Sales forecasting for Co-op's in-store bakery SCC.460

Ramdev Padinhatta Murali (36257699)

Lancaster University

I. INTRODUCTION

Co-op Food ("Co-op") is the supermarket convenience brand of the Co-op Group, consisting of over 2,500 stores throughout the UK. The majority of Co-op supermarkets contain an in-store bakery, allowing stores to sell savoury products freshly baked each day, on the premise.

Products are baked at the start of each day to be sold the same day, so it is in the interest of the Co-op that the numbers of products baked each day exactly meet customer demand. That is, loss resulting from under-supply (missed sales opportunities) or over-supply (products being discounted or thrown away) be avoided. Both scenarios are similarly detrimental to the business.

The objective of this project is therefore to predict customer demand for each product, at a given store on a given day of the year, so that each store may be informed of the optimal quantities of products to bake each day.

Predictions of bakery product demand were made using a dataset of sales records spanning 12 stores in the North West, from May 2021 to August 2022 (n = 51,962). Each instance in the data is attributed to a unique product-store-date combination, and the response variable in question is the sales quantity of each product at each store on each day. Potential predictor variables provided in the dataset include the product sub-category; store opening and closing times; the day of the week; weekday/weekend; bank holidays; various weather metrics; and reduced-to-clear sales quantities and turnovers. Research shows that factors such as weather can influence retail spending. In particular, Tian [1] shows that both higher and lower temperatures have a significant negative impact on consumers' desire and on the average price per customer, and argues that this is due to the impact of sunny weather, which stimulates consumer desire to buy. Murray [2] also suggests that exposure to sunlight decreases negative sentiment among customers, leading to increased spending. Moreover, it may reasonably be assumed that holidays and weekends are linked to spending patterns in some way. This project also explores the extent to which sales predictions can be made using this type of data alone.

We approached forecasting using three primary methods: (1) linear regression and Poisson generalised linear models (GLMs), (2) extreme gradient boosting (XGBoost), and (3) AutoRegressive Integrated Moving Average (ARIMA) time series analysis. For the GLM approach, an individual model was developed for each product-store combination. For ARIMA, an individual model was developed for each product at store 1 ("Petrol Three Peaks") only. For the linear regression, GLM, and XGBoost methods, an individual model was built in which store number and product code were included as categorical predictor variables. We find the amount of variability that can be explained using the data to inform linear regression models and GLMs is limited in most cases, returning root mean squared errors (RMSE) of 3.00 and 2.86 respectively, with the results for GLMs declining to an average RMSE of 3.87 when product-store combinations were modeled individually. A full XGBoost model returned an RMSE of 2.43: a reasonable improvement. We found ARIMA analysis, preformed on individual product-store combinations, to be overall the most appropriate and effective approach to sales forecasting, returning an average RMSE of 2.08.

II. METHODS

A. Overview

The project was carried out in three stages. Firstly, as the data was provided in raw unprocessed form, they were examined using and basic statistical and visual analysis. This step enabled identification of necessary data cleaning and preprocessing tasks, and discovery of data characteristics that would aid selection of appropriate forecasting methods. This stage also involved communication with the Co-op contact to address uncertainties regarding product sales history and missing data. Secondly, research into statistical and machine learning methods was carried out to understand what the most appropriate methods might be, given the data and research question. Approaches were tested with exploratory model building. Finally, model building, selection, and evaluation was carried carried using the chosen methods.

B. Exploratory Analysis

Exploratory analysis was composed of visual and quantitative checks to identify missing values, anomalies, and correlations. Visualisations of total quantities of products sold, categorised by store (Figure 1) illustrate the magnitude of differences between sales quantities between stores, and proportions of products sold at each store. These significant differences in the product composition of sales between stores imply that a generic model for all stores, in which the recommended quantities are scaled by store turnover, would be insufficiently detailed.
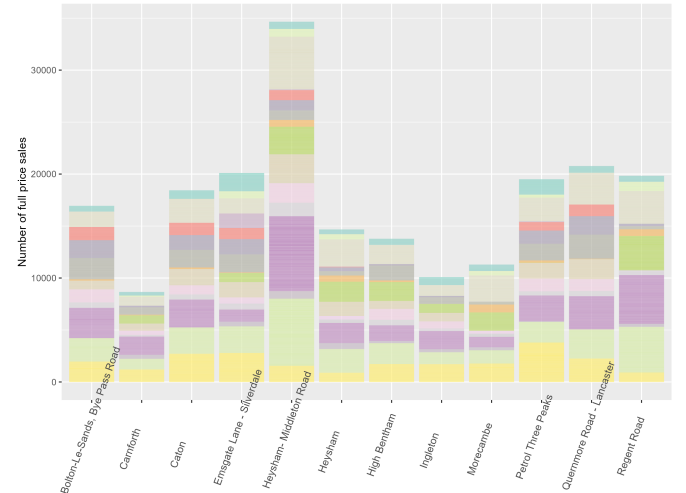

Fig. 1: Total sales quantities across all products. Colours correspond to individual product lines.

Time series of each product-store sales histories were plotted using R; an example is shown in (Figure 2). From this instance, and similarly in several other product-store combinations, it is clear that supply was not always continuous
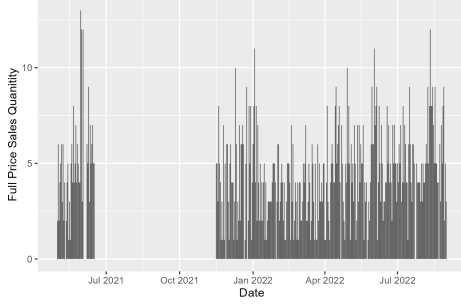
Fig. 2: Sales of brown scotch rolls at the *Petrol Three Peaks*

through the recorded period. This was addressed with the Co-op (section II.C. Preprocessing).

Total sales were also plotted for each date, with the bank holiday flag variable shown in red (Figure 3). From this figure it is apparent that some bank holiday flags were missing, and that identifying seasonal patterns may prove difficult, given the time span recorded is limited to about 18 months.
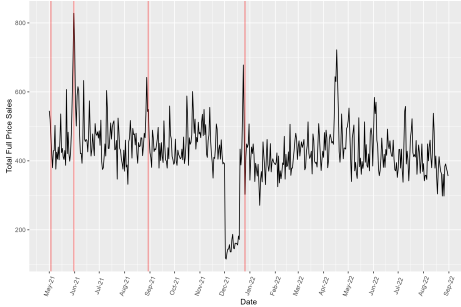

Fig. 3: Time series of total full price sales quantity. Bank holidays flagged in the data are shown in red.

Finally, a matrix of correlations between continuous numeric variables (Figure 4) reveals some small relationships between the full price sales quantities and some of the numeric predictor variables. Sales quantity and turnover of reduced-to-clear bakery products notably have small negative relationships with full price sales. This would be expected, as fewer full price sales leave a greater supply of reduced-to-clear items at the end of the day. However, reduced to clear sales were ultimately not included in the model, as they do not provide new information about customer demand for full price items.

C. Pre-processing

The operational process for data preprocessing is illustrated in Figure 5.

The client clarified that discontinuity, or absence, of sales for certain product-store combinations was usually due to either the seasonal nature of some products or due to ingredient supply issues, and therefore their interest was limited only to product-store combinations present in the final week of the data. As such, only data entries with these product-store combinations were included in the forecasting (n=46,287).

Further preprocessing was carried out in differing manners for the ARIMA analysis and for linear regression, GLMs, and
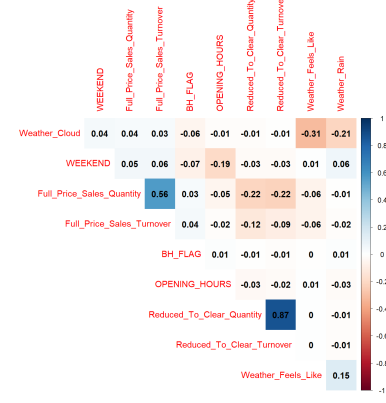

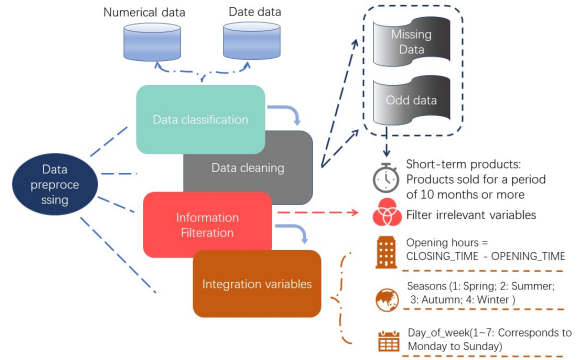Fig. 4: Correlation matrix for numeric variables in the dataset

Fig. 5: Schematic diagram of data pre-processing.

XGBoost. ARIMA analysis requires only a response variable (here, Full Price Sales Quantity) and the date, so for this method these two variables were extracted, and missing values handled as described in section *D. Analysis, 3) ARIMA*.

Preprocessing for linear modeling required several feature selection and engineering tasks. Variables known *a priori* to be uninformative or repeated were removed: "Product_Description", "Store_Name", and "SubSect_Description". Rather than retaining the two separate variables for "OPENING_TIME" and "CLOSING_TIME", these were transformed into a single variable quantifying the number of hours the store was open on the corresponding date - thus retaining a similar amount of information while reducing potential model complexity.

One-hot encoded variables (day of the week, weekend, weekday, and season) were each transformed into factor variables, which are more easily handled in linear modeling in R. Data points with negative values for the response variable (n=2) were removed, as these would be incompatible with a Poisson-based model.

In anticipation of potential polynomials or interaction with other variables in the modeling, the variable for temperature was standardised to the interval [0,1] using equation 1.

$$x_{ij} = \frac{(MAX(X_j) - x_{ij})}{(MAX(X_j) - MIN(X_j))} \quad (1)$$

Finally, we noticed flags for bank holidays missing for the

data spanning the year 2022 (Figure 3); this variable was corrected accordingly. XGBoost required categorical variables be one-hot encoded, and independent variables be in a matrix data type in R compatible with the library (e.g. using `Matrix::sparse.model.matrix()`).

D. Analysis

1) Linear Models and Poisson Generalised Linear Models: Since there is evidence for some linear relationships in the data, the first chosen method for forecasting was linear regression. Linear regression was also chosen on the basis that it is quick and efficient to implement, which is useful in the context of this project, where a relatively large number of models for each product and store may be required. It also produces easy-to-interpret models [3].

Models were created in R using the MASS and caret packages. An initial model were produced using simple linear regression, forward selection and five-fold cross validation [4]. This was a full model (all stores and products) trained on 80% of the preprocessed data.

As the response variable resembles a Poisson distribution (Figure 6), models were also built using the Poisson form of GLMs. First, a model was built using data containing all product-store combinations and five-fold cross-validation. Secondly, individual models were built, corresponding to each product-store combination, using backward selection. Predictions in continuous form were rounded to the nearest whole number.
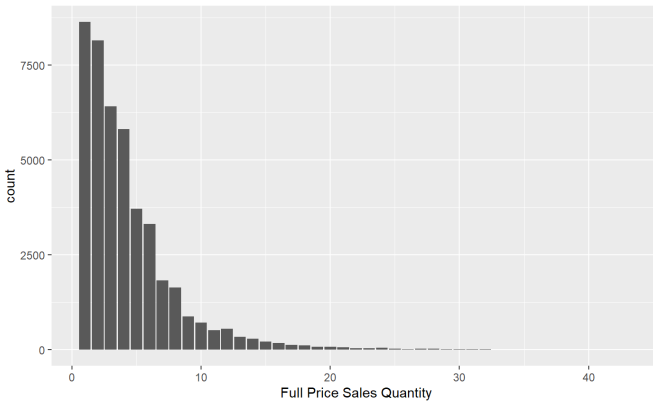

Fig. 6: Distribution of the target variable after removal of two negative values.
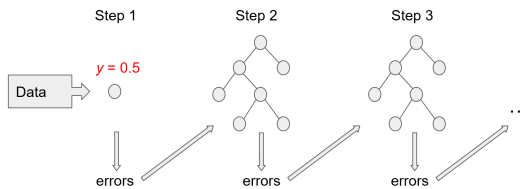

Fig. 7: Diagram of the XGBoost algorithm, simplified. The final model is the sum of all resulting trees.

2) XGBoost: XGBoost [5] is a library that provides gradient boosting tree algorithms featuring a large number of optimisation and regularisation methods. It was chosen on the basis that it is also straight-forward to implement, efficient and flexible: over and under fitting can be controlled using hyperparameters, and being a tree-based algorithm may be able to model non-linearities in the data that could not be captured in the GLMs. XGBoost is an iterative method that begins as a simple model (e.g. a single value, such as 0.5 for normalised data). At each iteration, the data is sub-sampled and fitted to the model, and the errors are returned; the model is then updated with the addition of a decision tree that is fitted to the errors on the previous iteration, thus improving the model each iteration, until a stopping criteria is met (Figure 7).

A predictive XGBoost model was built using first, default parameters, and secondly, an optimised model tuned using a random selection of 50 hyperparameter combinations, using the *mlr* package in R. In this case it was also necessary to round predictions to the nearest whole number.

3) ARIMA: ARIMA models, also known as Box-Jenkins models, are widely utilised forecasting technique characterised by their ability to account for both autoregressive (AR) and moving average (MA) components in the data, in addition to the presence of non-stationarity through differencing [6]. This makes ARIMA models well-suited for time series data that exhibit temporal dependencies, such as trends and seasonality [7].

ARIMA was therefore selected as a method of analysis due to the temporal format of the dataset and the possible presence of seasonal signals or other time dependent trends [8]. Moreover, as there was evidence (see Results) from the GLMs of autocorrelation in the data, i.e. correlation between the values of the response variable and those prior in time. ARIMA analysis is a form of linear regression that uses this lag as a predictor (i.e., "autoregressive").

As discussed in the previous sections, the existence of missing data points in some instances (due to missing ingredients or the seasonality of the product) can affect the accuracy of the models, therefore any remaining missing values were handled using the "backfill" method [9], resulting in a continuous time series of each product during the 487 day period. These time series were analysed evaluated to find the best-fitting ARIMA model parameters based on the Akaike Information Criterion (AIC), RMSE, and the mean absolute percentage error (MAPE), using the the `forecast::auto.arima` function in R [10].

ARIMA models were trained on 80% of each product data subset for store 1. The parameters used for each model is shown in the results section, in table II. The "AR" parameter corresponds to the *lag order (p)*; the "I" parameter corresponds to the *degree of differencing (d)* necessary to make the time series stationary, and "MA" corresponds to the *order of moving average (q)*.

III. RESULTS

A. Linear Regression and Poisson Generalised Linear Models

Results for the full simple linear model (all products and stores) were poor: r-squared = 0.36 (on both the training

and testing sets), MAE = 2.11, RMSE = 3.00; and models for individual product and store combinations generally preformed significantly worse. Model evaluation revealed some autocorrelation for small lags (Figure 8 and greater variability for larger response values. Overall linear models could be considered invalid, as the distribution of the predicted values did not resemble the true distribution of the response variable.
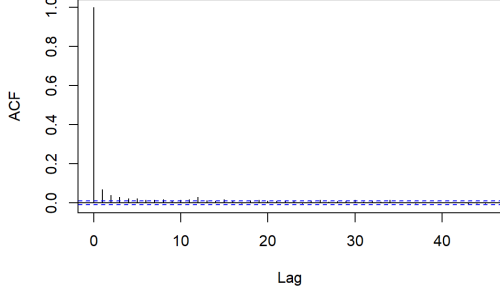

Fig. 8: Evidence in the residuals of the full simple linear model of autocorrelation for small lags.

The full Poisson GLM returned an r-squared value of 0.42 on training data, and 0.39 on the testing set. The RMSE was 3.02 on the training set, and 2.86 on the testing set. The dispersion parameter was 1.42. The model also poorly predicted larger sales quantities >15, and this is also apparent from the Normal Q-Q plot (Figure 9). The resulting model also failed to some degree on assumption of no autocorrelation (Figure 10), suggesting there exists a relationship in the time series between values and those prior in time. However, the model preformed significantly better than the simple linear regression, and while slightly over-dispersed, could reasonably be considered valid (Figure 11).

The final full GLM contained 40 features, the most significant being products with codes 710651, 693258, 698186, 800544, and 787306, followed by variables for "Tuesday", "Monday", "Weather_Rain", "Sunday", and "Wednesday".
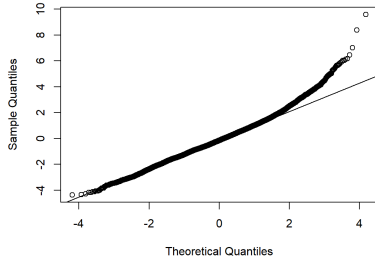

Fig. 9: Normal Q-Q Plot for the Poisson GLM.

Individual GLMs corresponding to each product-store combination, in the majority of cases, resulted in r-squared values consistently lower: median = 0.05, maximum = 0.21; and a mean RMSE of 3.87; median RMSE = 4.09. The results for two example stores are shown in table I. Model building in R returned warnings indicating the number of instances were too small relative to the number of independent variables and overly sparse information contained in the independent variables, indicating that many of these individual product-store models may be not be valid.
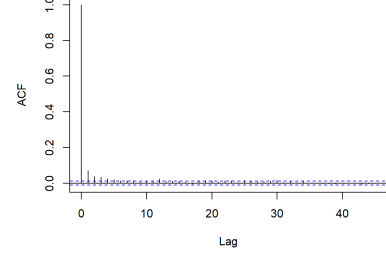

Fig. 10: Evidence in the residuals of the Poisson GLM of autocorrelation for small lags.

Store Number	Product Code	r-squared	RMSE	AIC	n
3446	673134	0.036	4.06	1920.76	380
3446	673159	0.023	3.93	1739.24	389
3446	673160	0.015	3.86	1728.16	350
3446	735944	0.103	3.81	948.85	231
3446	740282	0.010	3.77	1370.91	328
3446	748307	0.183	6.19	2087.80	398
3446	755255	0.061	3.84	1307.90	298
3446	799917	0.087	3.93	1293.73	381
3446	800543	0.073	3.77	1382.91	372
5147	673134	0.048	14.11	3618.66	401
5147	673158	0.043	4.92	2206.89	344
5147	673159	0.015	8.17	2372.09	410
5147	673160	0.040	12.05	2969.83	390
5147	693258	0.012	3.84	1382.13	360
5147	698186	0.004	4.37	1044.97	330
5147	710651	0.043	4.32	942.83	314
5147	735944	0.090	3.98	1419.53	330
5147	740282	0.052	5.47	1978.43	336
5147	748307	0.021	3.77	1509.26	361
5147	755255	0.011	4.32	943.27	281
5147	799917	0.036	4.19	1220.74	384
5147	800543	0.009	4.24	1144.13	363
5147	800544	0.044	4.09	1126.21	349
5147	853872	0.236	4.51	112.70	32

TABLE I: Results for GLMs corresponding to each unique product sold at Store 3446, "Petrol Three Peaks" and Store 5147, "Heysham- Middleton Road". n indicates the number of records available to build each model.
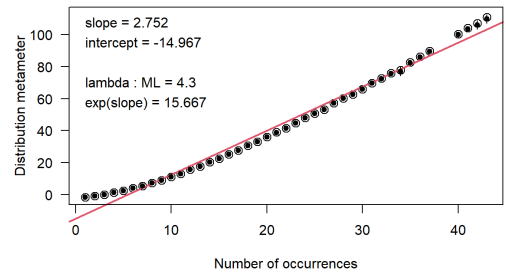

Fig. 11: "Poisonness" plot to show the fit of the response variable to the Poisson distribution.

B. XGBoost

The XGBoost model using default hyperparameters returned and r-squared value of 0.53 on the testing data set. Using optimised hyperparameters (`max_depth = 8`, `min_child_weight = 7.351867`, `gamma = 0.4565207`, `eta = 0.2364165`, `lambda = 6.683715`, `nrounds = 59`) returned an r-squared value of 0.579. The RMSE of the default model was 2.51, and the RMSE of the optimised model was 2.43.

The significance (importance) of features included in the final XGBoost model is shown in Figure 12. Generally, these are not comparable to the most significant features in the Poisson GLM.

The XGBoost model also appears to handle extreme values (high sales quantities) better than linear type models, returning a maximum prediction of 25. It is noted that time taken to tune the XGBoost hyperparameters was around 20 minutes; however, given the improvement in performance this was worthwhile.
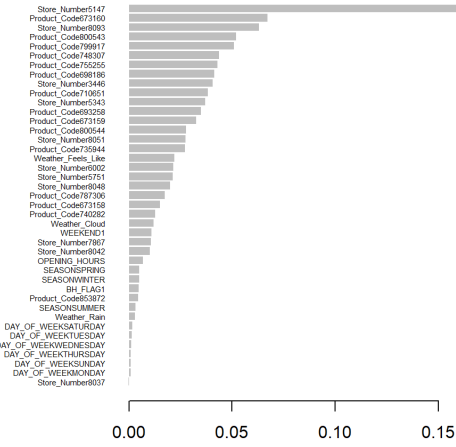

Fig. 12: Feature importance in the XGBoost model.

C. ARIMA

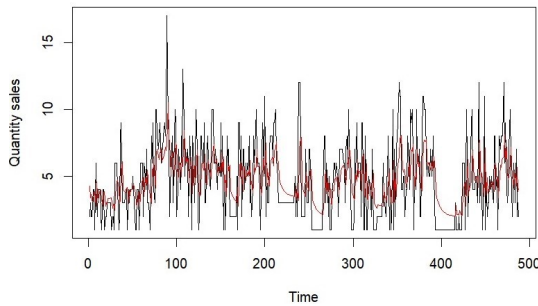Figures 14 and 13, and table II summarise the results of the ARIMA time series analysis.

Fig. 13: The fitted ARIMA model for product 673160 (Co-op White Rustic Roll) at the "Petrol Three Peaks" store, across the full time span of the data.

It should be noted that the mean percentage absolute error (MPAE) results for certain products were found to be infinity
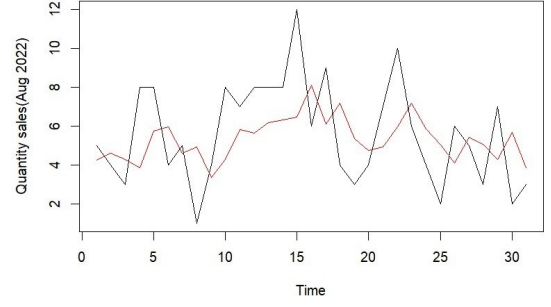

Fig. 14: The fitted ARIMA model for product 673160 at the "Petrol Three Peaks" store, showing only August 2022.

Product	Coefficients (AR,I,MA)	AIC	RMSE	MAPE
673160	(2,0,1)	2150.23	2.51	75.06
755255	(1,1,2)	1879.88	1.88	INF
673134	(2,1,1)	2223.59	2.76	62.06
673159	(1,0,0)	2052.81	2.22	49.29
735944	(2,1,2)	1669.38	1.50	INF
740282	(2,1,1)	1876.69	1.86	53.32
748307	(1,1,1)	2423.29	3.43	INF
799917	(0,1,2)	1385.87	1.10	INF
800543	(3,1,1)	1617.87	1.42	INF

TABLE II: Results of the ARIMA analysis for products sold at the "Petrol Three Peaks" Co-op. Note that MAPE values of INF result when predicted values are equal to 0.

(Table II). This is due to the forecasted value being zero, which causes division by zero and results in infinity.

The ARIMA models performed significantly better than the linear regression and Poisson GLMs in many instances, and although there was a large variability in the RMSE, the average RMSE for the Petrol Three Peaks store was 2.08.

D. Results Summary

A summary of the results is shown in Table III.

Model	Type	RMSE
Linear Regression	Full	3.00
Poisson GLM	Full	2.86
Poisson GLM	Product-Store	3.87 (mean)
XGBoost	Full	2.43
ARIMA	Product-Store	2.08 (mean)

TABLE III: Summary of results for all forecasting methods using a held-out testing set.

IV. DISCUSSION

Linear dependencies in the data given do not provide enough information to predict future quantity sales using simple linear regression or GLMs. There appears to be a large amount of noise in the data, and it is likely that unaccounted events (for example, demand influenced by COVID lockdowns) have influenced the data over the relatively short period on record. The predictive power is diminished significantly where training data is restricted to specific product-store examples, likely due to having far fewer data points by which to inform the model.

Moreover, the linear models failed on several assumptions for linear regression. The linear model was quickly found to

be unsuitable as the response variable more closely follows a Poisson distribution than a Normal distribution. There was evidence of auto-correlation in both the linear regression and the GLMs, suggesting that time series analysis techniques in general may be more appropriate. Many of the variables were categorical, resulting in a large number of dummy variables in the final model. Those treated as continuous ("Weather_Rain"; "Opening_Hours") while continuous in nature, were sparsely described in the data, and more closely resembled three categories in each variable. Therefore, any linear relationships were relatively weakly exhibited.

The linear model including store number and product code as dummy variables, also by design fails to capture the variability due to specific product-store combinations. It is possible that there are also interactions between weather and dates, e.g. warm and sunny weather could amplify sales on bank holidays more than it may on weekdays. Therefore, it is possible that the model could be improved by interaction terms between product and store dummy variables or various other predictor variables, however, this could result in a much larger and complex model.

Weather is of particular interest, as, for example, in the XGBoost model "Feels Like" temperature and cloudiness were the most significant variables without considering products and stores. As mentioned before, it has been shown that ideal temperatures for consumer spending may be found at moderately warm temperatures, and decrease at temperatures too high or too low [1]. This could indicate that a polynomial term, rather than the linear terms used in this project, would better describe the variability in sales resulting from temperature. However, the study cited was carried out in China, where climate and consumer spending habits may not be comparable.

XGBoost produced a reasonably useful model, considering the noise in the data. As tree-based models are able to capture non-linear factors, this is likely the reason for its superiority over linear models and GLMs. Some caution may be required around the optimised XGBoost model. While cross-validation was used, as there appears to be a large amount of random variability in the response variable, it is still possible that the model is overfit on season variables, given that only 18 months of data were available.

ARIMA analysis handled the time series relatively well. This method is more appropriate for harnessing information from recent sales, allowing unknown shorter term influences to influence near term predictions. Although the ARIMA models have shown better performance on average, the variation in RMSE and AIC could be high between models, and therefore some products (e.g. code 748307) produced predictions poor relative to others.

In such cases, it could be valuable to compare the ARIMA model forecast with predictions made using the full GLM or XGBoost model. Further research could also be made into mixed ensemble models that include both GLMs or XGBoost and ARIMA analysis. The ARIMA model is not able to capture variability arising from weather, which can exhibit large relatively hard-to-predict fluctuations in the UK climate, yet it is apparent from the feature importance of the XGBoost model that such variables can be informative. Longer time series spanning several years may be required for time series analysis to adjust for variability occurring on annual/bank holidays.

V. CONCLUSIONS

We conclude that the number of products to be baked at each store each day may be recommended using a corresponding ARIMA model. However, a degree of error must be expected due to the random nature of sales, and it could be necessary that unusual events such as lockdowns, holidays, or extreme weather, be considered in addition.

The best preforming models (using ARIMA) were produced solely using information from previous sales in time, without further input from variables such as weather or public holidays. However, GLMs and XGBoost showed that such variables contain some information to inform full price sales. In some cases, predictions made using a GLM or XGBoost model fitted to data from all stores and products may be more reliable. It is possible that a combination of these methods be developed - for example, an ensemble of ARIMA, a Poisson GLM, and XGBoost models, so as to capture variability arising from both the independent variables recorded and from recent previous sales.

Reduced-to-clear sales were not evaluated in the project. Where no reduced items have been sold, it is unknown whether it is due to under-supply of full price items (leaving none remaining to be reduced) or that supply has been perfectly met, or that some items were reduced-to-clear, and still not sold (over-supply). However, there is potential for the presence of reduced to clear sales to be used to flag the upper limits of sales. If data were collected quantifying unsold reduced items and sold-out full price items, this could also be investigated further.

REFERENCES

[1] X. Tian, S. Cao, and Y. Song, "The impact of weather on consumer behavior and retail performance: Evidence from a convenience store chain in china," *Journal of Retailing and Consumer Services*, vol. 62, p. 102583, 2021.

[2] K. B. Murray, F. Di Muro, A. Finn, and P. P. Leszczyc, "The effect of weather on consumer spending," *Journal of Retailing and Consumer Services*, vol. 17, no. 6, pp. 512–520, 2010.

[3] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*. Cambridge university press, 2013, vol. 53.

[4] P. Dalgaard, *Statics and Computing Introductory Statistics with R*. Springer, 2008.

[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[6] D. A. Dickey and W. A. Fuller, "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica: journal of the Econometric Society*, pp. 1057–1072, 1981.

[7] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer science & business media, 2009.

[8] K. Ord and S. Lowe, "Automatic forecasting," *The American Statistician*, vol. 50, no. 1, pp. 88–94, 1996.

[9] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Linear methods for regression," *The elements of statistical learning: Data mining, inference, and prediction*, pp. 43–99, 2009.

[10] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for r," *Journal of statistical software*, vol. 27, pp. 1–22, 2008.