Time Series Forecasting - Store Sales

Danica Ivkovic, Matea Lukic and Milos Korda

Faculty of Organizational Sciences, University of Belgrade

1. INTRODUCTION

The ability to accurately forecast data is highly desirable in a wide variety of fields, including sales, stock markets, sports performance, and natural phenomena.

This study focuses on several time series forecasting methods applied to retail sales data, specifically for Favorita stores in Ecuador. By examining different predictive techniques and their effectiveness, this research aims to enhance the accuracy of sales forecasts, thereby improving inventory management and overall retail operations.

Accurate sales forecasting is crucial for optimizing inventory levels. It helps retailers avoid stockouts, which can lead to missed sales opportunities and dissatisfied customers, as well as overstocks, which can increase storage costs and result in unsold inventory. By ensuring that products are available when customers want them, retailers can enhance the shopping experience, thereby increasing customer loyalty and repeat business. Precise sales predictions enable better planning of logistics and supply chain operations, reducing unnecessary expenses related to urgent restocking or excess inventory handling. Forecasting helps retailers plan and assess promotional campaigns more effectively, ensuring that promotions align with actual consumer demand and do not lead to stock imbalances. Better forecasts allow for more strategic allocation of resources, from staffing to store layout decisions, improving overall operational efficiency.

In conclusion, the ability to accurately forecast retail sales is of paramount importance to a wide range of stakeholders within the retail industry. This study aims to address the challenge of sales prediction for Favorita stores in Ecuador, providing insights and tools that can significantly improve retail operations and customer satisfaction.

2. DATASET AND FEATURES

The dataset is composed of sales for every product family recorded for the years 2013 through 2017 for 54 different stores. Each sample has the following features: id, date, store number, sales and on promotion indicating the the total number of items in a product family that were being promoted at a store at a given date. Apart from main dataset we are provided with datasets for oil prices through the years, transactions at a given store on a given date and a dataset containing metadata about each individual store.

2.1. Metrics Overview

For this problem, we chose Root Mean Squared Logarithmic Error as an evaluation metric. Using Root Mean Squared Error could lead to an explosion of the error term due to the presence of outliers. Since our sales have a mean of about 358 and a maximum value of 124,717, we want to avoid this situation. Additionally, RMSLE tends to reduce the impact of outliers compared to metrics like Mean Squared Error (MSE) or Mean Absolute Error (MAE). This is because the logarithmic transformation compresses the scale, so very high or very low values have a reduced effect.

2.2. Data Visualization

The yearly sales trend shows a consistent increase, but 2016 stands out with a spike in outliers. This anomaly aligns with the earthquake that struck Ecuador on April 16, 2016, disrupting normal business operations and causing irregular sales patterns. The outliers in the month-wise box plot for April reinforce this impact. Additionally, outliers with zero sales may signify days when stores were closed due to holidays or other reasons. (Bachmann)

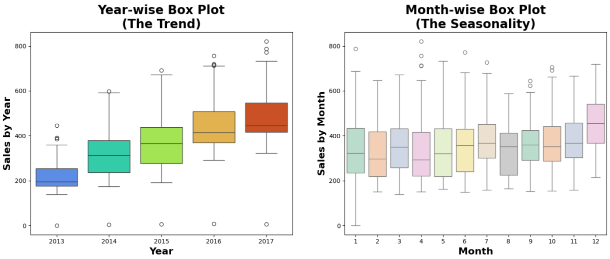

Fig. 1: Year-wise and month-wise box plots

When we plot sales by day in a month, we can see two dips around the 7th and the 24th, which can be explained by the fact that wages in the public sector are paid every two weeks, on the 15th and on the last day of the month.

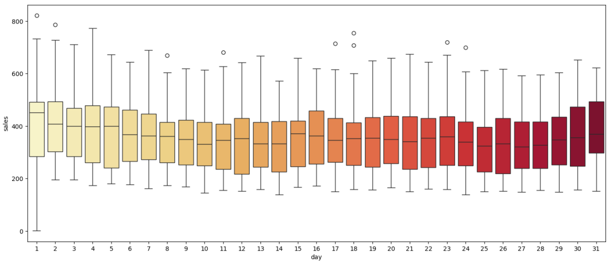

Fig. 2: Sales by day in a month

Also, looking at sales on every day of the week, we can see significantly more sales on weekends compared to weekdays.

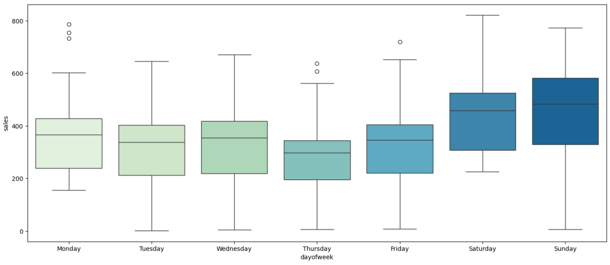

Fig. 3: Sales by day of week

Since Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices, we can look at the correlation between sales and oil prices. However, this analysis yields no significant correlation. It is only when we examine the correlation between sales of separate product families and oil prices that we see significant correlations.

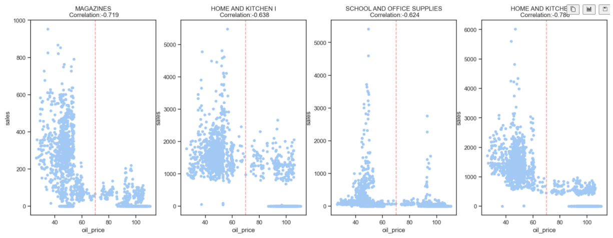

Fig. 4: Correlation between some product family sales and oil prices

2.3. Missing Values

In our dataset, conventional indicators of missing data, such as NaN or null values, are not present. However, it is important to note that four specific dates are missing, all of which correspond to Christmas. This pattern suggests a systematic omission, potentially due to operational or observational gaps during this holiday. Regarding the additional datasets mentioned earlier, the oil dataset has 43 missing values associated with oil prices. Additionally, no data entries concerning oil prices are available for weekends within that dataset. In the transactions dataset, 22 dates are missing, indicating a lack of transaction records for any of the 54 stores. Furthermore, there are 1564 dates with absent transactions data for specific stores.

3. DATA PREPROCESSING AND FEATURE ENGINEERING

3.1. Transformations

As previously mentioned, the dataset is missing 4 dates. To address this, we included these dates, ensuring that the dataset now encompasses the entire range of dates for each store across product families. In the newly added rows, missing values for sales and promotions were estimated using linear interpolation. For stores closed on Christmas and New Year's Day, the missing values were set to zero. To continue, in both the oil and transactions datasets, we employed linear interpolation to fill in the missing oil prices and number of transactions, respectively. This method ensured a smooth integration of the added dates.

3.2. New Attributes

From the additional datasets, we included data on oil prices, transactions and store metadata. We have also incorporated the following attributes:

3.2.1. Time-Based Attributes

In our dataset, we introduced several time-based attributes, including the month, day of the month, year etc. Additionally, we engineered more specific features, such as binary column indicating whether the date is the end of the month (1 if true, 0 otherwise).

3.2.2. Work Related Attributes

Furthermore, we created work-related features, including a binary column indicating whether a given day is a workday and another attribute specifying if it is a wage day. This indicators can influence sales as shown in the Fig. 2.

3.2.3. Lag and Moving Average Attributes

We observed that incorporating close target lags, such as those from the previous few days or weeks, can be particularly helpful in capturing short-term dependencies and immediate trends. Additionally, we included longer-term lags, specifically 365-day and

Model	lag 14	lag 120	lag 365	lag 730	ensemble
Linear Regression	0.39030	-	-	-	-
LightGBM	0.34984	0.33653	0.33520	0.34556	0.33367
XGBoost	0.34810	0.35070	0.33354	0.34094	0.33439

Table 1: Validation results - default parameters

Model	lag 14	lag 120	lag 365	lag 730	ensemble
LightGBM	0.33467	0.34526	0.33465	0.34124	0.33246
XGBoost	0.34601	0.34617	0.33354	0.34011	0.33278

Table 2: Validation results - optimized parameters

730-day lags, to account for annual and bi-annual seasonality. Lags for both future and past covariates were incorporated, along with 7 and 28 day moving averages for oil prices and the ”on promotion” attribute. (Bachmann)

3.2.4. Holidays Related Attributes

Using the holidays dataset, we conducted an A/B test over the holiday periods (Bayar), identifying 13 significant national holidays like Christmas and New Year’s Day. A binary column was added for each of these national holidays, indicating whether it was a holiday on that date or not. We also included numerical attributes indicating the number of days to the nearest future holiday and the number of days since the most recent past holiday. These attributes can significantly influence sales patterns, as consumer behavior often changes around holidays. For instance, sales may increase in the days leading up to a holiday due to preparations and gift purchases, and decrease immediately after a holiday as the demand subsides.

4. EARLY TRAINING

We trained three regression models with default configurations on the dataset as follows: the first model is linear regression with 14 days target lag that we used as the baseline model, while the other two are ensembles.

We achieve ensemble averaging by training multiple models and combining their forecasts through averaging. The first ensemble consists of four LightGBM models, and the second ensemble consists of four XGBoost (histogram-based) models. Each ensemble differed in the number of target lags included: the first model incorporated lag columns up to 14 days, the second model up to 120 days, the third model up to 365 days, and the fourth model up to 730 days. Also, for all validation tests, a subset of the dataset starting from January 1, 2015 was used. Additionally, for all tests on the test dataset, two ensembles were employed: one trained on the entire dataset and another trained on the aforementioned subset. The final result was computed as the average of their predictions. (Jie)

The results are presented in the Table 1. from which we can see that both ensembles outperformed the baseline model, with LightGBM ensemble being slightly better. After conducting this evaluation, LightGBM ensemble was tested on the test dataset, following previously described methodology, yielding an RMSLE of 0.38059. We should note that LightGBM ensemble was tested instead of XGBoost ensemble due to considerations of computational efficiency.

5. FEATURE SELECTION AND HYPERPARAMETER OPTIMIZATION

In addition to the before mentioned A/B test conducted on columns related to holidays, we also applied backward elimination method for feature selection. During this process, we removed combinations of time-related columns and holiday-related columns, which consistently resulted in poorer performance on the test set. As a result, we decided to retain all remaining columns for our analysis.

We employed an iterative approach to optimize hyperparameters for each model within an ensemble. For the LightGBM model, we selected parameters in-

Model	Lag	estimators
LightGBM	14	150
	120	300
	365	200
	730	100
XGBoost	14	350
	120	350
	365	200
	730	100

Table 3: Chosen no. of estimators

cluding number of estimators and learning rate. Likewise, for XGBoost, these parameters were fine-tuned, along with the subsample ratio of columns used in constructing each tree and subsample ratio of training instances. It is important to note that LightGBM was better optimized compared to the XGBoost model, primarily due to the longer training time required for the latter, which led to testing fewer parameter combinations. Table 3 presents the selected number of estimators. Across all models, an optimal learning rate of 0.05 was determined. Specifically for XGBoost models, the lag 14 model utilized 80% of the columns and rows, while the lag 120 model utilized 90%.

The results obtained using models with tuned parameters are presented in Table 2. Optimized parameters resulted in a slight improvement in results, with LightGBM ensembles still slightly outperforming XGBoost ensembles.

Finally, we evaluated models configured in this manner on the test dataset, and the results are presented in Table 4.

Model	RMSLE
Ens4LightGBM	0.37946
Ens4XGBoost	0.38059

Table 4: Test Results - Optimized parameters

It is evident that LightGBM ensembles outperformed XGBoost ensembles again, achieving score of 0.37946 which was ranked 9th in the store sales time series forecasting competition on Kaggle. This outcome underscores the effectivness of our approach.

6. CONCLUSION

In this paper, we evaluated three well established regression models - linear regression, LightGBM ensembles and XGBoost ensembles, in the context of store sales time series forecasting competition on Kaggle. Through adequate preprocessing, feature selection and hyperparameter optimization, we improved models performance, achieving 9th place finish on the competition leaderboard. We found that the LightGBM ensemble models outperform the rest. That is likely due to LightGBM's efficient training capabilities that enabled better model optimization. For future work, exploring store-level grouping instead of product family grouping, along with expanding the range and scope of parameter optimization, could be beneficial.

REFERENCES

Bachmann, J. M. 2021a, Time Series I An Introductory Start

Bachmann, J. M. 2021b, Time Series II Feature Engineering Concepts

Bayar, E. 2021, Store Sales TS Forecasting - A Comprehensive Guide

Jie, C. Z. 2023, Ecuador Store Sales — Global Forecasting LightGBM