Increasing Brand Penetration by Acquiring New Customers: A Prediction of New Product Purchases in FMCG

Author: Gabriele Stoeckl
Mentor: Cagdas Yetkin
Coding Nomads Data Science and Machine Learning
July 07, 2023

Abstract
According to the Ehrenberg Bass marketing philosophy, attracting new customers is crucial for brands' sustained business success. In the highly competitive FMCG market often shopper marketing is leveraged to this end. The objective of this project was to develop a model that predicts product switching within the yoghurt category based on transaction data from 800 households collected over a period of two years for a specific retailer as well as the corresponding shopper marketing data. Primarily, Random Forest, XGBoost and LightGBM classification models were used. Best-performing models were fed into a StackingClassifier which provided the basis for explaining feature importance via SHAP values. The resulting model's predictive power is satisfactory, especially given the limited nature of available features for this project. Pricing and household demographics were identified as key levers for triggering new product purchases. Further research building on this work ought to aim at including features covering a broader set of marketing strategies to further increase model performance.

Table of Contents

Introduction and Domain Background

The market for fast-moving consumer goods (FMCG) is highly competitive and many products are fighting for market shares. Purchase cycles in FMCG categories are usually short, from a few days to a few months. Moreover, most FMCG categories are low-involvement categories: They do not have a high emotional meaning for consumers, and consumers are not likely to put a considerable amount of research and effort into the selection process.

The key insights from Ehrenberg Bass' marketing philosophy, therefore, are especially relevant for the FMCG industry. Ehrenberg Bass lead to a fundamental paradigm shift in marketing research and thinking in recent years, based on their extensive empirical studies that challenged prevalent marketing assumptions. They emphasize the importance of targeting a broad consumer base, rather than focusing on loyal customers. Most importantly for our purposes, the underlying research revealed that a large portion of a brand's sales come from a large number of infrequent buyers, rather than a small group of highly loyal customers.

This means that marketeers ought to focus on

- Increasing brand penetration, i.e., the number of households buying a brand within a given time period, as a crucial driver of long-term growth. Expanding the customer base by acquiring new buyers is essential for sustainable business success.
- Consistent and continuous marketing efforts to reach a broad customer base. The notion of exclusively targeting loyal customers is challenged in favor of expanding brand visibility to attract non-customers.

So how do FMCG brands tackle this holy grail of acquiring new customers? Increasing visibility and reaching consumers as close to their purchase moment as possible is a promising strategy. A number of dedicated shopper marketing activities such as additional placements, price discounts, or direct-to-consumer marketing is implemented in cooperation with retail partners.

If it was possible to successfully predict new purchases based on shopper marketing activities, brands could benefit threefold:

- Improved marketing ROI: By targeting potential purchasers more accurately, marketing campaigns can yield higher returns on investment.
- Enhanced customer acquisition: The predictive model would help identify new customers who are more likely to make new (infrequent) purchases, thereby expanding the customer base in the most efficient way.
- Data-driven decision-making: The insights derived from such a predictive model would enable data-driven marketing strategies, resulting in more effective and efficient resource allocation.

Project Objective and Scope

The primary objective of this project was to develop a model that predicts new product purchases based on Machine Learning classification models. As a category, we chose yoghurt as a typical example of a fast-moving, low-involvement FMCG product category with a large enough number of transactions to yield sufficient data for the modeling.

In line with Ehrenberg Bass's paradigm of increasing brand penetration, because a large proportion of a brand's sales comes from a large number of infrequent buyers, we defined new purchases as either genuine first (trial) purchases or purchases from infrequent purchasers who have not bought the product in the recent past.

The modeling was based on a dataset comprising actual transaction data from a two-year period, encompassing approximately 800 households, and incorporating data on shopper marketing. Shopper marketing activities include direct price discounts, discounts based on retailer loyalty programs, increasing in-store visibility due to additional placements, and increasing awareness via mailers and other coupon campaigns. Additionally, household demographics were taken into account so as to understand how to target the most promising demographics.

A secondary objective consisted in understanding which shopper marketing activities drive new product purchases most. These insights help increase the business impact in that marketeers can derive actionable insights how to optimize their shopper marketing mix for acquiring new buyers.

It is important to bear in mind that due to the limitations of the data available for this project, several potentially relevant factors driving new product purchases could not be included in the modeling. Most notably, the data does not contain any information on above-the-line marketing which without any doubt does has a significant impact on purchase decisions. Also, though covering the full transaction history within one retailer, the data does not capture potential yoghurt purchases of the observed households at other retailers during the respective period.

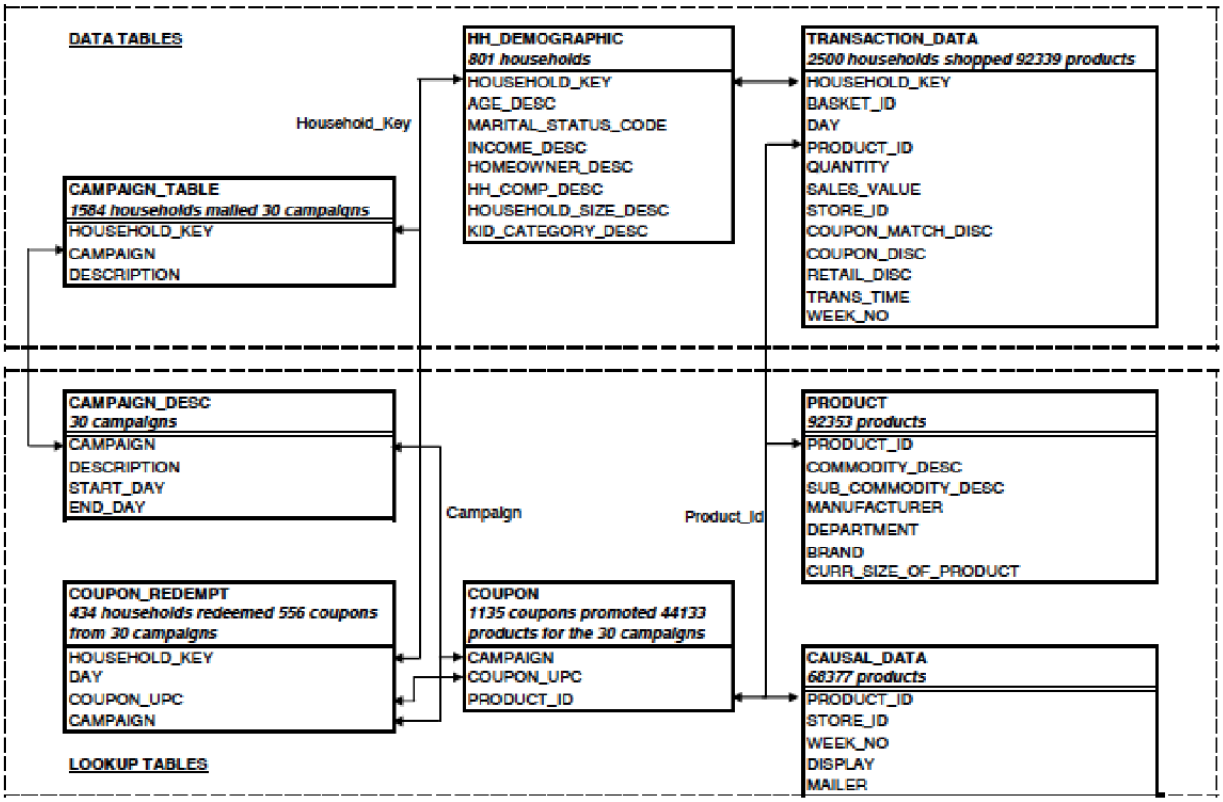
The resulting model, therefore, inevitably will not be able to cover all levers for increasing brand penetration and its predictive power will have natural limits.

The Dataset

Data from „The Complete Journey“ by dunnhumby (https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey) was used. It contains household-level transactions over two years from a group of 2,500 households who are frequent shoppers at one specific retailer. All of each household's purchases are contained across all product categories available at the retailer's stores.

All data is anonymized, including some of the shopper marketing variables. As a caveat for business practitioners, for some shopper marketing activities, we can determine their role in predicting new purchases, but we cannot tell how the activity was executed in detail.

The structure of the dataset:



Methodological Approach and Implementation

For the primary objective of predicting new purchases, tree-based classification algorithms were used. The predicted variable is binary: purchase of a new product vs purchase of a product that has recently been purchased already.

In the absence of previous modeling endeavors on the data for this objective, a dummy benchmark model was used. It was based on the actual distribution of new vs. repeat purchases found in the dataset.

When it comes to model evaluation metrics, we have to consider two fallacies which both have to be avoided:

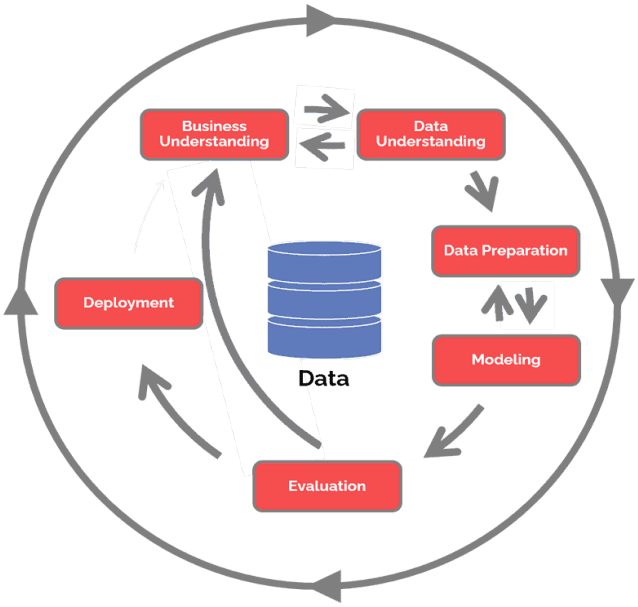
- False Positives: If a brand is spending money on shopper marketing, they want to make sure that it works. False Positives would mean money is not spent wisely and should be minimized.
- False Negatives: On the other hand, False Negatives indicate a risk that the marketing has effects but they are not correctly attributed to the marketing efforts. They come with a risk of stopping spending on effective marketing and, hence, should be minimized as well.

Consequently, the f1 score was chosen as the core evaluation metric, providing a balance

between minimizing False Positives and False Negatives or, in other words, between minimizing precision and recall. Additionally, also precision and recall were tracked in order to monitor each of these two fallacies individually.

For the secondary objective, measuring which variables drive new purchases most, SHAP (SHapley Additive exPlanations) was used. It explains the impact of individual features on the predictions made by a machine learning model. SHAP values provide a quantitative measure of the contribution of each feature towards the model's output, thereby enabling feature ranking and interpretability for business execution planning.

Overall, the project followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) model as a guiding framework. CRISP-DM is a widely used methodology for data science projects, providing a structured and iterative process and allowing for revisiting and refining phases based on insights gained:

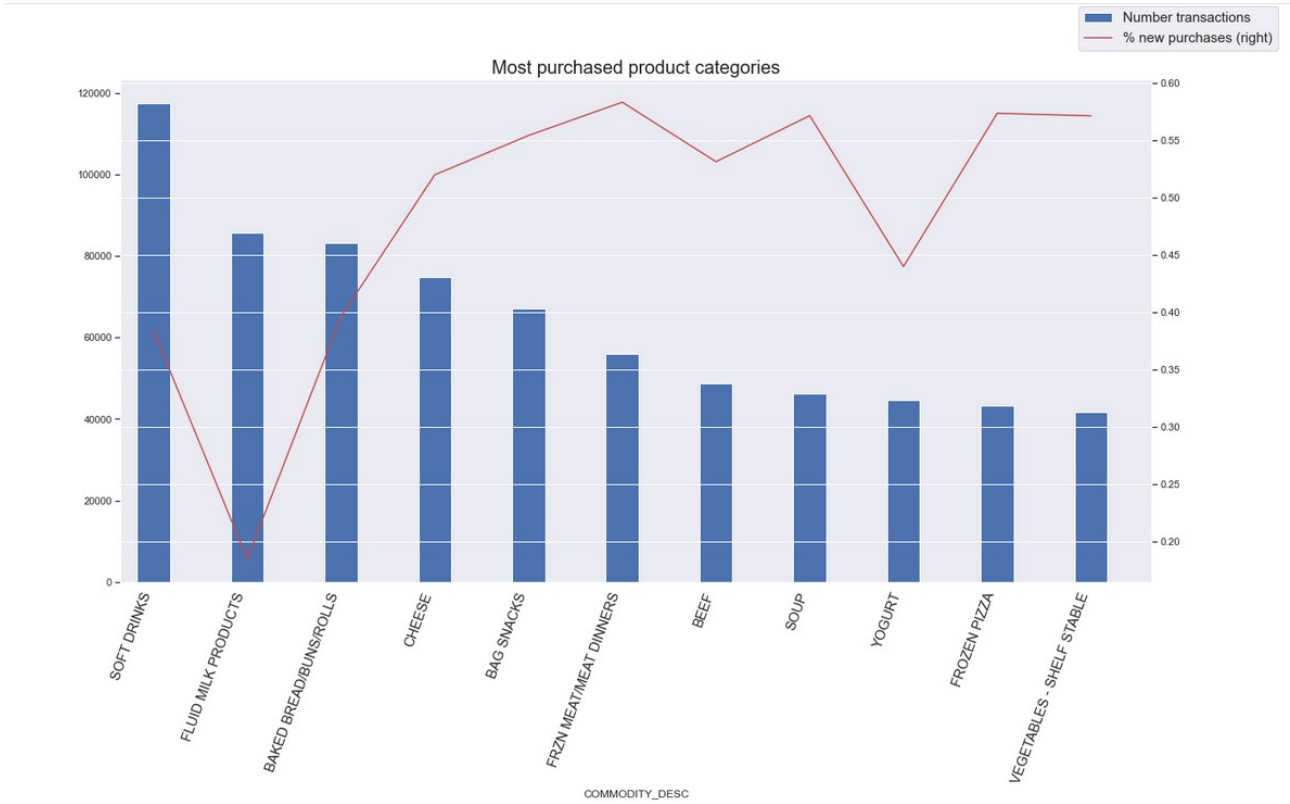


Anaconda Jupyter Lab was used as the local development environment. All steps and contents are accessible in the corresponding GitHub repository.

Selection of Product Category and Features

The original dataset is massive. For example, the largest table in the dataset, TRANSACTION_DATA, contains more than 12 million rows. Using the entire dataset would have gone way beyond the scope of this project – and the available computational power.

An obvious solution was selecting only one product category and focusing the project on this category. The following graph provides an overview of the overall number of available transactions/samples and the % of new purchases for the largest product category in the initial transactions dataset:



Eligible product categories for analysis and modeling should to meet the following criteria:

- a sufficiently high purchase frequency, or in other words sufficient number of transactions/samples
- a good distribution of the target variable
- from a domain perspective, be neither too price- and private label-driven nor characterized by too high consumer involvement

The yoghurt category meets all of these three criteria and hence was chosen for this project.

The selection of features was guided by the following rationale:

- **Price and discounts** are particularly relevant in the context of purchasing a new product, as they mitigate risk, influence value assessment, act as incentives, and differentiate the product within the market.
- **Direct-to-shopper marketing** plays a significant role in new product purchases by enhancing product awareness and visibility and providing customized trial offers/coupons. These marketing approaches effectively engage potential buyers and make purchasing a new product more likely.
- **Household demographics:** Household demographics influence purchases of new products by determining affordability. They also guide purchase priorities and can be viewed as a proxy for different shopper mindsets (e.g. different age segments may differ in novelty seeking, i.e. their openness towards trying out something new).

All available features referring to these three topics were extracted from the dataset.

Data Preparation

In order to build the dataset for the modeling, relevant variables from the different tables in the dataset were joined and transformed. Samples for yoghurt purchases were extracted (TRANSACTIONAL_DATA, PRODUCT). Variables relating to corresponding shopper marketing activities (CAMPAIGN_DESC, CAMPAIGN_TABLE, CAUSAL_DATA, COUPON) as well as household demographics (HH_DEMOGRAPHIC) were joined.

The target variable „first_purchase“ was extracted for each household-by-product-combination.

Features were engineered for shelf price, paid price, and discount levels.

One-hot-encoding was applied for the following variables:

- Marketing activities: display options, mailer options, campaign, campaign description (i.e. type of campaign)
- Household demographics: marital status, homeownership, household composition, kids category, age of household head, income, household size

The dataset was cleaned of unnecessary variables. Duplicates as well as samples with negative and infinite prices, both resulting from data inconsistencies, were removed. Samples without household demographic information were dropped as well.

The first 12 weeks were removed from the dataset. They contained systematically lower overall sales than the average, indicating a measurement issue at the beginning of the observation period.

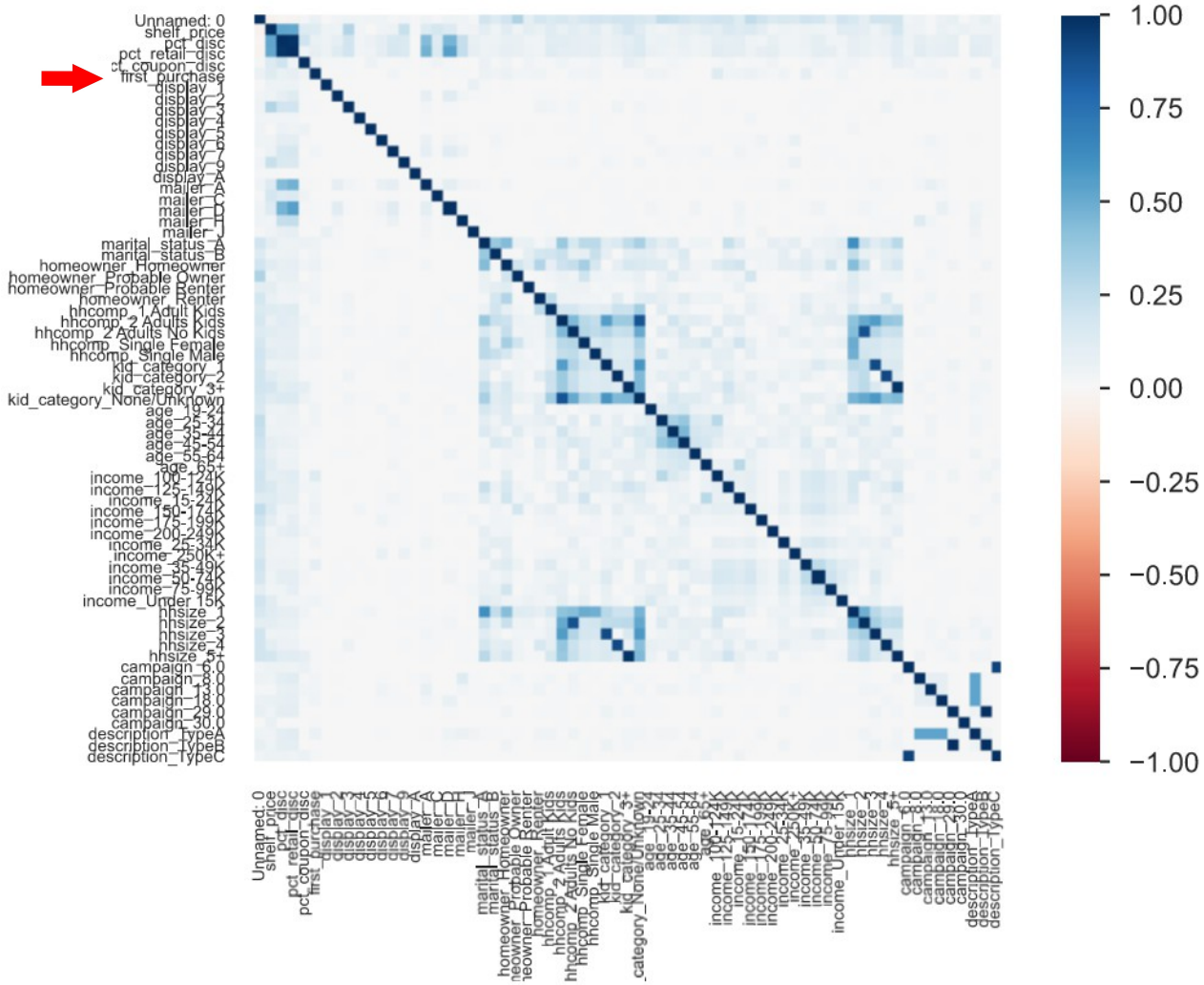
Additionally, weeks up to week 30 were removed. They represent two average purchase cycles for the yoghurt category on top of removing the first 12 weeks. This is to ensure we captured genuine new purchases rather than them simply being the first purchases of the respective products occurring during the observation period.

The dataset ready for modeling contains 21,251 samples and 59 features (see table on next page). The final data does not contain any missing values.

Data preparation was implemented via a module with functions that can be easily applied to create different datasets for different product categories in order to effectively support potential future modeling endeavors for further categories.

Name	Type	Topic	Description
shelf_price	Float	Price	Regular shelf price
pct_disc	Float	Price	Total discount vs regular shelf price in %
pct_retail_disc	Float	Price	Retailer loyalty card discount vs regular shelf price in %
pct_coupon_disc	Float	Price	Manufacturer coupon discount vs regular shelf price in %
display_2	Dummy	In-store display	Store rear
display_3	Dummy	In-store display	Front end cap
display_4	Dummy	In-store display	Mid-aisle end cap
display_6	Dummy	In-store display	Side-aisle end cap
display_7	Dummy	In-store display	In-aisle
display_9	Dummy	In-store display	Secondary location display
mailer_A	Dummy	Email marketing	Interior page feature
mailer_C	Dummy	Email marketing	Interior page line item
mailer_D	Dummy	Email marketing	Front page feature
mailer_H	Dummy	Email marketing	Wrap front feature
mailer_J	Dummy	Email marketing	Interior page coupon
marital_status_A	Dummy	Household demographic	Married
marital_status_B	Dummy	Household demographic	Single
homeowner_Homeowner	Dummy	Household demographic	Home owner
homeowner_Probable Owner	Dummy	Household demographic	Probable home owner
homeowner_Probable Renter	Dummy	Household demographic	Probable renter
homeowner_Renter	Dummy	Household demographic	Renter
hhcomp_1 Adult Kids	Dummy	Household demographic	One adult with kids
hhcomp_2 Adults Kids	Dummy	Household demographic	Two adults with kids
hhcomp_2 Adults No Kids	Dummy	Household demographic	Two adults, no kids
hhcomp_Single Female	Dummy	Household demographic	Single female
hhcomp_Single Male	Dummy	Household demographic	Single male
kid_category_1	Dummy	Household demographic	One child
kid_category_2	Dummy	Household demographic	Two children
kid_category_3+	Dummy	Household demographic	Three or more children
kid_category_None/Unknown	Dummy	Household demographic	No children (known)
age_19-24	Dummy	Household demographic	Estimated age range 19-24 years
age_25-34	Dummy	Household demographic	Estimated age range 25-34 years
age_35-44	Dummy	Household demographic	Estimated age range 35-44 years
age_45-54	Dummy	Household demographic	Estimated age range 45-54 years
age_55-64	Dummy	Household demographic	Estimated age range 55-64 years
age_65+	Dummy	Household demographic	Estimated age range 65+ years
income_100-124K	Dummy	Household demographic	Household income 100-124k
income_125-149K	Dummy	Household demographic	Household income 125-149k
income_15-24K	Dummy	Household demographic	Household income 15-24k
income_150-174K	Dummy	Household demographic	Household income 150-174k
income_175-199K	Dummy	Household demographic	Household income 175-199k
income_200-249K	Dummy	Household demographic	Household income 200-249k
income_25-34K	Dummy	Household demographic	Household income 25-34k
income_250K+	Dummy	Household demographic	Household income 250k+
income_35-49K	Dummy	Household demographic	Household income 35-49k
income_50-74K	Dummy	Household demographic	Household income 50-74k
income_75-99K	Dummy	Household demographic	Household income 75-99k
income_Under 15K	Dummy	Household demographic	Household income under 15k
hhsize_1	Dummy	Household demographic	One person
hhsize_2	Dummy	Household demographic	Two persons
hhsize_3	Dummy	Household demographic	Three persons
hhsize_4	Dummy	Household demographic	Four persons
hhsize_5+	Dummy	Household demographic	Five or more persons
campaign_8.0	Dummy	Coupon campaign	Household received coupon for product campaign no.8
campaign_13.0	Dummy	Coupon campaign	Household received coupon for product campaign no.13
campaign_18.0	Dummy	Coupon campaign	Household received coupon for product campaign no.18
campaign_29.0	Dummy	Coupon campaign	Household received coupon for product campaign no.29
description_TypeA	Dummy	Coupon campaign	Campaign type A
description_TypeB	Dummy	Coupon campaign	Campaign type B

Inspecting correlations during exploratory data analysis indicated no strong bi-variate correlations of individual features with the target variable first_purchase:



Nevertheless, there is a number of features where we found a small difference in means afer splitting them by the target variable.
This indicates at least weak directional relationships at a bi-variate level in the data.

(only displayed if difference >= 0.02)

	first_purchase		
	False	True	DIFF
shelf_price	1,071	1,325	0,254
marital_status_A	0,507	0,453	-0,054
marital_status_B	0,116	0,143	0,027
homeowner_Homeowner	0,705	0,650	-0,054
hhcomp_2 Adults No Kids	0,324	0,294	-0,031
hhcomp_Single Female	0,141	0,161	0,020
hhcomp_Single Male	0,131	0,106	-0,025
age_25-34	0,222	0,193	-0,029
income_100-124K	0,104	0,034	-0,071
income_15-24K	0,047	0,069	0,021
income_150-174K	0,098	0,061	-0,037
income_35-49K	0,157	0,226	0,069
hhsize_2	0,381	0,361	-0,020
description_TypeA	0,249	0,193	-0,055

In summary, bi-variate relationships with the target variable are relatively low. The modeling was going to show whether the available features provide a sufficient basis for a prediction of the purchase of new products.

Modeling

In this phase, various modeling techniques were applied to the prepared dataset. This involved selecting and applying appropriate algorithms or models based on the project's objectives. The models were trained, evaluated, and fine-tuned to achieve the best possible results:

- **DummyClassifier (benchmark model)**
- **DecisionTreeClassifier**
- **RandomForestClassifier** („hist“ was chosen as a tree method in order to keep computing time within reasonable limits)
- **XGBoostClassifier**
- **LightGBMClassifier**

The data has a certain structure based on households, products, and moments in time during the observation period. In order to avoid the risk of systematic effects, StratifiedShuffleSplit was selected for cross-validation instead of StratifiedKFold.

RandomSearch and GridSearch as well as careful inspection of cross-validation curves for individual hyperparameters were conducted for hyperparameter tuning.

Since the target variable is unbalanced, weighting the target variable proved to be an effective tuning step in order to increase the primary metric f1. This improvement is driven by higher recall, however, goes hand in hand with a decrease in prediction.

As a concluding modeling step, the best-performing individual models were combined to train a **StackingClassifier.**

Stacking benefits from diverse base models in terms of their underlying algorithms, architectures, and learning approaches. Therefore, three different algorithms were chosen. Moreover, the LightGBM unweighted model does not fall far behind the corresponding LightGBM weighted model when it comes to their f1 score. The trade-off between False Positives and False Negatives, or in other words, precision and recall, is induced by weighting. Therefore, it was decided to feed in the unweighted LightGBM model into the StackingClassifier in the spirit of diversity of base models improving the overall performance of the ensemble model.

Again, GridSearch and careful inspection of cross-validation curves both for the StackingClassifier parameters and the final estimator were conducted. The tuning was carried out with a fixed random state to facilitate comparability. The validation against the test set was done without a random state.

Last but not least, the resulting StackingClassifier model was explained with **SHAP values** in order to quantify the impact of the different features on the target variable.

Results

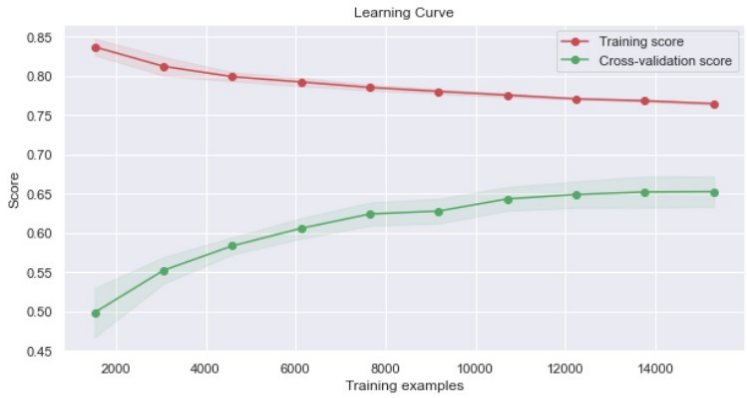
1. Individual models

Overview of evaluation metrics for best-performing individual models on test and training data:

	Dummy / Benchmark	**Decision Tree**	**Random Forest**	**XGBoost**	**LightGBM**	
Model name		*fav model balanced*	*rf_3*	*Xgb_2*	*lgbm_5*	*lgbm_unbalanced*
Weighting		*yes*	*yes*	*yes*	*yes*	*no*
Test						
f1	**0,3342**	**0,6420**	**0,6660**	**0,6609**	**0,6511**	**0,6441**
precision	**0,3364**	**0,5995**	**0,6224**	**0,6105**	**0,6100**	**0,7175**
recall	**0,3319**	**0,6911**	**0,7162**	**0,7204**	**0,6980**	**0,5844**
Train						
f1		0,7670	0,7645	0,7692	0,7588	0,7618
precision		0,7181	0,7211	0,7093	0,7030	0,8444
recall		0,8232	0,8134	0,8403	0,8242	0,6939

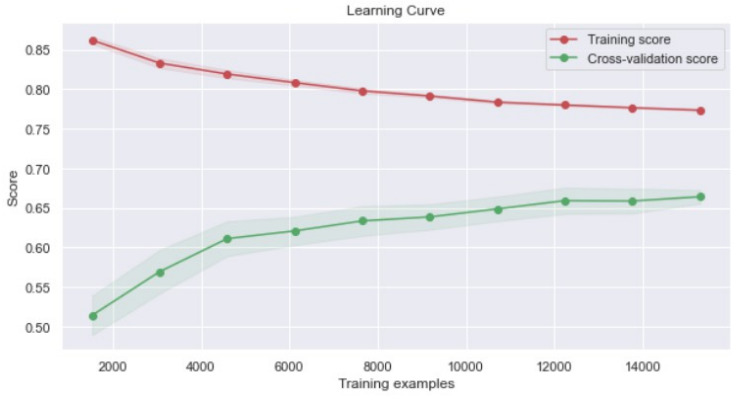
Learning curves for three models that were proceeded to the StackingClassifier:

- rf_3 (RandomForestClassifier)



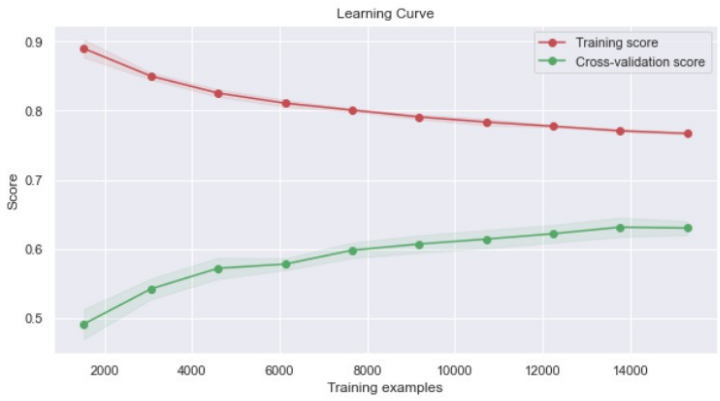

```
rf_3 = RandomForestClassifier(n_jobs = -1, class_weight = "balanced",
                              max_depth = 20, max_samples = 0.7, min_samples_split = 6, n_estimators = 250)
```

- xgb_2 (XGBoostClassifier)




```
xgb_2 = xgb.XGBClassifier(scale_pos_weight = 1.965, colsample_bytree = 0.6, colsample_bynode = 0.8, colsample_bylevel = 0.5,
                          max_delta_step = 1, tree_method = "hist", objective = "binary:logistic",
                          alpha = 1, reg_lambda = 2, learning_rate = 0.2, max_depth = 14, min_child_weight = 3,
                          n_estimators = 300, subsample = 0.8)
```

- lgbm_unbalanced (LightGBMClassifier)




```
lgbm_unbalanced =  lgb.LGBMClassifier(objective = "binary", max_depth = 9,
                                      boosting_type = 'gbdt', colsample_bytree = 0.8, learning_rate= 0.5,  max_bin= 40,
                                      min_child_samples= 50,  n_estimators= 400, num_leaves= 70,  reg_alpha= 0)
```

All models significantly outperform the benchmark Dummy Classifier. Still, the level the evaluation metrics reached in our models is only satisfactory and the predictive power of the models is clearly limited. Moreover, gaps between test and train scores indicate overfitting.

More data might help, however, the learning curves show a tendency of becoming flatter with more samples. This indicates that we would rather need more features than more data in order to improve model performance. In view of the limited range of features available in our dataset, this intuitively makes sense.

One could even argue the models predict the purchase of new products surprisingly well, given that the available dataset misses a number of crucial factors:

- Information on marketing activities other than the direct shopper marketing efforts connected to our specific retailer
- Information about the observed households' purchases at other retailers. This is likely to act as a source of noise for our dataset.

2. StackingClassifier ensemble model

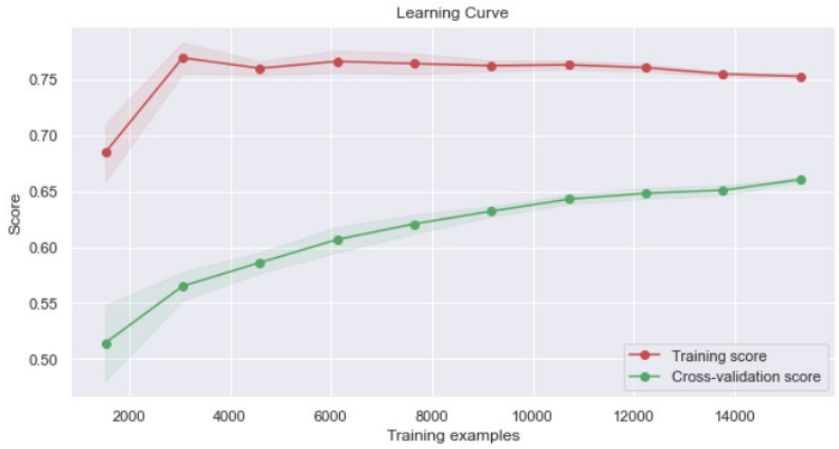
Evaluation metrics for StackingClassifier versus base models on test and training data:

	Dummy / Benchmark	Random Forest	XGBoost	LightGBM	Stacking Classifier
Model name		*rf_3*	*Xgb_2*	*lgbm_unbalanced*	*stack_21_no_rand*
Weighting		*yes*	*yes*	*no*	*yes*
Test					
f1	**0,3342**	**0,6660**	**0,6609**	**0,6441**	**0,6741**
precision	**0,3364**	**0,6224**	**0,6105**	**0,7175**	**0,6088**
recall	**0,3319**	**0,7162**	**0,7204**	**0,5844**	**0,7552**
Train					
f1		0,7645	0,7692	0,7618	0,7505
precision		0,7211	0,7093	0,8444	0,6821
recall		0,8134	0,8403	0,6939	0,8343

The StackingClassifier represents a further small improvement in evaluation metrics as compared to the base models.

Notably, the level of overfitting could be reduced and the learning curve trajectory of the ensemble model looks more reassuring.

Learning curve for StackingClassifier:

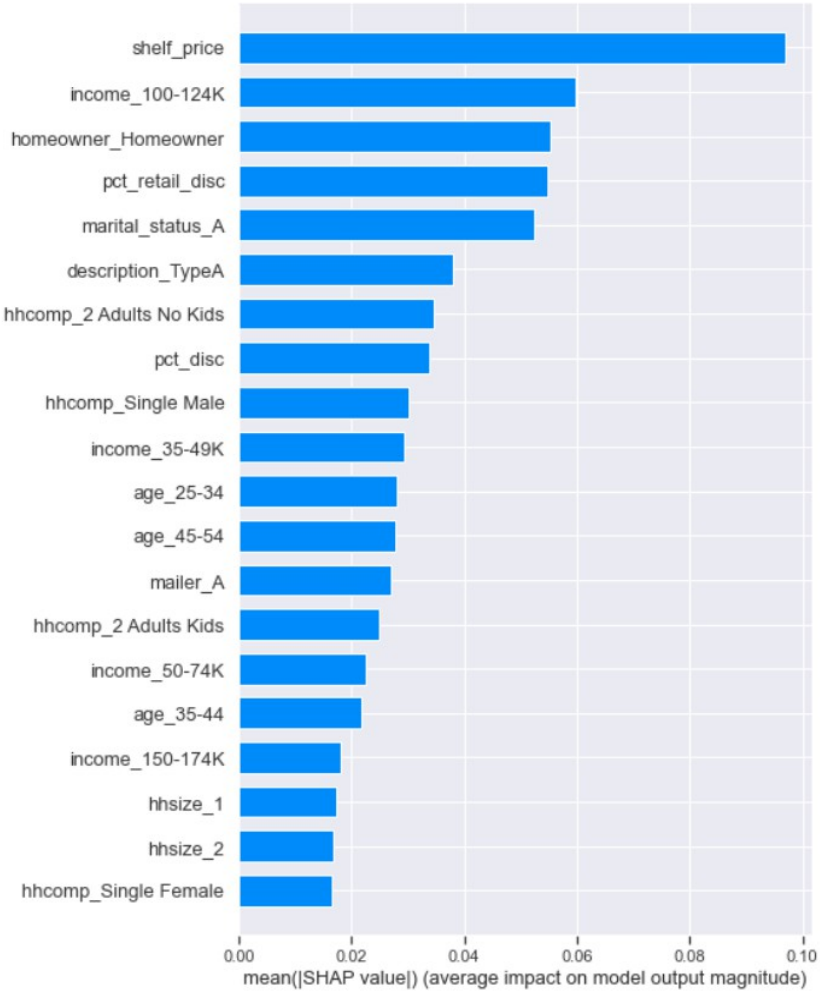



```
stack_21_no_rand = StackingClassifier(estimators = estimators_no_rand, n_jobs = -1, cv=5,
                        final_estimator= lgb.LGBMClassifier(scale_pos_weight = 1.965, boosting_type = "goss", learning_rate = 0.1,
                                                            max_bin = 45, max_depth = 5, min_child_samples = 10,
                                                            n_estimators = 100, reg_alpha = 5),
                        passthrough = True, stack_method = "predict")
```

3. Feature importance

The SHAP values were computed based on 1,000 samples from the test set with KernelExplainer which was required for the StackingClassifier.

The bar chart summarizes the global effect of the features.

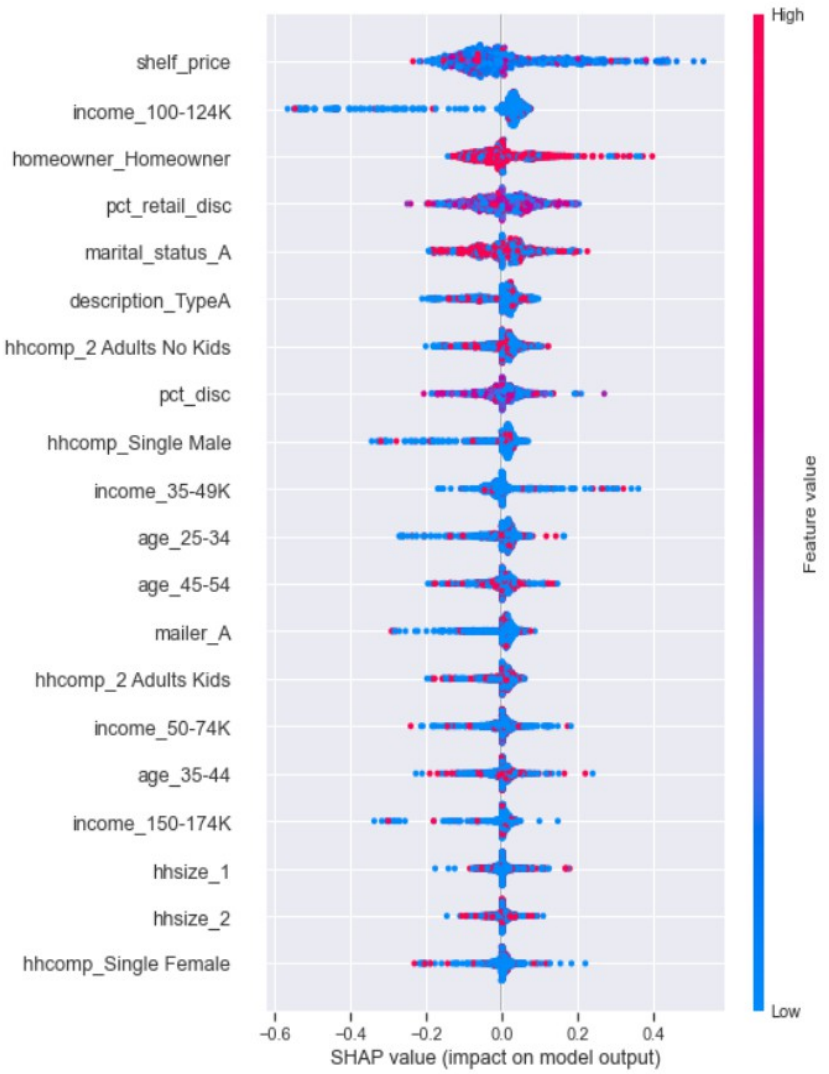


Shelf price clearly is the number one driver of new product purchases, followed by household income group 100-124k, homeownership, the retailer loyalty card discount, and being married.

Among all shopper marketing features, only coupon campaigns of type A and interior page features on the weekly marketing campaign (mailer_A) have a relevant contribution to predicting the target variable.

In the light of the only satisfactory model performance, the insights on feature importance have to be treated with caution in general. Strictly speaking, the model can be still described more as an educated guess than a good prediction.

The beeswarm plot is supposed to show the direction of each feature's effect, but actually does not provide a very clear picture for many of the features. Besides the limited model performance also multicollinearity in the features may be contributing to this.



All in all, we can derive some directional insights from both the SHAP analysis and the exploratory analysis carried out during the data understanding and preparation stages. However, these should be treated as pre-liminary and directional only and require further substantiation in the context of a more comprehensive future model before informing actual business decisions.

- Low shelf prices have a negative impact on new product purchases. In conjunction with the importance of (loyalty card) discounts and their positive correlation with shelf price, a likely explanation is that it is most attractive to pick a new product if it wouldn't be cheap generally, but a discount offer is making it look like a deal. Additionally, low prices often are associated with private labels which are typically less often on promotion and less attractive to try out (unless due to budgetary reasons).
- Higher household incomes of 100-124k, 150-175k, but also 50-75k come with lower propensity of purchasing new products. The picture for the income band 35-49k is less clear, even more so bearing in mind the positive bi-variate relationship of this feature with the target variable. A possible rationale behind this might be that for

higher income households the preferred products are more affordable and they are less prone to hunt for deals. This hypothesis would also be in line with the tendency among homeowners, another proxy of a stable financial situation, to purchase new products less often.

- Similar to homeownership with which it is positively correlated, being married (marital_status_A) seems to have a negative impact.
- Households with two adults/no kids and single men tend to buy fewer new products. Again, this might be connected to a higher discretionary income or a mindset that reduces the likelihood to shop for deals or novelty seeking.
- Interestingly, coupon campaigns of type A reduce the propensity of buying new products. Since we do not know anything about the nature of this campaign type, one possible explanation is that this may have been a dedicated loyalty campaign. Another possibility is that coupon campaigns, which engage the shopper outside the shopping environment, tend to give an advantage to regularly bought products rather than create a lasting enough awareness for new products to influence the purchase decision at a later stage in the store.

These directional learnings and hypotheses indicate three areas worthy of further investigation in order to drive brand penetration via shopper marketing: 1) Identifying the sweet spot for the interaction of regular shelf price and discounts. 2) Validating shopper demographics who are most prone to buy new products in order to optimize targeting. 3) Researching shopper marketing activities on a broader scale – some campaigns definitely do have an effect, however, based on the dataset in hand these effects cannot be understood sufficiently.

Discussion and Conclusion

Increasing brand penetration by acquiring buyers is a cornerstone of ongoing brand success according to the work of Ehrenberg Bass. This project aimed to develop a predictive model for new product purchases in the FMCG industry, with a specific focus on the yoghurt category, by leveraging transaction and shopper marketing data.

While the developed models are outperforming the simplistic benchmark model by far, they do not reach sufficiently good performance to be considered ready for deployment or fit for deriving any real-world action planning for businesses.

In this context, it is crucial to acknowledge the limitations of the dataset in hand which lacks information on marketing strategies outside of direct-to-shopper marketing, including any above-the-line activities, as well as the complete transaction history of the observed households across all retailers. These factors, though highly influential in the case of other marketing strategies and a source of noise in the data in the case of transactions at other retailers, could not be considered in our models and are limiting the models' performance potential. Future research, thus, ought to aim to incorporate a broader range of marketing strategies and retailers to develop a more comprehensive and accurate understanding of new product purchases in the FMCG market. Incorporating these factors is likely to enhance model accuracy and effectiveness.

Having said that, one could also argue that in view of the obvious limitations of the dataset, the models do a surprisingly good job in predicting – or, at the performance levels they reached, rather „educated guessing“ - new product purchases. This clearly underlines the notion that investing in direct-to-shopper marketing close to the moment-of-purchase represents an effective way for brands to acquire buyers.

Last but not least, expanding the research scope beyond the yoghurt category to include other FMCG product segments would be a worthwhile future step to understand the nuances across various product categories and help brands into new opportunities in their respective categories.