Luis F. C. Costa

The Bakery Report

Introduction

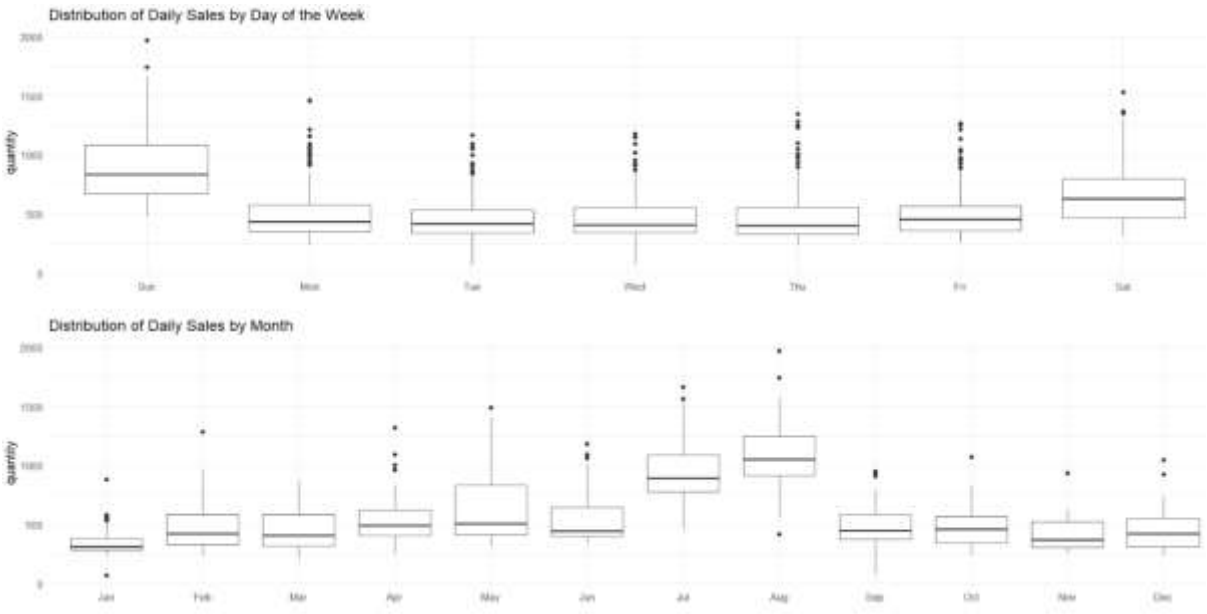
The bakery seeks to balance production to meet customer demand while minimising waste by using sales and weather data. The objective is to develop a forecasting model that incorporates these data to predict daily sales. There are various approaches to this problem, and I have chosen to analyse aggregate sales by day rather than examining each product individually.

Data Exploration and Preprocessing

Sales Data: Quantity is the dependent (target) variable. Some articles identified as 'dot' were excluded due to zero price. Articles with negative quantities, potentially representing discarded products, were also excluded from this analysis.

Weather Data: he 'tsun' data are entirely NA. Snow data has 15 days recorded, with NA values filled as zero. 'wspd' and 'wdir' had two NA values, interpreted as no-wind days and filled with zero. 'wpgt' values were zero for no-wind days, with other NAs replaced by their average values. Similarly, 'press' NAs were replaced by their average values.

Feature Engineering : Features for the day of the week and the month were created to capture seasonal patterns, as shown in the sales distribution figure below.



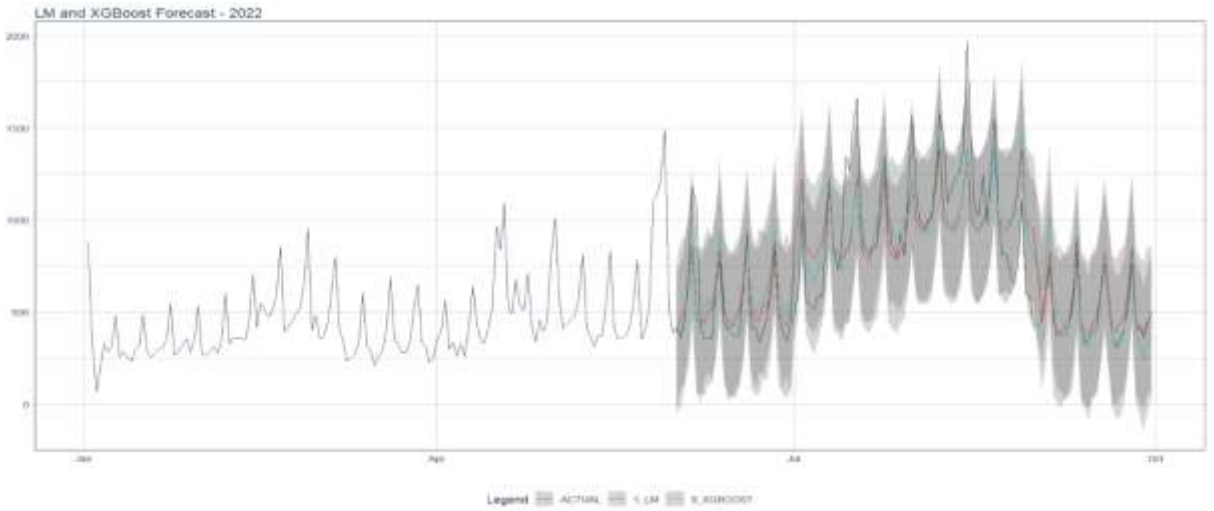
There are clearly two main patterns: sales are higher on Saturdays and Sundays, and there is also an increase in sales during July and August. This seasonal trend may be positively correlated with higher temperatures during these months.

Modelling Approach

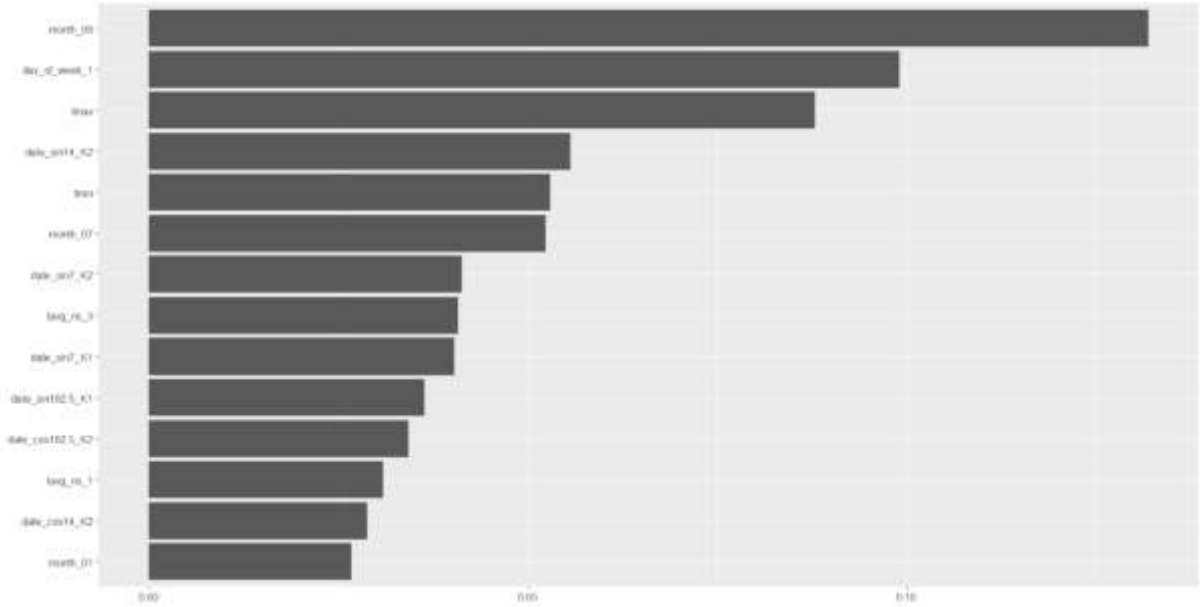
A time-series approach was employed to extract additional information and predict future sales. Extra features, such as Fourier series and natural splines of average temperature, were created to account for specific data patterns caused by seasonality.

The models tested included linear regression, elastic-net (regularisation), ARIMA, random forest, and XGBoost. The data was split into training and testing sets with an 80-20 proportion. XGBoost and Random Forest models were fine tunned and cross-validation was used to ensure the models' generalizability. The model accuracy was evaluated on the test set using MAE, RMSE and other relevant metrics.

Among the nine different models tested, XGBoost proved to be the most accurate with an RMSE of 197, followed by the linear regression model with an RMSE of 202.7. Notably, the linear regression model had the lowest MAE at 142.26. The chart below illustrates the predictions of these two models over the test set. As observed, both models capture the seasonal patterns indicated by peaks in the data to some extent. However, further fine-tuning could enhance the accuracy of sales forecasts.



By analyzing the XGBoost importance plot, we can see how the features influence sales. The most important features explaining sales are the month of August, Sundays, and the temperature.



Therefore, the bakery could use weather forecasts to plan its production and increase output during the summer months.

Limitations

Additional information, such as discounts, promotions, and holidays, can be included to enhance the model's accuracy. Considering interactions between features and including lagged sales and moving averages could also improve predictions. While this analysis focused on aggregate sales rather than individual products, assessing the impact of each product separately could provide deeper insights.