfeld, D. L. (1988). *Econometric Models and Economic Forecasts*. McGraw-Hill, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Timmermann, A. and Zhu, Y. (2019). Comparing forecasting performance with panel data. *Journal of Econometrics*, 211(1):69–83.

Tukey, J. W. (1977). *Exploratory Data Analysis*, volume 2. Addison-Wesley, Reading, MA.

Wedel, M. and Zhang, J. (2004). Analyzing brand competition across subcategories. *Journal of Marketing Research*, 41(4):448–456.

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.