Unpacking Consumer Preferences: Econometric and Machine Learning Models for Cereal Brand Selection

Anja Petric (598370), Dominik Baus (614414),
Iryna Matsiuk (598762), and Boris Rine (613746)

ERASMUS UNIVERSITEIT ROTTERDAM

Abstract

It is of great importance for supermarkets to be able to understand purchasing behaviour of customers. This paper aims to describe the influence of a broad set of variables on the purchase behaviour of households and to predict future purchasing decisions. The study uses a dataset consisting of all supermarket visits by 250 households spanning a time period of two years. We use the general and mixed logit model as econometric models to describe purchasing behaviour of households. To predict future purchases, additionally to the econometric models; Random Forest, XGBoost and AdaBoost machine learning models are used. We find that variables capturing past purchasing behaviour and marketing-mix variables have high significance when describing and predicting cereal purchases. While all models outperform the general logit model in terms of several scores, machine learning models calibrated with isotonic regressions yield the best predictions for this dataset.

Date: June 20, 2024

1 Introduction

Advances in digital technology have had significant benefits and implications for consumer behaviour analysis. Supermarkets now leverage these technological advancements to collect and analyze customer data more efficiently through scanner checkouts and loyalty cards, which provides insights into customer behavior and enable targeted commercial activities.

This report aims to compare the performance of traditional econometric models and state-of-the-art machine learning models in describing and predicting customer purchasing behaviour. Specifically, we examine logistic regression models, including the general and mixed logit models. To explore advanced methodologies, specifically for purchasing predictions, we incorporate machine learning approaches, including Random Forest, XGBoost, and AdaBoost. The motivation for using these models lies in their flexibility and ability to handle complex, non-linear relationships within the data, which traditional econometric models might not effectively capture (Varian, 2014).

While logit models are favored in econometrics for their tractability and interpretability, enabling analysts to understand the underlying mechanics and derive insights from model parameters, they rely on certain assumptions and constraints that may limit their adaptability. In contrast, machine learning models operate as 'black boxes', potentially outperforming logit models due to their lack of constraints regarding interpretability, allowing for superior predictive performance (Clements, 2016).

All models are applied to a consumer dataset containing purchasing information for six cereal brands from 250 households. This study aims to describe customer purchasing behaviour and determine which model provides the most accurate predictions. By evaluating these models, we seek to provide insights into the most efficient methods for consumer behavior description and prediction.

The structure of the report is as follows: Section 2 describes the data. Section 3 presents explains the methodology used. Section 4 presents the results and lastly, Section 5 summarizes the findings and concludes the report.

2 Data

We investigate the consumer purchase behaviour of households for six cereal brands in one supermarket chain: General Mills, Kellogg's, Philip Morris, Quaker Oats, Ralston Purina and Nabisco. The data are collected from a supermarket chain itself, covering a time period of around two years. The database contains information from 250 households, specifically, information about all trips to this supermarket chain for that time period for these households.

2.1 Basic Variables

- **Price ($\mathbf{price}_{ijt}$)**: The price of brand j during visit t by household i. Price is a crucial factor influencing consumer choice, as it directly affects the perceived value and affordability of the product.

- **Display ($\mathbf{dis}_{ijt}$)**: A dummy variable indicating whether brand j is on display during visit t by household i. Displays can attract consumer attention and increase the likelihood of purchase through enhanced visibility.

- **Feature Promotion (feat$_{ijt}$)**: A dummy variable indicating whether brand j is featured in advertisements during visit t by household i. Feature promotions can influence consumer behavior by highlighting specific brands and their benefits.
- **Lagged Choice (lag$_{ijt}$)**: A dummy variable indicating whether household i purchased brand j on the previous visit. This variable captures brand loyalty and the likelihood of repeat purchases based on past behavior.
- **Household Size (hhsize$_{it}$)**: The size of household i during visit t. Larger households may have different purchasing patterns compared to smaller ones due to different consumption needs.
- **Income (Income$_{it}$)**: The annual income of household i. Income can influence purchasing power and brand preferences, with higher-income households potentially opting for premium brands.
- **Time Since Last Purchase (weekslast$_{it}$)**: The number of weeks since the last purchase of cereal by household i. This variable can help understand the frequency of cereal purchases and potentially identify stockpiling behavior.
- **Day of the Week (day$_{it}$)**: The day of the week when the purchase was made. Purchase patterns can vary across different days due to shopping habits and routines.
- **No Purchase (nopurchase$_{it}$)**: A dummy variable indicating whether no purchase was made during visit t. This variable helps in modeling the decision to purchase or not.
- **Store Identifier (storeid$_{it}$)**: The identifier of the store where the purchase was made. Different stores may influence purchasing behavior differently due to location, store layout, and marketing strategies.

2.2 Additional Variables

Based on our dataset, we also create additional variables. First, we create a **Discount**$_{jt}$ denoting the possible discount for brand j during trip t, as we believe that deals often influence consumers.

To compute it, we group our data by brand, and for each brand, we calculate baseline price (P_{base}) as a mean of all prices for that brand when it is not on display or featured on an advertisement. We then calculate $P_{display}$ and $P_{feature}$ as average prices for this brand's serial when the product is displayed or featured in an ad respectively. We take averages over groups of 12 weeks to account for possible price changes over time. We then calculate the discount rates for feature and display as:

$$\delta_{D(j,t)} = \frac{P_{base(j,t)} - P_{display(i,t)}}{P_{base(j,t)}}, \ \delta_{F(j,t)} = \frac{P_{base(j,t)} - P_{feature(j,t)}}{P_{base(j,t)}}$$

Use them to get the effective price of the cereal:

$$P_{effective(j,t)} = price_{j,t} * (1 - \delta_{D(j,t)} * dis_{j,t}) * (1 - \delta_{F(j,t)} * feat_{j,t})$$

$$Discount_{jt} = P_{base(j,t)} - P_{effective(j,t)}$$

To allow for state dependence, we include a **lagged_choice**$_{ijt}$ variable in our models, as we believe that current decisions of the household are influenced by past ones. It will be included in the form of a dummy variable: lagged_choice$_{ijt} = I[y_{i,t-1} = j]$. Its value equals 1 if the household i chose brand j in the previous period and 0 otherwise. If the household did not buy cereal on trip $t-1$, we take their last purchased cereal into account and not the "no cereal" option. For the first trip of each household in our dataset, this variable is set to 0.

We also include a **brand_loyalty**$_{ijt}$ variable to account for households' preferences and repetitive purchasing behavior, defined as an exponentially weighted average of past purchase decisions:

$$b_{i,j,t} = \sum_{l=0}^{t-1} w_{i,j,l} I[y_{i,l} = j]$$

Where $b_{i,j,t}$ is the brand loyalty that household i has to brand j on trip t and $w_{i,j,l} = exp(-(\frac{t-l}{50}))$ are the weights of each observation: more recent trips (larger l) have bigger weights.

2.3 Summary Statistics

The summary statistics of all variables that are not specific to the cereal brand are presented in Table 1. All other summary statistics for variables that are cereal-specific are presented in Appendix A.

Table 1: Summary statistics

Variable	Mean	Std. Dev.	Min	Max
week	664.961	29.724	614	717
day	4.301	1.879	1	7
brandbought	2.294	1.339	1	6
trips	1.090	0.331	1	6
brandlag	0.622	1.283	0	6
FMYSize	2.530	1.426	1	6
Income	6.319	3.209	1	11
NoIncome	0	0	0	0
non_zero_brandlag	2.175	1.394	0	6

Overall, the households in our sample purchase one of the six cereals in 14.77% of the trips made to the supermarket.

3 Methodology

As mentioned in Section 2, during most of the trips no cereals are bought. As a consequence, we will model and estimate two different stages. First, we predict the probability that a household will purchase any cereal using a binary logit model. Second, we predict which cereal households that bought any cereal will purchase. We will focus our analysis on multi-class classification models and therefore compare the accuracy of the models using only the second stage. The compared models are the econometric models general and mixed logit and machine learning models random forest, XGBoost and ADABoost, which all will be introduced in this section.

3.1 First-Stage Binary Logit Model

The Binary Logit model (BL) can be used to model the probability of households purchasing a cereal brand or not, as this can be represented by a binary variable. The model allows us to estimate the effect of various explanatory variables on the probability of purchasing a cereal. We can analyze the marginal effects of the coefficients β to interpret how changes in the explanatory variables affect the purchasing decision of households. The model can be represented as follows:

$$P(Y_t = 1|X_t) = \frac{\exp(x_i'\beta_j)}{1 + \exp(x_i'\beta_k)} \tag{1}$$

3.2 Second-Stage Econometric Models

3.2.1 General Logit Model

The General Logit (GL) model, as described in Franses and Paap (2010) is an appropriate choice for modeling the probability of household cereal brand choice based on various marketing-mix variables. It is ideal for handling choice data where brands aren't ranked and explanatory variables are of two types: individual-specific and individual and choice-specific.

The GL model allows us to estimate the effect of various explanatory variables on the probability of choosing a particular brand. By analyzing the marginal effects of coefficients β, we can interpret how changes in the marketing-mix variables influence the likelihood of a household choosing a specific brand. Furthermore, we can evaluate if households with particular traits prefer specific brands.

This model offers a detailed understanding of consumer behavior, quantifying the impact of different factors on brand selection and enabling forecasting of these choices.

The goal is to model the probability that on a supermarket visit t cereal brand j is chosen. The GL model assumes that the utility $U_{i,j,t}$ that household i derives from choosing brand j at visit t is given by:

$$U_{i,j,t} = x_{i,t}'\beta_j + w_{i,j,t}'\gamma_j + \epsilon_{i,j,t}$$

Where:

- $x_{i,t}$: is a vector of household-specific explanatory variables during trip t that includes the 1 for the intercept.
- $w_{i,j,t}$ – a vector of choice-specific explanatory variables.
- $\epsilon_{i,j,t}$ – error terms
- β_j and γ_j – alternative-specific coefficients.

Coefficients named this way will maintain consistent meaning and interpretation throughout subsequent models.

Since the current utility derived from choosing brand j likely depends on the past utility, we also introduce an AR term in our utility specification, such that it becomes

$$U_{i,j,t} = x'_{i,t}\beta_j + w'_{i,j,t}\gamma_j + \epsilon_{i,j,t} + \rho u_{i,j,t-1}$$

The probability that a household i chooses brand j at visit t is then modeled as:

$$P(Y_{i,t} = j|X_{i,t}, W_{i,t}) = \frac{\exp(x'_{i,t}\beta_j + w'_{i,j,t}\gamma_j + \epsilon_{i,j,t} + \rho u_{i,j,t-1})}{\sum_{k=1}^{m} \exp(x'_{i,t}\beta_k + w'_{i,k,t}\gamma_k + \epsilon_{i,k,t} + \rho u_{i,k,t-1})} \tag{2}$$

The parameters of this model are estimated by maximizing the log-likelihood function:

$$log(L) = \sum_{t=1}^{n}\sum_{j=1}^{m} y_{t,j} * log(P(Y_{i,t} = j)) \tag{3}$$

where $y_{i,j,t} = 1$ if household i chose brand j was chosen on trip t and 0 else.

We will select the most frequently purchased brand as a reference category for identification purposes. This model will be used for forecasting and interpretation and to do the latter, we will use odds ratios:

$$\frac{P(Y = j \mid X, W)}{P(Y = k \mid X, W)} = \exp\left((\beta_{0,j} + \beta_{1,j}X_1 + \ldots + \beta_{p,j}X_p) - (\beta_{0,k} + \beta_{1,k}X_1 + \ldots + \beta_{p,k}X_p)\right.$$
$$\left. + \gamma_{1,j}W_1 + \ldots + \gamma_{p,j}W_p - (\gamma_{1,k}W_1 + \ldots + \gamma_{p,k}W_p)\right)$$

3.2.2 Mixed Logit Model

The mixed logit model is a generalization of the standard logit model that accommodates random taste variation, unrestricted substitution patterns, and correlations in unobserved factors over time Train (2009). It is particularly suitable for modeling household cereal brand choices as it accounts for heterogeneity in household preferences and allows for more flexible substitution patterns between brands, capturing more realistic consumer behavior compared to the General Logit model.

The general logit model assumes homogeneous preferences and independently distributed unobserved factors, which can be overly restrictive. The mixed logit model addresses these limitations by allowing coefficients to vary across individuals according to a specified distribution, thereby capturing household preference heterogeneity. It also accommodates correlations in unobserved factors over time, crucial for our panel data, where the same households are repeatedly observed. This flexibility helps us better understand household responses to marketing-mix variables such as price, display, and feature promotions.

In the mixed logit model with panel data, the utility U_{ijt} that household i derives from choosing alternative j (cereal brand) in choice situation t is given by:

$$U_{ijt} = \alpha_j + \beta'_i x_{ijt} + \gamma'_j z_{it} + \epsilon_{ijt} \tag{4}$$

where α_j is the alternative-specific intercept for brand j, x_{ijt} are the observed variables related to the alternative j and household i at time t, β_i is a vector of coefficients representing household i's tastes, which varies over decision-makers in the population with a density function $f(\beta \mid \theta)$, z_{it} are the household-specific variables with alternative-specific coefficients γ_j, ϵ_{ijt} is an error term that is independently and identically distributed (iid) extreme value.

The probability that household i chooses alternative j at time t is given by:

$$P_{ijt} = \int L_{ijt}(\beta) f(\beta \mid \theta) d\beta \tag{5}$$

where:

$$L_{ijt}(\beta) = \frac{\exp\left(\alpha_j + \beta' x_{ijt} + \gamma_j' z_{it}\right)}{\sum_j \exp\left(\alpha_j + \beta' x_{ijt} + \gamma_j' z_{it}\right)} \tag{6}$$

Given the panel data, the overall likelihood for household i making a sequence of choices over T time periods is:

$$L_{ij}(\beta) = \prod_{t=1}^{T_i} \left[\frac{\exp\left(\alpha_j + \beta' x_{ijt} + \gamma_j' z_{it}\right)}{\sum_j \exp\left(\alpha_j + \beta' x_{ijt} + \gamma_j' z_{it}\right)} \right] \tag{7}$$

The unconditional probability of household i choosing a sequence of alternatives is:

$$P_{ia} = \int L_{ia}(\beta) f(\beta \mid \theta) d\beta \tag{8}$$

In the mixed logit model, different distributions are applied to variables to better capture the heterogeneity in household preferences:

- **Log-Normal Distribution:** Coefficients for variables that consistently influence decision-making in the same direction follow a log-normal distribution. It is appropriate for marketing-mix variables, as their effect is consistently positive or negative across all consumers.

- **Normal Distribution:** Coefficients for variables that capture behaviors with more variability follow a normal distribution. This allows for a range of positive and negative effects, reflecting the diverse responses of consumers.

The estimation of the mixed logit model involves integrating the choice probability over the distribution of the random parameters, typically approximated using simulation methods. We use the Simulated Maximum Likelihood Estimation (SMLE) approach, where the simulated probability $\hat{P}_{ia}$ is the average of the logit probabilities over multiple draws of the random parameters:

$$\hat{P}_{ia} = \frac{1}{R} \sum_{r=1}^{R} L_{ia}(\beta^r) \tag{9}$$

The simulated log-likelihood function (SLL) is then given by:

$$SLL = \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ij} \ln \hat{P}_{ij} \tag{10}$$

where $d_{ij} = 1$ if household i chooses brand j and 0 otherwise. The parameters θ are estimated by maximizing the simulated log-likelihood function.

For implementation, we utilized the `xlogit` package in Python, which handles the simulation and estimation processes efficiently.

3.3 Machine Learning Models

In our study, we employ several ensemble machine learning algorithms—namely Random Forest, XGBoost, and AdaBoost—to predict household cereal purchases and brand choices during their next supermarket visit. These algorithms effectively handle large, complex datasets with many features, making them ideal for analyzing consumer purchase behavior. Additionally, to add interpretability to the models, we graphed the Feature importance plot that clearly shows the key variables influencing purchasing decisions. Consequently, for variables with the highest importance scores, we include Partial Dependence plots to visualise the marginal effect these features have on the predicted outcome of a machine learning model.

To further improve our models' predictive accuracy and reliability, we implemented advanced calibration techniques using CalibratedClassifierCV from sklearn.calibration in Python. We specifically focused on isotonic regression and Venn predictors to enhance the probability outputs of our classifiers after their initial predictions. Isotonic regression adjusts the model's probability predictions to fit the observed outcomes better by adjusting probability predictions in non-decreasing order, ensuring that higher actual values correlate with higher predicted probabilities. This method is helpful because it directly aligns the predicted probabilities with observed outcomes in a logical sequence, effectively addressing overfitting and underfitting issues without assuming a parametric relationship. On the other hand, Venn predictors offer a complementary approach by recalibrating probabilities non-parametrically based on empirical frequencies. This method divides predictions into subsets and adjusts probabilities to match the observed frequencies within each group without assuming a predefined relationship between predicted and observed probabilities. Both techniques ensure that our model outputs reflect the true likelihood of events and maintain consistency across different data samples.

3.3.1 Random Forest

The random forest technique, developed by Breiman (2001), is a robust ensemble learning method that builds multiple decision trees and aggregates their predictions for regression tasks. This method employs randomness by using bootstrap sampling and selecting random feature subsets at each tree split, reducing tree correlation and minimizing overfitting for enhanced prediction stability. Trees grow to their maximum depth without pruning, potentially overfitting their respective bootstrap samples. The model then averages the predictions from all trees to form the final prediction, ensuring the reliability and robustness of our results. We applied the RandomForestClassifier from the sklearn.ensemble library, optimal for handling panel data due to its efficacy in managing categorical and continuous variables.

3.3.2 Extreme Gradient Boosting Model

XGBoost, a gradient-boosting algorithm created by Chen and Guestrin (2016), is particularly effective in handling large-scale datasets with high-dimensional feature spaces (Hastie, Tibshirani, & Friedman, 2000). It builds a series of decision trees sequentially, each attempting to correct the errors of its predecessors, gradually improving predictive accuracy. Starting with an initial prediction, typically the mean of the target variable, new trees are added to predict residual errors of the existing model, thereby refining the overall prediction. To control model complexity and prevent overfitting, XGBoost applies regularization techniques, incorporating both L1 (lasso) and L2 (ridge) penalties, which help ensure the model generalizes well to new data.

Let $\mathbf{X}$ be the feature matrix with n samples and m features, and let $\mathbf{y}$ be the corresponding response vector of size n.

XGBoost can be expressed with the following equation:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{11}$$

where $\hat{y}_i$ is the predicted value, K is the total number of trees, f_k represents the k-th tree model, and x_i is the i-th input sample. Each base learner is a decision tree, and the final prediction is the sum of predictions from all the trees.

We implement the model with the help of the xgboost package in Python.

3.3.3 Adaptive Boosting Model

AdaBoost is a robust ensemble machine learning algorithm primarily used for binary classification but can also be extended to multi-class problems. It is particularly well-suited for predicting household cereal purchases because it handles imbalanced data effectively, focusing on problematic cases and enhancing overall model accuracy. By dynamically adjusting the weights of misclassified instances, AdaBoost ensures that the model pays adequate attention to minority classes and irregular purchase patterns, common in consumer purchase data. Moreover, AdaBoost mitigates the risk of overfitting by aggregating the modest contributions of individual weak classifiers, ensuring robust generalization to new data. We again used sklearn.ensemble library, here specifically the AdaBoostClassifier.

Prediction and Model Comparison

To see whether our additions to the available variables indeed improve the model's performance, we will estimate the 5 specifications described in **??** with the General Logit model and choose the best one based on their R^2 and AIC scores. we will then use the same variables for the Mixed Logit estimation.

Following the estimation of parameters for our models, including the General Logit (GL) model, Mixed Logit (ML) model, Random Forest, AdaBoost and XGBoost, we proceed to utilize these models to predict household cereal brand choices. The effectiveness and accuracy of these predictions are crucial for evaluating the performance of each model.

To ensure a rigorous evaluation, we will employ 5-fold cross-validation. This involves partitioning the dataset into five subsets. Each econometric model is trained on four subsets and validated on the remaining subset. For the machine learning models, we allocate 25% of the training set to the calibration process which refines the probability estimates of the models, improving accuracy and reliability. This process is repeated five times, with each subset used once as the validation set. The results from each fold are averaged to provide a reliable estimate of the model's performance. Cross-validation mitigates the risk of overfitting and ensures that our findings are generalizable across different samples of the data.

We use several metrics to comprehensively evaluate and compare the predictive performance of the models: hit rate (average accuracy), Brier Score, and confusion matrices.

Firstly, the hit rate, or average accuracy, measures the proportion of correctly predicted brand choices. This metric provides a straightforward assessment of model performance by comparing the predicted brand choice to the actual choice for each household and shopping trip. It is calculated as follows:

$$\text{Hit Rate} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{Y}_i = Y_i) \tag{12}$$

where $I(\cdot)$ is an indicator function that equals 1 if the condition is true and 0 otherwise, and N is the number of observations. Higher hit rates indicate better predictive accuracy.

Secondly, the Brier Score offers a more nuanced evaluation by quantifying the mean squared difference between the predicted probabilities and the actual outcomes (Clements, 2016). This score ranges from 0 to 1, with lower values indicating more accurate predictions. The Brier Score is calculated as:

$$\text{Brier Score} = \frac{1}{N}\sum_{t=1}^{N}\sum_{c=1}^{R}(f_{t,c} - o_{t,c})^2 \tag{13}$$

where N is the number of observations, R is the number of categories, f_{tc} is the predicted probability for category c at observation t, and o_{tc} is the actual outcome for category c at observation t. A lower Brier Score indicates better predictive performance, as it accounts for the accuracy of probabilistic predictions, penalizing both overconfident and underconfident predictions, thus providing a comprehensive measure of model calibration.

We also compute F1 score – widely used metric, that balances the trade-off between precision and recall in classification tasks. These values are computed in the following way:

Precision:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{14}$$

Recall:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{15}$$

F1 Score :

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

Lastly, confusion matrices provide detailed insights into the model's predictive performance by showing the frequency of correct and incorrect predictions for each brand. Each cell in the confusion matrix represents the number of times a particular brand was predicted as another brand, allowing us to identify specific patterns of misclassification and understand the nature of prediction errors.

4 Results

4.1 Binary logit

The binary logit model has an accuracy of 0.86713. Considering that around 85% of the households do not purchase cereal in our dataset, this accuracy is not very impressive. As a consequence, we will estimate the second stage of the model separately from these predictions, to be able to better compare the various econometric and machine learning models directly with each other. The coefficients are presented in Table 2.

Table 2: Coefficients of the binary logit model

Variable	**Estimate**	**Std. Error**	**Significance**
(Intercept)	-1.329	0.659	*
price:1	-0.840	2.022	
price:2	-2.686	2.673	
price:3	4.427	2.875	
price:4	-3.033	3.373	
price:5	-2.524	1.652	
price:6	-5.690	3.945	
feat:1	0.007	0.071	
feat:2	0.336	0.076	***
feat:3	0.124	0.088	
feat:4	0.010	0.116	
feat:5	-0.179	0.086	*
feat:6	0.037	0.089	
dis:1	0.028	0.047	
dis:2	-0.073	0.063	
dis:3	0.081	0.068	
dis:4	0.053	0.082	
dis:5	0.151	0.067	*
dis:6	-0.014	0.120	
lag:1	-0.685	0.080	***
lag:2	-0.606	0.069	***
lag:3	-0.599	0.133	***
lag:4	-0.391	0.129	**
lag:5	-0.347	0.185	.
lag:6	-0.362	0.191	.
weekslast	-0.014	0.003	***
FMYSize	-0.033	0.017	*
Income	0.026	0.006	***
cumulative:bought	0.006	0.002	**
proportion:bought	6.478	0.140	***
day:2	0.066	0.073	
day:3	-0.015	0.075	
day:4	0.088	0.071	
day:5	0.107	0.070	
day:6	0.159	0.068	*
day:7	0.119	0.075	
store.id:1420	-0.002	0.134	
store.id:1422	0.110	0.145	
store.id:1423	0.092	0.258	
store.id:1424	0.164	0.277	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

We can see that household-specific information such as the family size, income and weeks since last purchase, for example, are important to describe the purchasing decision of households. We find that the lags for buying different brands in the past are negatively associated with purchasing cereal at a certain trip while the total amount of cereal bought in the past by one specific household increases the likelihood of purchasing cereal.

4.2 General logit

Table 3: Variable selection

	S1	**S2**	**S3**	**S4**	**S5**
FMYSize	✓	✓	✓	✓	✓
Income	✓	✓	✓	✓	✓
dis	✓	✓	✓	✓	✓
feat	✓	✓	✓	✓	✓
price	✓	✓	✓	✓	✓
lagged choice		✓	✓	✓	✓
discount			✓	✓	✓
loyalty				✓	✓
AR utility					✓
AIC	14162.74	12965.48	12960.18	11974.64	11972.08
McFadden R2	0.025783	0.10918	0.11024	0.17901	0.17988

Brand 2 is the most popular choice in our sample, therefore we will now refer to it as a baseline category when estimating econometric models. To select the final specification, we start by investigating whether our additions described in subsection 2.2 and subsubsection 3.2.1 indeed improve our model. The AIC and McFadden R^2 of different specifications are described in Table 3. We indeed see that the AIC decreases with each added variable, therefore, we will use Specification 5 for forecasting and interpretations.

Table 4: General Logit estimation results

Variable	**Estimate**	**Std. Error**	**Significance**	**Odds Ratio**
(Intercept):1	-1.709	0.461	***	0.181
(Intercept):3	-4.109	0.888	***	0.016
(Intercept):4	-4.921	1.109	***	0.007
(Intercept):5	3.473	2.047	.	32.281
(Intercept):6	-3.989	2.763		0.018
price:1	1.547	2.013		4.698
price:3	10.356	5.398	.	28877.970
price:4	17.475	6.973	*	269794.303
price:5	-33.417	9.542	***	0.000

Continued on next page

Table 4 – continued from previous page

Variable	Estimate	Std. Error	Significance	Odds Ratio
price:6	6.233	15.348		512.694
dis:1	0.199	0.082	*	1.219
dis:3	0.104	0.176		1.111
dis:4	-0.223	0.212		0.800
dis:5	-0.280	0.295		0.757
dis:6	0.435	0.505		1.547
feat:1	0.483	0.107	***	1.620
feat:3	0.937	0.220	***	2.555
feat:4	0.465	0.271	.	1.589
feat:5	0.687	0.349	*	1.989
feat:6	-0.184	0.420		0.832
FMYSize:1	0.075	0.029	**	1.078
FMYSize:3	0.073	0.039	.	1.076
FMYSize:4	0.011	0.045		1.011
FMYSize:5	0.051	0.063		1.052
FMYSize:6	-0.282	0.095	**	0.756
Income:1	0.001	0.011		1.001
Income:3	0.015	0.017		1.015
Income:4	0.021	0.018		1.022
Income:5	0.053	0.029	.	1.055
Income:6	0.046	0.030		1.047
lagged_choice_non_zero:1	0.143	0.088		1.154
lagged_choice_non_zero:3	0.433	0.148	**	1.543
lagged_choice_non_zero:4	0.914	0.155	***	2.496
lagged_choice_non_zero:5	0.801	0.290	**	2.231
lagged_choice_non_zero:6	-0.860	0.383	*	0.423
discount:1	1.675	1.456		5.333
discount:3	4.833	4.994		125.729
discount:4	22.958	6.201	***	10260000000.000
discount:5	-19.305	8.795	*	0.000
discount:6	13.637	16.064		768300.000
brand_loyalty:1	2.607	0.150	***	13.562
brand_loyalty:3	3.515	0.303	***	33.594
brand_loyalty:4	2.877	0.296	***	17.748
brand_loyalty:5	4.787	0.487	***	119.520
brand_loyalty:6	6.813	0.489	***	911.123
lagged_utility:1	-0.659	0.211	**	0.521
lagged_utility:3	0.263	0.433		1.301
lagged_utility:4	-0.775	0.593		0.459
lagged_utility:5	0.589	0.719		1.804
lagged_utility:6	-0.222	0.635		0.801

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

The General Logit estimation results are presented in Table 4. From the table, we can observe that all variables, except for Income, significantly influence individuals' choice probabilities for some brands. Notably, the brand loyalty variable is significant at the 0.1% level for all alternatives. We can interpret the odds ratios provided in this table. For instance, an increase in the price of brand 5 significantly reduces the likelihood of a person preferring brand 5 over brand 2, assuming all other factors remain constant. Interestingly, our model indicates that an increase in the price of brand 4 actually increases the odds of consumers preferring it over the reference brand. Consistent with our expectations, we also find that cereals featured in promotions positively affect consumer preference over the reference category. Additionally, larger households tend to prefer brand 1 over brand 2 and are less likely to choose brand 6. Logically, having previously purchased a cereal and brand loyalty positively influence the likelihood of choosing it over cereal 2 for most brands. However, intriguingly, our results suggest that a larger discount on brand 5 and prior purchases of brand 6 decrease the odds of consumers preferring them over brand 2. Contrary to our expectations, an increase in past utility from buying brand 1 seems to lower the likelihood of customers preferring it over brand 2, all else being equal.

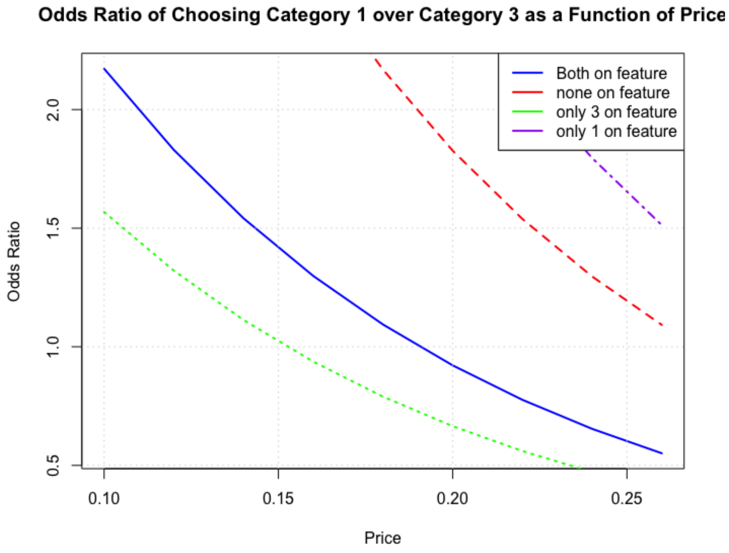


Figure 1: General Logit

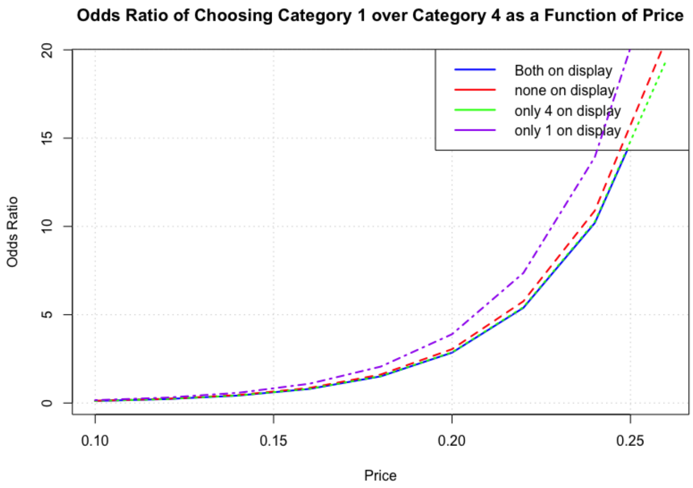


Figure 2: General Logit

To analyze the marketing mix variables in greater detail, we also plot the odds ratios as a function of price for some of the brands in our dataset. Since brands 1 and 3 both have a significant impact at the 0.1% level when featured in promotions, we compare the odds of them being preferred over each other depending on the feature variable value. As expected, Figure 1 shows that the highest probability of choosing brand 3 over brand 1, across all price levels within our dataset range and with other explanatory variables set to zero, occurs when only brand 1 is promoted and the lowest probability occurs when only brand 3 is promoted. Similar results are observed when comparing these brands with brand 5 (see Figure 9, Figure 10 in Appendix B). On Figure 2 we plot odds ratios of choosing brand 1 over brand 4 as a function of price. Consistently, the highest odds ratio for this event occurs when brand 1 is the only one on display. We find that odds ratios are roughly the same when only brand 4 is displayed and when both brands are displayed. This is supported by the observation that being on display does not significantly affect the probability of preferring brand 4, as shown in Table 4. Therefore, we expect similar plots for the other brands, as the display variable does not have a significant coefficient for them either.

4.3 Mixed Logit

In Table 5, the mixed logit estimation results are presented. In this analysis, the coefficients for price, display, feature, and discount variables follow a log-normal distribution, while the coefficients for lagged choice and brand loyalty follow a normal distribution. The AIC is 11889.9 with an associated McFadden R^2 of 0.17292.

Table 5: Coefficients of the Mixed Logit model

Variable	Estimate	Std. Error	Significance
(Intercept):1	-0.791	0.167	***
(Intercept):3	-1.078	0.151	***
(Intercept):4	-1.215	0.161	***
(Intercept):5	-2.837	0.303	***

Continued on next page

Table 5 – continued from previous page

Variable	Estimate	Std. Error	Significance
(Intercept):6	-1.566	0.225	***
FMYSize:1	0.085	0.027	**
FMYSize:3	0.014	0.040	
FMYSize:4	0.033	0.041	
FMYSize:5	0.060	0.064	
FMYSize:6	-0.373	0.087	***
Income:1	0.006	0.012	
Income:3	0.028	0.018	
Income:4	0.035	0.018	.
Income:5	0.089	0.028	**
Income:6	0.066	0.033	*
price	1.139	0.580	*
dis	-2.553	1.255	*
feat	-0.932	0.219	***
discount	1.243	0.594	*
lagged_choice_non_zero	0.300	0.066	***
brand_loyalty	2.004	0.068	***
sd.price	0.787	0.258	**
sd.dis	-0.968	1.109	
sd.feat	-0.788	0.163	***
sd.discount	-0.426	1.015	
sd.lagged_choice_non_zero	0.509	0.128	***
sd.brand_loyalty	1.037	0.080	***

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

The mixed logit model reveals significant insights into consumer heterogeneity. All intercepts are negative and significant, indicating a strong baseline preference toward brand 2. Furthermore, household size and income levels are significant indicators of brand choice for certain brands as well, contrasting with the general logit model where income was not significant.

Notably, the significant standard deviations for price, feature promotions, lagged choice, and brand loyalty underscore considerable heterogeneity in consumer responses. This finding demonstrates that the mixed logit model captures the variability in consumer preferences across households, a nuance missed by the general logit model. Unlike the general logit, which assumes uniformity in behavior, the mixed logit recognizes that consumers do not respond uniformly to changes in price, marketing efforts, or past purchases. This ability to accommodate heterogeneity makes the mixed logit model superior in providing a more accurate and realistic depiction of consumer choice behavior.

The price variable shows a positive and significant effect, contrary to the expectation that higher prices reduce brand choice. This suggests higher prices may indicate higher quality or prestige, possibly due to an underlying correlation between price and perceived quality or a segment of consumers prioritizing quality or brand prestige over price sensitivity.

Marketing efforts such as displays and feature promotions exhibit counterintuitive effects. The display variable is significant and negative, suggesting displays might decrease brand choice. Similarly, the feature promotion variable is highly significant and negative, indicating feature promotions tend to reduce brand choice. These findings may result from aggressive use of promotions for lower-performing brands, leading to negative associations, or frequent promotions signaling lower quality. Endogeneity could also be a factor, with less popular brands being more frequently promoted, rather than promotions directly reducing brand choice.

4.4 Machine Learning

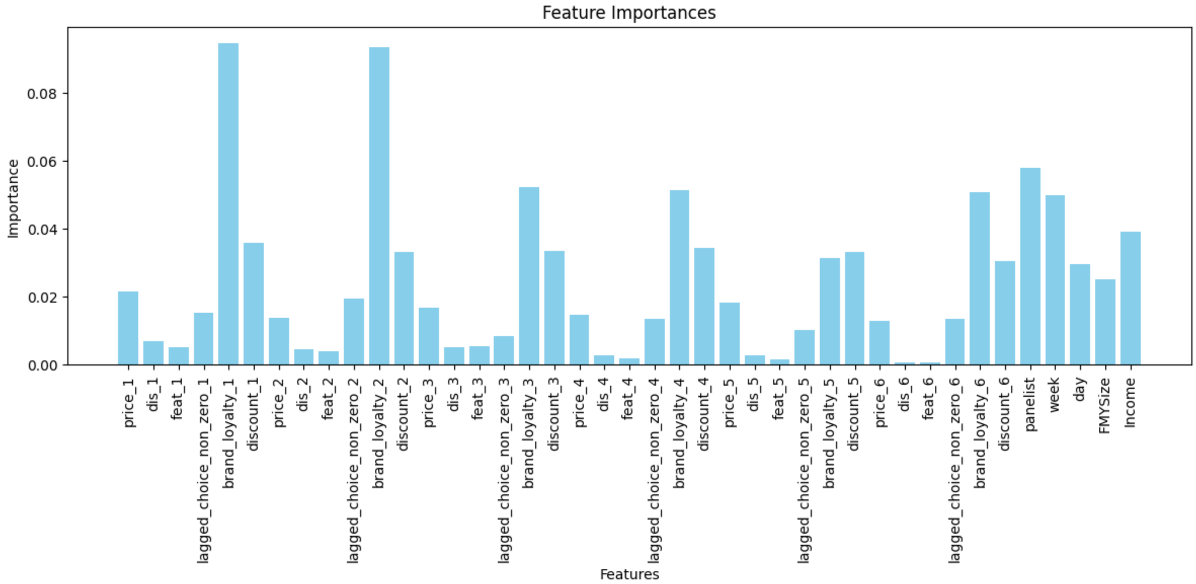


Figure 3: Random forest Feature importance

From Figure 3, we observe that income, brand loyalty, discount, last purchased brand, and price have the highest importance scores. We will now explore how these features relate to the outcome variable.

Figure 4 reveals that as consumer income increases, the likelihood of choosing brands 5 and 4 also rises. Brands 1 and 2 are more frequently preferred by middle and upper-middle-class consumers, while brands 3 and 6 are more popular among lower-class consumers.

Figure 5 shows that higher discounts are positively correlated with the choice of every brand except for brand 6. Notably, for brands 1 and 5, this relationship takes a U-shape.

As expected, there is a negative relationship between price and preference for brands 1, 2, and 5, as displayed in Figure 6. Interestingly, brands 3 and 4 exhibit a U-shaped relationship between price and choice, while brand 6 shows a positive correlation with price.

Naturally, higher brand loyalty and this brand being the previous purchase, positively impact consumers' preferences towards all brands (see Figure 11 and Figure 12 in Appendix C).

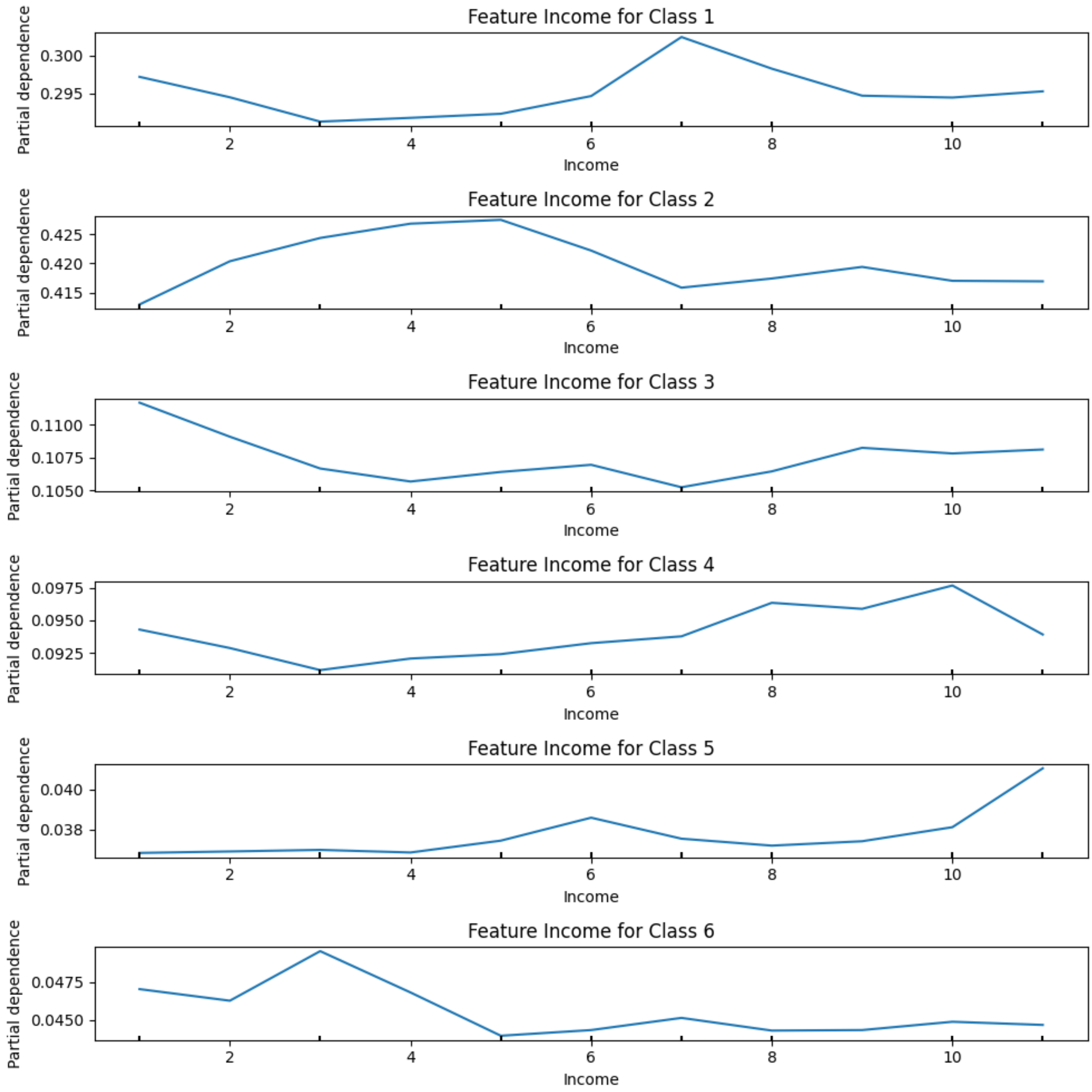


Figure 4: Income PDP

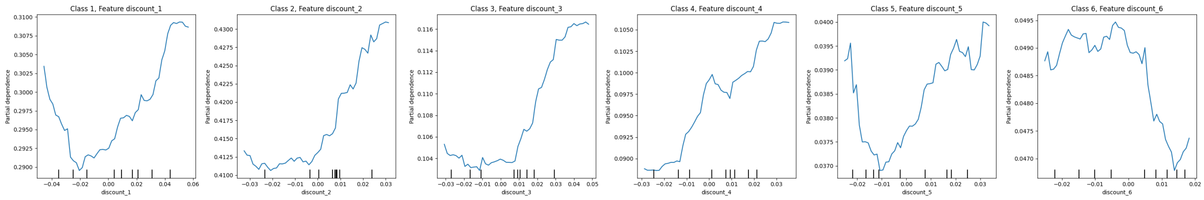


Figure 5: Discount PDP

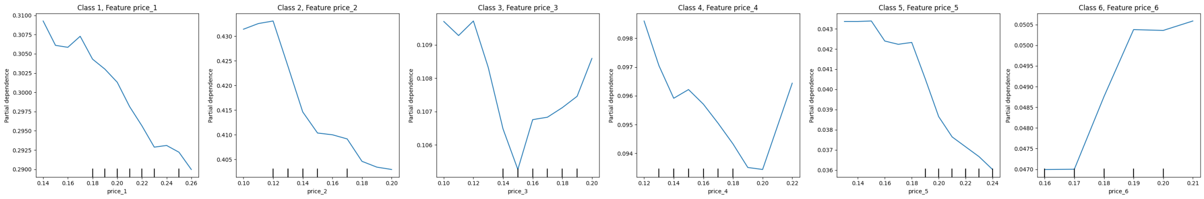


Figure 6: Price PDP

4.5 Predictive performance comparison

Having interpreted the effects, we now move to forecasting. The machine learning models—random forest, XGBoost, and AdaBoost — are evaluated using cross-validation, and their performance metrics are summarized in Table 6. The models were calibrated using isotonic regression and Venn predictors to improve the accuracy of probability estimates. The results indicate that among the evaluated ML models, the random forest combined with isotonic regression calibration is the best-performing model for predicting household cereal purchases. It achieved the highest accuracy (52.044%) and a low Brier Score (0.10294), suggesting that it accurately predicts household cereal purchases and provides reliable probability estimates. The XGBoost model with isotonic regression also demonstrated strong performance, achieving an accuracy of 50.613% and a Brier Score of 0.10360, making it an effective but slightly less reliable alternative. Interestingly, using Venn predictors generally resulted in lower performance across all models, particularly in accuracy, F1 score and Brier Score. This may be attributed to the nature of Venn predictors, which prioritize providing well-calibrated probabilities over improving raw classification accuracy.

Table 6: Model comparison with cross validation

Model	Accuracy	Brier-Score	F1-Score	ROC-AUC
General Logit	0.52520	0.61128	0.37700	-
Mixed Logit	0.53983	0.16917	0.50231	-
Random Forest and Isotonic Regressions	0.52044	0.10294	0.48278	0.77965
Random Forest and Venn Predictors	0.49263	0.10645	0.44848	0.76445
XGBoost and Isotonic Regressions	0.50613	0.10360	0.46771	0.77609
XGBoost and Venn Predictors	0.48957	0.10813	0.44943	0.75346
ADABoost and Isotonic Regressions	0.49897	0.10291	0.47532	0.78567
ADABoost and Venn Predictors	0.44192	0.11245	0.40859	0.73988

Figure 7 and Figure 8 show the confusion matrices of the general logit and mixed logit predictions, respectively. Confusion matrices for the machine learning models with both isotonic regressions and Venn predictors can be found in Appendix D. Our model is very inclined towards predicting brands 1 and 2, which can be explained by the fact that they are most prominent in our sample. The accuracy of general logit model is 52.5% on average, however the Brier score is high (0.61128), as shown in Table 6. This relatively high accuracy, coupled with a very high Brier score, indicates that while the model often predicts the correct outcomes, its probability estimation is poor. The model frequently predicts brands 1 or 2 correctly, not because it estimates probabilities accurately, but because these brands are more frequent. Consequently, the Brier score heavily penal-

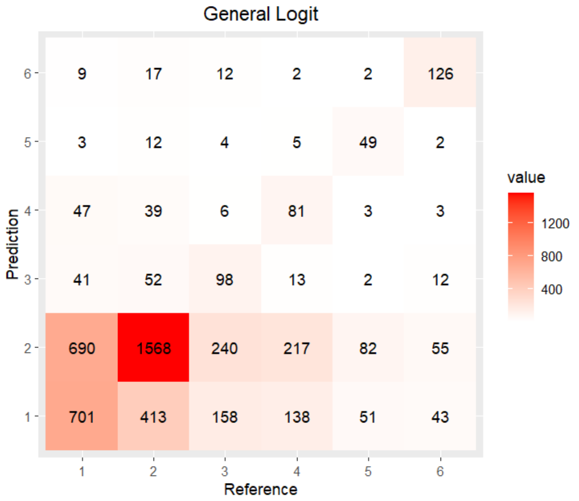


Figure 7: Confusion matrix General Logit

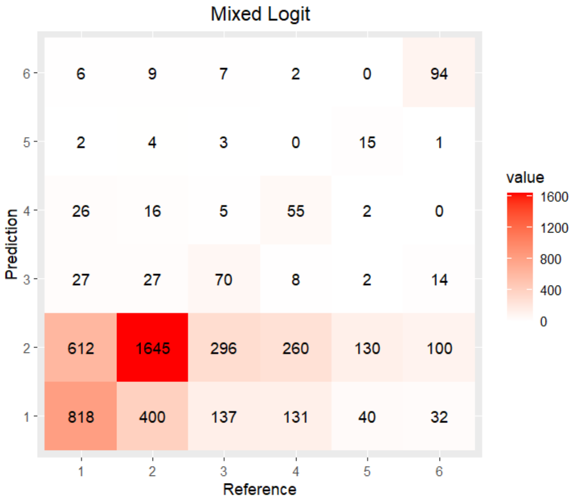


Figure 8: Confusion matrix Mixed Logit

izes these cases, highlighting that the general logit model is the worst-performing model overall.

The mixed logit model provided better predictions than the general logit model due to its flexibility in accounting for individual-specific random effects. It showed a marginally higher accuracy of 53.98% and significantly better-calibrated probability predictions with a Brier score is 0.1692. The mixed logit model also achieves a higher F1 score, indicating balanced performance in terms of precision and recall. However, as shown in Table 8, the mixed logit McFadden's R^2 is slightly lower than thatfor the General Logit (0.17292 vs. 0.17988), suggesting it captures more nuanced consumer behaviors but introduces complexity that doesn't always enhance explanatory power. Its lower AIC (11889.877 vs. 11972.0) indicates a better overall fit, balancing model complexity and goodness of fit more effectively.

Therefore, machine learning approaches, particularly Random Forest and XGBoost with isotonic regression, provided the most accurate and reliable predictions. These findings suggest that advanced machine learning techniques, when properly calibrated, can significantly outperform traditional econometric models in predicting consumer behavior.

5 Conclusion

In this report, we described and predicted purchasing behaviour of 250 households for six cereal brands over a span of two years. We employed various econometric models, including the general logit and mixed logit, as well as machine learning models such as random forests, XGBoost, and AdaBoost.

Our findings indicate that household-specific variables, like family size and income, along with marketing mix variables, significantly influence purchasing decisions. Incorporating additional variables derived from the dataset, such as discounts and brand loyalty measures, further improved the performance of our models. As expected, the presence of cereal on display or its promotion positively impacts the likelihood of its purchase.

Through rigorous 5-fold cross-validation, we found that basic econometric models like the general logit were the least effective. The mixed logit model, which better captures the complexity of consumer behaviour, showed improved accuracy. However, the most notable performance came from machine learning models, particularly when calibrated with isotonic regressions, which consistently yielded the highest accuracy.

For future research, exploring additional variables such as total spending per trip or segmenting consumers based on their spending behaviour could provide further insights. Furthermore, since our dataset is relatively unbalanced with regard to the purchased cereal, sample balancing techniques or bootstrapping could improve model performance.

References

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Clements, M. P. (2016). Evaluating the accuracy of forecasts from non-linear models. *Journal of Forecasting*, *35*(1), 1–19.

Franses, P. H., & Paap, R. (2010). *Quantitative models in marketing research.* Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, *28*(2), 337–407.

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge University Press.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Appendix

A Summary Statistics

Table 7: Summary statistics

Variable	Mean	Std. Dev.	Min	Max
price:1	0.208	0.024	0.140	0.260
price:2	0.139	0.017	0.100	0.200
price:3	0.158	0.020	0.100	0.200
price:4	0.152	0.018	0.120	0.220
price:5	0.212	0.020	0.130	0.240
price:6	0.178	0.015	0.160	0.210
dis:1	0.325	0.469	0	1
dis:2	0.209	0.406	0	1
dis:3	0.173	0.378	0	1
dis:4	0.084	0.277	0	1
dis:5	0.117	0.322	0	1
dis:6	0.030	0.170	0	1
feat:1	0.151	0.358	0	1
feat:2	0.157	0.364	0	1
feat:3	0.126	0.332	0	1
feat:4	0.046	0.210	0	1
feat:5	0.078	0.268	0	1
feat:6	0.056	0.230	0	1
lagged_choice:1	0.078	0.268	0	1
lagged_choice:2	0.102	0.303	0	1
lagged_choice:3	0.023	0.151	0	1
lagged_choice:4	0.027	0.162	0	1
lagged_choice:5	0.014	0.117	0	1
lagged_choice:6	0.016	0.124	0	1
lagged_choice_non_zero:1	0.285	0.451	0	1
lagged_choice_non_zero:2	0.400	0.490	0	1
lagged_choice_non_zero:3	0.098	0.297	0	1
lagged_choice_non_zero:4	0.086	0.280	0	1
lagged_choice_non_zero:5	0.037	0.188	0	1
lagged_choice_non_zero:6	0.045	0.207	0	1
brand_loyalty:1	0.294	0.287	0	1
brand_loyalty:2	0.405	0.314	0	1
brand_loyalty:3	0.086	0.155	0	1
brand_loyalty:4	0.081	0.157	0	1
brand_loyalty:5	0.036	0.115	0	1
brand_loyalty:6	0.048	0.166	0	1
discount:1	0.006	0.032	-0.073	0.106
discount:2	0.003	0.019	-0.080	0.044

Continued on next page

Table 7 – continued from previous page

Variable	Mean	Std. Dev.	Min	Max
discount:3	0.004	0.023	-0.060	0.089
discount:4	0.002	0.020	-0.104	0.073
discount:5	0.001	0.022	-0.048	0.108
discount:6	0.000	0.015	-0.046	0.036

B Odds ratios

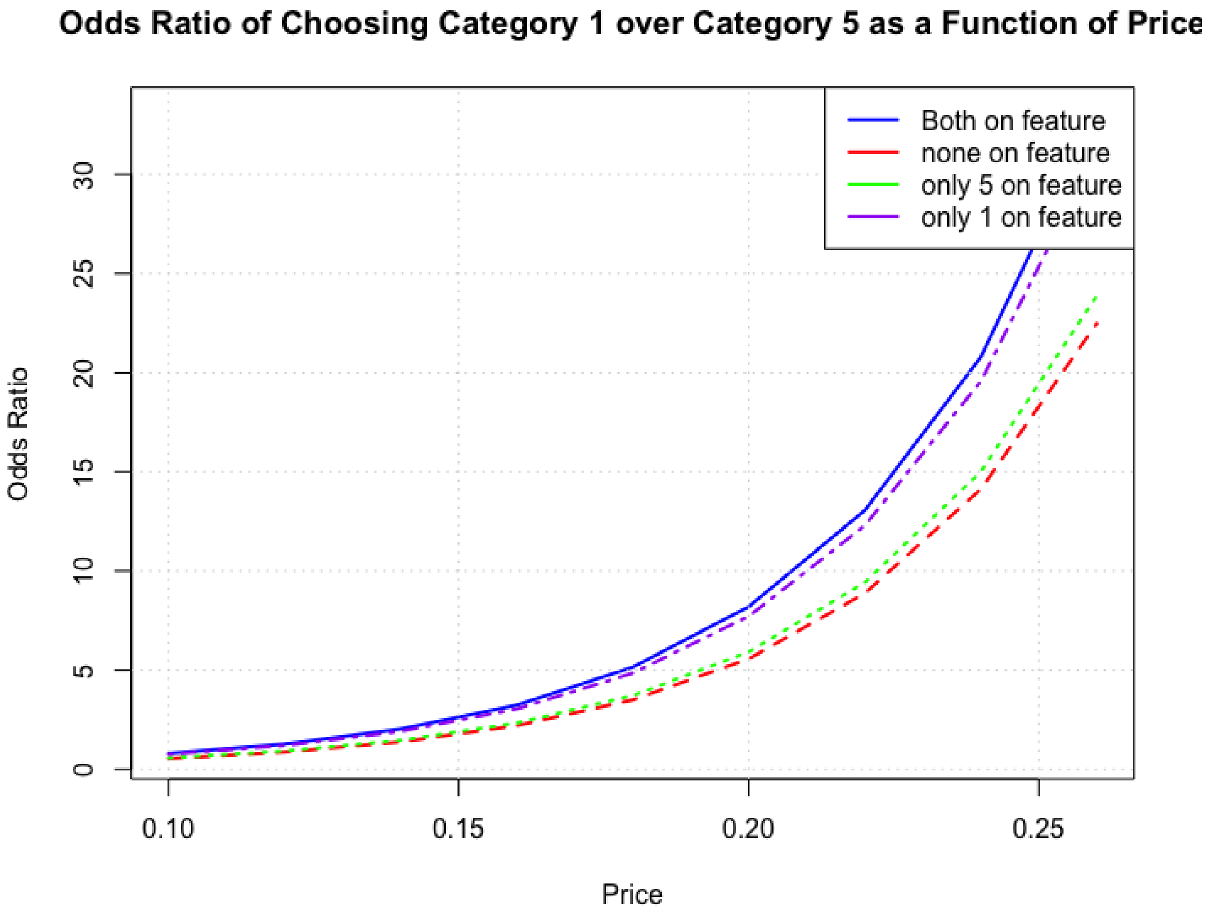


Figure 9: General Logit

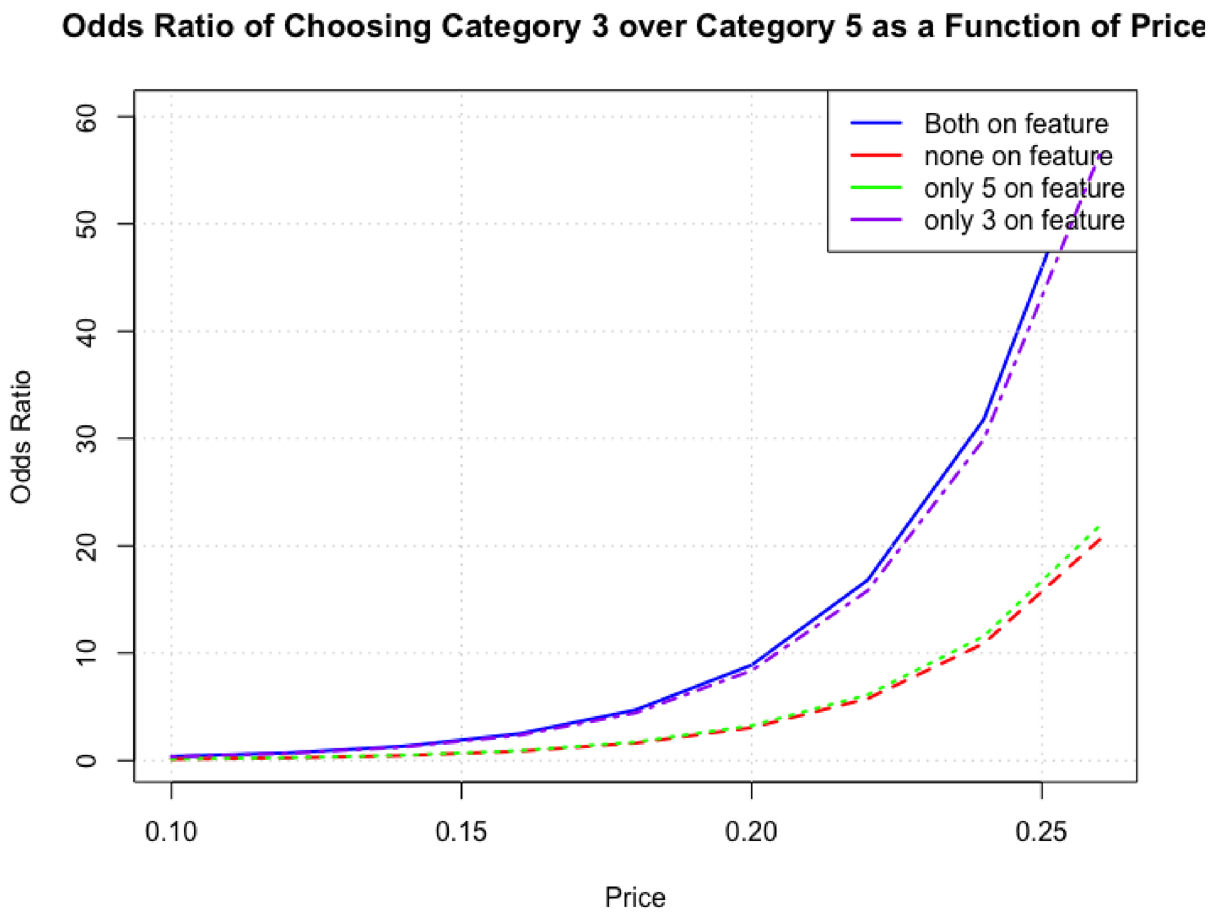


Figure 10: General Logit

C PDP plots

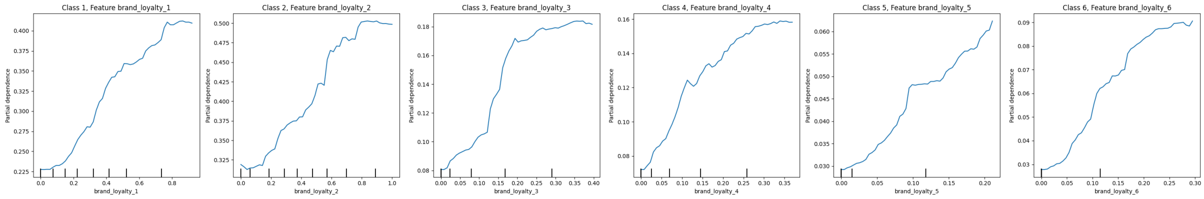


Figure 11: Brand Loyalty PDP

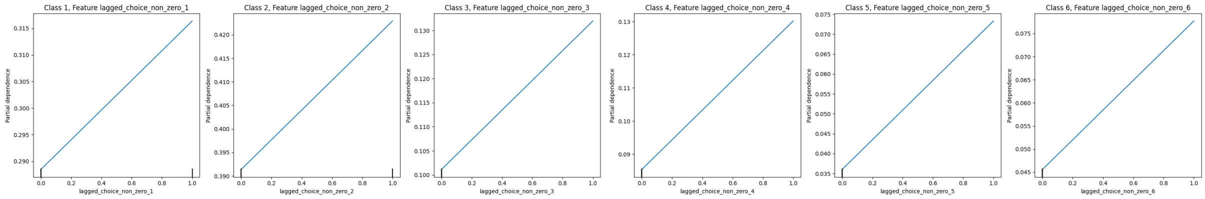
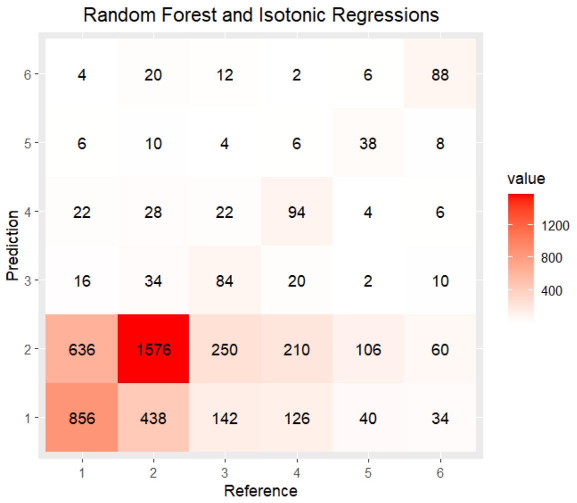


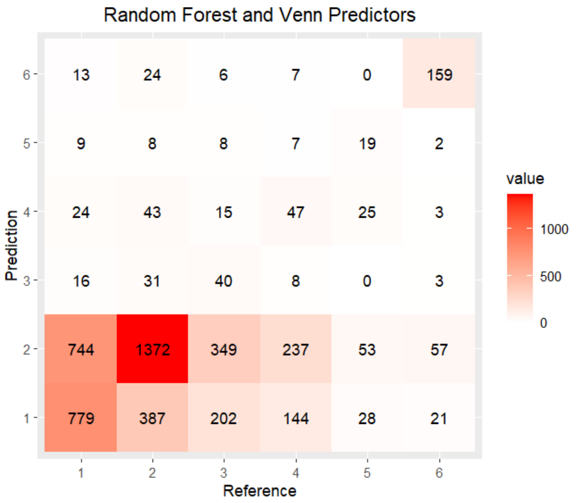
Figure 12: Brand choice lag PDP

D Confusion matrices of machine learning models


Random Forest and Isotonic Regressions

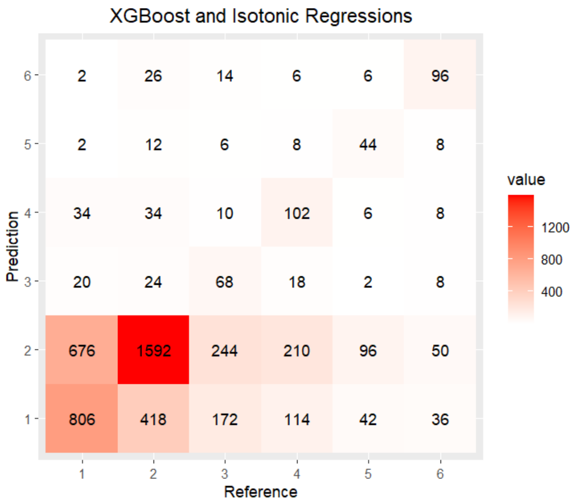
Prediction \ Reference	1	2	3	4	5	6
6	4	20	12	2	6	88
5	6	10	4	6	38	8
4	22	28	22	94	4	6
3	16	34	84	20	2	10
2	636	1576	250	210	106	60
1	856	438	142	126	40	34

Figure 13: Confusion matrix Random Forest and Isotonic Regressions


Random Forest and Venn Predictors

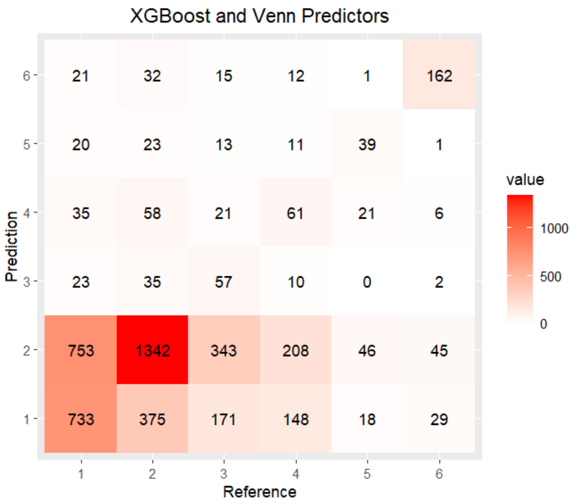
Prediction \ Reference	1	2	3	4	5	6
6	13	24	6	7	0	159
5	9	8	8	7	19	2
4	24	43	15	47	25	3
3	16	31	40	8	0	3
2	744	1372	349	237	53	57
1	779	387	202	144	28	21

Figure 14: Confusion matrix Random Forest and Venn Predictors


XGBoost and Isotonic Regressions

Prediction \ Reference	1	2	3	4	5	6
6	2	26	14	6	6	96
5	2	12	6	8	44	8
4	34	34	10	102	6	8
3	20	24	68	18	2	8
2	676	1592	244	210	96	50
1	806	418	172	114	42	36

Figure 15: Confusion matrix XGBoost and Isotonic Regressions


XGBoost and Venn Predictors

Prediction \ Reference	1	2	3	4	5	6
6	21	32	15	12	1	162
5	20	23	13	11	39	1
4	35	58	21	61	21	6
3	23	35	57	10	0	2
2	753	1342	343	208	46	45
1	733	375	171	148	18	29

Figure 16: Confusion matrix XGBoost and Venn Predictors

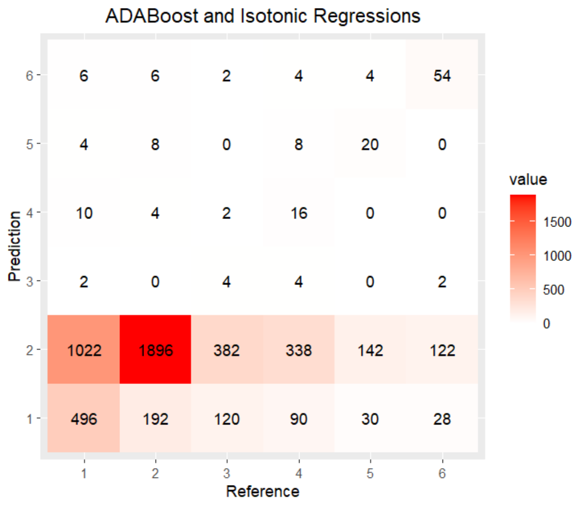


Figure 17: Confusion matrix ADABoost and Isotonic Regressions

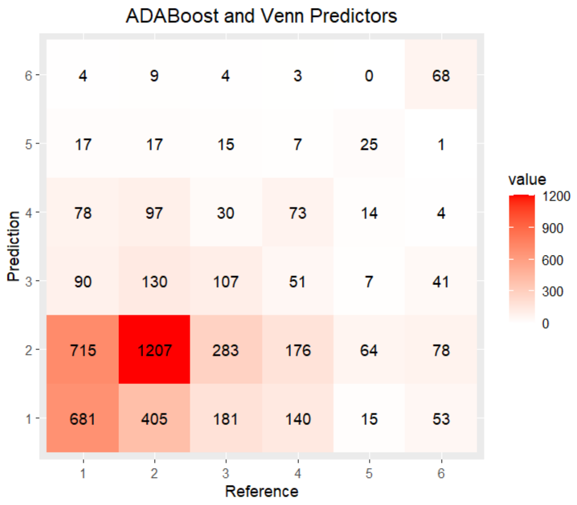


Figure 18: Confusion matrix ADABoost and Venn Predictors