TASK 1: THE BAKERY
NICD NEWCASTLE TECHNICAL INTERVIEW

MIHAIL HURMUZOV

GITHUB REPOSITORY

1. Data preparation

As the main aim of the bakery is to better predict how much goods it needs to produce each day and minimise waste, the first step we take here is to categorise all items. Every item has been coded as one of:

- Bread,
- Pastry,
- Packaging,
- Sandwich,
- Cake,
- Drink,
- Sweet, or
- Prepared meal.

We assign a *quantity* to each item as part of its category. An obvious illustration of this is that "ROYAL 4P" and "ROYAL 6P" refer to the number of pieces of the layered chocolate mousse Royal cake the customer has bought; these will be assigned 4 and 6 in the coding, respectively. Next, the sales data is transformed so that for each day and each category, we have a number of units in that category sold.

The idea of this approach is that a Fraisier cake and a Trois Chocolat are more likely to be interchangeable for both the bakery and for its customers than a Palet Breton and a baguette. Moreover, a Banette might need roughly the same amount of ingredients as two Banettine, so they are given appropriate weights in the coding. As predicting each individual product is both unfeasable and potentially misleading, our method performs aggregated product sales forecasting that is more actionable without losing too much predictive power (see [1]).

The full codebook can be found here.

2. Modelling

For each product category we include as predictors:

- the day of the week,
- the month,
- the number of days the bakery will be closed from tomorrow,
- the number of days the bakery was closed until yesterday,
- the sales in the product category for the last two open days,
- the average and maximum temperature, the precipitation, the wind speed, and the snow coverage for the previous two days, the current day, and the following day.

Previous studies support the lagging of weather data, alongside the influence of future weather expectations on customer behaviour ([1]).

We fit 8 random forest models, one for each category, on the first 80% of the data and test them on the last 20%. On the most popular category – bread – the model achieves mean absolute percentage error (MAPE) of just under 13% (see Figure 1 for a graph of the predictions).

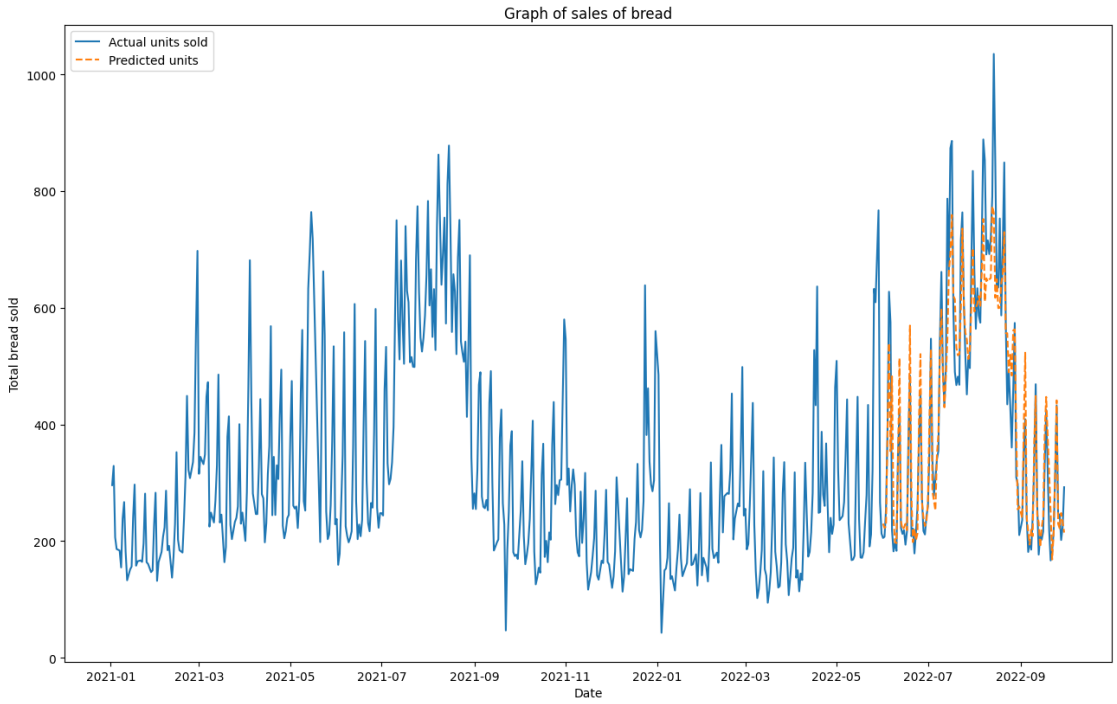

FIGURE 1. Bread units actual vs predicted.

The second most popular category, pastry, has MAPE of ∼20% (see top of Figure 2), while the smaller categories like cake and sweet achieve MAPE of 69% and 54%, respectively. When the 8 predictions are aggregated, the overall unit sales forecast has MAPE of 14.7% (bottom of Figure 2).

3. Summary and limitations

The codebook would ideally be developed in collaboration with the client to match their processes more closely. Additionally, information on promotional campaigns, special events, etc., could improve the model's performance, particularly during very busy periods.

While a time-series approach seems more appropriate for this problem, the intermittent nature of the sales data makes it difficult to directly account for seasonality. Even Meta's top-of-the-line *prophet* module gave significantly worse results. Some of the literature also suggests that similar predictive tasks are better addressed with random forests ([2]).

The results used in this report can be reproduced with the code supplied at the github repository https://github.com/mhurmu/technical_assignment.

Bibliography

[1] H. Chan and M.I.M. Wahab, *A machine learning framework for predicting weather impact on retail sales*, Supply Chain Analytics **5** (2024), 100058, DOI https://doi.org/10.1016/j.sca.2024.100058.

[2] Stuti Raizada and Jatinderkumar Saini, *Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting*, International Journal of Advanced Computer Science and Applications **12** (2021), DOI 10.14569/IJACSA.2021.0121112.

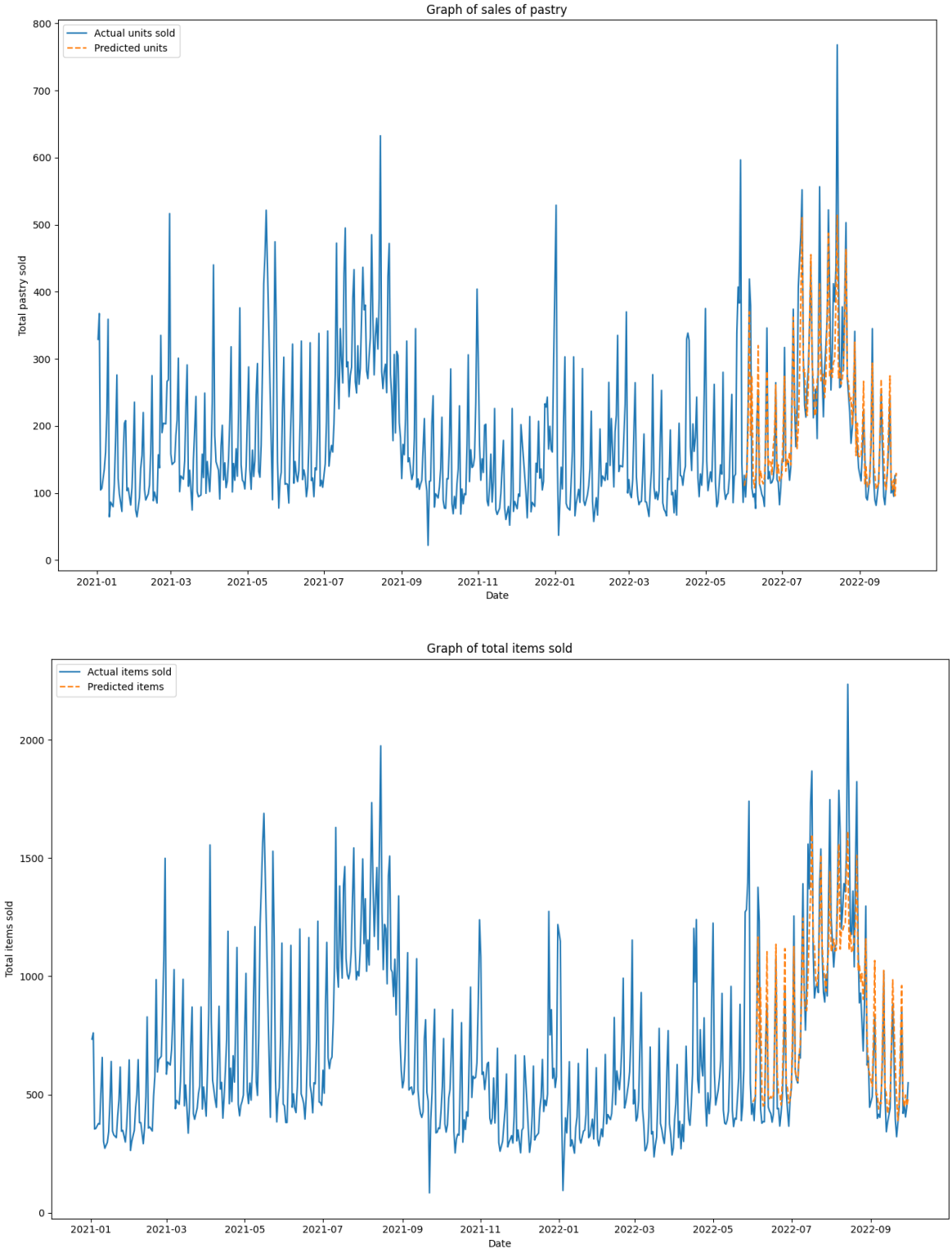

FIGURE 2. Total units actual vs predicted.