JOURNAL OF
EMERGING INVESTIGATORS

Article

# Analyzing market dynamics and optimizing sales performance with machine learning

**Sumedh Kamat[1], Joseph Shamsian[2]**

[1] Evergreen Valley High School, San Jose, California

[2] Carnegie Mellon University, Pittsburgh, Pennsylvania

## SUMMARY

**In today's rapidly evolving business landscape, understanding the market is crucial for maximizing sales. However, the overwhelming amount of market factors to consider makes it extremely difficult for companies to pinpoint the key drivers of sales performance. Focusing on the most influential market factors is crucial, as it enables businesses to concentrate on what truly drives sales. This precision can make or break a business, and failure to understand sales trends can leave companies struggling to compete in a dynamic market. In this research market study, we used machine learning regression models of sales data from Corporación Favorita, the largest Ecuadorian grocery franchise, to analyze which factors influence sales in grocery retail. We hypothesized that macroeconomic factors have a larger impact on sales compared to geographic and seasonal features. In this project, we used the sales data for training lasso and ridge regression models, which were subsequently examined through Shapley analysis. We found that macroeconomic features, particularly the size of Ecuador's labor force, exert the strongest effect on sales. However, we also found that other select features, such as city altitude and holiday proximity, also have a high impact on sales performance and should be incorporated along with macroeconomic conditions when developing business strategies. This research applies interpretable machine learning models for market analysis to improve profits and provide a competitive advantage to businesses in grocery retail.**

## INTRODUCTION

Understanding the factors that drive sales is crucial for businesses to succeed. The ability to comprehensively understand market dynamics offers significant advantages, such as optimized marketing strategies and potential to boost sales. These advantages lead to a competitive edge in the market, enabling firms to improve profit margins and outperform competitors. However, as globalization advances, the business landscape has become more competitive with foreign competition seen in almost every product market worldwide (1). It is becoming harder for businesses to stay ahead of the competition, especially in developing countries like Ecuador. Achieving business success in such countries becomes especially challenging due to limited access to capital and financial resources, restricting the ability of firms to invest in growth. Even when capital is available, it is often loaned at a higher cost in developing countries (2). These constraints mean that efficient management and smarter business strategies are of greater importance in developing countries so that firms can optimize their limited resources (3).

In Ecuador, the grocery retail sector is one of the strongest-performing sectors and a major component of the overall retail industry, accounting for nearly 62% of total retail sales (4). This is significantly higher compared to developed countries where the ratio is usually below 50 percent (4). Consequently, the performance of businesses operating in Ecuador's grocery retail market are crucial for the country's citizens and overall economy. In addition to supporting Ecuador's population, the grocery retail industry has the potential to catalyze the country's development due to its significant role in Ecuador's economy. Business growth in developing countries has been shown to significantly boost economic development, leading to improvements in living standards (5). Therefore, the success of Ecuador's grocery retail franchises is essential to driving the country's broader social and economic development.

Previous research in sales forecasting has prioritized the application of deep learning techniques such as artificial neural networks, long short-term memory networks, and recurrent neural networks (6). Such complex neural network models inherently possess a "black box" nature, meaning they cannot be fully interpreted to understand why they produce a certain output (7). As a result, these models are not as useful in the context of business management.

Ultimately, the primary goal of a business is to sell at scale and generate a profit. Therefore, interpretable models that provide insights about the factors driving sales become more valuable than predictive models with a black-box nature. Our research takes a novel approach by applying interpretable machine learning algorithms, specifically regression models, to craft customized business strategies based on the trends found in the data. Regression models provide a unique advantage for interpreting sales trends through the analysis of variable coefficient signs (8). The signs of the coefficients in these models indicate the relationship between each variable and sales, showing whether an increase in a particular variable is likely to boost or reduce sales (8). This insight into the direct relationships between market features and sales can help businesses understand exactly how external market conditions impact sales performance. By identifying these relationships, businesses can pivot to changing market conditions by adjusting sales strategies, ultimately maximizing overall profitability.

Here, we investigate which specific market variables most significantly impact grocery retail sales performance in

https://doi.org/10.59720/24-076

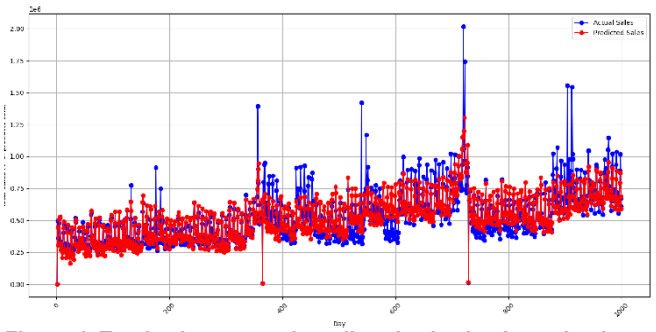


**Figure 1: Total sales vs. total predicted sales by day using lasso regression.** The actual daily total sales (blue) compared to the lasso regression predicted total sales (red) across all Corporación Favorita stores from January 1, 2013 (Day 1) to September 28, 2015 (Day 1000). Sales data from all stores were summed and grouped by day to display the franchise's actual and predicted sales performance.

Ecuador. We hypothesized that macroeconomic conditions influence sales performance to a greater extent than geographic and seasonality features. If macroeconomic factors significantly impact sales, businesses could focus on economic trends to adjust marketing and inventory strategies. Conversely, if seasonality or geographic features significantly drive sales, businesses could efficiently boost revenue by incorporating these consumer preferences in promotional campaigns. Through machine learning and data analysis, we concluded that macroeconomic features overall have the strongest impact on sales in grocery retail. Using the model's findings, we formulated marketing and sales strategies designed to efficiently optimize the profits of Ecuadorian grocery retail businesses.

## RESULTS

In this study, we fit the sales data to a lasso regression model to test our hypothesis. The lasso regression model prediction curve indicates that the model accurately predicted the overall sales trends of Corporación Favorita (**Figure 1**). The model achieved a mean absolute percentage error (MAPE) of 12.60%, which is expected considering the natural volatility of sales on a day-to-day basis. To assess feature importance, we used Shapley analysis to assign SHAP values to each feature in the model. There are three types of market features being investigated: macroeconomic, geographic, and seasonality indicators (**Table 1**). The magnitude of these SHAP values quantifies the impact of each feature on the model's prediction, allowing us to determine their relative importance. Among the 77 encoded features in the model, we focused on the top 10 features with the highest SHAP scores to test our hypothesis. Our analysis revealed that many of the most important features were indeed macroeconomic data, accounting for 5 of the top 10 predictors of sales. In addition, four of the top predictors were related to seasonality, while only one was a geographic feature (**Table 2**).

Using the lasso regression model, we used the coefficient signs for each of the top 10 features to evaluate their relationships with sales. The size of Ecuador's labor force had the greatest impact on sales, as indicated by the highest SHAP score, and it exhibited an inverse relationship with sales, as shown by the negative coefficient sign (**Table 2**). This suggests that as the size of Ecuador's labor force increases, sales tend to decline. City altitude, the only geographic

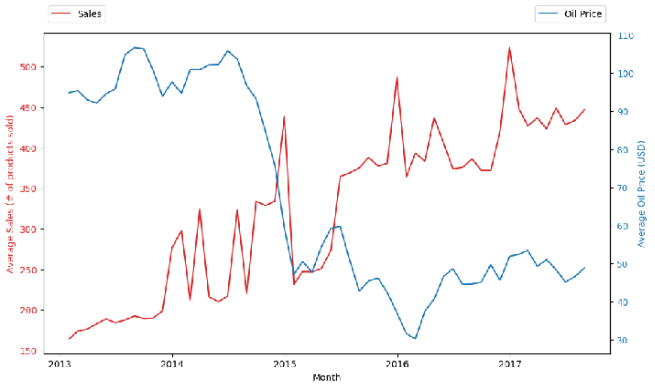


**Figure 2: Average sales vs. average oil price by month.** This graph shows the average monthly sales (red) for Corporación Favorita and the average monthly oil prices (blue) for Ecuador's oil exports. Sales data of the franchise and oil prices were averaged by month to illustrate their inverse relationship.

feature in the top 10 predictors list, had the second highest impact on sales. Similarly, the negative coefficient indicates its inverse relationship, meaning that stores located at higher altitudes tend to have lower sales performance than stores at lower altitudes (**Table 2**). Additionally, daily oil price emerged as the third most impactful feature (**Table 2**). We found that as oil prices gradually decrease, sales correspondingly increase (**Figure 2**).

The remaining top-performing macroeconomic features (Gross Domestic Product (GDP) growth, inflation rate, and unemployment rate) all had positive relationships with sales (**Table 2**). Notably, sales and unemployment rates tend to rise and fall in unison, which aligns with the negative correlation found between sales and the size of Ecuador's labor force (**Figure 3**). Both variables suggest that as the number of workers in the labor force decreases, the resulting unemployment in Ecuador corresponds to an increase in sales performance. This is an interesting pattern to recognize and can be strategically advantageous when developing sales strategies.

The seasonality indicators day of the week (Sunday and Saturday) showed positive correlations with sales, indicating that store performance generally improves on weekends, as

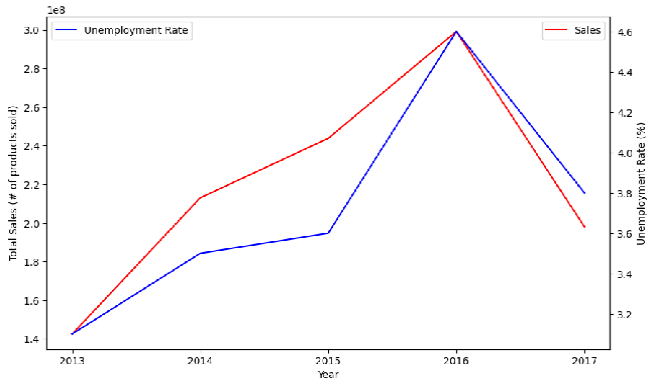


**Figure 3: Total annual sales vs. unemployment rate.** The graph shows the total yearly sales (red) for Corporación Favorita and the corresponding unemployment rate (blue) in Ecuador. Sales data from all stores were summed and grouped by year to show the direct relationship with the national unemployment rate.

https://doi.org/10.59720/24-076

| Feature | Data Description | Data Type | Data Category |
|---|---|---|---|
| Month | Indicates month of year using numbers 1-12 | Integer | Seasonality |
| Year | Indicates year | Integer | Seasonality |
| Day of Week | Indicates the day of the week | Object | Seasonality |
| Day of Month | Indicates the date number of the month | Integer | Seasonality |
| Holiday | Indicates whether a day is a holiday | Boolean | Seasonality |
| Daily Oil Price | Indicates daily oil price in the global oil market. | Float | Macroeconomic |
| City | Indicates the city in which the specific store is in | Object | Geographic |
| State | Indicates the state in which the specific store is in | Object | Geographic |
| Days Until Holiday | Indicates the days until the next holiday (local/regional/national holidays included) | Integer | Seasonality |
| Sales | Indicates the number of products sold by a given store on a given day | Integer | Target Variable |
| Region | Indicates the geographic region (North/South/East/West/Central) in which a specific store is in | Object | Geographic |
| Coastal | Indicates whether the store is in a coastal state | Boolean | Geographic |
| Location | Indicates whether the store is in an urban or rural area | Object | Geographic |
| City Altitude | Indicates the elevation of the city the store is in (ft) | Float | Geographic |
| Ecuador GDP Growth | Indicates the annual percent change in Ecuador's GDP | Float | Macroeconomic |
| Ecuador Inflation Rate | Indicates the annual inflation rate in Ecuador | Float | Macroeconomic |
| Ecuador Labor Force | Indicates the annual size of Ecuador's workforce | Integer | Macroeconomic |
| Ecuador Unemployment Rate | Indicates the annual unemployment rate in Ecuador | Float | Macroeconomic |
| Ecuador Dependency Ratio | Indicates the annual dependency ratio between workforce and dependents in Ecuador | Float | Macroeconomic |
| Ecuador CPI | Indicates the annual CPI (percent of which prices fluctuate) in Ecuador | Float | Macroeconomic |
| Ecuador Median Age | Indicates the annual median age among people in Ecuador | Float | Macroeconomic |
| Season | Indicates whether a day is during a dry or rainy season | Object | Seasonality |
| Weekend | Indicates whether a day is on a weekend | Boolean | Seasonality |

**Table 1: Corporación Favorita dataset features.** This table details the features in the sales dataset, providing information about what each feature measures, its data type, and its data category.

https://doi.org/10.59720/24-076

| Feature | SHAP Value (Magnitude) | Regression Coefficient Sign | Data Category |
|---|---|---|---|
| Ecuador Labor Force | 7.53e+10 | Negative | Macroeconomic |
| City Altitude | 2.90e+09 | Negative | Geographic |
| Daily Oil Price | 2.39e+07 | Negative | Macroeconomic |
| Day of Month | 9.03e+06 | Negative | Seasonality |
| Ecuador GDP Growth | 2.51e+06 | Positive | Macroeconomic |
| Ecuador Inflation Rate | 1.89e+06 | Positive | Macroeconomic |
| Day of Week Sunday | 1.02e+06 | Positive | Seasonality |
| Ecuador Unemployment Rate | 8.87e+05 | Positive | Macroeconomic |
| Day of Week Saturday | 4.14e+05 | Positive | Seasonality |
| Days Until Holiday | 3.77e+05 | Negative | Seasonality |

**Table 2: Top 10 strongest predictors in the lasso model.** This table presents the top 10 features from the lasso model with the highest SHAP values from Shapley analysis, their correlation with sales based on the regression coefficient sign, and their data category.

expected. In addition, the day of the month and days until the holiday served as important seasonality predictors (**Table 2**). We observed that sales were highest at the beginning of the month and generally decreased as the month progressed (**Figure 4**). In the last couple of days of the month, sales begin to rise again as they approach the peak in the first few days of the following month, which contributes to this feature's negative correlation with sales (**Table 2**). Additionally, the days until holiday seasonality feature's negative correlation with sales highlights how sales tend to spike as the holidays approach, likely due to increased consumer spending in preparation for holidays (**Table 2**).

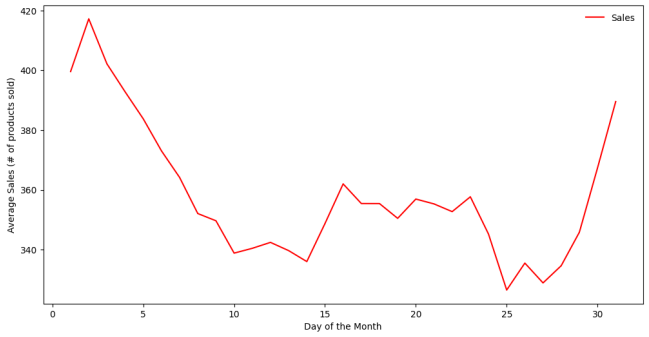


**Figure 4. Average daily sales for Corporación Favorita by day of the month.** This graph shows the average number of products sold each day of the month for Corporación Favorita. Sales data for the franchise was averaged by day to highlight the typical sales fluctuation throughout the month.

To validate our results from the lasso model, we fit the sales data to a ridge regression model, which also achieved a MAPE of 12.60%. Using Shapley analysis on the ridge model, we found that among the top ten predictors for sales, six were macroeconomic features, three were related to seasonality, and one was geographic (Table 3). These results are consistent with the findings from the lasso model, further supporting our hypothesis that macroeconomic conditions have the greatest impact on sales. The top five performing features remained the same, with Ecuador's labor force, city altitude, and daily oil price continuing to be the top three predictors in order (**Table 3**).

Another insight from our analysis using lasso regression is the removal of geographic features from the model due to their lack of predictive power. The lasso regression model automatically performs feature selection by eliminating insignificant variables, with the purpose of focusing on more impactful features in its predictions. Of the 17 features removed from the model, 15 were geographic features of the stores, while only 2 were related to macroeconomic or seasonal factors (**Table 4**). This indicates that these geographic features of store locations generally have minimal influence on sales compared to macroeconomic and seasonal predictors.

Notably, Ecuador Consumer Price Index (CPI) emerged as a new impactful macroeconomic feature in the ridge model, showing a negative correlation with sales (**Table 3**). Additionally, the day of the week Friday feature appeared as an important seasonality predictor and interestingly has

https://doi.org/10.59720/24-076

| Feature | SHAP Value (Magnitude) | Regression Coefficient Sign | Data Category |
|---|---|---|---|
| Ecuador Labor Force | 3.23e+12 | Negative | Macroeconomic |
| City Altitude | 7.96e+08 | Negative | Geographic |
| Daily Oil Price | 5.10e+08 | Negative | Macroeconomic |
| Ecuador GDP Growth | 2.78e+07 | Positive | Macroeconomic |
| Day of Month | 1.95e+07 | Negative | Seasonality |
| Month | 8.99e+06 | Positive | Seasonality |
| Ecuador CPI | 3.59e+06 | Negative | Macroeconomic |
| Ecuador Inflation Rate | 3.57e+06 | Positive | Macroeconomic |
| Ecuador Unemployment Rate | 1.65e+06 | Positive | Macroeconomic |
| Day of Week Friday | 7.71e+05 | Negative | Seasonality |

**Table 3: Top 10 strongest predictors in the ridge model.** This table presents the top 10 features from the ridge model with the highest SHAP values from Shapley analysis, their correlation with sales based on the regression coefficient sign, and their data category.

a negative correlation with sales. This insight, coupled with the positive relationships from the lasso model for sales on weekend days, reveals valuable specifics of consumer spending patterns to help businesses optimize sales strategies. We also found that the month feature had a positive correlation with sales, indicating better sales performance towards the end of the year holiday season (**Table 3**).

Both models show that macroeconomic features collectively have the greatest impact on sales, followed by seasonality and then geographic features. These findings support our hypothesis that macroeconomic conditions play the most significant role in determining business performance in the grocery retail sector and should be primarily utilized when making business decisions.

## DISCUSSION

Upon analyzing the results, we found that macroeconomic conditions had the strongest effect on sales. In both models, the size of Ecuador's labor force emerged as the strongest predictor of sales and exhibited a negative correlation with sales (**Tables 2–3**). This insight, along with the strong positive correlation between unemployment rates and sales, highlights exactly how economic conditions impact business performance (**Tables 2–3**). When more people are out of jobs, there appears to be an increased demand for essential products, resulting in higher sales for grocery retailers. This is likely due to unemployed individuals prioritizing essentials offered in grocery retail, such as food supplies, over non-essential purchases in other markets. Despite financial uncertainties, consumers appear to purchase larger quantities of essentials, providing an opportunity for businesses to boost their profits. Government support mechanisms like the Bono de Desarrollo Humano (Human Development Voucher) in Ecuador provide financial assistance to unemployed individuals (9). This aid helps sustain consumer spending on essentials, specifically benefiting the grocery retail industry. Businesses can leverage this economic correlation by tailoring their marketing campaigns towards individuals who are not actively employed. By doing so, they can capitalize on consumer behavior during financial uncertainties and generate more revenue than competitors by establishing a strong brand connection with these unemployed individuals. Additionally, businesses could consider raising retail prices during periods of high unemployment to take advantage of the increased demand for essential goods. Businesses can utilize these unexpected trends, often overlooked by competitors, to increase sales and expand market presence.

Another macroeconomic feature found to have a strong impact on sales is the daily oil price. In both models, daily oil price was the third highest predictor of sales and exhibited a negative correlation (**Tables 2–3**). Ecuador's crude oil production is the country's most important export, accounting for 27% of total export value (10). Consequently, Ecuador's economy is highly vulnerable to fluctuations in oil prices, linking oil commodity exports to grocery retail sales. As oil prices decrease, less revenue enters the country, leading to

https://doi.org/10.59720/24-076

| Feature | Data Category |
|---|---|
| Ecuador Median Age | Macroeconomic |
| Day of Week Wednesday | Seasonality |
| City Bahahoyo | Geographic |
| City Cayambe | Geographic |
| City Esmeraldas | Geographic |
| City Quito | Geographic |
| State Chimborazo | Geographic |
| State Esmeraldas | Geographic |
| State Guayas | Geographic |
| State Manabi | Geographic |
| State Pichincha | Geographic |
| State Tungurahua | Geographic |
| Region Central | Geographic |
| Region East | Geographic |
| Region North | Geographic |
| Region South | Geographic |
| Region West | Geographic |

**Table 4. Features removed from the lasso model via L1 regularization.** This table displays the features that were eliminated from the lasso model. Through L1 regularization, the lasso model performed feature selection by shrinking these features' coefficients to zero, effectively removing them from the model.

financial losses for the government and private businesses in the oil industry. To mitigate these losses, companies often reduce costs and make layoffs, which contribute to a decrease in workers in the labor force and an increase in unemployment. The strong inverse relationship between daily oil price and sales further emphasizes the correlation between grocery retail sales and unemployment, as also observed from the Ecuador labor force and unemployment features. This recurring trend across multiple high-performing features highlights the value brought by unemployed consumers and suggests targeting this market segment can boost sales performance.

Ecuador GDP growth emerged as another strong macroeconomic predictor of sales. In both models, GDP growth demonstrated a positive correlation with sales, indicating that higher economic activity leads to increased consumer spending in grocery retail (**Tables 2–3**). This relationship is expected, as GDP growth generally signifies improved economic conditions and greater overall consumer spending, resulting in higher sales volumes. Businesses can take advantage of periods of GDP growth by expanding product offerings and investing in promotional activities to capture increased consumer spending.

The final macroeconomic predictors with a strong impact on sales are the inflation rate and CPI. The lasso and ridge models both revealed that the inflation rate has a strong positive correlation with sales, while the ridge model showed that CPI has a strong negative correlation with sales (**Tables 2–3**). Initially, the different relationships seem counterintuitive since both measures are related to price changes. However, understanding their differences allowed us to interpret these results in a business context. CPI measures the average change in prices relative to a base year, reflecting long-term price trends and indicating how price levels affect consumer purchasing power over time. The negative correlation between CPI and sales shows that as prices increase over time, sales gradually decrease, which is an expected relationship as the purchasing power of consumers diminishes. In contrast, the positive correlation between the inflation rate and sales is intriguing. The inflation rate represents the annual percentage change in CPI, indicating short-term price changes. This positive correlation suggests that rapid, short-term price increases drive consumers to make more purchases as they anticipate further price increases. This pattern is particularly beneficial for the grocery retail industry, as higher inflation seems to push consumers to prioritize buying essential goods in the short term. Businesses can leverage this insight by anticipating budget-conscious consumer behavior during periods of high inflation. By offering discounts and highlighting affordability during these times, businesses can strengthen their brand connection with consumers concerned about rising prices. This strategy not only helps boost sales volume in the short term but also establishes long-term customer loyalty, positioning the brand favorably in the market if prices surge again in the future.

Among the seasonality predictors, the day of the month feature exhibited a strong negative correlation with sales in both models (**Tables 2–3**). This inverse relationship explains why sales are highest at the beginning of the month and generally decrease as the month progresses. In the last couple of days of the month, sales begin to rise again, peaking in the first few days of the following month (Figure 4). Interestingly, there is also a notable spike in sales following the 15th day of the month, lasting for about ten days, where sales performance is relatively higher. These sales trends correspond with wage payment schedules in Ecuador, where payroll cycles are typically monthly or bimonthly. Monthly payments are made at the end of the month, while bi-monthly payments occur on the 15th and the last day of the month. In the bimonthly payroll cycle, it is common practice to pay 40% of the salary by the middle of the month and the remaining 60% by the end of the month (11). Ecuador's payroll cycle closely matches the observed sales pattern, with sales spiking after the 15th day and again at the end of the month. This insight into consumer behavior reveals that people tend to spend more right after receiving their wages, creating a valuable window of opportunity for businesses. By strategically stocking up on inventory and offering promotions during this period of increased spending, businesses can fully capitalize on this behavior and maximize sales volume.

https://doi.org/10.59720/24-076

In both models, the days until holiday feature showed a strong negative correlation with sales, indicating that sales tend to increase as the holidays approach (**Tables 2–3**). Additionally, in the ridge model, the month feature emerged as a strong predictor with a positive correlation with sales (Table 3). Together, these results reveal that holidays, particularly those at the end of the year, have a relatively large impact on sales and business performance. This period includes major holidays in Ecuador, such as Navidad (Christmas) and Año Viejo (New Year's Eve), which drive increased demand for goods. These holidays include preparing traditional foods like tamales and empanadas, which require various grocery products such as meat, chilies, cheese, wheat flour, and vegetables (12). The need to prepare for these large family meals and festive gatherings during the end-of-year holiday season leads to a surge in grocery retail sales. Businesses can capitalize on this period by stocking up on ingredients commonly used for traditional dishes. Additionally, offering discounts on these specific ingredients allows stores to undercut competitors and drive higher sales volumes. This strategy can help businesses build a stronger brand connection with consumers, potentially boosting their year-round sales performance.

Another intriguing insight from this research is the strong positive correlation between grocery retail sales and weekend days, contrasted with the strong negative correlation for sales on Fridays (**Tables 2–3**). While higher sales on weekends are expected, as consumers have more free time to shop, the decrease in sales on Fridays reveals an interesting and unexpected trend. One might anticipate that sales would be higher on Fridays since it marks the start of the weekend when workers finish their week and are likely to have more time to shop. However, the results suggest otherwise. This unexpected trend indicates that after a long work week, consumers prefer to avoid "chore-like" activities such as grocery shopping. Instead, they opt to relax, unwind, or engage in more exciting social activities. In Ecuador, Friday nights are particularly popular for socializing and parties, reflecting a cultural behavior where consumers prioritize fun and leisure over routine tasks on Fridays (13). Therefore, taking advantage of this sales trend is crucial. By recognizing that consumers are less inclined to shop for groceries on Fridays, businesses can optimize their strategies by reducing marketing expenditures on this day. Instead, they can focus their efforts on weekends when consumer traffic is higher. This insight not only helps in cost-saving but also aligns marketing campaigns with consumer behavior, ultimately optimizing business efficiency.

The sole geographic predictor that emerged as a high-impact feature in both models was city altitude, which had a negative correlation with sales (**Tables 2–3**). This result indicates that sales performance is better in lower-elevation cities and highlights how store location impacts sales. Lower elevation areas in Ecuador have a more temperate climate, making them favorable for agriculture and capable of supporting a wide variety of agricultural products. Additionally, these areas have less challenging terrain compared to high-altitude regions in the Andes, which supports large-scale commercial agriculture. In Ecuador, the agriculture sector employs about 32% of the total workforce, making it the second highest employing sector behind the services sector (14). Importantly, the agriculture sector provides the highest number of unskilled jobs, offering a lower barrier of entry compared to the technical service sector. This attracts a significant portion of the labor force to lower-elevation areas where agricultural opportunities are abundant. Consequently, the majority of Ecuador's population resides in these lower-elevation regions. This higher population creates a larger market for grocery retail businesses, offering businesses greater potential for sales growth. The model's results suggest that business franchises should focus expansion efforts into lower-elevation areas with a larger market.

While our results provide valuable insights into Ecuador's grocery retail market, they also open avenues for further research and expanded applications of these models. Future studies could incorporate company data on customer demographics to analyze purchasing trends across different age groups, offering deeper insights for targeted marketing strategies. Additionally, integrating more granular data on store locations—such as proximity to highways, residential neighborhoods, or commercial districts—could enhance the analysis of sales performance within cities. Beyond Ecuador, these modeling techniques could be applied to other markets by incorporating region-specific factors and training models on local business datasets. This would allow businesses globally to apply the analytical frameworks from this research to grow sales and refine marketing strategies.

We propose expanding research on the application of machine learning in business to shift from sales forecasting to sales analysis, as businesses benefit more from understanding and growing their sales rather than merely predicting them. We also propose conducting this research in various global markets to identify universal market trends and collectively deepen our understanding of sales patterns. By utilizing machine learning to analyze sales trends, businesses can gain a deeper understanding of the key features influencing sales, allowing them to make more informed decisions driven by data. This approach enables companies to optimize sales strategies, expand their market dominance, and stay ahead of competitors in an increasingly dynamic business landscape.

## MATERIALS AND METHODS

### Loading data

The Corporación Favorita sales data used in this research is obtained from Kaggle (15). The data is indexed by store and date, providing daily sales for 54 stores from January 1, 2013, to August 15, 2017. The original dataset included the features month, year, day of the week, day of the month, holiday, daily oil price, city, state, and weekend. A 'days until holiday' feature was created by utilizing 103 local, national, and regional holidays in Ecuador from the existing holiday feature. The feature 'days until holiday' measures the number of days from any given day to the next upcoming holiday to indicate holiday proximity. Ecuador's annual GDP growth, inflation rate, unemployment rate, size of labor force, median age, dependency ratio, and consumer price index data were imported from the World Bank to incorporate macroeconomic features into the dataset. The regional, coastal, location, and city altitude features were created using the existing city and state features to provide additional geographic data to the model. The season feature was generated using the existing month feature. The final dataset consisted of 23 unique features (**Table 1**).

https://doi.org/10.59720/24-076

### Modeling

For modeling, one-hot encoding was applied to categorical variables, resulting in 77 encoded features. To improve computational speed, the dataset was trimmed to cover the period from January 1, 2013, to September 28, 2015 (1000 days). This period was chosen as it spans over two and a half years and reduces the computational challenges posed by the large number of encoded features. Additionally, a considerable increase of null values in sales and other features were observed beyond the 1000-day mark, which would complicate the sales analysis.

Real-world business environments are influenced by a wide range of external factors, but only a subset of these factors substantially impacts sales. The 77 encoded features included in our model present a challenge for standard linear regression models due to their inability to perform feature selection. Linear regression treats all features equally and lacks the ability to filter through the data to identify the most impactful predictors. This can lead to overfitting and significant computational inefficiencies when dealing with many features. This limitation hinders the model's usefulness for sales analysis in the context of this research. To address this issue, lasso regression was used as our primary method.

Lasso regression begins by fitting initial coefficient values to variables, determining the relationship between each predictor and the target feature. It then applies L1 regularization, adding a penalty proportional to the absolute value of the coefficients. This penalty forces less significant coefficients to zero, effectively removing variables that are not important from the model (16). As a result, lasso regression filters through the provided market data and focuses its predictions on the key factors that significantly impact sales. Lasso's feature selection not only highlights important predictors but also identifies the features that do not substantially impact sales, guiding businesses on where to focus their efforts. To provide a comparison with lasso, ridge regression was also used. Ridge regression uses L2 regularization, which reduces the magnitude of less significant coefficients without forcing them to zero, allowing all features to contribute to the model's predictions to some extent (17). The results of both models were used to evaluate the hypothesis and offer strategic sales insights to businesses.

### Data analysis

To analyze our fitted models, the coefficient signs were examined, which indicate the relationship between each variable and sales. A positive or negative coefficient sign shows whether a variable increases or decreases sales. However, regression models assume that predictor variables are independent and not correlated (18). This assumption does not hold true in the context of market data, where our features exhibit multicollinearity. For example, economic features such as the CPI and the inflation rate are inherently correlated, as the CPI is used to calculate the inflation rate. Due to this multicollinearity, the data does not meet the independence assumption required by regression models, and thus, the coefficient values themselves may not accurately represent the individual relationship of each predictor to sales. To address these limitations, Shapley analysis was utilized, a method that provides a more reliable interpretation of feature importance. In Shapley analysis, the model's predictions were simulated both with and without each feature to establish a baseline (average) prediction. From these simulations, SHAP values were assigned, which quantify how much the model's predictions change from this baseline when including each feature. SHAP values also have signs (+/-) to indicate whether a feature increases or decreases the prediction. The model's predictions are expressed as the sum of the baseline prediction and the SHAP values of every feature, indicating the feature's contribution to the model's prediction (19). For our study, the absolute value (magnitude) of the SHAP values was calculated, and the features were ranked accordingly to determine their importance. Out of the 77 encoded features, the top 10 highest-impact features in each model were focused on to assess our hypothesis. This method allowed for the evaluation the relative impact of each market feature on sales and the interpretation of the results in a business context. The full code for this analysis is provided in the references (20).

## ACKNOWLEDGMENTS

The authors would like to thank Veritas AI for helping in the publication process.

**Received:** March 10, 2024
**Accepted:** July 3, 2024
**Published:** May 31, 2025

## REFERENCES

1. Govil, S. K., and Rashmi Jain. "Globalization of Markets." *Advances in Management*. www.econbiz.de/Record/globalization-of-markets-govil/10010659976. Accessed 18 Mar. 2025.
2. Bartz-Zuccala, Wiebke, et al. "The Role of Innovation and Management Practices in Determining Firm Productivity." *Comparative Economic Studies*, vol. 60, 2016, https://doi.org/10.1057/s41294-018-0075-3.
3. McKenzie, David, and Christopher Woodruff. "Business Practices in Small Firms in Developing Countries." *Kauffman: Large Research Projects*, 2015, https://doi.org/10.1287/mnsc.2016.2492.
4. "Retail Foods." *USDA Foreign Agricultural Service*, www.fas.usda.gov/retail-foods. Accessed 12 Apr. 2024.
5. Mohamed, Maha Mohamed Alsebai, et al. "Causality between Technological Innovation and Economic Growth: Evidence from the Economies of Developing Countries." *Sustainability*, vol. 14, 2022, https://doi.org/10.3390/su14063586.
6. Aguiar-Pérez, J. M., et al. "An Insight of Deep Learning-Based Demand Forecasting in Smart Grids." *Sensors*, vol. 23, 2023, https://doi.org/10.3390/s23031467.
7. Aurelle, D., et al. "Illuminating the 'Black Box': A Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks." *Ecological Modelling*, vol. 155, 2002, https://doi.org/10.1016/S0304-3800(02)00064-9.
8. Walizada, Sayeeda. "Significance of Correlation in Statistics." *International Journal of Multidisciplinary Research and Growth Evaluation*, vol. 2, no. 6, 2021, https://doi.org/10.54660/IJMRGE.2021.2.6.317-318.
9. Paxson, C., and N. Schady. "Does Money Matter? The Effects of Cash Transfers on Child Development in Rural Ecuador." *Economic Development and Cultural Change*, vol. 58, no. 1, 2010, https://doi.org/10.1086/655458.

https://doi.org/10.59720/24-076

10. "Country Analysis Brief: Ecuador." *US Department of Energy*, www.eia.gov/ecuador-analysis. Accessed 20 May 2024.
11. "Hiring Talent and Managing Payroll in Ecuador." *Tarmack*, www.tarmack.com/hiring-ecuador. Accessed 25 May 2024.
12. Sánchez, C., et al. "Rediscovering the Original Recipe for the 'Empanada Sampedrana.'" *Community and Interculturality in Dialogue*, vol. 32, 2022, https://doi.org/10.56294/cid202232.
13. "Ecuador Nightlife." *Go Backpacking*, www.gobackpacking.com/ecuador-nightlife. Accessed 5 June 2024.
14. "Employment in Agriculture - Ecuador." *International Labour Organization*, www.ilo.org/ecuador-agriculture-employment. Accessed 8 June 2024.
15. "Store Sales Time Series Forecasting." *Kaggle*, https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data. Accessed 18 Mar. 2025.
16. Tibshirani, Robert. "The Lasso Method for Variable Selection in the Cox Model." *Statistics in Medicine*, vol. 16, no. 4, 1997, https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4%3C385::AID-SIM380%3E3.0.CO;2-3.
17. Hoerl, Arthur E., and Robert W. Kennard. "Ridge Regression: Applications to Nonorthogonal Problems." *Technometrics*, vol. 12, no. 1, 1970, https://doi.org/10.1080/00401706.1970.10488635.
18. Liang, Kung Yee, and Scott L. Zeger. "Regression Analysis for Correlated Data." *Annual Review of Public Health,* vol. 14, 1993, https://doi.org/10.1146/annurev.pu.14.050193.000355.
19. Aas, Kjersti, et al. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence*, vol. 298, 2021, https://doi.org/10.1016/j.artint.2021.103502.
20. "Analyzing Market Dynamics and Optimizing Sales Performance with Machine Learning." *GitHub*, https://github.com/Sumedh-Kamat/test/blob/main/Analyzing_Market_Dynamics_and_Optimizing_Sales_Performance_with_Machine_Learning.ipynb. Accessed 18 Mar. 2025.