Exploring the impact of customer demographics and coupon redemption on the accuracy of XGBoost demand forecasting model

MSc Research Project
MSc in Data Analytics

Boryana Mihaylova
Student ID: x23295546

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Boryana Mihaylova
Student ID:	X23295546
Programme:	MSc in Data Analytics **Year:** 2
Module:	MSc Research Project
Supervisor:	Mohammed Hasanuzzaman
Submission Due Date:	11/12/2025
Project Title:	Exploring the impact of customer demographics and coupon redemption on the accuracy of XGBoost demand forecasting model
Word Count:	5,957 …………………… **Page Count** 17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Boryana Mihaylova

Date: 11/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploring the impact of customer demographics and coupon redemption on the accuracy of XGBoost demand forecasting model

Boryana Mihaylova
X23295546

Abstract

Demand forecasting in the food and retail sector is essential for efficient supply chain management, yet it is a highly challenging task due to a variety of internal and external factors affecting demand. Promotions are one of the factors that complicate forecasting. Customer engagement is crucial for business growth and retailers must have a good understanding of their consumer base to tailor their marketing efforts accordingly. Customer demographics and coupon redemption data can be incorporated into the demand forecast to capture their effect. This paper investigates the impact of these data on model performance and compares a baseline XGBoost model with 9 features and an XGBoost model with 19 features, 10 of which are derived from demographic and promotional data. The enhanced model led to a drop of 10% in both RMSE and MAE and 18% reduction in MSE. Homeowner description and household composition ranked in top 10 predictive features. The findings have implications for both demand forecasting and marketing. Grocery retailers can achieve better forecast and reduce the risk of stockouts or overstocking by incorporating demographic and coupon data. The findings can be used to tailor marketing strategies to household profiles to further strengthen customer satisfaction and engagement.

1 Introduction

1.1 Background and industry context

The food and grocery retail is a sizeable and steadily growing industry with a global annual value of over €10 trillion in 2023 and a compound annual growth rate (CAGR) of 5.3% in the five-year period leading up to 2023 (MarketLine, 2025). Main factors in this growth are the increase in disposable income, rapid advances in technology, and limited direct alternatives. Additionally, changes in consumer behaviour, such as a notable switch to healthier choices and home meal preparation during and post the Covid pandemic, have also played a part.

The food and grocery retail is highly competitive due to strong buyer power and easy switching to alternative retailers. Brand loyalty among customers is weak in favour of high price sensitivity, which is even more pronounced in periods of financial instability and inflation. The result is tighter profit margins, therefore, efficient supply chain management is essential to attract and retain customers and achieve sustained growth. The foundation of good supply chain management is accurate demand forecasting. This paper examines the role of coupons and demographic data in improving demand forecast predictions.

1.2 Importance

The high market competitiveness in the industry makes forecasting more challenging. Demand fluctuations could be caused by both external factors such as seasonality, and internal factors such as marketing campaigns. Marketing campaigns are a form of price competition strategy in which some product prices are reduced for a fixed period to attract higher sales volumes. Retailers aim to avoid both understocking and overstocking for various reasons. Lack of stock forces customers to look for alternative retailers which leads to lost sales. Overstocking is challenging from operational and cost perspective, as it leads to increase in storage and labour expenditure and is especially detrimental for perishable goods with short shelf life (Joseph et al. 2022). Retailers rely on demand forecasting to plan the necessary shelf space for the ever-growing assortment of products (Klepo and Novoselnik, 2024). Finding the optimum shelf space allocation is essential to minimise operational and storage costs and maximise profits.

Promotions also make demand forecasting complex. Even though demand is expected to surge in periods of promotions, the effect is not certain. Previous successful campaign does not necessarily translate in future success or the same increase in demand. In addition to coupons and discounts, various other factors drive demand surges during promotions, such as the promotional channels, frequency and duration of the campaign, parallel promotions at competitors. There is also the possibility of overall sales cannibalisation for the same product or products in the same category. For example, the customer might purchase larger quantities during the promotional week but will not buy the product in the following weeks or months due to overstocking. It is also possible that the promoted product cannibalised the sales of similar products. Frequent promotions create variability in demand which could lead to prediction errors and stockouts as a result (Ali et al. 2009). This makes it even more important to keep track of promotions to understand and account for their impact in demand modelling.

Retailers face various challenges in its supply chain management such as demand impacted by promotions, different demand patterns in online and in-store retail which complicate inventory operations, and the coordination of price reductions for their online and offline stores (Deng et al., 2023). To address these, companies often look to integrate their demand forecasting, inventory management, and price and markdown optimisation aiming at improvements in supply chain operations and customer service, profit and sales growth and operational and inventory cost reduction.

Grocery retailers need to understand their audience well to refine their marketing strategy and foster customer engagement. Thorough analysis of existing demographics data, such as customer type, marital status, income, education and so on can help retailers build a clear view of their customer base and develop tailored and relevant marketing strategies.

1.3 Research question and objective

A core business objective in grocery retail is to reduce both overstock and understock through better demand prediction. The size of the grocery retail market and the interconnectedness between supply chain management, promotions and demographics necessitate demand forecasting which goes beyond statistical time series to capture the effects of promotions and demographics.

Research Question. The challenges highlighted above motivate the following research question: How does the addition of customer demographics and coupon redemption features impact the accuracy of the demand forecasting model?

Research Gap. Most modern research on demand forecasting focuses on the comparison of machine learning methods and their performance. Even when marketing and/or demographic factors are included, their incremental effect on the model is not in focus (McRae et al. 2022). This research fills this gap by measuring the incremental effect of these factors on forecast performance.

Objectives. To address the above research question and identified gap, two XGBoost models are developed. The first model uses features derived from transactional data and serves as a baseline, whereas the second model incorporates customer demographics and coupon redemption in addition to the beforementioned features. The performance of the two models is evaluated and compared to capture the effect of the added data. In addition, the importance of features is analysed to determine the factors affecting the demand prediction. The study uses two years purchase history of frequent customers at a US-based grocery retailer with available demographic characteristics. Four of the available eight tables are utilised - transactions, products, demographics, and causal data.

1.4 Structure of the paper

Section 2 reviews literature on demand forecasting in retail and the impact of promotions and demographics on customer responsiveness. Section 3 outlines the selected methodology, section 4 provides details of the implementation, section 5 evaluates the models. Section 6 concludes the paper.

2 Related Work

2.1 Demand forecasting in retail

Joseph et al. (2022) explore deep learning approaches for demand forecasting in retail as an alternative to traditional machine learning models. The authors developed a hybrid model combining convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM), optimised with the Lazy Adam algorithm for efficient handling of updates and improved model training. The study aimed at short term forecast, 3 months, and uses transactional data over 60 months period, covering 10 stores and 50 stock items, and containing almost a million features. The baseline models developed for the study were K-nearest neighbour, linear regression, bagging, random forest, stochastic gradient descent (SGD), support vector regressor (SVR), XGBoost and a hybrid of CNN and LSTM with Adam optimiser. The resulting model outperformed the benchmark models based on MAE, MAPE and R-squared metrics. Since this model is based on neural networks, it involves extended training time, rendering it computationally expensive. However, the authors note that the prediction time was less compared to the baseline, making it a suitable candidate for real-world scenarios.

Klepo and Novoselnik (2024) address the shelf space allocation problem in retail developing two machine learning demand forecasting models, XGBoost and LSTM. The authors used two years of transactional data for the Favorita retail chain based in Ecuador, enriched with features for holiday type and daily oil price. Both models performed better compared to the statistical ARIMA approach. Based on weighted MAE, the XGBoost model outperformed LSTM by 13.78%. The authors note the relatively small size of the dataset as a limitation in terms of generalisability. The paper also states that although LSTM requires more computing power, it is better than XGBoost at capturing seasonality and temporal dependency in the data. Based on their findings, the authors recommend the exploration of a hybrid between machine learning and statistical approaches for future research.

Nguyen et al. (2024) use data provided by the multinational retailer Walmart to develop separate LSTM models to predict demand for each product category. Similarly to neural networks, LSTM is computationally expensive, hence the analysis was constricted to 300 stock items out of the original 3,049, and 600 out of the total 1,913 days to avoid lengthy training time. The selected items belonged to 3 categories. The data included transactional history from 10 stores across 3 US-states, product categories, promotional activities, and pricing, which was enriched with external macroeconomic data – unemployment rate, consumer price index (CPI), and consumer sentiment index (ICS). The addition of external data brought minor improvement in model performance. The authors indicated higher performance for the separate models compared to a model trained on all categories. They also discovered that model performance varied for the various categories due to underlying differences in demand. In their systematic literature review, Vlachos and Reddy (2025) mention LSTM models in the context of food/retailing and indicate their suitability for goods with relatively constant demand which might not be the case where extensive promotional campaigns are present. The authors also note the computational intensity, but better model accuracy compared to other techniques.

Ali et al. (2009) emphasise the difficulties caused by promotions on grocery demand forecasting and explore various techniques in addition to time series to improve model accuracy. The authors use the more traditional exponential smoothing as a benchmark and in addition experiment with stepwise regression, and with machine learning alternatives such as support vector regression (SVR) and regression tree. The implementation costs and model complexity were the main factors in evaluating the different techniques. The authors indicate that the more advanced the model, the better the prediction, however, at the expense of higher complexity and data preparation. Furthermore, the presence of promotions impacted model accuracy. In periods with stable demand and no discounts or coupons present, the exponential smoothing technique was performing well, and the more advanced techniques did not improve forecast accuracy. However, in promotional weeks, the benchmark performed poorly, while the other models significantly improved performance, with the regression tree method outperforming all techniques with an accuracy improvement of 65.17%. Possible limitation of this study is the limited focus on a single product category from a number of stores belonging to a retailer. The findings might not be applicable to other categories, or regions. In addition, even though the authors conclude that model accuracy improves by incorporating data from multiple stores and categories, the study uses smaller data than a typical retailer, hence pooling boundaries require further research.

2.2 Effect of promotional campaigns on consumer behaviour

Ali et al. (2009) indicate much higher sales during promotional weeks which is to be expected due to the influence of price and discount on customer behaviour. Being price sensitive, customers would react to higher discounts. Demand spikes can trigger stockouts if the demand forecast is inadequate, which in turn will result in lost sales and possible customer dissatisfaction.

Deng et al. (2023) note that customer demand is sensitive to promotions and during a marketing campaign an ecommerce might experience increase in demand overall and not only for items with reduced prices. The authors also note differences between substitute and complementary products. For substitutes, reduced price for one item might result in increased demand for it but reduced sales for another similar item. For complimentary products however, reduced price for one item might increase both the sales of this and another related item.

Arce-Urriza et al. (2017) examine the impact of promotion on consumer behaviour online and offline, looking for nuances between frequent and infrequent buyers. For the purpose, the authors analyse data from a Spanning grocer, focusing on a single product category. The study concludes that customer response to promotions in physical stores depends on purchase frequency, i.e. regular customers tend to be more sensitive to coupons and discounts than irregular buyers (Cortinas et al. 2010, Arce-Urriza et al. 2017). The same effect however was not observed in online shopping where convenience and time saving were found to be the main drivers for customer purchase rather than the price or purchase frequency. In addition, the authors noted homogenous customer behaviour online, compared to heterogeneous reaction to discounts in-store. It is important to note that these findings are based on research prior to the pandemic and the subsequent cost of living crisis. Consumer behaviour has changed significantly and the observed differences between online and in-store shopping might not be applicable any longer.

Promotions are a powerful marketing tool for attracting new and retaining existing customers (Alexandru and Andrei-Dorian, 2024). Many companies use loyalty programs to reward their existing customers to strengthen customer engagement and perception of the business. Whether discounts, flash sales or offers of the type buy-one-get-one-free (BOGO), promotions are offered for a fixed period to create the psychological sense of good bargain and urgency which could lead to impulsive buying. Huch and Loewenstein (1991) were the first to look at the impulse purchase behaviour and its causes. The authors note that promotions are external stimuli that affect consumer perception of the product and create a feeling of loss or the fear of missing out (FOMO) if they do not purchase the product. Driven by their perception and the feeling of instant satisfaction, consumers may buy products impulsively, justifying the behaviour as necessary to avoid a loss.

2.3 Customer demographics and promotional responsiveness

Customer purchase history could be analysed to derive individual preferences and purchase patterns in order to personalise a marketing campaign to achieve a higher conversion of offers to sales (Alexandru and Andrei-Dorian, 2024). Personalised strategies are achieved by analysing the available customer demographics information and the difference in demand in the different customer groups. Analysing customer demographics is instrumental for efficient demand forecasting and for devising of targeted promotions to improve the effectiveness of marketing.

Tang (2024) proposes demand forecasting model with recency, frequency, monetary value (RFM) customer segmentation for better decision making in retail. Customers were segmented in four categories based on the level of their engagement. The author notes that demographic variables provide a better understanding of the patterns in purchasing behaviour exhibited based on age or gender, which can facilitate tailored promotional campaigns. Furthermore, future studies could include more demographics such as income for better segmentation and could explore more advanced techniques such as deep learning.

Virvilaite and Saladiene (2012) indicate that in addition to marketing stimuli, demographics such as gender, age, profession and income are determining factors in consumer behaviour and their responses to promotions in the shopping environment. The effect of self-perception and gender on consumer behaviour was explored in the model of impulse buying developed earlier by Diettmar et al. (1996). According to the authors, consumers tend to make a purchase to improve their self-perceived image. The higher the gap between the perceived and desired image, the higher the frequency in impulse purchases.

The impulse buying theory also touches on the aspects of materialism – the psychological perception that possessions equate with happiness. Consumers have different levels of materialism which is heavily dependent on demographic characteristics such as gender, age, income and education. The self-perceived image is also highly dependent on gender. For example, for women, physical appearance is central in one’s projection of image, hence purchases such as cosmetics, makeup, clothes, and jewellery, are often considered important for one’s appearance and are frequently bought on impulse – a behaviour which would be amplified further by promotions of these products. Men on the other hand are attracted to items expressing their independence and hobbies, such as tools, electronic gadgets, sports and leisure equipment, and as a result are more sensitive to promotional activities involving these products. Understanding these nuances, companies can tailor their marketing strategies to take in consideration demographics and self-perception to achieve higher campaign success in the targeted groups.

3 Research Methodology

The research methodology follows the steps of the CRISP-DM process – business understanding, data understanding, data preparation, modelling, evaluation and deployment. The dataset used for the project was obtained from Kaggle[1] and consists of the below eight tables which are described in detail in the Data Understanding section:

- Transactions
- Products
- Household demographics
- Coupons
- Coupon redemption
- Campaigns
- Campaign description
- Causal data

3.1 Business Understanding

This project aims to evaluate the effect of customer demographics and coupon redemption features on the accuracy of product demand forecast in grocery retail. Customer demographics are not readily available in transaction data and require merging of tables and feature engineering to include in the forecasting model. Developing two models, the first using transactional data, and the second using additional demographic and coupon data will allow to determine if this addition is beneficial for the model accuracy. The success criteria would be the improvement in model performance based on evaluation metrics. In addition, the SHAP plot will be used to interpret the predictive power of the features used in the model.

3.2 Data Understanding

Figure 1 shows the database schema and table variables prior to data exploration and table merging. In addition, code snippets from the exploration are included in the configuration manual. The transactional table contains 2,595,732 line items from 276,484 customer transactions over a two-year period. A total of 92,339 products were purchased by 2,500 customers in 582 stores. Other than household key, no personally identifiable information (PII) is included in the table, hence no further anonymisation or pseudonymisation is required. The preliminary correlation matrix which is included in the configuration manual shows high correlation between basket ID, day and week. Even though day and week numbers are included, they are not temporal variables but rather index identifiers representing the number of days in the table – 711 and number of weeks – 102.

---

[1] Dunnhumby - The Complete Journey

The product table contains 92,353 product IDs which are grouped in 44 departments, 308 commodity descriptions, 2383 sub-commodity description, 6476 manufacturers, 2 brand, 4345 product sizes.

The demographics table contains 801 unique customer IDs with their demographic characteristics – age bracket, marital status code, income bracket, homeowner description, household composition, kid category.

The coupon table consists of 119,384 entries after removing duplicates and shows coupons sent to customers during 30 marketing campaigns. A coupon can belong to only one campaign but is valid for more than one product. A product can be part of many campaigns.
The coupon redemption table lists the coupons redeemed by customers. The data are incomplete – in transactions, coupons are redeemed from day 1 to 705, while this table lists coupons from day 225 to 704.

The campaigns table contains 7208 entries and uniquely identifies 30 campaigns and their types and the household ID that received coupons. 1584 households were recipients of campaigns. The campaign description table lists the start and end day of a campaign. It is important to note that campaigns can overlap.

The causal data table contains 36,786,524 entries and details whether a product is part of a weekly mailer, and if it was featured in-store. The size of the table is due to the unique combination of product ID, store ID and week number. The table is incomplete – there is information for 68,377 products out of the total 92,339 unique products sold during the period available in transactions. Causal data are available for only 115 stores out of the 582 for which there is transaction data. Finally, 93 weeks are available out of the total 102 part of the transactional data.

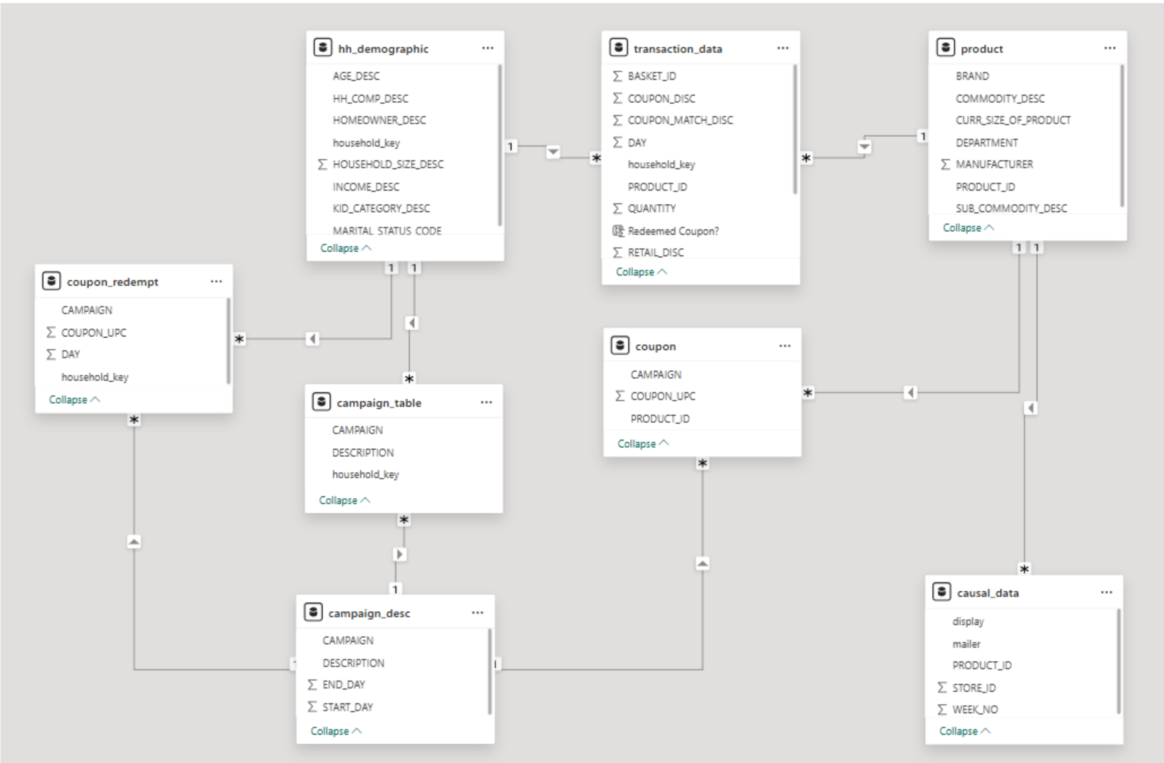

Figure 1 Database schema

3.3 Data Preparation

After exploring the available variables and the potential and usefulness of merging tables, four tables were merged: demographics, transactions, products, causal data. The following columns were merged into transactions from products table: commodity description, brand, department, current size of product. Upon examining the resulting data frame, whitespaces were encountered in current size of product. These were replaced with "Unknown." Demographics data frame was merged with transactions on household key. Only transactions with demographics are kept. Since the causal data table had records for 115 stores out of the total 582 stores in transactions, I used the fillna() method during merge to replace missing values with "U" short for "unknown".

After merging, the number of line items were reduced to 1,427,303 to include transactions for the subset of customers with available demographic information. This number was reduced to 1,383,748 after data cleaning and grouping. The count of line items with redeemed coupons was only 24,549 which is less than 2% of the total.

3.4 Feature Engineering

Recency, Frequency and Monetary value (RFM)

From a data privacy perspective, household key is considered PII and should be excluded from model development. To be able to derive customer characteristics, recency, frequency and monetary value (RFM) were calculated and ranks from 1 to 5 were assigned to each. The higher the value in frequency and monetary value, the higher their ranks. For recency, it is the opposite, the lower the value, i.e. the more recent a transaction is, the higher the rank. The three ranks were then combined in a final RFM score variable for each customer, with 555 being the highest and 111 being the lowest rank.

Unit Price

As mentioned earlier, correlation matrix was produced for the transactions table before merging which shows high correlation between sales value and quantity, since sales value is derived from quantity multiplied by price. To avoid data leakage, sales values were not used in the model. A unit price feature was created instead.

Basket size and weekly number of transactions

Although Store ID integer variable was present in the merged table, it is just an identifier and meaningless for the model. To derive some store information, weekly basket size and weekly transaction by store were calculated and added to the merged data frame.

Redeemed Coupon

A binary variable was created to indicate whether a coupon was redeemed for a product.

Received Loyalty

A binary variable was created to indicate whether a discount was applied due to loyalty card. The basis for this feature was the retail discount variable which was not used in the model.

Ordinal variables

Income, age, household size and kid category are ordinal variables because there is a ranking in each of them. For those, ordinal encoding was applied to convert to integer array.

Nominal variables

The below variables are nominal, i.e. they are just categories with no order or ranking. They were cast as type "category" to avail of XGBoost native category handling:

- Brand, department, current size of product, marital status, household composition description, homeowner description.

Finally, the cleaned data were grouped by week, store ID, household key and product ID.

3.5 Modelling

This step involved forecasting quantity for the last four weeks of the available 102 weeks for 801 customers and 582 stores. The dataset was first grouped by week, store, customer and product, and then was split into training, validation and test set. The XGBoost gradient boosting library was used to train two models - a baseline model with features derived from transaction data, and a second model incorporating coupon redemption and demographic information.

3.6 Evaluation

The choice of evaluation metrics depends on the type of the task – whether it's a regression or classification problem (Theobald, 2024). The metrics below are applicable for regression tasks.

- **Mean squared error (MSE)** is calculated as the squared sum of the differences between predicted and actual values (Gallatin, 2023).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

  In the above formula, n is the number of observations, $\hat{y}_i$ is the predicted value of the target, while y_i is the actual target value. Lower MSE indicates better model and smaller distance between predicted and actuals. This metric has penalising effect for a few larger errors and could result in a higher value for a model that otherwise has low distance between predicted and actual values.
- **Root mean squared error** (RMSE) takes the root of MSE thus removing the penalising effect of squaring. It indicates the overall accuracy of the model.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

- **Mean absolute error** (MAE) calculates the average difference between the predicted and actual data in absolute terms.

$$MAE = \frac{1}{n}\sum |\hat{y} - y_i|$$

- **R-squared** is calculated as the sum of the squared differences between the actual target and the predicted value is divided by the sum of the squared differences between the actual and mean of the target, which is then subtracted from 1. R-squared captures how well the model explains the actual demand variance, and it takes a value between 0 and 1, with higher metric meaning that the model captures most of the factors influencing it.

MAE and RMSE are most commonly used in practice due to easier interpretation. This paper provides the results for all four metrics.

4 Implementation

XGBoost was utilised for this project. The implementation used the XGBRegressor class with objective: "reg:squrederror" specifying a regression learning task and loss function. The remaining parameters are as follows:

- n_estimators=1000 – this determines the number of trees or boosting rounds. Higher number increases training time.
- tree_method = 'hist' – this is a learning type parameter specifying the speed of modelling and memory use. "Hist" is used with higher speed requirements.
- enable_categorical = True – this parameter was used to handle categorical variables that are nominal.
- max_depth=7 – indicates the maximum tree depth with a range from 0 to infinity. The default for this parameter is 6. Higher value increases model complexity and the possibility to overfit, but lower value could result in underfitting.
- eta=0.01 – this is the learning rate which is used to avoid overfitting of the model. The learning rate is an important boosting parameter that reduces feature weights. It can have values from 0 to 1, with a lower rate ensuring a better fit but needing longer training.
- subsample=0.7 – this is also a boosting parameter ranging from 0 - 1 which indicates what portion of the training data is used per tree.
- colsample_bytree=0.8 – this boosting parameter ranges from 0 - 1 and indicates the portion of the features per tree.
- random_state=42 – random state is used for code reproducibility.

4.1 Model 1

The baseline model included 9 features indicating week number, brand, department, product size, RFM score (Table 1). The data were split as follows:

- Training set – from week 1 to week 94
- Validation set - from week 95 to week 98

- Test set - from week 99 to week 102

The XGBoost regressor was initialised and native categorical handling was enabled for the features brand, department and product size. The same parameters were used for both model 1 and 2 and are listed in section 4 Implementation.

The MAE for the training set shows that the forecast is off by 37 units on seen data (Table 2). The R-squared value is 86% which is relatively balanced and does not indicate over- or underfitting. The training RMSE indicates a typical error of 539 units which requires a follow-up.

The MAE for the validation set is slightly worse than the training, at 42 units, and the RMSE value increases by 11%, indicating some outliers. Overall, the model is generalising well during development.

4.2 Model 2

The second model included 19 features as outlined in Table 1.

Table 1 Features Overview

Table	Feature	Model 1	Model 2
Transactional	Week number	X	X
	Unit price	X	X
	Loyalty discount	X	X
	RFM score	X	X
	Redeemed coupon		X
	Weekly basket size	X	X
	Weekly transactions	X	X
Product	Brand	X	X
	Department	X	X
	Product size	X	X
Causal data	Display		X
	Mailer		X
Demographics	Marital status code		X
	Household composition description		X
	Homeowner description		X
	Income description encoded		X
	Age description encoded		X
	Household size description encoded		X
	Kid category description encoded		X

The MAE for the training set shows that the forecast is off by 32 units on seen data which is 12% lower than model 1 (Table 2). The training set R-squared is 89% and is not indicative of overfitting. The training RMSE shows a typical error of 472 units which is 12% lower than model 1 but still high enough to require further check.

The MAE for the validation set shows that the forecast is off by 39 units, and the R-squared of 85% does not indicate overfitting during the development.

5 Evaluation

The RMSE, MAE, MSE and R-squared for the two untuned models are summarised in Table 2. Both models performed well on unseen data, and with R-squared values of 84% and 87% for model 1 and 2 respectively, capture the variance in product demand well and the different factors influencing demand are accounted for. On average, the first forecasting model is off by 39 units weekly, while the second model is off by 35 units weekly from the actual quantities. However, the RMSE scores for both are relatively high, indicating that there are some large individual errors in the forecast. The typical error for unseen data for model 1 is 549 units, while for model 2, the average error is 496 units. This prediction errors could be due to variability in demand caused by promotions (Ali et al. 2009).

Table 2 Model Evaluation Metrics

Metric	XGBoost Model 1			XGBoost Model 2 incl demographics and coupons		
	Train	Validation	Test	Train	Validation	Test
RMSE	539.1474	596.9801	549.2254	472.8298	554.4226	496.9819
MAE	37.1650	42.3507	38.5158	32.7213	39.2004	34.6737
MSE	290,679.875	356,385.25	301,648.5	223,568	307,384.4375	246,991.0156
R^2	0.8558	0.8313	0.8443	0.8891	0.8545	0.8725

Figure 2 shows the features that are used more often across all the trees in the model. The top three features are department, loyalty and unit price.

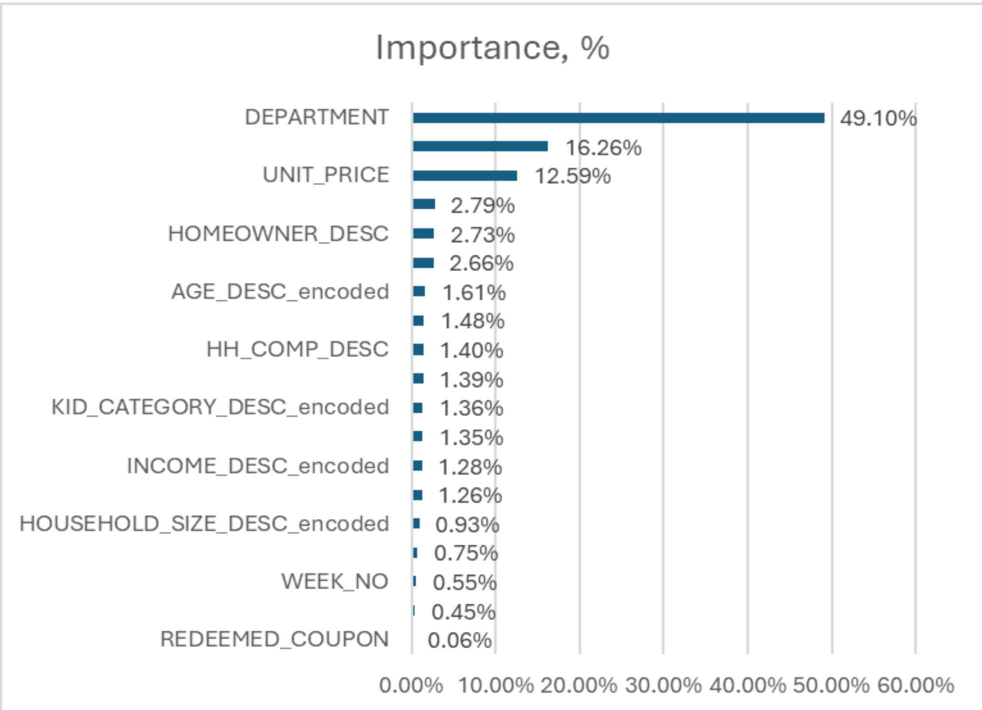

Figure 2 Feature use in the model

SHAP plots helps to interpret how a feature increases or decreases an individual prediction or globally for the model (Awan, 2023). The mean SHAP values plot can be used to visualise

the feature importance based on their impact on predictions overall (Figure 3). Department is the most influential feature impacting product demand for the given dataset, followed by unit price. In third place is the weekly basket size which is a proxy feature for store. It means that demand vary across stores. Looking at demographics, the homeowner description came 7th, and the household composition ranked 10th before income.
The impact of promotions ranked last with all three promotional features at the bottom of the plot. This needs to be interpreted with caution, since the number of redeemed coupons was low and the data for mailer and store display was available for only 115 out of the 582 stores. Perhaps, more data could change the predictive power of promotions. In addition, more features could be engineered by merging other tables, however, this requires significant effort in terms of data preparation and cleaning.

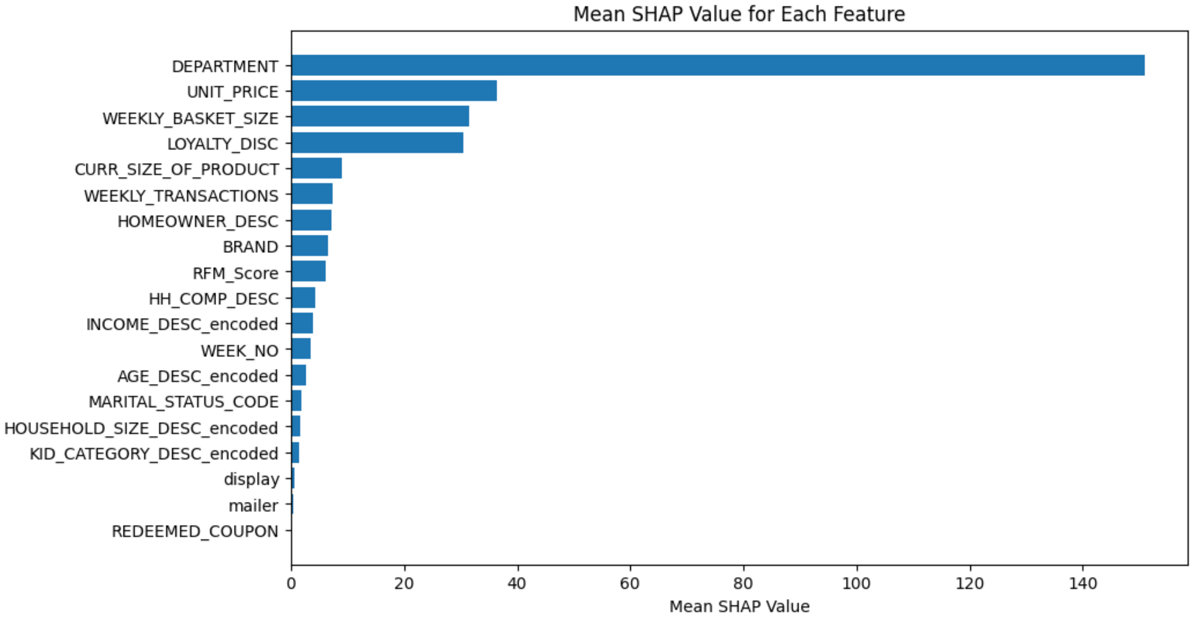

Figure 3 Mean SHAP Value for Each Feature - Model 2

5.1 Discussion

The incorporation of demographic and coupon features in the XGBoost model resulted in better model performance with a reduction of 10% in both RMSE and MAE and 18% decrease in MSE (Table 3). This is quite significant for better inventory management. The RMSE of 496 units should be addressed further by examining the large errors in the data to avoid significant under- or overstocking for some products. The R-squared value also improved but by a smaller percentage – 3%, and this could be due to noise being introduced in the model with the inclusion of the week number variable which is an integer from 1 to 102 and not an actual year week.

Table 3 Percentage improvement in Model 2 over baseline

Metric	XGBoost Model 1	XGBoost Model 2	Change (%)
RMSE	549.2254	**496.9819**	-10%
MAE	38.5158	**34.6737**	-10%
MSE	301,648.5	**246,991.0156**	-18%
R^2	0.8443	**0.8725**	+3%

Three indicators of promotions were included – coupon redemption, mailer, and display. The first is readily available in the transaction data, however the data for the other two were incomplete and required merging from a large dataset. Data completeness might result in a better model.

The data preparation and feature engineering to be able to include demographic and promotional featured is time-consuming and further complicated by missing data. There is room for improvement in this regard. In addition, feature engineering requires more than one round. Overall, the improvement of the model is significant, but in practice it requires evaluation of the effort to include demographic and promotional factors in forecasting.

5.2 Limitations

This study has some limitations. The causal data are incomplete, with display and mailer information available for only 115 out of the total 582 stores. The transactional dataset is partial, which might not represent customer behaviour at the company overall, and the same apply for the subset of 801 households whose demographics were used to engineer features for the model. Subsequently, this has the potential to introduce bias. Demographic records are self-reported by customers during account creation, and as such, might not be fully accurate. External economic indicators and competitor promotion data, which could improve demand forecasting, are not available. Week and day numbers are provided but there are no actual dates to evaluate seasonality or to experiment with other approaches such as LSTM. Including actual temporal features could improve the implemented XGBoost model further.

5.3 Ethical Considerations

The data utilized in this study is anonymised and the only available PII was the household key which was used as an identifier but is not a feature in the model. The demographic information cannot be linked back to specific customers, therefore, their use for the improvement of demand forecast does not raise ethical concerns. However, the results of the model can be used for customer profiling and tailoring marketing strategies to target more engaged households.

6 Conclusion and Future Work

The goal of this paper was to measure the impact of customer demographics and coupon redemption data on demand forecasting. This was achieved through model development using the XGBoost technique and evaluating the performance.

The inclusion of demographics and coupon data improved model performance, leading to a reduction of 10% in both RMSE and MAE and 18% reduction in MSE. The R-squared value showed that 87% of demand variance is captured by the model and most factors affecting demand are accounted for. In terms of importance for predictability, two of the demographic features made it to top 10 – homeowner description and household composition. These two features were more important for the predictions than household income. These findings are significant for marketing and could be used further to tailor marketing strategies. Promotion features did not impact predictions significantly.

The two models require tuning which would further improve performance. Future work could also focus on business users – the findings could be combined with the original data and imported in Power BI to create a dynamic report for further data exploration and visualisation. This research did not distinguish between different types of promotions, such as discounts or BOGO, which could be refined in future work. The study of the demand of 801 customers is a simplification due to the number of transactions for the selection and unavailability of account information for all customers and therefore cannot fully emulate a real-world scenario. The work could be improved further by a larger sample which is a better representation of the customer base.

References

'MarketLine Industry Profile: Food & Grocery Retail in Global' (2025) *Food & Grocery Retail Industry Profile: Global*, pp. 1–47. Available at: https://research.ebsco.com/linkprocessor/plink?id=70c2e23b-f085-3af7-8a38-cea1e14ea990 (Accessed: 19 September 2025).

Ali, Ö.G. *et al.* (2009) 'SKU demand forecasting in the presence of promotions', *Expert Systems With Applications*, 36(10), pp. 12340–12348. doi:10.1016/j.eswa.2009.04.052.
Cortinas, M., Elorz, M. and Andreeva, M. (2010). 'The role of purchase regularity and purchase frequency in price and discount sensitivity', *Journal of Marketing Trends 1961-7798*. 1. 57-73.

Arce-Urriza, M., Cebollada, J. and Tarira, M. (2017) 'The effect of price promotions on consumer shopping behavior across online and offline channels: differences between frequent and non-frequent shoppers', *Information Systems & e-Business Management*, 15(1), pp. 69–87. doi:10.1007/s10257-016-0310-2.

Joseph, R.V. *et al.* (2022) 'A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting', *Computers and Electrical Engineering*, 103. doi:10.1016/j.compeleceng.2022.108358.

McRae, C., Nie, J.-Y. and Dube, L. (2022) 'Heterogeneity in feature importance and prediction performance for sales at the market and store levels: the case of branded yogurt products in Quebec', 2022 IEEE International Conference on Big Data (Big Data), Big Data (Big Data), 2022 IEEE International Conference on, pp. 4088–4096. doi:10.1109/BigData55660.2022.10020567.

Tuan Ngoc Nguyen *et al.* (2024) 'Product Demand Forecasting with Neural Networks and Macroeconomic Indicators: A Comparative Study among Product Categories', *Journal of Business & Management Studies (2709-0876)*, 6(2), pp. 170–175. doi:10.32996/jbms.2024.6.2.17.

Vlachos, I. and Reddy, P.G. (2025) ‘Machine learning in supply chain management: systematic literature review and future research agenda’, *International Journal of Production Research*, 63(16), pp. 5987–6016. doi:10.1080/00207543.2025.2466062.

Klepo, M. and Novoselnik, B. (2024) ‘Product Demand Forecasting for Shelf Space Allocation in Retail via Machine Learning’, *2024 47th MIPRO ICT and Electronics Convention (MIPRO), MIPRO ICT and Electronics Convention (MIPRO), 2024 47th*, pp. 146–151. doi:10.1109/MIPRO60963.2024.10569517.

Deng, Y. et al. (2023) ‘Alibaba Realizes Millions in Cost Savings Through Integrated Demand Forecasting, Inventory Management, Price Optimization, and Product Recommendations’, *INFORMS Journal on Applied Analytics*, 53(1), pp. 32–46. doi:10.1287/inte.2022.1145.

Alexandru, S.F. and Andrei-Dorian, P. (2024) ‘The Impact of Promotions on Consumer Behavior’, *Annals of ‘Constantin Brancusi’ University of Targu-Jiu. Economy Series / Analele Universităţii ‘Constantin Brâncuşi’ din Târgu-Jiu Seria Economie*, (6), pp. 296–300. Available at: https://research.ebsco.com/linkprocessor/plink?id=4a78d35f-b3b1-3b4b-a3ac-cd44f4f74167 (Accessed: 26 October 2025).

Tang J. (2025)‘Unlocking Retail Insights: Predictive Modeling and Customer Segmentation Through Data Analytics’, *Journal of Theoretical and Applied Electronic Commerce Research*, 20(2):59. https://doi.org/10.3390/jtaer20020059

Hoch, S.J. and Loewenstein, G.F. (1991). ‘Time-Inconsistent Preferences and Consumer Self-Control’, *Journal of Consumer Research*, 17:4, 492-507. https://doi.org/10.1086/208573
Virvilaitė R. and Saladienė V. (2012) ‘Models Investigation of Factors Affecting Consumer Impulsive Purchase Behaviour in Retail Environment’, *Economics & Management*, 17(2):664-670. doi:10.5755/j01.em.17.2.2196

Dittmar, H., Beattie J. and Friese S. (1996) ‘Objects, decision considerations and self-image in men’s and women’s impulse purchases’, *Acta Psychologica*, 93(1-3):187-206. doi:10.1016/0001-6918(96)00019-4

Datasets (Dunnhumby - The Complete Journey) – Source: https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey/ [Accessed 14 September 2025]

Theobald, O. (2024). *Machine Learning with Python.* Available at: https://learning.oreilly.com/library/view/machine-learning-with/9781835461969/ [Accessed 04 Dec 2025]

Gallatin, K. and Albon, C. (2023). *Machine Learning with Python Cookbook*. 2nd Edition: Available at: https://learning.oreilly.com/library/view/machine-learning-with/9781098135713/ [Accessed 04 Dec 2025]

Awan, A.A. (2023) An Introduction to SHAP Values and Machine Learning Interpretability, Available at: https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability [Accessed 06 Dec 2025]