Contents

[Problem identification 2](#_Toc64038464)

[Problem Statement 2](#_Toc64038465)

[Context 2](#_Toc64038466)

[Criteria for Success 2](#_Toc64038467)

[Scope of Solution Space 2](#_Toc64038468)

[Constraints 2](#_Toc64038469)

[Stakeholders 3](#_Toc64038470)

[Dataset Description 3](#_Toc64038471)

[Customer Transaction Data 3](#_Toc64038472)

[Customer Demographics 3](#_Toc64038473)

[Item Data 4](#_Toc64038474)

[Coupon Item Mapping 4](#_Toc64038475)

[Campaign Data 5](#_Toc64038476)

[Coupon Data 5](#_Toc64038477)

[Data Wrangling 5](#_Toc64038478)

[EDA 6](#_Toc64038479)

[Univariate Analysis 6](#_Toc64038480)

[Bivariate Analysis 10](#_Toc64038481)

[Feature Engineering 12](#_Toc64038482)

[Modeling 12](#_Toc64038483)

[Hyperparameter Tuning 12](#_Toc64038484)

[Feature Importance 13](#_Toc64038485)

# Problem identification

## Problem Statement

How can ABC company develop a more precise and targeted marketing  strategy to increase their customer base by 20% by the end of 2021, by using previous campaign data to identify customer coupon redemption behavior based on factors like customer profile, type of items and seasonality.

## Context

ABC is an established Brick & Mortar retailer that uses marketing campaigns to attract new customers and retain existing ones. These campaigns include offering coupon discounts on a specific product or a range of products. ABC wants to use machine learning to get the ability to predict coupon redemption. This will help their marketing team to better design these coupon campaigns and develop a more precisely targeted marketing strategy that results in an increase in coupon redemption, thereby resulting in an increase in the customer base.

## Criteria for Success

* 20% increase in overall customer base by end of 2021
* 15% improvement in customer retention
* 25% increase in new customers

## Scope of Solution Space

Analysis is restricted to data from the last 28 campaigns and will identify the top 10 features that affect coupon redemption by customers.

## Constraints

* Data only available for 1.5-year span
* No insight into campaign channels like email, in-store, etc.
* Limited information about customer profile.
* Items are anonymized - they only have item IDs and no names

## Stakeholders

* VP Sales and Marketing
* VP Finance

# Dataset Description

The data is taken from Kaggle and consists of the following datasets:

## Customer Transaction Data

Transaction data for all customers for duration of campaigns in the train data. It consists of the following columns:

|  |  |
| --- | --- |
| Variable | Definition |
| date | Date of transaction |
| customer\_id | Unique id for a customer |
| item\_id | Unique id for item |
| quantity | quantity of item bought |
| selling\_price | Sales value of the transaction |
| other\_discount | Discount from other sources such as manufacturer coupon/loyalty card |
| coupon\_discount | Discount availed from retailer coupon |

## Customer Demographics

Customer demographic information for some customers containing the following columns:

|  |  |
| --- | --- |
| Variable | Definition |
| customer\_id | Unique id for a customer |
| age\_range | Age range of customer family in years |
| marital\_status | Married/Single |
| rented | 0 - not rented accommodation, 1 - rented accommodation |
| family\_size | Number of family members |
| noofchildren | Number of children in the family |
| income\_bracket | Label Encoded Income Bracket (Higher income corresponds to higher number) |

## Item Data

Item information for each item sold by the retailer:

|  |  |
| --- | --- |
| Variable | Definition |
| item\_id | Unique id for itemv |
| brand | Unique id for item brand |
| brand\_type | Brand Type (local/Established) |
| category | Item Category |

## Coupon Item Mapping

Mapping of coupon and items valid for discount under that coupon:

|  |  |
| --- | --- |
| Variable | Definition |
| coupon\_id | Unique id for a discount coupon (no order) |
| item\_id | Unique id for items for which given coupon is valid (no order) |

## Campaign Data

Campaign information for each of the 28 campaigns:

|  |  |
| --- | --- |
| Variable | Definition |
| campaign\_id | Unique id for a discount campaign |
| campaign\_type | Anonymized Campaign Type (X/Y) |
| start\_date | Campaign Start Date |
| end\_date | Campaign End Date |

## Coupon Data

Data containing the coupons offered to the given customers under the 18 campaigns:

|  |  |
| --- | --- |
| Variable | Definition |
| id | Unique id for coupon customer impression |
| campaign\_id | Unique id for a discount campaign |
| coupon\_id | Unique id for a discount coupon |
| customer\_id | Unique id for a customer |
| redemption\_status | (target) (0 - Coupon not redeemed, 1 - Coupon redeemed) |

# Data Wrangling

1. We start by converting the date columns in transaction and campaign datasets into datetime objects.
2. We group customer transaction data by customer and merge it with customer demographics and item data sets into a transactions-by-customer dataset. We note that we have a large number of missing values for marital\_status and no\_of\_children columns and will drop these variables later.
3. We extract information from the date column of transactions dataset into 5 new columns – day, month, year, weekday, week (of the year).
4. Lastly, we map the integer weekdays to strings from Monday to Friday, and integer months to strings from January to December.

We end up with 20 columns/variables in transactions-by-customer and save this dataset in our folder.

# EDA

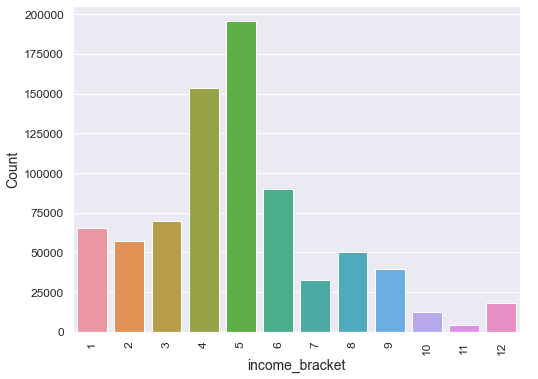
We split our exploratory data analysis into 2 main categories – univariate, looking at distributions of individual variables, and bivariate, exploring behaviors between 2 variables.

## Univariate Analysis

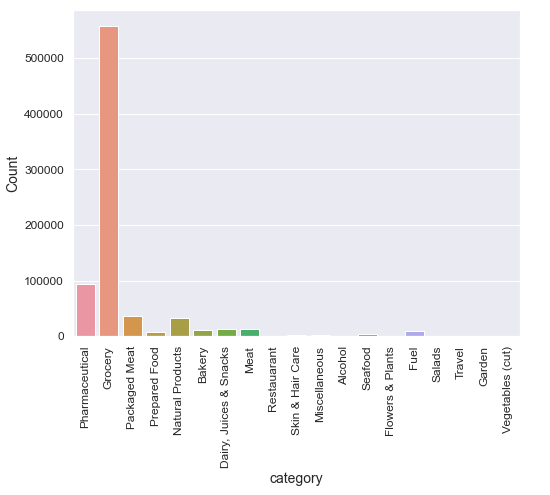
Below are some interesting insights we found during our univariate analysis:

### Most customers belong to middle income range:

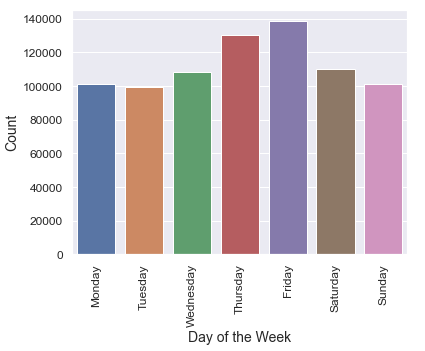
Here higher income corresponds to higher number



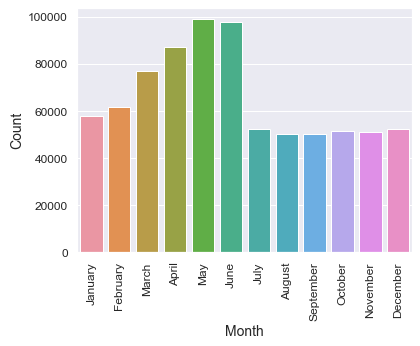
### A vast majority of items fell under Grocery category:



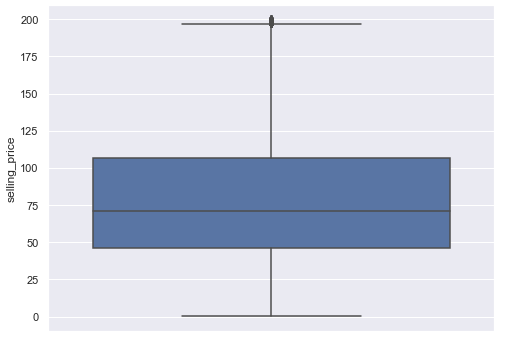
### Thursday and Friday were the busiest weekdays for sales:



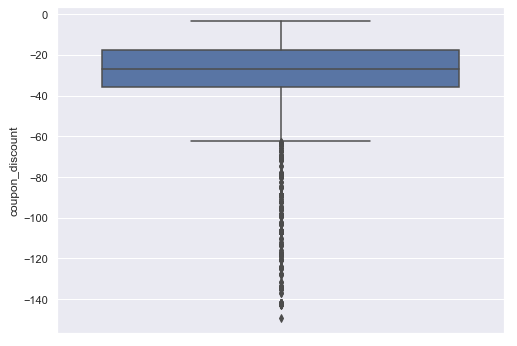
### June and July were the busiest months for sales:



### Selling price in half the transactions falls roughly between $45 and $110:

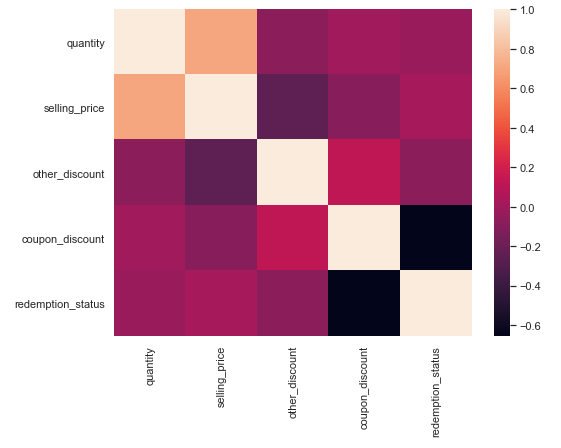


### Coupon discount in half the transactions falls roughly between $15 and $35:



## Bivariate Analysis

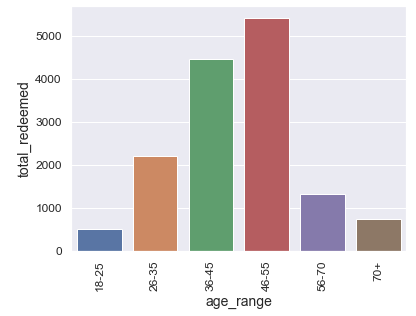
### Correlation Matrix:



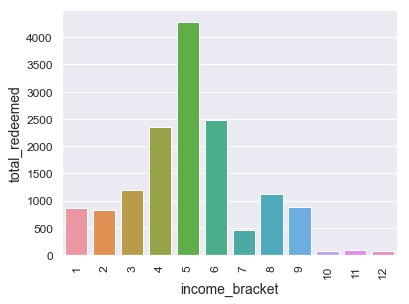
There seems to be a relationship between the amount of coupon discount and the redemption status. We also see weak relationships between selling price and coupon discount, as well as, other discount and coupon discount.

### Features vs. Redemption Status

### People between 36 and 55 years of age redeemed most coupons:



### Income bracket 5 redeemed most coupons:



# Feature Engineering

1. First, we label-encoded the ordinal columns - age range, family size, income bracket.
2. Then, we one-hot encoded the nominal columns - day, month, year, weekday, brand\_type, rented.
3. We condensed the number of categories and the one-hot encode the category column.
4. We dropped redundant and unwanted columns – no of children, marital status, brand, quantity, customer id, item id.
5. Lastly, we add dependent variable – redemption status.

# Modeling

We started by separating the dataset into train and test sets. We checked multicollinearity between all features and removed redundant columns.

We then fitted the test data to three models and compared their performances. Our metric of choice was AUC-ROC curve. Here are our calculations:

|  |  |
| --- | --- |
| Model | Area under ROC curve |
| Random Forest | 0.61 |
| Logistic Regression | 0.65 |
| XGB Classifier | 0.71 |

We decide to go with the XG boost model and proceed with tuning the model hyperparameters.

# Hyperparameter Tuning

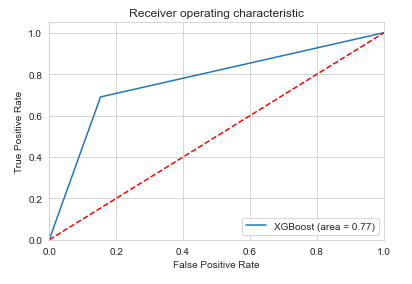
We selected the following parameter grid to use with GridSearchCV:

|  |  |
| --- | --- |
| eta | 0.01, 0.05, 0.1 |
| max\_depth | 9, 10, 11 |
| scale\_pos\_weight | 49, 50, 51 |
| max\_delta\_step | 4, 6, 8 |
| grow\_policy | depthwise, losswise |

This gave us a best score of 0.86 for the following parameters:

|  |  |
| --- | --- |
| eta | 0.1 |
| max\_depth | 10 |
| scale\_pos\_weight | 51 |
| max\_delta\_step | 4 |
| grow\_policy | depthwise |

We used these hyperparameters in our XGB Classifier model, fit it to our training data and predict the test data labels. We get a roc\_auc score of 0.77:

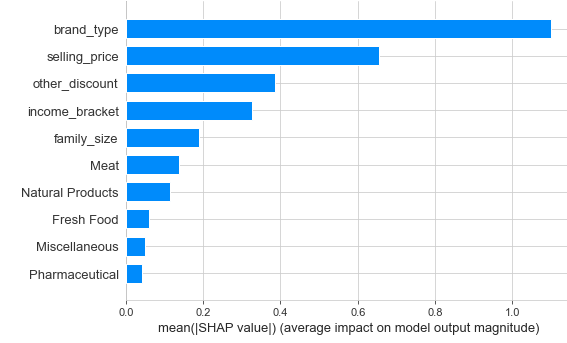


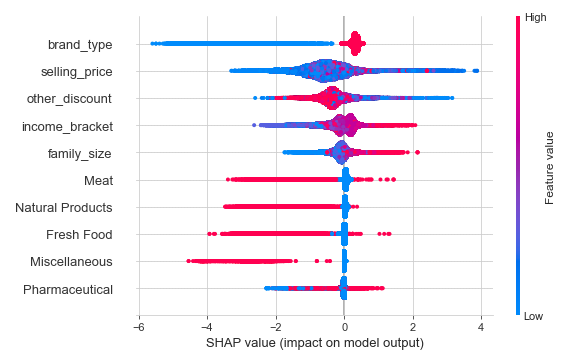
This is the confusion matrix we obtained:



# Feature Importance

At the end, we performed feature importance analysis using the SHAP module:





The two graphs above show us that the most important features that contribute to prediction of coupon redemption are:

1. **Brand Type (Established vs. New)** The summary plot shows that people are more inclined to redeem a coupon if an established brand offers it. This is expected as customers tend to place more trust in bigger, long running brands.
2. **Selling Price** This is the second most important feature. According to the summary plot, there is no clear linear relationship between the selling price of an item and the coupon redemption rate. This can also be seen in the histplot between these columns in the EDA notebook. Most coupons are redeemed between selling price of 50 and 100. Coupon redemption falls off for higher priced items. This could be because fewer coupons are offered for high priced items.
3. **Other Discount** The higher the value of other discount, the lower the coupon redemption. This makes sense as brands tend to not offer double discounts on the selling price.
4. **Income Bracket** There is no linear relationship between income bracket and coupon redemption. As can be seen from the bar plot in the EDA notebook, highest coupon redemption takes place in mid-range income brackets. This could be because the lower income households do not have enough resources to look for coupons and the higher income households don't care about spending some extra dollars.
5. **Family Size** As the family size increase, coupon redemption increases. This makes sense since families with higher number of members still have 1 or 2 earning heads and would look for every opportunity to save money.
6. **Meat** People tend to redeem more coupons on meat since meat items are high priced.
7. **Natural Products** The same is true for natural products. They tend to be higher priced, so people are prone to redeeming coupons on these.
8. **Fresh Food** Again, these are higher priced, so more coupons get redeemed on them.
9. **Miscellaneous** Again, these are higher priced, so more coupons get redeemed on them.
10. **Pharmaceutical** Again, these are higher priced, so more coupons get redeemed on them.

# Conclusion

The biggest predictors of coupon redemption are brand type, selling price and other discount. Customers are most likely to redeem coupons if established brands offer them or selling price of the item is high or no other discount is being offered on the item.

## Looking Forward

### Extracting Relevant Transactions

Currently, we are considering all transactions that fall within the dates that marketing campaigns were being offered. This includes all transactions regardless of whether a coupon was offered on that item or not. Because of complex data relationships, we did not get enough time to formulate a way to extract transactions with actual coupon discounts. Having this data will make our prediction model more accurate.

### Adding More Features

Additional details about customers could improve our model. This information could include their geographic information (city, state, zip code, etc.), payment method (card, cash, apple pay, etc.), buying behavior (first time buyer, frequency, etc.), occupation and so on.

### Time Series Analysis

Performing a time series analysis would give us more insights into coupon redemption behavior regarding trend, seasonality and the impact of holidays on the same.