SUPERMART MARKETING CAMPAIGNS

What Drives Customer Responses?

Arturo, Marin, Clement, Boyang

Group 4

THE CHALLENGE

- 📢 **We plan to launch a new marketing campaign…**
- But **who should we target?**
- Not all customers respond the same way to promotions.
- We want to **invest our marketing budget efficiently** by focusing on the right audience.

OBJECTIVES

Customer Segmentation – *"Understanding Customer Groups"*

Use clustering to group customers based on spending habits and engagement, helping tailor marketing strategies for different segments.

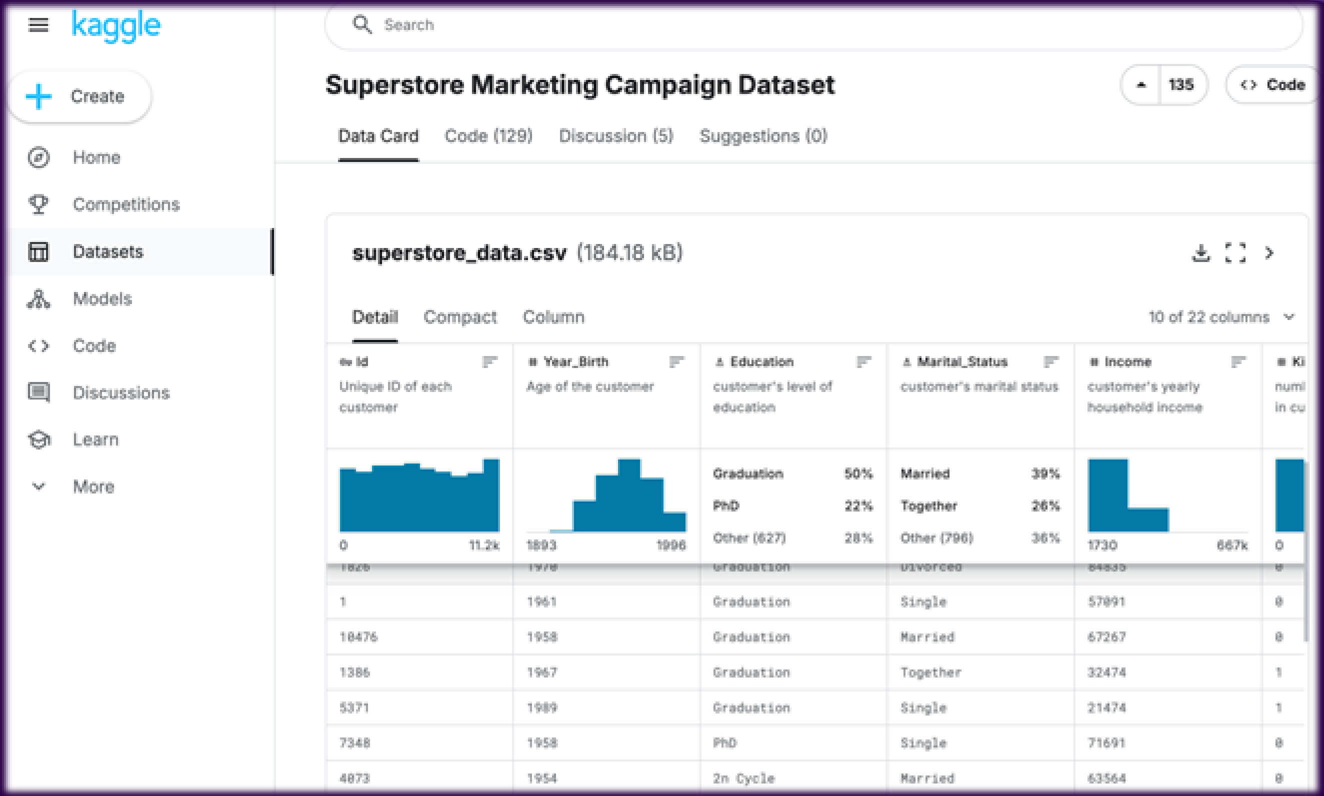
Predictive Modeling – *"Targeting the Right Customers"*

Use machine learning to predict which customers are most likely to respond, so campaigns can be targeted more effectively.

Causal Inference – *"Measuring True Marketing Impact"*

Use causal inference to measure the true impact of marketing efforts.

DATASET



https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset

DATA DICTIONARY

Customer Information

- Customer ID
- Enrollment Date

Demographics

- Year of Birth
- Income
- Family Structure
- Education & Marital Status

Purchasing Behavior

- Amount Spent
- Number of Purchases
- Discount Usage

Marketing Engagement

- Campaign Response
- Recency
- Website Visits

SOLUTIONS

Customer Segmentation

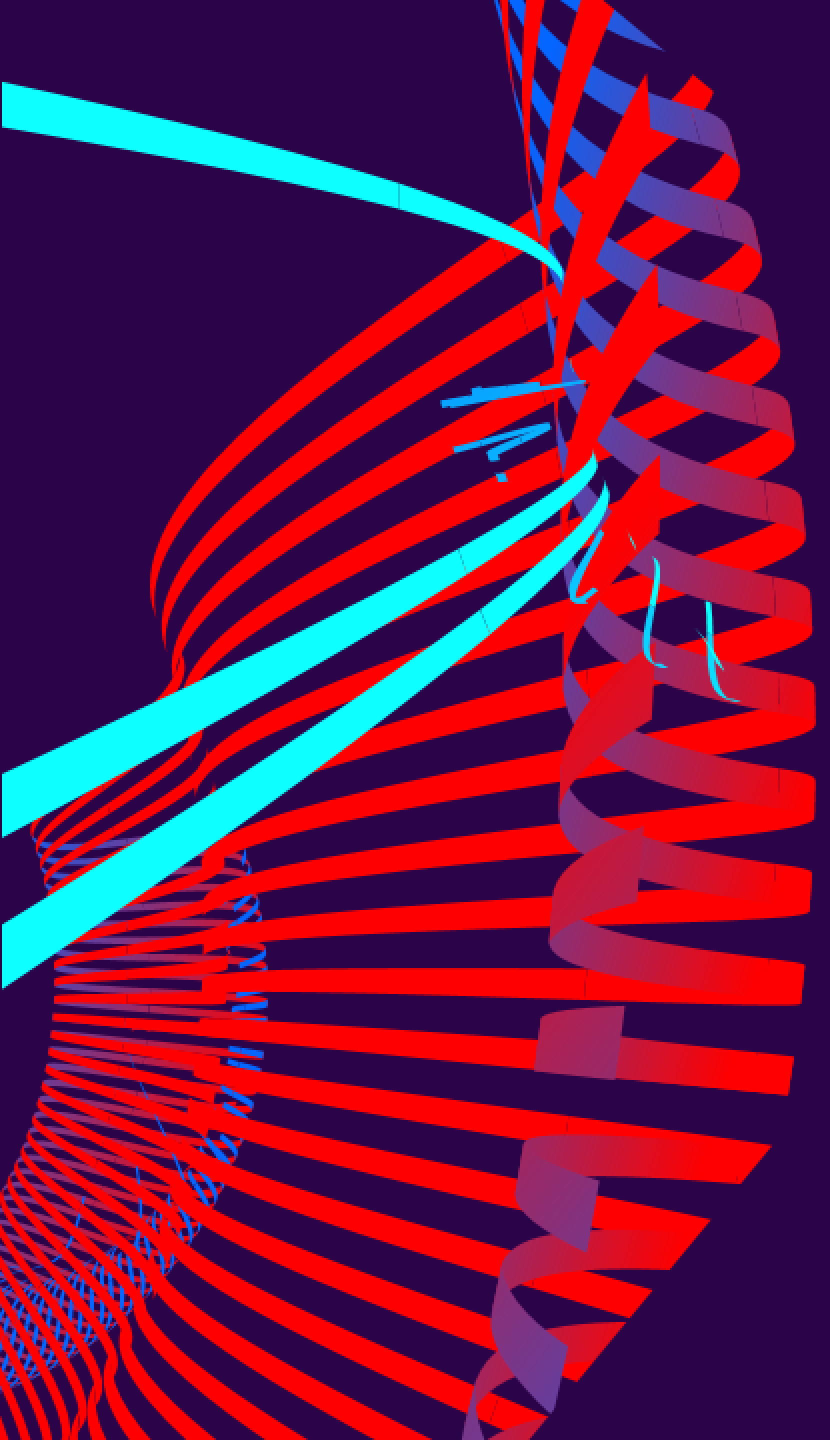
- Segmented customers using K-Means based on spending and demographics.
- Identified groups to improve marketing targeting.

Predictive Modeling

- Built models (Logistic Regression, Random Forest, XGBoost, LightGBM) to predict customer response.
- Evaluated performance using ROC-AUC and classification metrics.

Causal Impact Analysis

- Used **R-Learner with LightGBM** to estimate marketing impact.
- Measured **Average Treatment Effect (ATE)** to find the best target audience.



DATA PREPROCESSING

DATA PREPROCESSING

HANDLING MISSING VALUES

Median imputation for missing income

ENCODING CATEGORICAL DATA

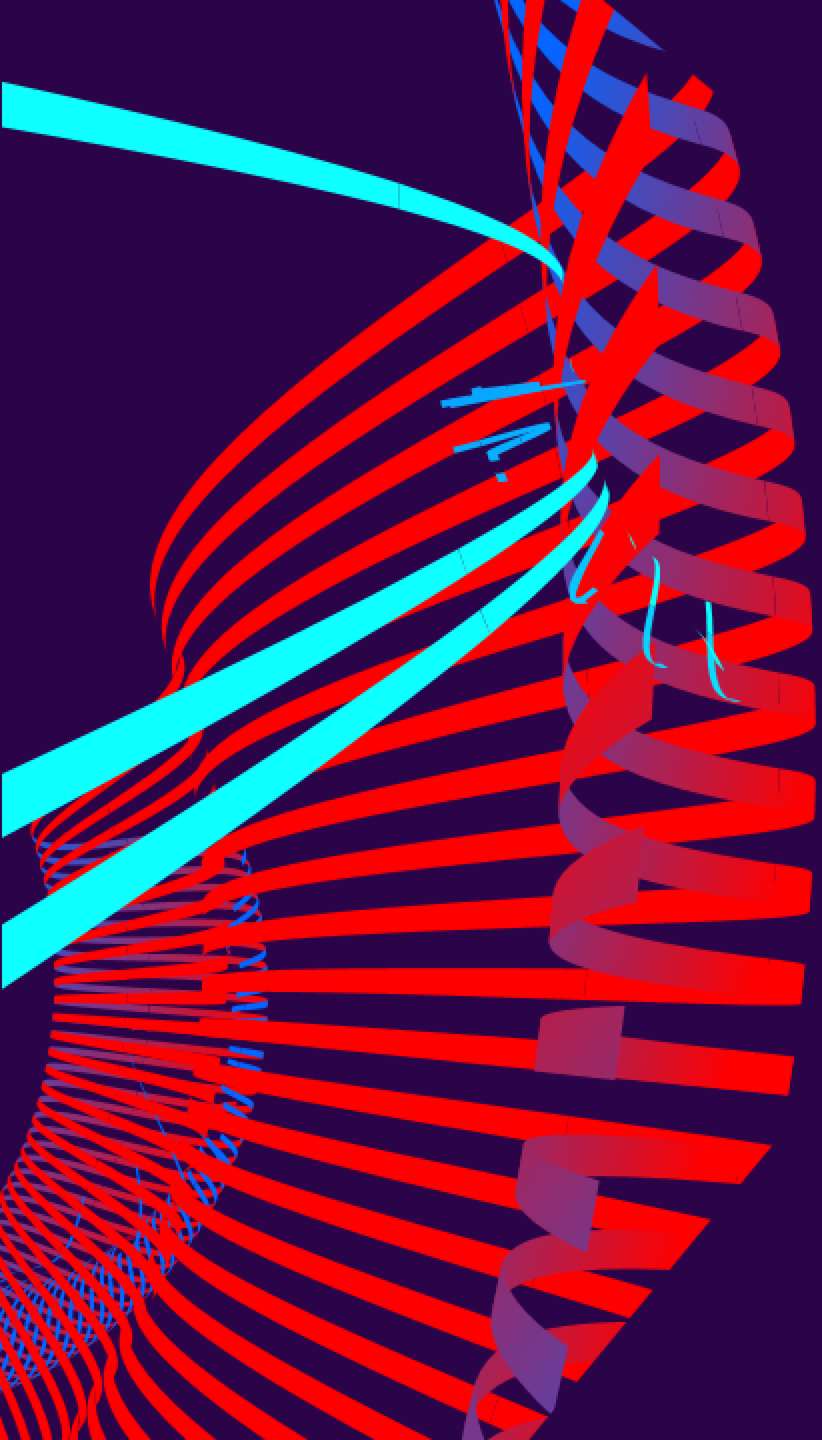
Encoded categorical features to more equally distributed groups

FEATURE ENGINEERING

Log transformation for columns with skewness above 1

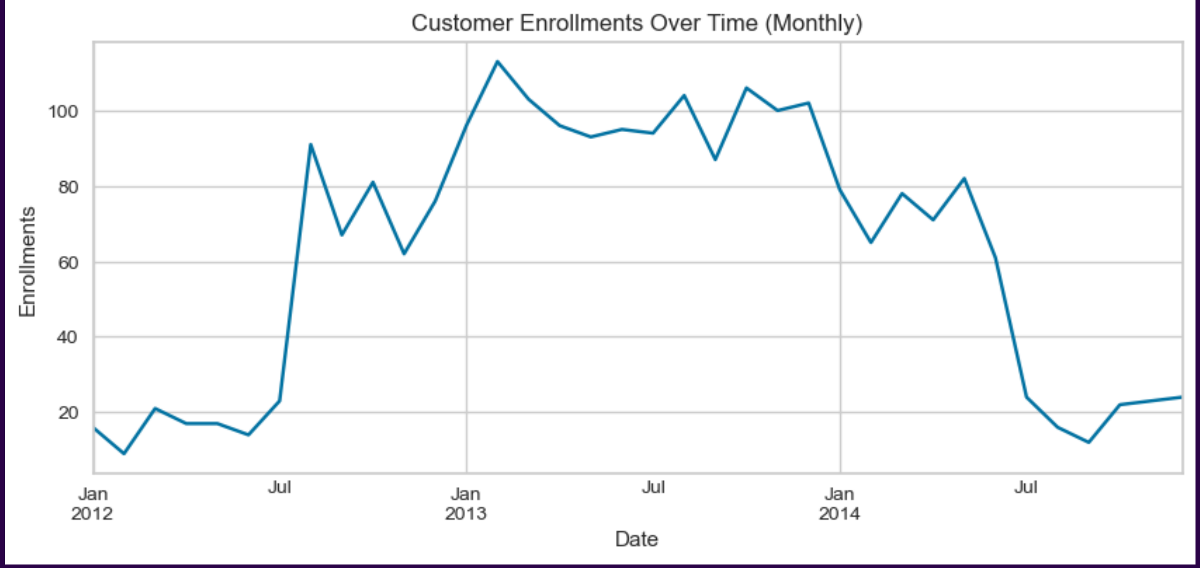
DATA SPLITTING

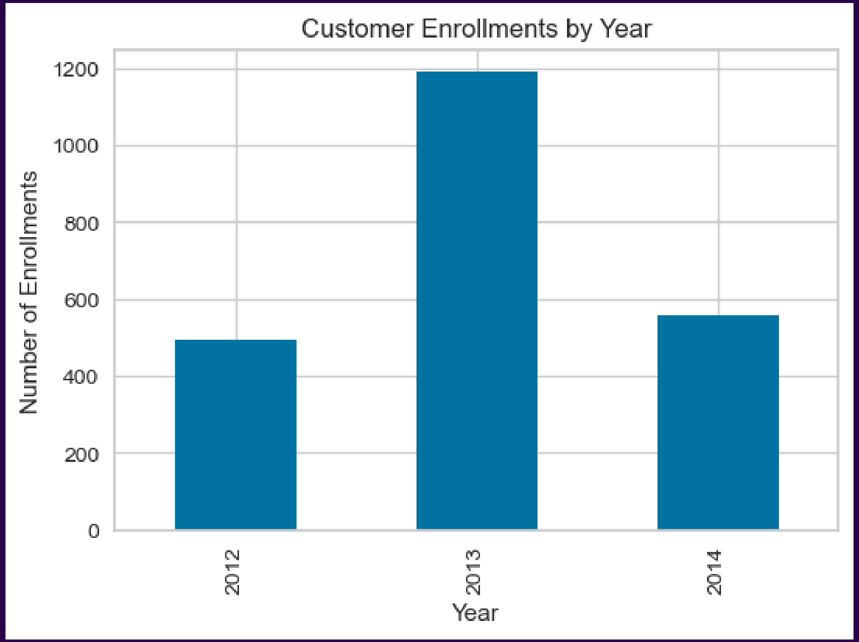
Created training, validation, and test sets with stratification

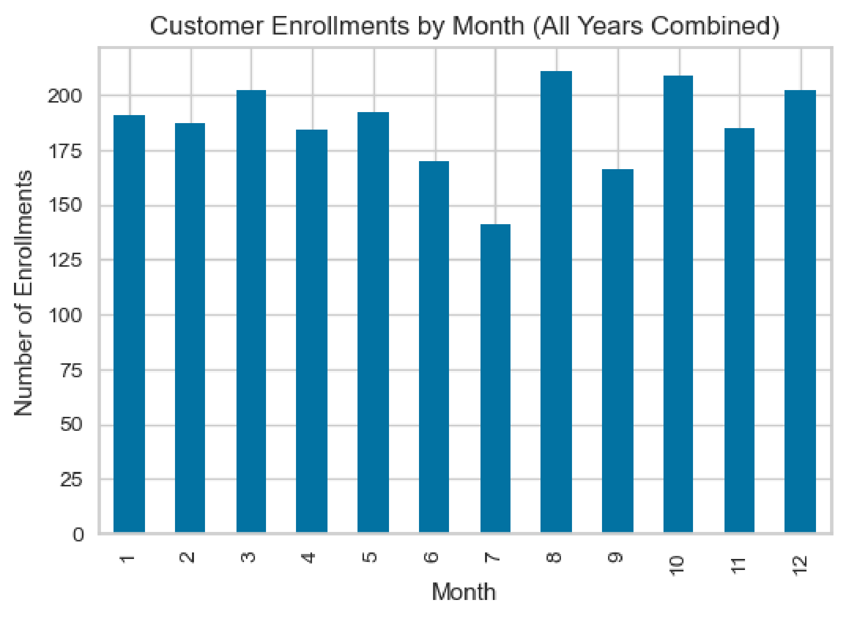


EXPLORATORY DATA ANALYSIS

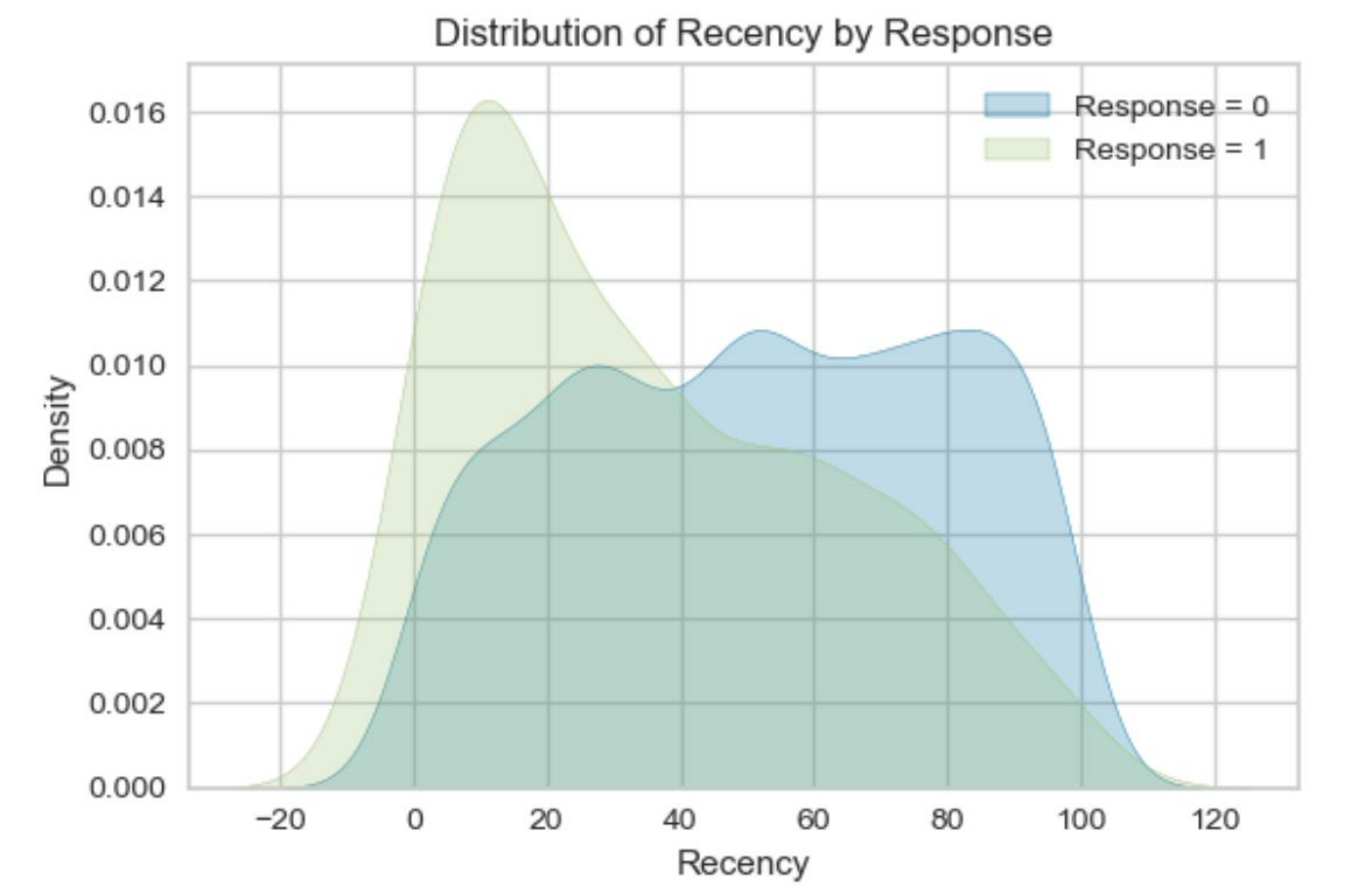
CUSTOMER ENROLLMENT TRENDS OVER TIME







RECENCY & RESPONSE



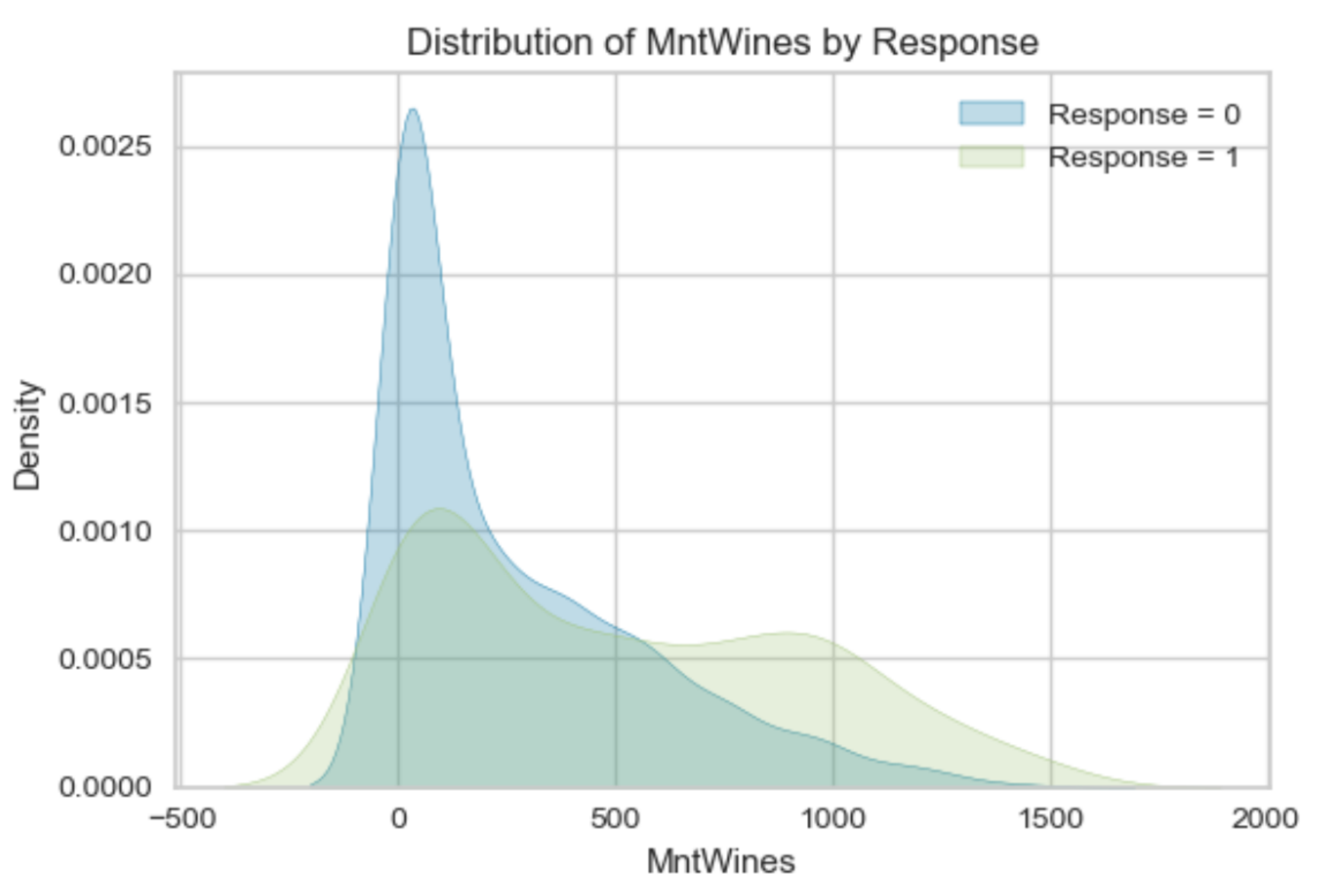
Customers who **purchased recently** are **more likely to engage** with campaigns.

WEB PURCHASES & RESPONSE

Customers who make more web purchases are more likely to engage with marketing campaigns, suggesting that online shoppers are a key target audience for digital promotions.

WINE BUYERS& RESPONSE

Customers who spend more on wine are significantly more likely to engage with marketing campaigns, indicating that high-spending wine buyers are a valuable target for promotions.



CUSTOMER SEGMENTATION

"UNDERSTANDING CUSTOMER GROUPS"

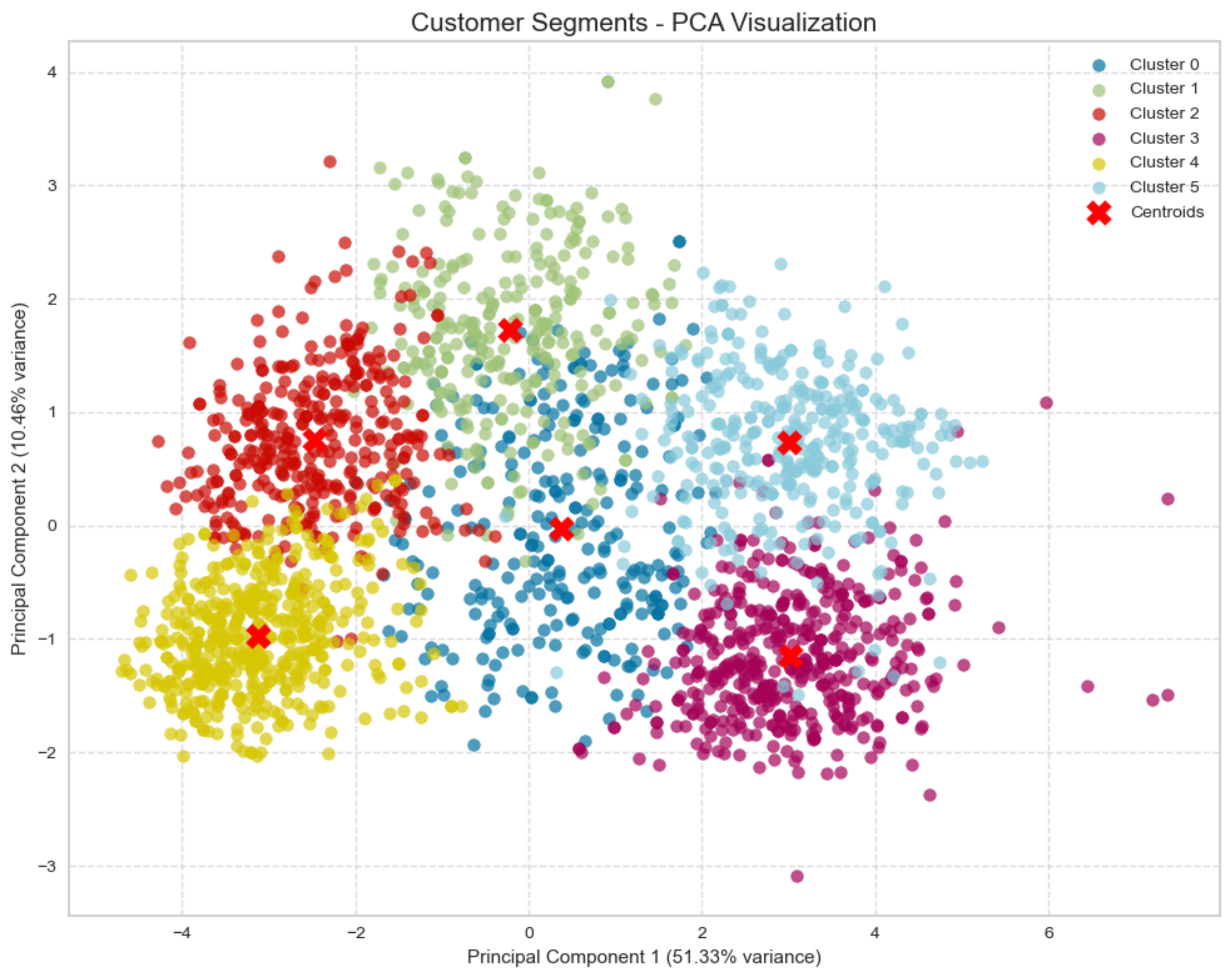
CUSTOMER SEGMENTATION ANALYSIS

WHAT WE HAVE DONE

Applied **K-Means Clustering** to group customers based on spending, demographics, and engagement.

- Selected key features
- Standardized data
- Elbow Method
- Create 6 Clusters

CUSTOMER SEGMENTATION ANALYSIS



CUSTOMER SEGMENTATION ANALYSIS

INSIGHTS

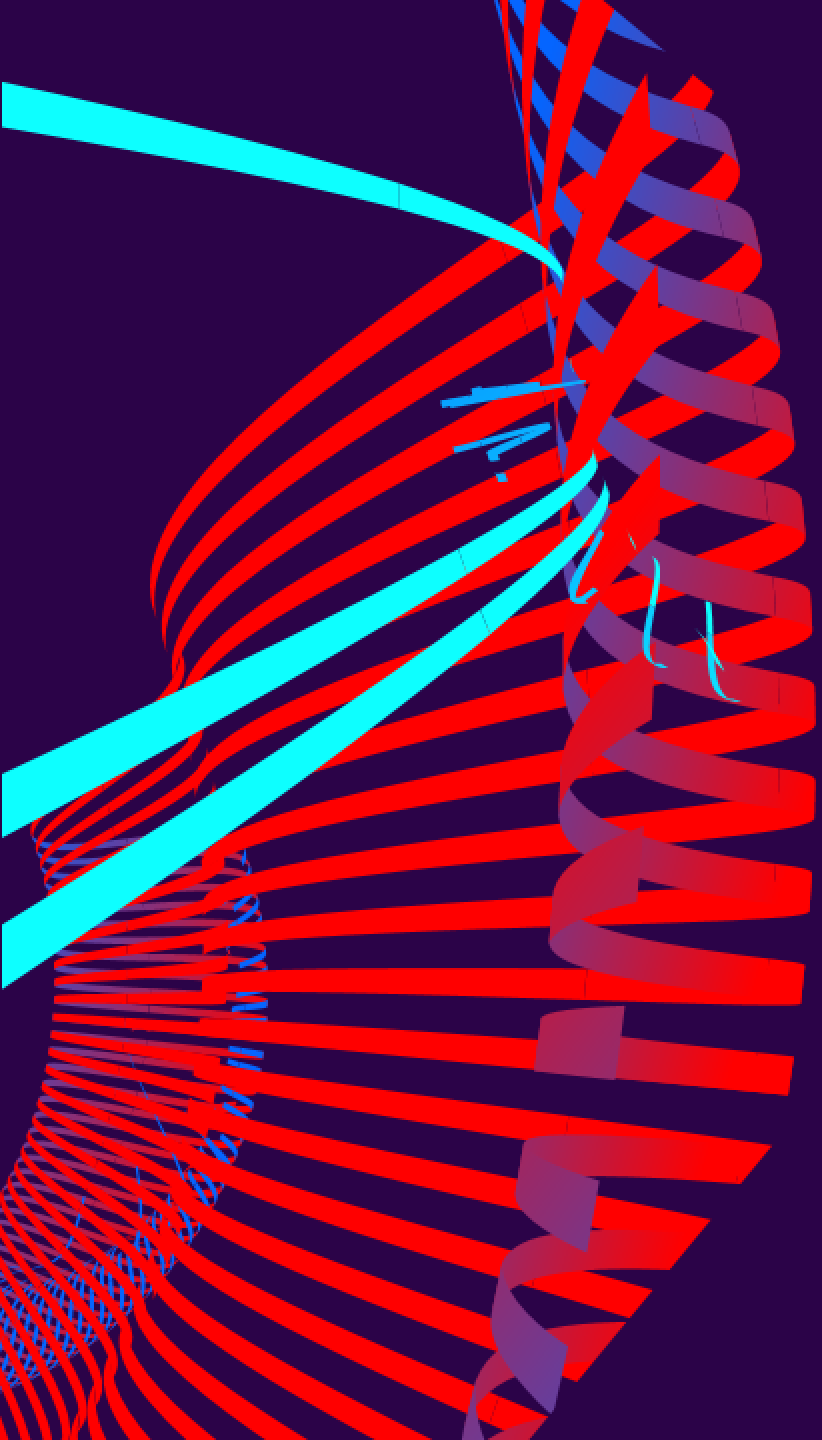
💡 **Segment 4 has the highest campaign response rate (~30%) →**

These customers spend the most on premium products (wines, meat) and are highly engaged.

They are the best target for marketing campaigns.

💡 Segment 3 & 5 have the lowest engagement (~11% and ~2.9%) →

These customers spend the least and rarely engage in online purchases, meaning they require a different strategy, such as personalized offers or loyalty incentives.



PREDICTIVE MODELING

"TARGETING THE RIGHT CUSTOMERS"

WHAT ARE WE PREDICTING?

Predict if a customer will respond to a marketing campaign.

1

Responded

0

Did not Respond

Engaged with the campaign

e.g. Made a purchase

Ignored the campaign

BUILDING PREDICTION MODEL

HANDLED CLASS IMBALANCE

Used SMOTE (Synthetic Minority Over-sampling Technique) to balance positive and negative responses.

WHAT MODELS DID WE TRAIN?

Baseline Model

Logistic Regression

Ensemble Model

LightGBM Random Forest
CatBoost XGBoost
AdaBoost LightGBM
Bagging Extra Trees

Others

KNN
SVM

HYPERPARAMETER TUNING

GRID SEARCH WITH CROSS-VALIDATION (CV=5)

- Exhaustively tested different hyperparameter combinations.
- Used cross-validation to ensure robust performance across different data splits.

OPTIMIZED PARAMETERS FOR EACH MODEL TYPE

- Tree-Based Models (Random Forest, XGBoost, LightGBM, etc.) → Tuned number of trees, max depth, learning rate.
- Logistic Regression & SVM → Adjusted regularization strength (C), solver, and kernel types.
- KNN → Tuned the number of neighbors and weighting method.
- Bagging Models → Experimented with different numbers of base learners.

MODEL PERFORMANCE COMPARISON

ROC-AUC Score

If we want a balanced model that performs well across all response probabilities

Confusion Matrix

If we want to avoid targeting the wrong customers with marketing efforts

Precision-Recall Curve

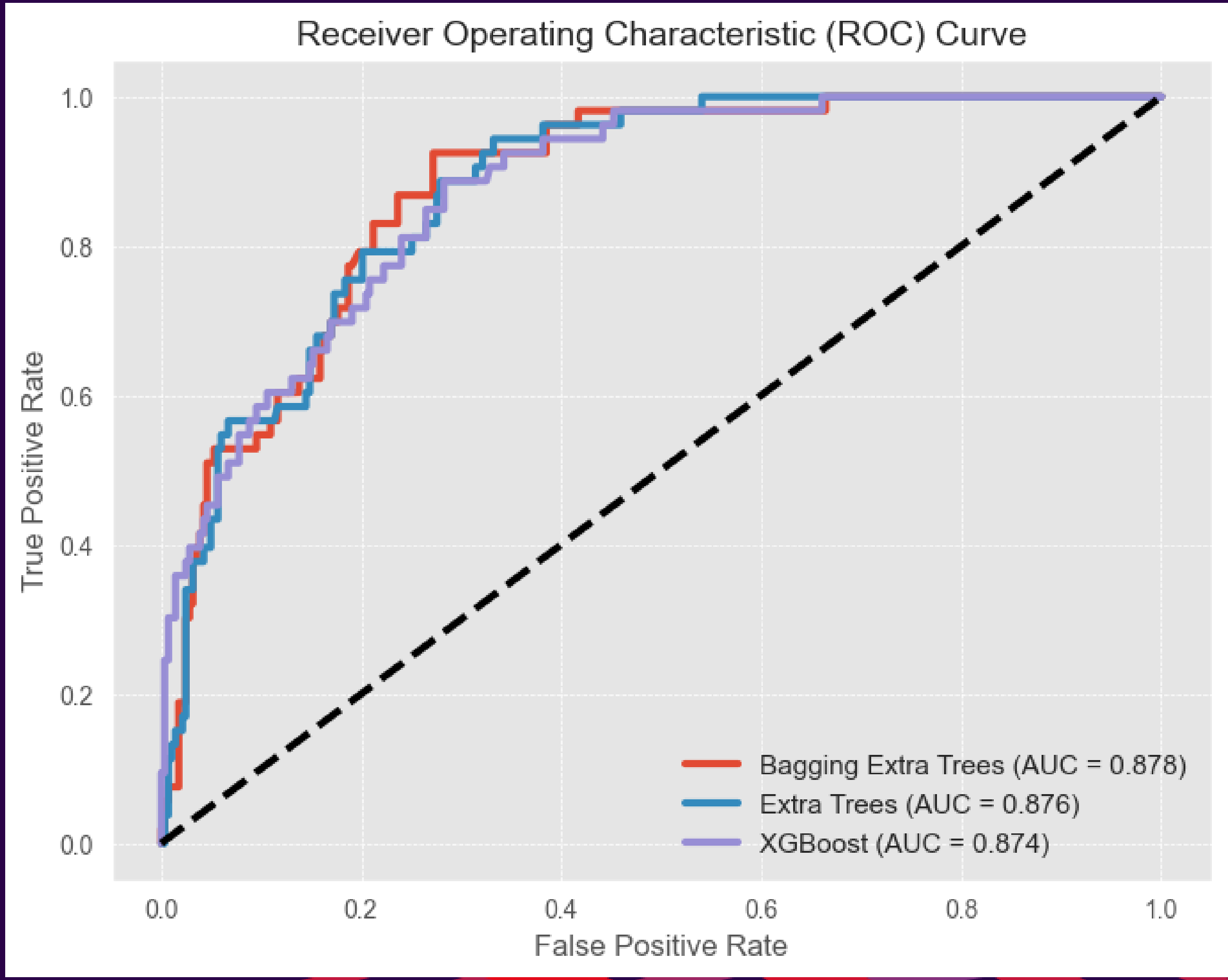
If we care more about correctly identifying responders rather than avoiding false positives

ROC CURVE

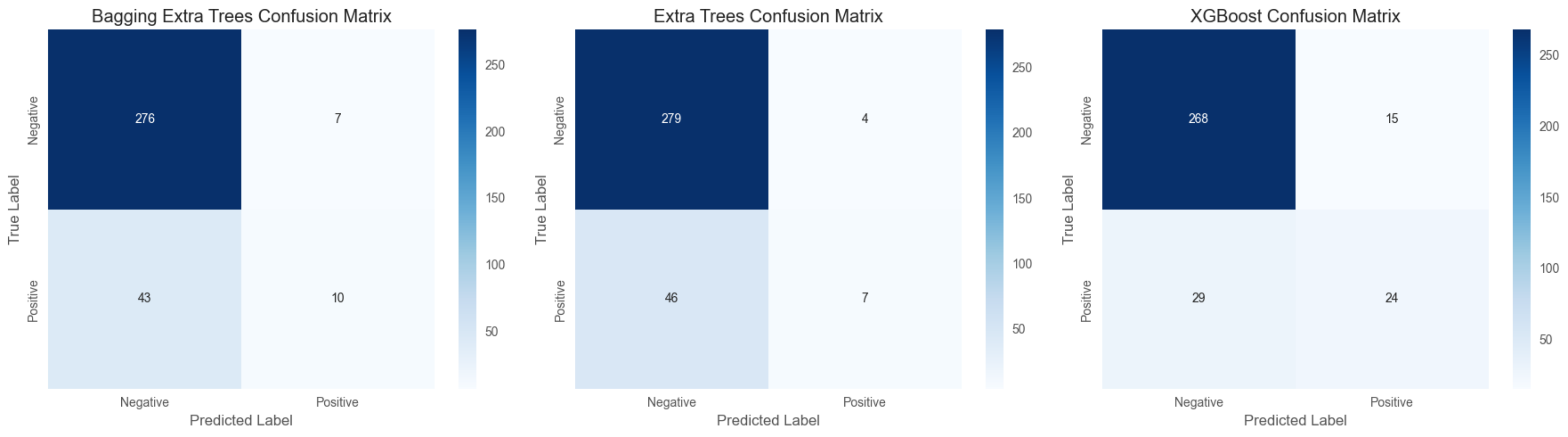
- The closer the curve is to the top-left, the better the model.
- AUC (Area Under Curve) values:
  - Bagging Extra Trees (0.878)
  - Extra Trees (0.876)
  - XGBoost (0.874)

CONCLUSION

All three models perform well, with Bagging Extra Trees slightly ahead



CONFUSION MATRICES



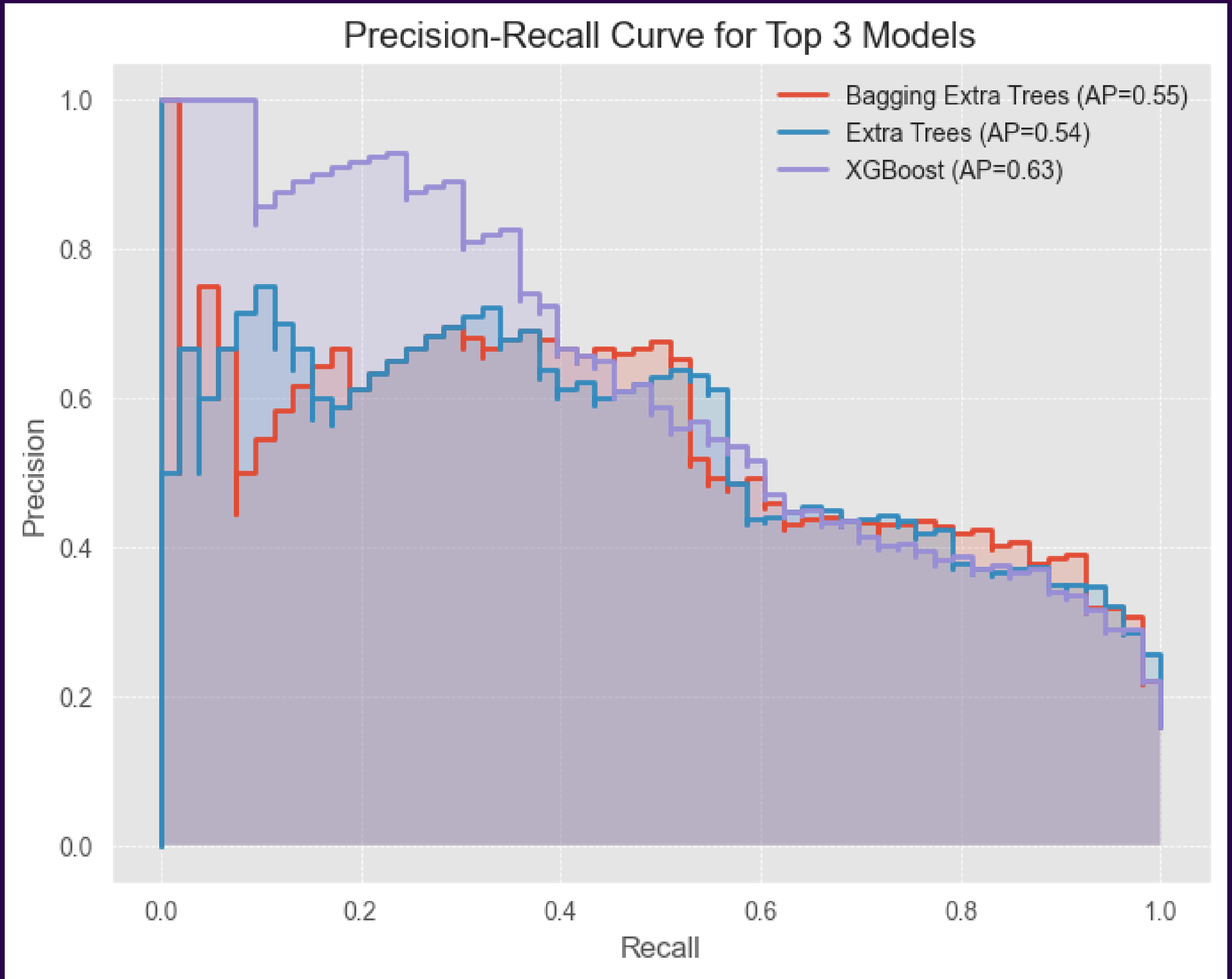
- Bagging Extra Trees minimized false positives.
- XGBoost captured more positives but had slightly more false positives.

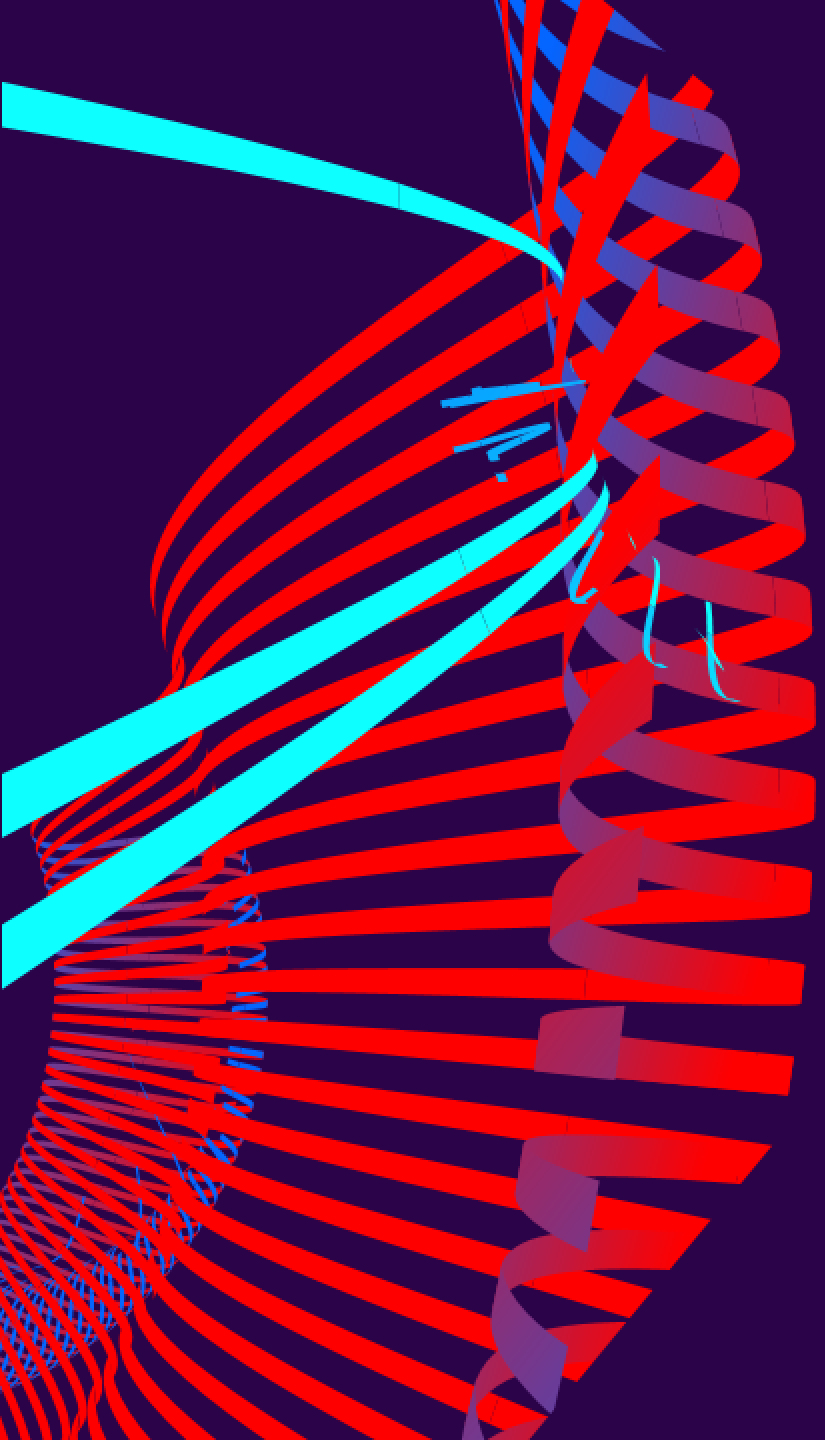
PRECISION-RECALL CURVE

- XGBoost had the highest average precision (AP = 0.63).
- Extra Trees and Bagging Extra Trees had similar scores (AP ≈ 0.55).

CONCLUSION

If we care more about identifying responders than avoiding false positives, XGBoost is better.





CAUSAL INFERENCE

"MAKING TRUE MARKETING IMPACT"

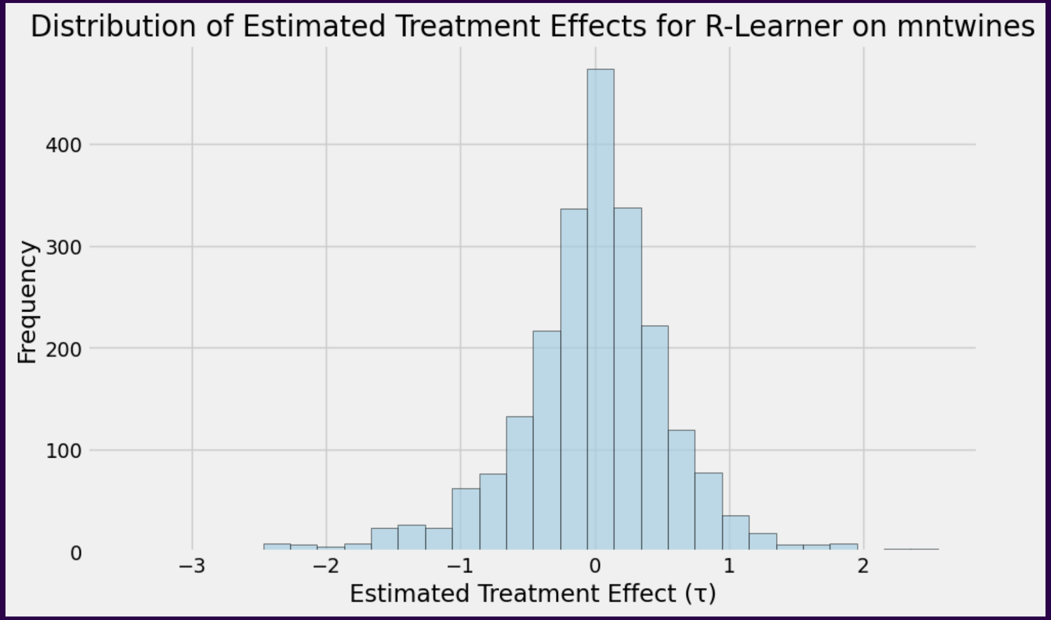
BUILDING A CAUSAL MODEL

Investigating Customer Responses to Marketing Campaign:

- Understand the impact that the campaign has on the customer purchases in each of the product categories:
    - Wine, Fruit, Meat, Fish, Sweet
- Used R-Learner to estimate the Conditional Average Treatment Effects (CATES), with integrated LGBM model for each product category
    - Y = Product Category
    - T = Response to the Marketing Campaign
    - X = The rest of the variables
- Analyzed the CATEs, as well as the feature importance and the SHAP values

CATE

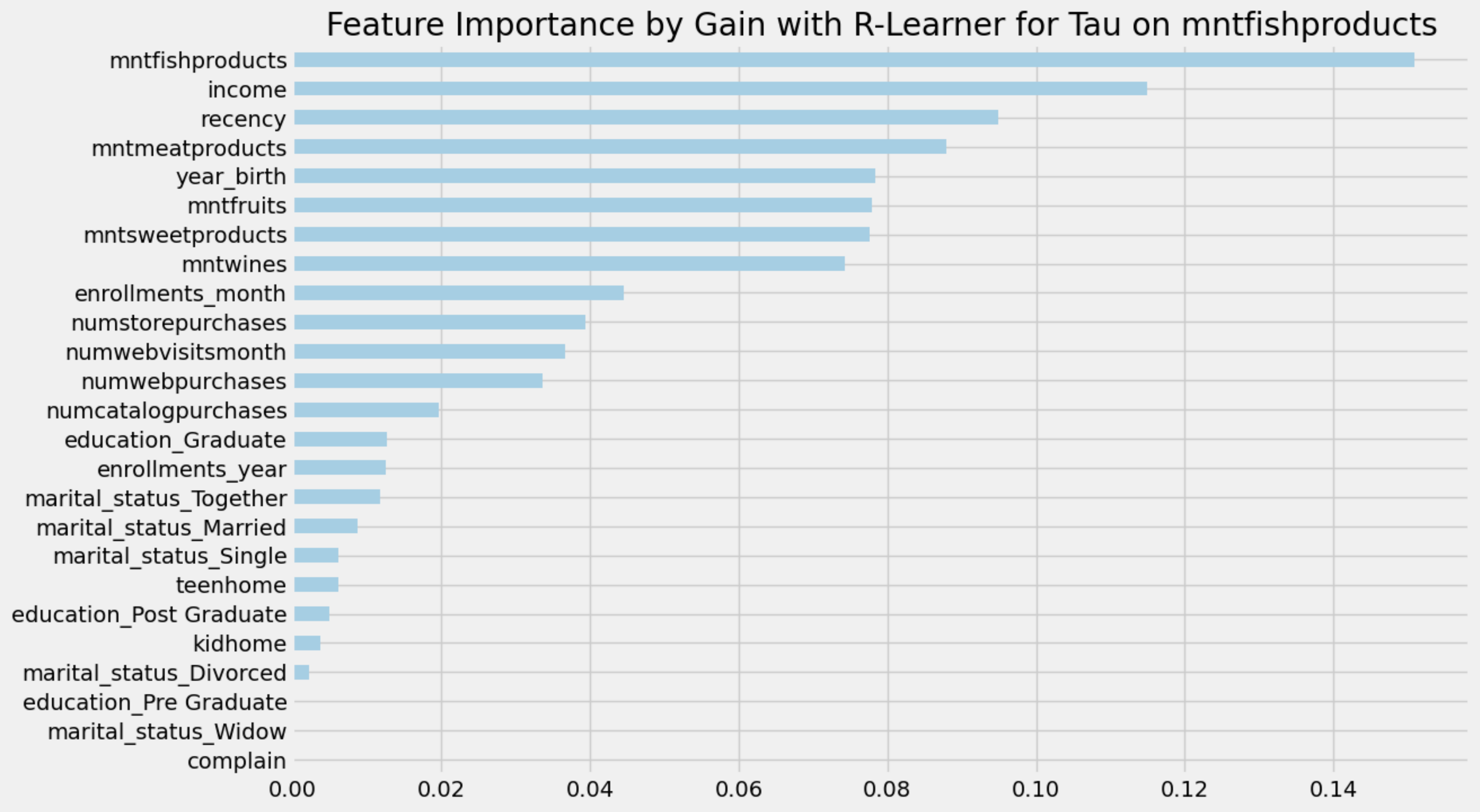
	target var	ate	lb	ub
0	mntwines	-0.073664	-0.074577	-0.072752
1	mntfruits	-0.165210	-0.166978	-0.163441
2	mntmeatproducts	0.004689	0.003714	0.005664
3	mntfishproducts	-0.304885	-0.306766	-0.303005
4	mntsweetproducts	-0.007372	-0.009301	-0.005444
5	mntgoldprods	0.120612	0.118654	0.122571



- Mostly negative effects which are near 0
- A very large distribution of τ
    - Normal and centered around 0

FEATURE IMPORTANCE

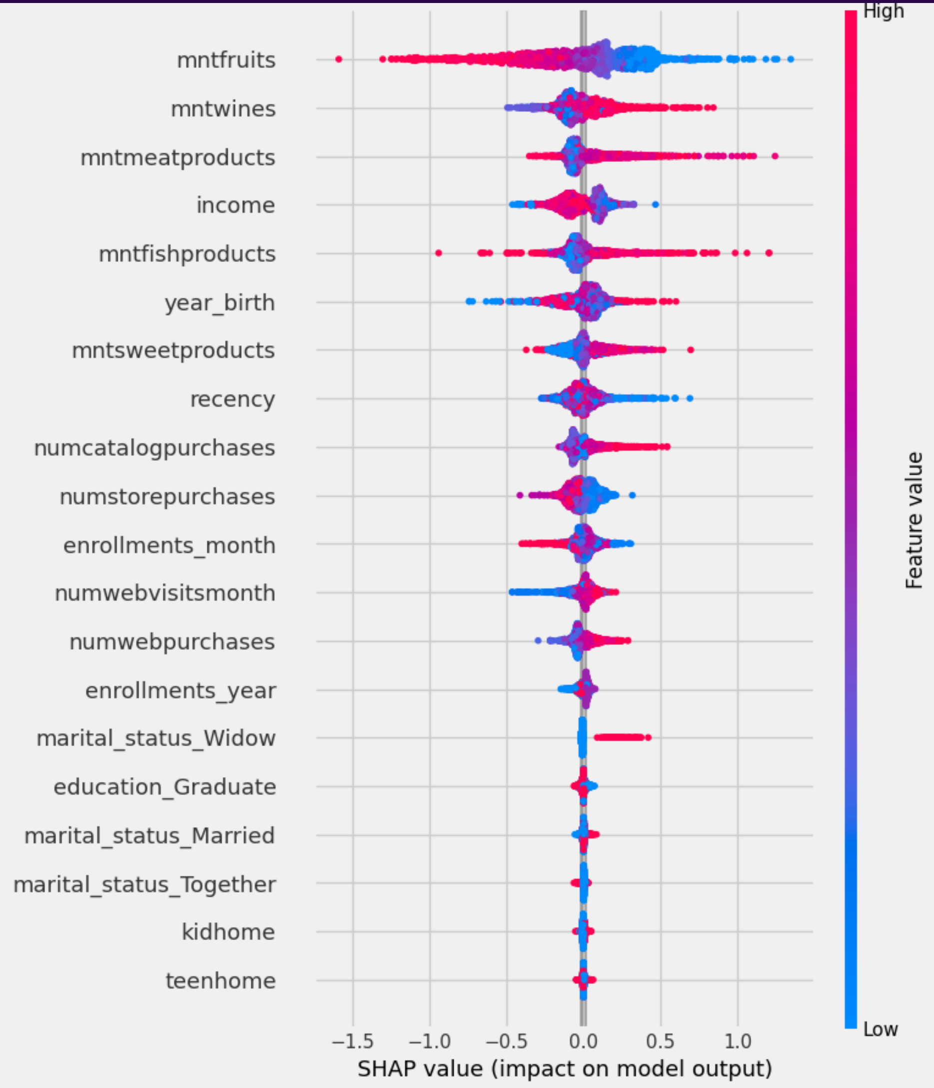
- Generally, the most important feature is the target variable
- Other product category amounts for different targets also have lots of importance
- The control variables with the most importance are: income and recency



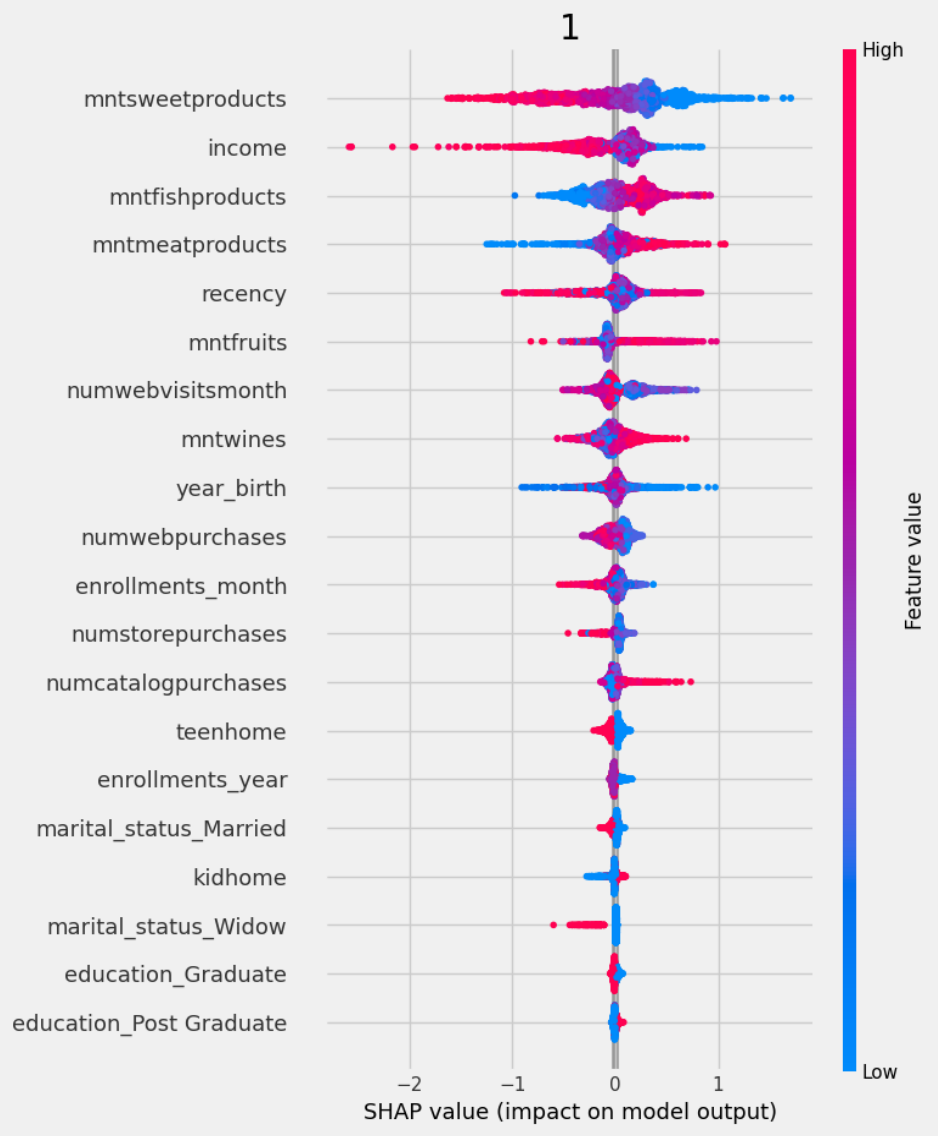
SHAP ANALYSIS

General takeaways form SHAP analysis:

- The SHAP values of the target variables share the same pattern:
  - The customers who don't buy much of the target variable are associated with larger positive SHAP values
  - The opposite is true for the customers who buy a lot of a category
- Interpretation:
  - The campaign makes customers try out products they don't usually purchase



SHAP ANALYSIS



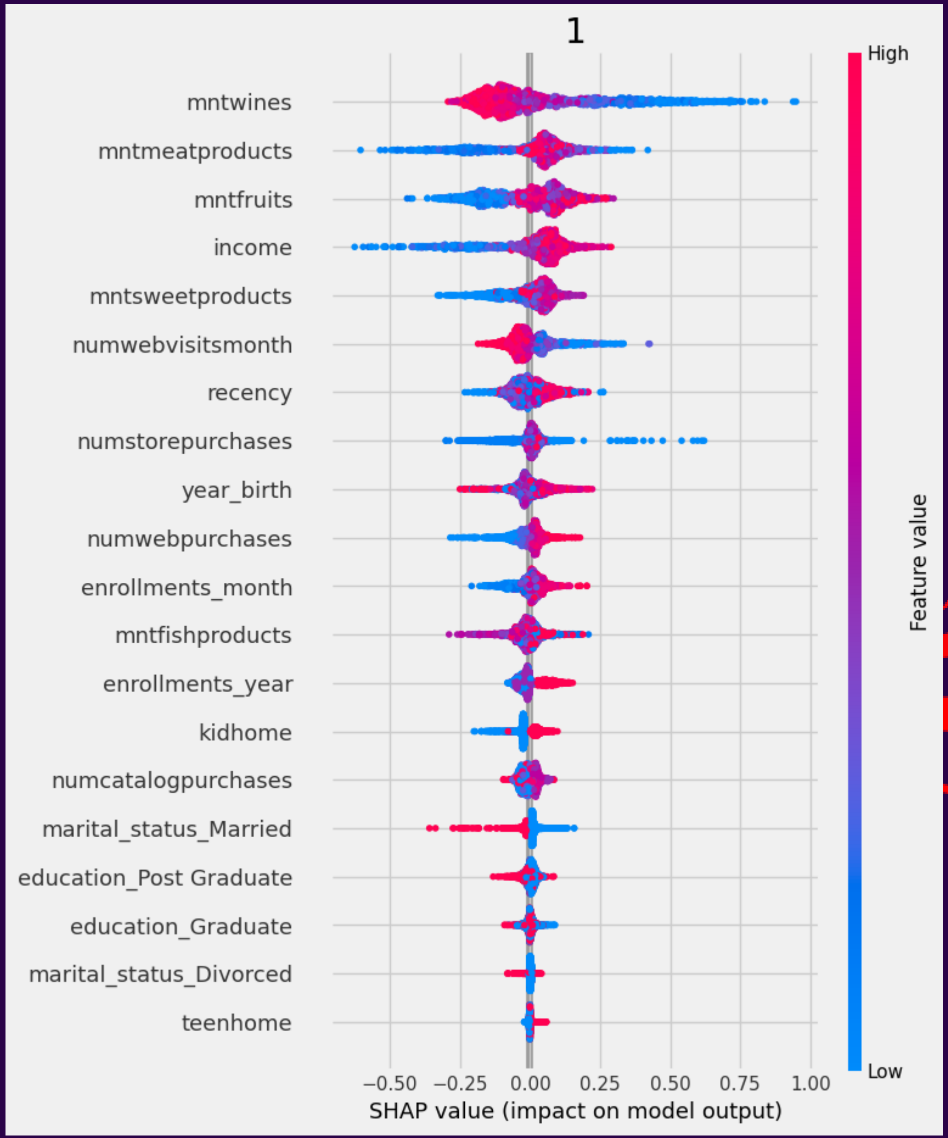
General takeaways form SHAP analysis:

- The SHAP values of the amount of product control variables:
  - Have the opposite direction of the SHAP values of the target variable
  - Has smaller magnitude relative to the target variable
- Interpretation:
  - There is a strong correlation between each of the amount variables, those who buy a lot of a category tend to buy more of another
  - The campaign has a stronger effect on the customers who buy a lot of products

SHAP ANALYSIS

Specific takeaways form SHAP analysis:

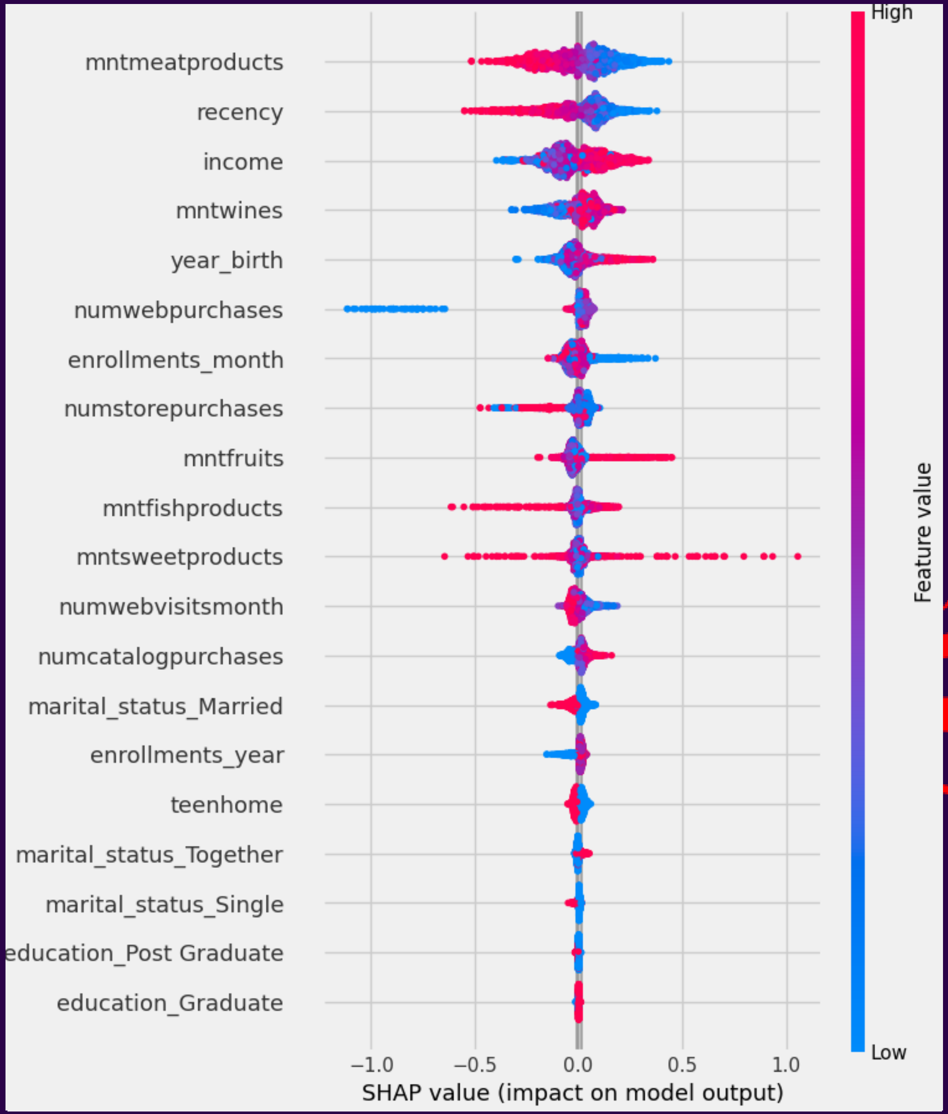
- SHAP for wine as target variable:
  - Same directional pattern for the effect of the amount of wine on the treatment effect
  - No strong negative effect for the customers which the most wine
- Interpretation:
  - The campaign's effect of changing the purchasing habits isn't as strong for the wine category
  - A harder habit to break



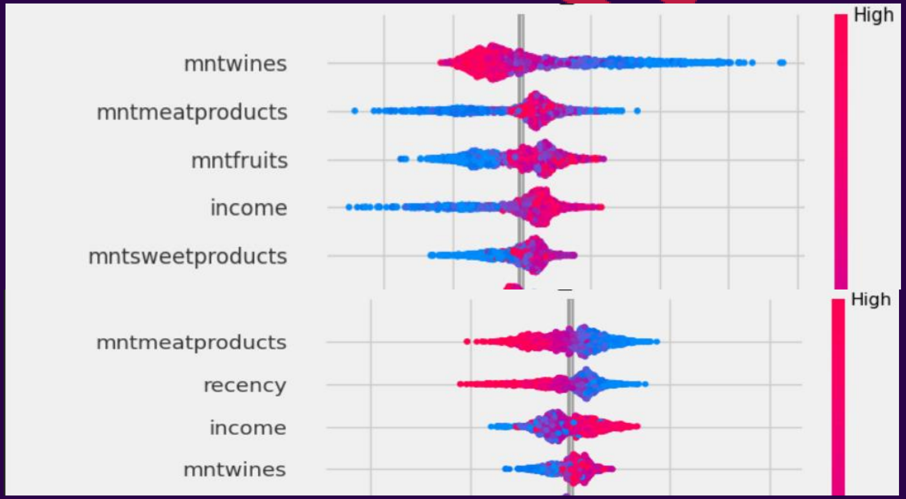
SHAP ANALYSIS

Specific takeaways form SHAP analysis:

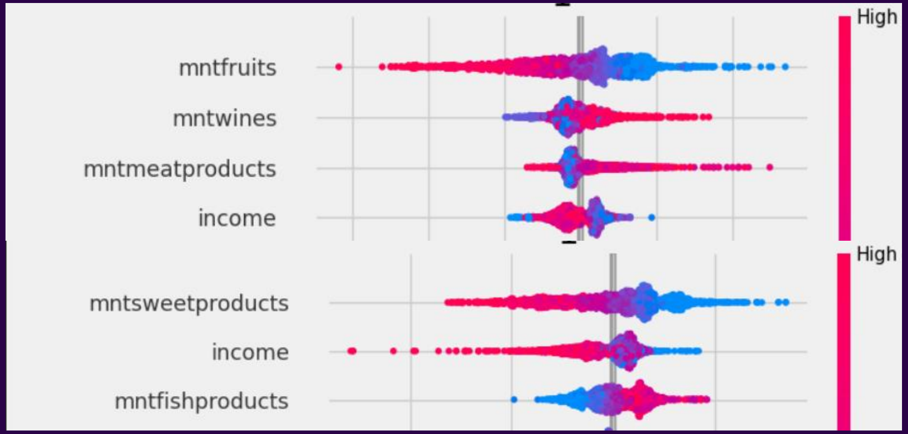
- SHAP for meat as target variable:
  - Same directional pattern for the effect of the amount of meat on the treatment effect
  - Overall, magnitude across variables is much smaller
  - Strictly negative effect of those who don't have web purchases
- Interpretation:
  - The campaign's effect is smaller on the meat
  - The campaign shouldn't target meat purchases for people who don't buy online



SHAP ANALYSIS



Exprensive



Cheap

Interpretation:
For the customers with the larger income see the marketing campaign as an occasion to buy "Wants" and those with a lower income "Needs"

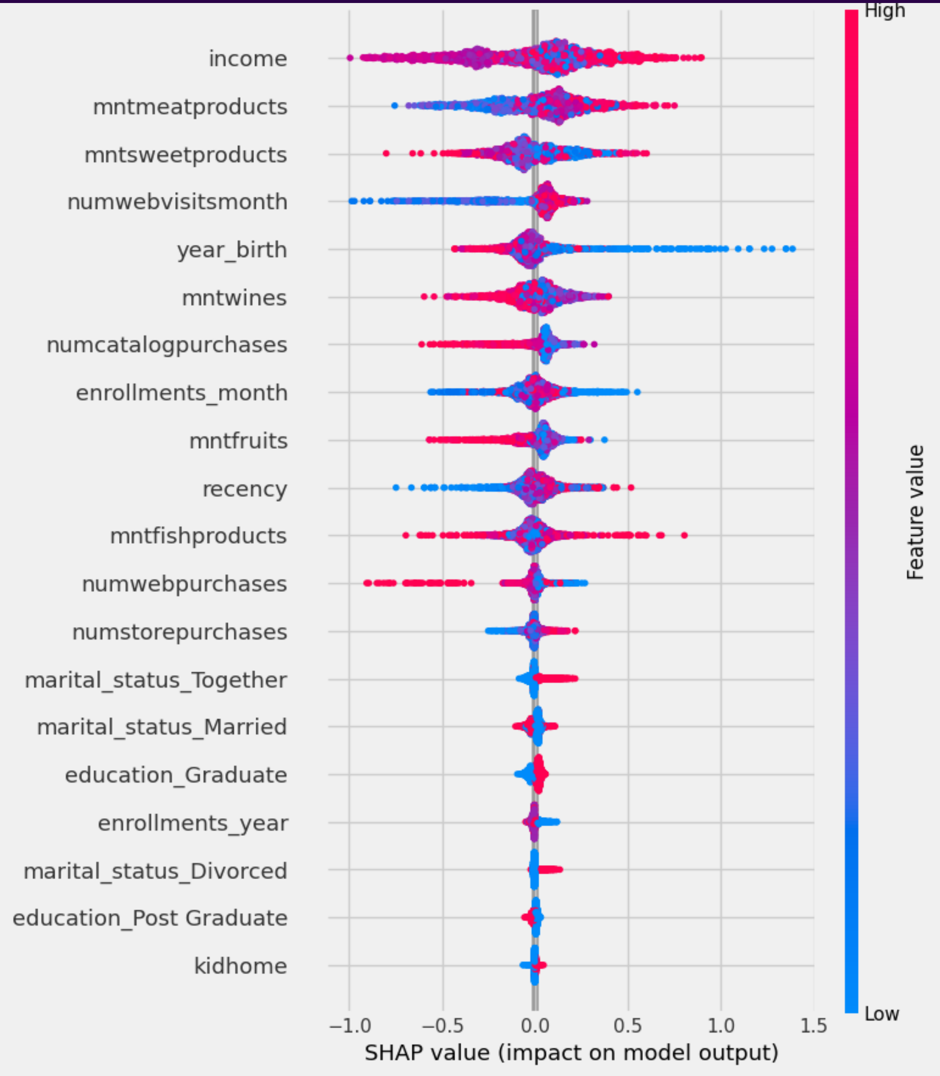
THE OUTLIER

There is one product category which was not covered, Amount of Gold Porducts, because of its atypical behavior:

- Larger CATE
- Feature importance didn't include the targer variable
    - Both gain and permutation
- Lots of variability in the SHAP values in comparaison to the others

Interpretation:

- Assume that the goal products are special promotional products across multiple other categories
- The data is aggregated across 2 year of purchases
- Due to how this data in entangled temporaly and across multiple categories, we lack the necessay specific information to determine the causal realtionship between the gold products and the campaign



BUSINESS VALUES

Optimized Marketing Spend

Increased Customer Engagement

Higher ROI on Promotions

Scalable and Repeatable

THANK YOU