Bakery Sales - interview pre-assessment report

Dr Peter Michalák

February 11, 2024

1 Executive Summary

The presented challenge is to optimise production of bakery products by leveraging sales and weather datasets. This report presents an analysis, reveals key insights, and compares two forecasting approaches. The sales dataset contains anomalies, such as negative 'sold' quantities and missing receipt IDs. The weather dataset required imputation for missing values of peak wind gust and revealed missing sunshine measurements.

2 Sales dataset analysis

The sales dataset captures detailed information on bakery sales totalling €558,640.85. It contains approximately 149 unique articles.

In this dataset, 234,005 items were sold across 136,451 receipts. Notably, 1,295 receipts indicate negative quantities, presumably denoting refunds. The largest refund recorded involved 200 items (ticket number 179932). Additionally, 32 items were sold at a unit price of €0.

An examination of receipt numbers revealed 2,423 missing receipt IDs. Assuming sequential incrementation, this suggests potential gaps in the data (e.g., missing IDs such as 150047, 150057, 150226) and should be investigated by the client.

The 'TRADITIONAL BAGUETTE' (117,463 items), 'CROISSANT' (29,654 items), and 'PAIN AU CHOCOLAT' (25,236 items) led sales, as depicted in Figure 1 showcasing the monthly volumes for the top 10 products.

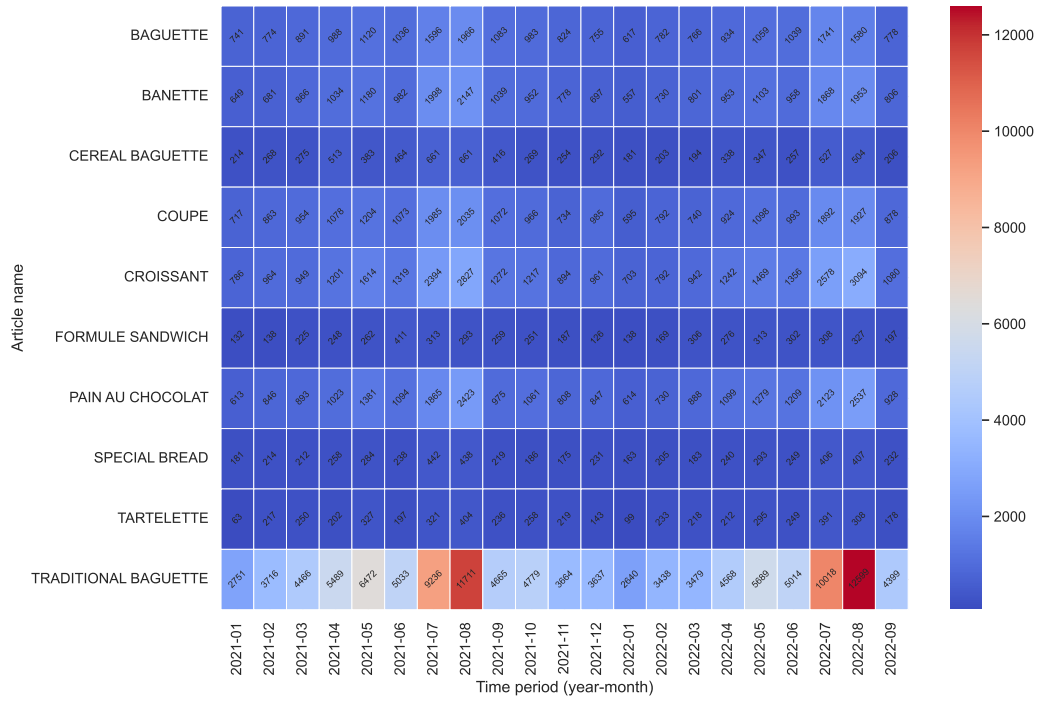

Figure 1: Sales volumes of top 10 most popular products.

3 Weather dataset analysis

The weather dataset was missing any sunshine information in 'tsun' column. Additional analysis involved imputing missing peak wind gust data through linear regression, leveraging its high correlation with the average wind speed.

4 Analysis

Combination of the two client-supplied datasets reveals a noticeable trend: a correlation between weather conditions and sales, which supports the client's empirical evidence. Furthermore, the Figure 2 illustrates the substantial impact of public holidays on sales, especially during the weekdays.

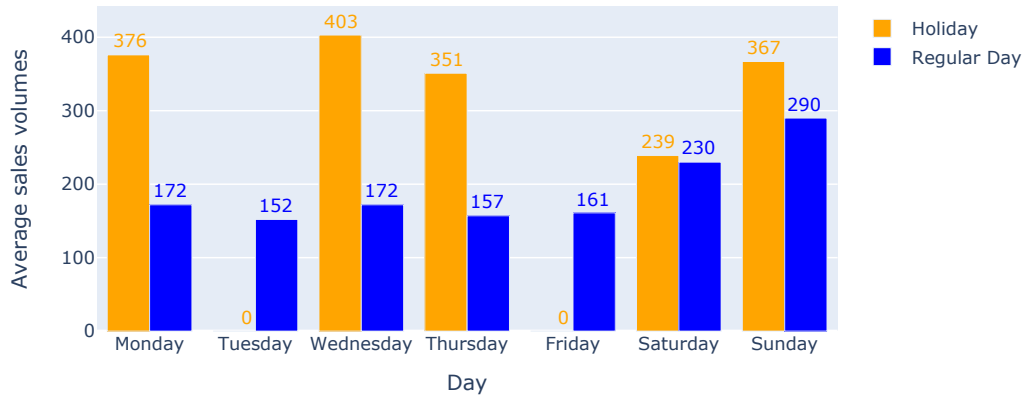

Figure 2: Average number of 'TRADITIONAL BAGUETTE' sold during holidays and regular days.

5 Forecasting tool

To address the bakery's challenge of lowering the food waste, two predictive models — Prophet [1] and Random Forests [2] — were implemented. The first 20 months of data from the combined dataset, enriched with French public holidays, were utilised for training. The data from the last month (September '22) were held out and used as the test dataset.

Random Forests outperformed Prophet, exhibiting a Mean Absolute Error of 51.4 compared to 198.0. Both models showed wide prediction intervals, suggesting room for improvement. Figure 3 visualises the predicted vs. actual sales for the most popular item.

Visit project's GitHub repository[3] to review the code base.

[1] Sean J Taylor and Benjamin Letham. Forecasting at scale. 2017. *URL: https://github.com/facebook/prophet*, 2017

[2] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001

[3] https://github.com/PetoMichalak/bakery-sales

6 Limitations and Next Steps

The analysis validates the bakery's intuition about weather impacting sales. A recommendation for the client is to incorporate additional data sources such as daily sunshine information and details on nearby events or competitors.

In addition, the Bayesian approach used by Prophet might benefit from further exploration to account for observed seasonality patterns. Similarly, the Random Forest model could be improved by conducting a grid search over the hyperparameter space.

It's important to note that the current analysis scope was limited to a select number of popular items and should be expanded to encompass all articles for a comprehensive understanding.

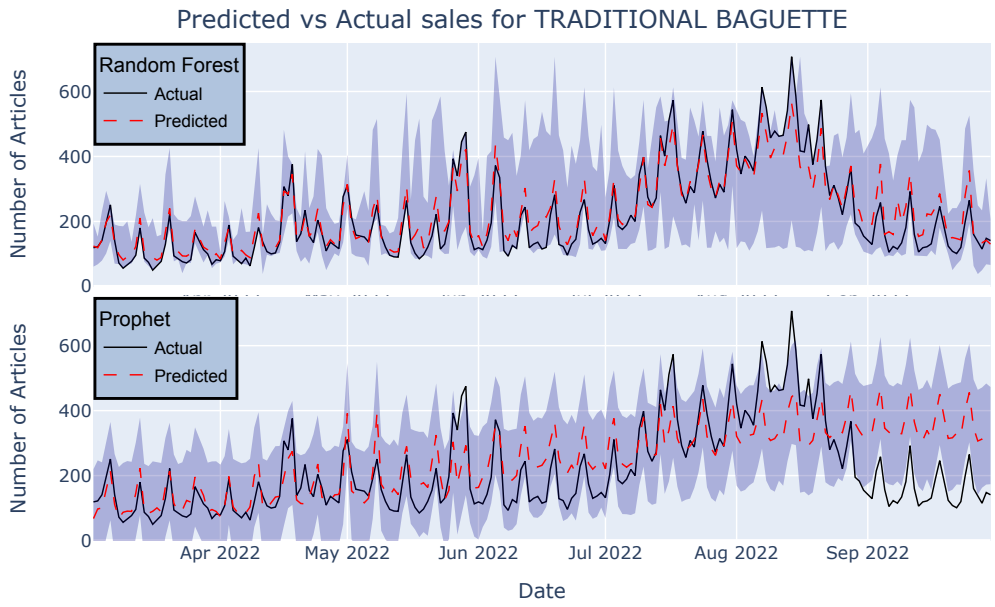

Figure 3: Predicted vs Actual sales for a single item with 95% prediction intervals.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] Sean J Taylor and Benjamin Letham. Forecasting at scale. 2017. *URL: https://github.com/facebook/prophet*, 2017.