ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis - Msc Data Science and Marketing Analytics

Optimizing Coupon Strategy in Fashion Retail: A Product Attributes Approach using Machine Learning

---

Author: Hang Nguyen

Student number: 700234

Supervisor: Dr. R. Karpienko

Second assessor: Dr. JBA Hemerik

Date of final version: 08/08/2024

Table of Contents

1. Introduction **1**
2. Literature Review **3**
2.1. Theoretical Background **3**
2.1.1. Redemption Rate 3
2.1.2. Product Characteristics 4
2.1.3. Seasonality 6
2.2. Conceptual Framework 7
3. Methodology **8**
3.1. Linear Regression **8**
3.2. XGBoost Model **10**
3.2.1. Decision Tree 11
3.2.2. XGBoost (Extreme Gradient Boosting) 13
3.3. Model Interpretation **15**
3.3.1. Feature Importance 15
3.3.2. Accumulative Local Effect (ALE) 16
4. Data Description **18**
4.1. Data Preparation **18**
4.2. Data Exploration **20**
5. Results **26**
5.1. Regression Model **26**
5.1.1. Model Performance 26
5.1.2. Model Interpretation 27
5.2. XGBoost Model **34**
5.2.1. Model Performance 34
5.2.2. Model Interpretation 36
5.3. Model Comparison **43**
6. Implications **46**
7. Limitations and Further Research **48**
Appendix **54**

Abstract

This research addresses the challenge of optimizing coupon allocation within outlet fashion stores, where stores still use an indiscriminate distribution of coupons across product attributes and seasonal periods. While this approach seems to be simple and convenient, it can cause huge drawbacks that erode profit margin over time. The thesis proposes a model leveraging advanced machine learning techniques to determine the optimal distribution of coupons by incorporating variables such as product characteristics, price and quantity throughout the year. By using Redemption Rate as an indicator for coupons necessity, findings reveal that the coupon distribution need to be varied by product category and season: Clothing has a high Redemption Rate in March-May and October-December, Kids' products peak in December and June-July, Sportswear and Accessories have high rates from April-June, and Shoes have stable rates year-round. In addition, Spring-Summer collections show low Redemption Rates during the peak season but increase from October-February, while Fall-Winter collections peak from May-July and are lower from November-February. For the Price group, the Discounts have the peak effect on Redemption Rate at 20%-30% discounts and decrease significantly when discounts exceed 30%, and even more so at over 80%. Higher product prices increase Redemption Rate, whereas items priced below 20 EUR show a reduction in this rate.

1. Introduction

Coupons are one of the most effective marketing tools influencing consumers (Shamout 2016). Oliver (2003) also highlights that coupons can significantly impact buyer perceptions during checkout, noting that the lack of coupons may lead to negative price satisfaction. Given their importance in online fashion retail, a strategic approach to using coupons is essential.

However, indiscriminate application of coupons across all products can lead to profit erosion, as noted by Osuna (2016). Mutius (2020) proposed a customer-centric category selection approach for promotions in loyalty reward programs. It highlights the importance of maximizing cross-category profits. Kawakatsu (2010) also points out the importance of seasonality collection of items because it allows retailers to maximize their overall profit by taking into account the seasonal factors such as seasonal collection. Those articles suggest the need to align closely with the specific product features and seasonal pattern, the sale has been boosted significantly.

Given the complex interplay between product attributes, it is crucial for fashion companies to develop a detailed coupon strategy that considers these factors. This thesis aims to build a predictive model to determine the Redemption Rate of coupons across various attributes, guided by the following research question:

Which product features drive the Redemption Rate of coupons in online retailing ?

This thesis will examine the effects of all product's features in three subgroups: Product Quantity, Characteristics, and Price. In addition, this research explores two key sub-questions related to the seasonal moderation effects:

- *Do the effects of Product Characteristics such as Main Season and Product Category differ across different times of year ?*

- *Do the effects of Product Price such as Discount and Sale Price differ across different times of year ?*

The main goal of this is to develop a model which can analyze the Redemption Rate on product features and recommend which products should offer discounts to increase sales. The reason behind using the Redemption Rate as an indicator to monitor the coupon usage is because this rate, which is the ratio of used coupons to issued coupons, can reflect how much sales rely on discounts. A high rate indicates strong consumer interest and suggests a need for more coupons to sustain or increase sales whereas a low rate may signal competitive pricing without needing extra incentives.

The model combines multiple factors such as: Product Quantity, Product Characteristics, Product price among different selling times which provides a framework for developing and applying predictive models in the context of online retail. Marketers can leverage the findings from the thesis to strategize more tailored and targeted promotional coupon campaigns. By understanding the interplay between all product attributes on Redemption Rate, marketing plans can be refined to appeal to specific products and seasonality.

Additionally, online fashion retailers can use the findings from this thesis to apply the customized coupons strategy based on the product features. By identifying which product features are most influential, retailers can tailor their coupon offerings to maximize profitability. This approach ensures that discounts are strategically applied to products that are likely to generate higher sales, thereby avoiding the profit erosion. Moreover, the predictive model in this research empowers retailers to make data-driven decisions, enhancing their ability to respond quickly to the market. By continuously analyzing Redemption Rate across different time periods, retailers can better forecast demand, schedule promotions to align with incoming periods, leading to more efficient and higher overall profitability. This level of strategic insight is crucial for maintaining competitiveness in the fast-paced online fashion industry.

2. Literature Review

2.1. Theoretical Background

2.1.1. Redemption Rate

The Redemption Rate is calculated by dividing the total number of used coupons by the total number of coupons. These ratios show how dependent product sales are by coupons. If a product's coupons have a high redemption rate, it suggests a strong consumer interest and a high demand for discounts on that product. This might indicate a need for more coupons to maintain or boost sales. Conversely, low redemption rates might indicate low interest or the product is priced competitively without needing additional incentives. Because this ratio is a relative number, it can still be informative regardless of product popularity. For instance, if products are sold in large quantities , it does not necessarily mean that the sales of them are good by themselves and they do not need coupons to attract customers. In fact, the redemption rate will show whether the majority of the sales are driven by coupons or not. If the redemption rate is low, indicating that customers are willing to buy products at offered price, without coupons. In contrast, if products have a high redemption rate, which means a large proportion of these sales are driven by coupons. This suggests that coupons are a significant factor in driving sales. This metric, therefore, provides informative insights in the process of coupons releasing.

The researcher Mutius (2021) built an experiment with redemption rates by analyzing promotional data from a leading German retailer. This experiment suggests that most profitable categories for printed promotion are those that achieve high, but not excessively high redemption rates. This requests a balance to ensure that the cost of contacting customers will not decline too much of the profit in each category. Coupons are beneficial when they strike a balance between attracting customers and maintaining profit margins. They are less effective or even detrimental when the redemption rate is either too low, indicating a lack of customer interest, or too high, suggesting that the

promotions are too generous and cutting into profits. However, in the digital context, when coupons can be easily sent to customers, this finding may not be entirely suitable and need more investigation. Nayal (2020) also conducted a meta-analysis to explore the factors effect on consumer intention to redeem digital coupons. The redemption rate is an essential metric for marketers because it reflects the effectiveness of coupons in attracting consumers and encouraging purchases.

Moreover, coupon redemption rate has a direct influence on consumer behavior and brand profitability. In the finding of Zhang (2020), design of coupons, in terms of duration and values can affect the likelihood of consumers using coupons. For instance, long duration coupons can increase seller profits and always increase consumer purchases. Research on redemption rates is vital in helping retailers optimize both customer engagement and profitability. This strategic targeting based on redemption rate can lead to more successful marketing campaigns and better financial outcomes for retailers.

2.1.2. Product Characteristics

In terms of product characteristics, there are several factors that are essential variables such as: Product Category, Main Season, Stock-tier, Size Range Type, etc. Research from Ignacio (2016) suggests a method to map the categories of brands that a retailer should promote depending on whether the objective is to increase customer loyalty rewarding clients for buying brands or entice them to buy in categories that they are not yet purchasing at the store. Additionally, Subhojit (2009) suggests that the right combination of product and promotion enhances sales more effectively. Therefore, it is essential to evaluate the promotion effectiveness by considering the product category and its features. Even though those papers do not suggest the direct effect of product category on coupon redemption rate, those still suggest the relationship between product category and coupon

campaign. Moreover, Wagner & Mokhtari (2000) found that Seasonal Collection can affect the marketing effectiveness in the fashion industry. However, this effect is moderated by the weather pattern or fashion events. For example, promoting Summer collection will have different marketing results in the hot months than the cold months. This suggests that the effect of the Main Season of the collection can significantly affect the Redemption Rate, and this effect can be found across seasonality.

Regarding Product Quantity, research by Guo (2021) finds out the impact of product display quantity on consumers' online purchase proneness. It suggests that product display quantity significantly increases online purchase intention. When consumers see a larger quantity of a product, they are more likely to buy. The research also suggests that seeing more of a product reduces the psychological discomforts or perceived cost associated with spending money on that product. This effect is crucial as it also impacts the redemption rate of coupons. When consumers see a low quantity of a product, they may be less likely to seek coupons because of the scarcity, leading to a lower redemption rate. Therefore, increasing the visible quantity of products can not only drive more purchases but also enhance the effectiveness of promotional strategies by encouraging more coupon use.

In the realm of pricing, Thomas (2003) points out that premium priced products can lead to increased spending by consumers. His study proves that more than one out of four coupon users actually increased their expenditures when using coupons compared to their spending without coupons. This phenomenon is because coupons alter the perception of price, making a high price seem like a mixed gain rather than a net loss and becoming more attractive to customers. Hsu (2011) also mentioned that promotions such as discounts or special pricing have a positive effect on consumer choice for high tier brands than low tier brands. The explanation for this can come from the perception of greater

value or a rare opportunity to purchase a premium product at an affordable price, which can be quite appealing to customers.

Additionally, according to Kim (2008), service coupons can significantly alter consumer perception of the trade-off between price and perceived quality. The coupons can enhance the perceived values of a service by making the price more attractive without changing the perceived quality. This means consumers may be affected by the coupons, making them believe that they have remained high quality service with lower prices. While this research is mainly in service sectors, it seems to be quite potential to apply this to the fashion industry. Offering coupons in the fashion industry could similarly affect consumer behaviors by making the prices more appealing while maintaining the perceived quality of the products. Hsu (2011) also discovers how promotion influences consumers' decisions on price decrease and perceived quality. This suggests that promotions need to be strategically planned to scope all these aspects.

2.1.3. Seasonality

In the fashion industry, selling time is a crucial factor affecting the Redemption Rate. This industry is known for short product life cycles, high demand volatility and low predictability, therefore, it is quite challenging in management (Shen et al., 2016). Nudell (2023) highlighted that the fashion calendar is meticulously organized by months, with events aligned with seasonal cycles. The commercialization of this calendar, including the monthly scheduling of fashion events, significantly influences consumer culture, as seen historically with American holidays and their impact on consumer behavior (Schmidt, 1991). These insights suggest that the months of the year are a critical variable affecting the relationships between other factors.

2.2. Conceptual Framework

The main focus on this part is the conceptual framework, which serves as the core of the thesis. By combining the theoretical background above, Figure 2.1 shows an overview of the concept as well as all key factors of this model. At its main part is Product, which concludes 3 main factors: Product Characteristics, Product Quantity and Product Price. Those factors have a significant effect on Redemption Rate, the central focus.

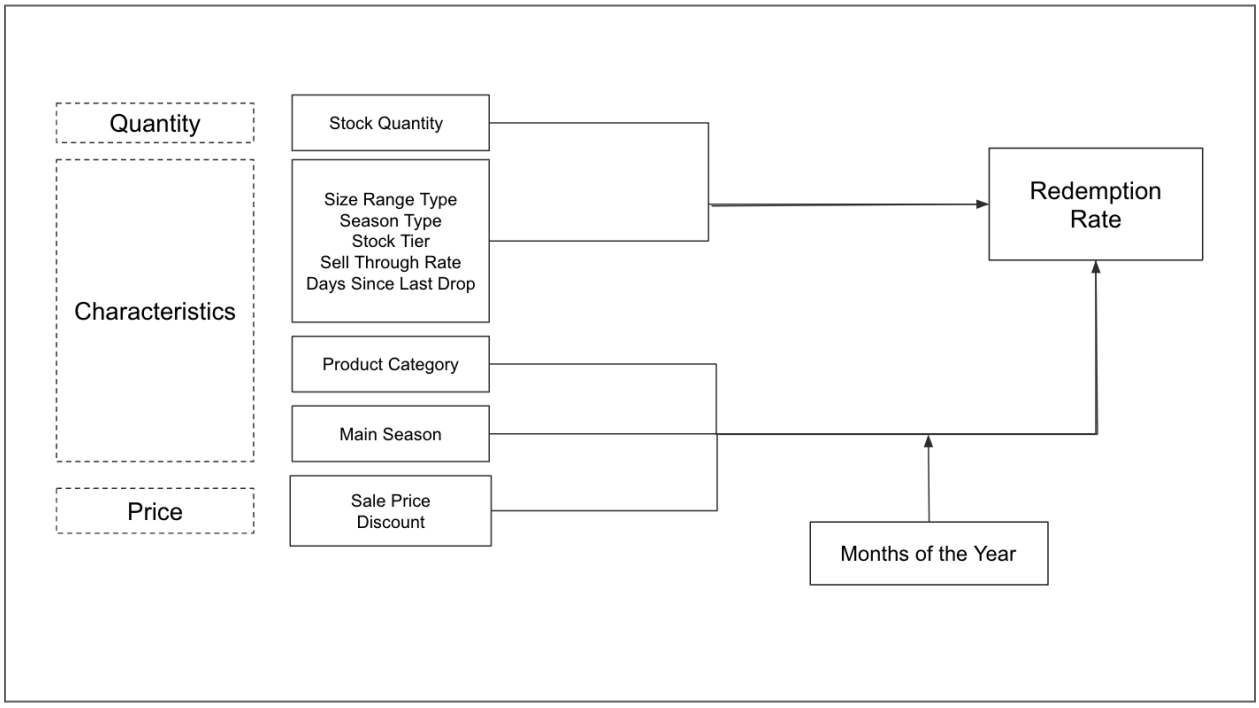

Figure 2.1. Conceptual Framework

In addition, the conceptual framework also marks the moderating role of seasonality (represented by months of the year). This variable acts as a moderator, affecting how other factors influence the Redemption Rate. This highlights the importance of timing in the fashion industry, where consumer demand is significantly influenced by seasonal cycles and monthly events. Therefore, the volume of effect from Product Category, Discount and Price to Redemption Rate can be different across months. Moreover, for Main Season variables, the effect on Redemption Rate from Spring Summer or Fall

Winter collection can be different across each month, typically aligning with the changing weather patterns or seasonal preferences.

3. Methodology

3.1. Linear Regression

Linear Regression is a widely used statistical technique for modeling the relationship between one dependent variable and one or multiple independent variables (Senter 2008). Depending on the number of independent variables, linear regression can be classified into two types: Simple Linear Regression and Multiple Linear Regression. Since this research examines multiple factors affecting the Redemption Rate, Multiple Linear Regression has been chosen for the application.

Many studies have utilized regression analysis in marketing, particularly in the area of coupon usage. Barat (2007) used a regression model to test hypotheses about the factors affecting consumers' intentions to redeem coupons. This study applied multivariate regression to assess the influence of psychological, socio-economic, and behavioral factors on coupon redemption intentions. Similarly, Mutius (2020) uses a regression model to estimate the return on marketing investment (ROMI) by considering the impact of promotions on cross-category profits and redemption rates. Reibstein (1982) also created a model to predict coupon redemption rates, employing regression analysis to identify key influencing factors. Those papers suggested that regression can be one of those methods suitable to discover the impact of factors in Redemption Rates.

The goal of Simple Linear Regression is to predict the value of one dependent by an independent variable. To be more precise, it tries to capture and explain the variance of the dependent variable by the change of the independent variable. The task of simple linear regression is to exactly determine the straight line which best describes the linear

relationship between 2 variables. The main attempt of this is to make the error in estimation as small as possible.

Unlike Simple Linear Regression, Multiple Linear Regression used more than two independent variables to train the model. The goal is to estimate the dependent variable based on several independent variables. The equation for this calculation of multiple regression is obtained k dependent variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

The formula is explained with:

y: is the Dependent Variable

$x_{1,} x_2, \ldots, x_k$: are the Independent Variables

β_0: is the Intercept

$\beta_1, \beta_2, \ldots, \beta_k$: are the Coefficients

The coefficients can now be interpreted similarly to the Linear Regression. If an independent variable changes by one unit, the associated coefficient indicates by how much the dependent variable changes. Additionally, the regression coefficients are used to compare the relative importance of each independent variable on a dependent variable, therefore, it is essential to standardize the deviation of each dependent variable before comparing them. Standardizing a regression model involves transforming the variables to have a mean of zero and a standard deviation of one. This process helps in comparing the impact of different variables by removing the units of measurement.

Besides coefficient interpretation to see the volume of impact from independent variables on target variables, it is also important to check the significance of them. This test is used

to rule out the possibility that regression coefficients are not just random and have completely different values on another sample. This test is based on t-distribution, which checks if the slopes (regression coefficients) differ from zero in the population. If one coefficient has the p-value is lower than 0.05, it has enough evidence to reject the null hypothesis that coefficient is zero (H_0: $\beta = 0$). In contrast, id the p-value is higher than 0.05, it can not reject the null hypothesis that coefficient is zero (H_0: $\beta = 0$).

Additionally, in order to check the model performance, Root Mean Squared Error (RMSE) is used to check the model accuracy. It measures the avenger difference between predicted values and their actual values. The lower the value of RMSE, the better the model is. It can also be mentioned as the standard deviation of the error since it is the square root of the error variance. Besides RMSE, R-squared and adjusted R-squared are also used to evaluate model performance. R-squared is a statistical measure that estimates the proportion of variance in the dependent variable that can be explained by the independent variables. It can be considered as an indicator showing how well the data fit the regression model (the goodness of fit). However, R-squared can only increase or stay the same when new predictors are added to a multiple regression model. This means that even irrelevant variables can cause R-squared to rise or remain unchanged. To address this issue, Adjusted R-squared is used, as it takes the number of predictors into account. Adjusted R-squared increases only if the new predictors enhance the model and decreases if they do not contribute meaningfully.

3.2. XGBoost Model

XGBoost has been widely used in studies related to coupon usage. For example, Yan (2018) created a model combining Random Forest and XGBoost for targeting e-commerce coupon users, performing well on the Alibaba Coupon Usage Forecast. Duan (2018) developed a personalized coupon usage prediction model using XGBoost. This model achieved a relatively high AUC value of 0.8496. Additionally, Ren (2021)

combined customer segmentation with XGBoost, using an improved RFM model (RFS) and K-means algorithm to predict coupon usage. Song (2018) introduced the Digital Coupon Use Prediction Model (DCUPM) based on XGBoost, which handled large-scale data and provided precise predictions, highlighting that high-scoring features can guide targeted coupon delivery. These studies suggest the effectiveness of XGBoost in marketing and coupon prediction.

3.2.1. Decision Tree

In order to fully grasp an in-depth understanding of XGBoost, it is essential to understand its original method, Decision Tree. The Decision Tree is a non parametric supervised learning algorithm, which can be used for both classification and regression tasks. The goal of Decision Tree is to create a training model by learning simple decision rules from training data.

The process of Decision Tree includes several components (Figure 3.1). In the beginning, the whole data set is considered as the Root Note, this entire sample gets divided into two or more homogeneous groups. This division process is called the Splitting process, this process continues to the next Decision Node, creating Branches or Sub-Tree from each Decision Node. Each Node in the tree acts as a test case for some attributes, and each edge descending from the node corresponds to the possible answers to the test case. Until the last Decision Nodes which can not split, those become Leaf (or Terminal Node).

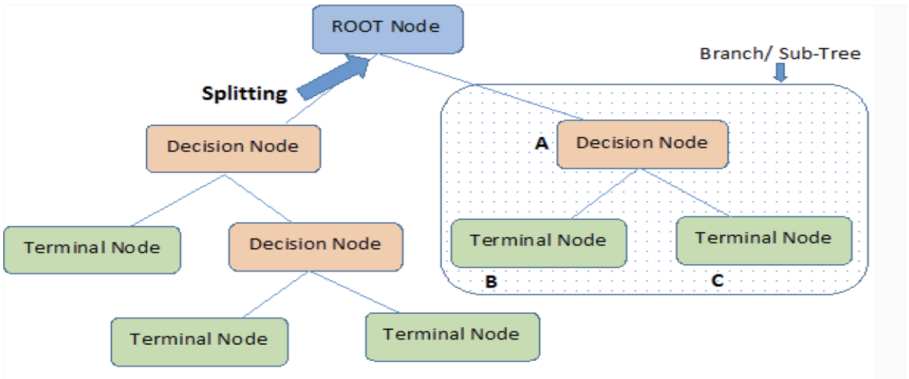

Figure 3.1. Decision Tree components

However, there is one key question for this process, which is to identify which attributes to consider as the Root Node and each level of that variable. Since there are multiple attributes that can be involved in the model, just randomly selecting any node to be the root can't solve the issue. If it follows a random approach, it may give us bad results with low accuracy.

In order to answer that question, Decision trees use various algorithms to decide to split a node into more and more sub-notes. The main goal of this process is to increase the homogeneity of the next sub-nodes. In other words, it tries to increase the purity of the nodes with respect to the target variable. There are multiple algorithms to access the purity in each Node, such as: Entropy (H), Information Gain, Gini Index, etc. However, since the target variable is continuous variables, it can not calculate those indicators. It needs a different measurement, which tells how much the predictors deviate from the original target and that's the entry-point of Mean Square Error.

Basically, in the Regression Tree algorithm, it does the same thing as the Classification trees. But, it tries to reduce the Mean Square Error at each child rather than the entropy. The approach of this algorithm is slightly different between continuous independent variables and categorical independent variables. For continuous variables, the method finds the value point in the independent variable to split the data-set into 2 parts, so that the MSE is minimized at that point. For categorical variables, this uses the binary approach, which converts categorical variables using one-hot encoding or label encoding. It considers each categorical term as 0 or 1 and calculates the MSE for each of them. The points that minimize the MSE are calculated for all variables in the dataset. Among those variables and the points calculated for them, the one that has the least MSE would be chosen as the first Root Node. This process continues for all next Decision Nodes until the MSE can not be improved after the split anymore.

However, this Regression Trees method is prone to be overfitting. In order to reduce the MSE, the decision tree needs to split the dataset into a large number of subsets to the point where a set contains only one row or record. Even though this might reduce the MSE to zero, this is obviously not a good thing. Therefore, this required a Max Depth parameter to control the maximum depth which decision is allowed to grow.

3.2.2. XGBoost (Extreme Gradient Boosting)

Before discussing XGBoost, it is essential to mention the Gradient Boosting Method, which is the origin of the XGBoost. Gradient Boosting Method is the method derived from the Ensemble Learning method. The main difference between Gradient Boosting and Decision Tree is that, Gradient Boosting is the form of an ensemble of weak prediction models (normally are Decision Trees). It starts by fitting the initial model to the data. Then a second model is built but only focuses on predicting the residual of the first model. The combination of these two models is expected to be better than either model alone. Then this process of boosting repeats multiple times. Each successive model attempts to correct for the shortcomings of the combined boosted ensemble of all previous models. Those decisions are created sequentially at a certain number of decision trees, with the main goal is to reduce Loss Function as shown below:

$$L(f) = \sum_{i=1}^{N} L(y_i, F_{m-1+\Delta_m}) \quad (1)$$

The function *(1)* shows the Total Loss Function, $L(f)$ represents the sum of all individual loss functions over all N trees. It measures the difference between true value y_i and the prediction F_m. The calculation for F_m is calculated as:

$$F_m = F_{m-1} + \eta\Delta_m \quad (2)$$

This equation *(2)* represents the update rule for the predictions. The new prediction is $F_m$obtained by adding a scaled version of the update Δ_m to the previous prediction F_{m-1}.

The scaling factor η is the Learning rate Eta, which controls the size of the step we take in the direction of the gradient to minimize the loss function.

Similar to the Gradient Boosting Method, the main goal of XGBoost is to minimize the Loss Function. But what makes it outstanding in comparison to Gradient Boosting Method is the ability of handling overfitting by adding a smart twist of penalty terms $\sum_{k=1}^{K} \Omega(f_k)$ to the Lost Function, which create a new function:

$$obj(\theta) = \sum_{i=1}^{N} L(f) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

While the first part $\sum_{i=1}^{N} L(f)$ is the Loss Function as in Gradient Boosting Method, the second part $\sum_{k=1}^{K} \Omega(f_k)$ is the Regularization Term. The Regularization Term is then defined by:

$$\Omega(f_k) = \gamma T + 0.5\lambda \sum_{j=1}^{T} w_j^2 \quad (4)$$

In the equation *(4)*, it has T as the number of the total leaves in each tree, *w* as the vector of scores on the leaves of a tree. In this equation, it includes 2 regularized parameters, namely Lambda (λ) and Gamma (γ). While Gamma represents the minimum loss reduction needed to split a leaf node for each tree, Lambda prevents model from fitting the training data by adding a penalty term,

In conclusion, there are five crucial hyperparameters in XGBoost that need to be focused on. The first is Maximum Depth, which originates from the Decision Tree algorithm and controls the complexity of the model. The next two important hyperparameters are the Number of Trees and the Learning Rate (Eta), both derived from Gradient Boosting, which help capture complex patterns within the data. Finally, the Regularization

Parameters, Lambda and Gamma, are vital for mitigating overfitting and enhancing the model's generalization capabilities.

To find the optimal combination of these four parameters, I employed the Cross Validation method for hyperparameter tuning. Specifically, I used K-fold cross validation, which splits the training data into k equal parts. In each iteration, one part is randomly selected as the test data set while the remaining k-1 parts serve as the training data set. The model is then fitted on the training set and evaluated on the test set. This process is repeated k times, ensuring that each part is used as the test set once. During each iteration, the Mean Squared Error (MSE) scores are recorded for various parameter combinations, enabling the identification of the best-performing parameters.

3.3. Model Interpretation

XGBoost model is considered as a high level of accuracy model, however, it is also considered as a Black Box Method in Machine Learning. In other words, this method only gives the results without the explanation of how it makes decisions. The internal processes used and the various weighted factors remain unknown. Therefore, in the scope of this thesis, I decided to use 2 black box interpretation methods: Feature Importance and Accumulative Local Effect to explain the results.

3.3.1. Feature Importance

In XGBoost, feature importance is considered as the significance of each in predicting the target variable. The XGBoost function itself provides several methods to evaluate the feature importance, helping in the model's explanation. There are 3 main approaches to evaluate the feature importance includes: Gain, Cover and Frequency (or Weight)

Cover measures the number of observations affected by a feature, averaged over all the splits where the feature is used. Features with higher cover values means they will have a

larger impact portion of the dataset, indicating their broader influence on the model. Beside Cover, Frequency counts the number of times a feature is used to split the data across all trees in the model. Features that are used more frequently are considered more important. This is because they contribute to more decisions within the model. Lastly, Gain measures the improvement in accuracy thanks to features to the branches it is on. It represents the average gain of the splits which use the feature. The higher gain values are, the more significance of them in reducing the error. Additionally, the Gain is the most relevant attribute to interpret the relative importance of each feature.

3.3.2. Accumulative Local Effect (ALE)

The Accumulative Local Effect (ALE) describes how each feature can influence the prediction of a machine learning model on average. However, what makes ALE outstanding in comparison to other Black Box Interpretation Methods is the ability to see the effect of separate independent features on target variables while taking into account the correlation between it with other variables in the dataset.

To estimate the local effects, the features are divided into many intervals and compute the differences in the prediction as formula below:

$$f_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{x_j \in N_j(k)} [f(z_{k,j}, x_{-j}^{(i)}) - f(z_{k-1,j}, x_{-j}^{(i)})] \quad (1)$$

As mentioned above, the investigating feature is divided into many ranges, which create different grid values z. The ALE model calculates the difference in predictions where the feature of interest is replaced by a different grid value z. The difference in prediction is the “Effect” the feature has for an individual observation in a certain interval. The second total in equation *(1)* adds up the effect of all observations within an interval as neighborhood $N_j(k)$. Then it divides this sum by the number of instances in this specific

interval which calculate the average difference of the predictors for this interval. The left total adds up the average effect across all intervals. Therefore, the (uncentered) ALE of a feature value that lies, for example, in the third interval is the sum of the effects of the first, second and third intervals.

After calculating the effect of the feature of all observations, it is requested to center the effect so that the mean effect is zero. This step is simply to calculate the difference between the ALE effect of each observation and the mean ALE effect of all observations. Therefore, the result can interpret as if the ALE affect = a at $x_i = b$ indicates that when the j-th feature takes on the value of b, the prediction is a unit higher than the average prediction.

ALE plots are considered to be unbiased, which means they still work when the features are correlated. Other plots such as PDP (Partial Dependence Plots) fail in this because they estimate some combinations of feature values which are unlikely or impossible to happen. Additionally, the ALE plots are also better in computer efficiency, since the largest possible number of intervals is the number of instances with one interval per instance. The best advantage of this model is that the interpretation of this is clear since the ALE plots are centered at zero. This makes their interpretation nice, because the value at each point of the ALE curve is the difference to the mean prediction.

However, ALE plots do not solve all issues for Black Box Interpretation. ALE plots can swiftly fluctuate with a high number of intervals. In this case, reducing the number of intervals makes the estimates smoother but also having the risk of losing some important relationships. There is no perfect number of intervals the model should have, if the number is too small, the ALE plots may not reflect all possible relationships. If the number is too high, the curve can be unstable.

4. Data Description

4.1. Data Preparation

The initial step in data preparation is collecting data. In order to have comprehensive data from various aspects, it is required to select data from multiple sources. This is also called Extract - Transform- Load (ETL) process, where data is extracted, transformed and loaded to the storage. This process helps to gather data from separate sources to one united data. The model of tables and the relationship between them are shown in Figure 4.1. Four main tables are mentioned, including: Products, Orders, Coupons and Customers.

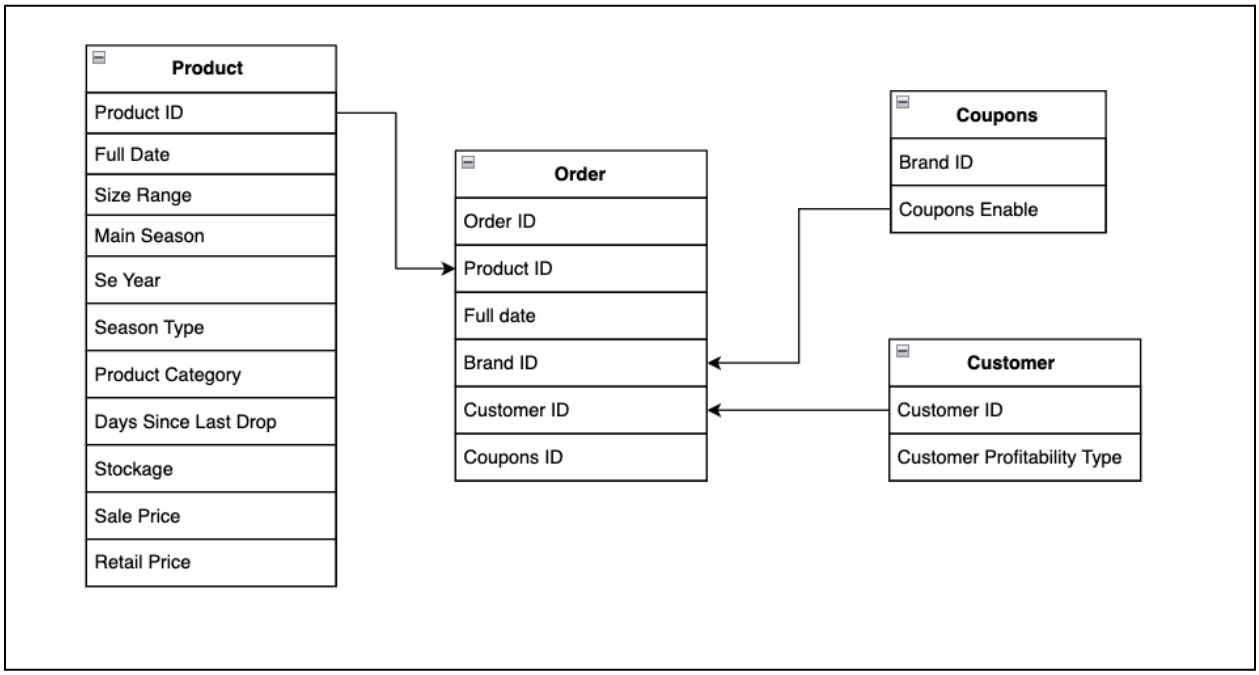

Figure 4.1. Entity relationship diagram

Once the data was gathered, it was cleaned by removing duplicates and correcting errors. The original data was daily data for each product, which caused a large amount of data, challenging the computer efficiency. To handle this issue, I decided to do aggregation and transform all the data from daily level to monthly level. This means that for each row,

data was collected and recorded once per month, capturing the attribute if it was during that month instead of once per day. This can reduce the size of the data to a more manageable level. Finally, after having all necessary features, all the data was loaded into the target storage which variables are explained in Operationalization Table (Table 4.1)

Table 4.1. Operationalization table

Variable	Operationalization	Data Type	Values	Missing values
(Y) Redemption Rate	Percentage of products sold via coupons in total sales Redemption Rate = Total products sold by coupons / Total Sales	Continuous	Min: 0 Max: 1	0%
(1) Product Characteristics				
Stockage	Weighted average days for current items on stock*	Continuous	Min: 0 Max: 1317	3.98%
Size Range Type	Type of size that the products available	Categorical	Full size range: Full size from XS to XXL, Last size range: Only one size left, Broken size range: Only have a few sizes	0%
Sell Through Rate	Percentage of inventory sold monthly**. If the Sell Through rate is high, it means the products are favored by consumers	Continuous	Min: 0 Max: 0.86	0%
Main Season	Main season of the products	Categorical	NOS: Never out of Season SS: Spring Summer FW: Fall Winter	0%
Season Type	Shows whether or not a product is still fashionable or not	Categorical	Fashionable, Old season, One season old	0%
Product Category	Type of fashion products	Categorical	Sportswear, Kids Maternity Wear Shoes, Clothing Accessories	1.4%
Stock Tier	Quality of the products	Categorical	Gold, Sliver, Bronze, Reproduction	0%

Days Since Last Drop	Days since products stocked or restocked on the website	Continuous	Min: 0 Max: 1310	1.84%
(2) Product Quantity				
Stock Quantity	Total number of products in the stock	Continuous	Min: 1 Max: 3605	0%
Total Sold 7 Days	Total number of sold products for the last 7 days	Continuous	Min: 4 Max: 1308	0%
Total Sold 30 Days	Total sold products for the last 30 days	Continuous	Min: 30 Max: 2974	0%
(3) Product Price				
Retail Price	The average monthly retail price of the products.	Continuous	Min: 6.14 Max: 1095	0.1%
Discount Live	The average monthly discount of the products.	Continuous	Min: 0.1 Max: 0.87	0%
Sale Price	The average monthly sale price of the products.	Continuous	Min: 2.98 Max: 431.9	2.8%
(4) Selling Time				
Month	Month of the year	Categorical	Min: 1 Max: 12	0%

** Stockage is calculated by Total (Quantity of Drop * Days Since Drop Lived) / (Total Days Since Drop Lives)*
*** Sell Through Rate is calculated by Total number of units sold / Total number of stocks on hand*

4.2. Data Exploration

The first step in the descriptive analysis was checking missing values. The initial data comprised 91,943 observations, with each observation representing the monthly status of a specific product. As illustrated in Table 4.1, there were several missing values in the data set after aggregation, however, the percentage of missing values in the dataset was relatively low. The highest percentage of missing values was 3.62%, indicating that removing those values will have an insignificant effect on the data integrity. After

cleaning the data and removing the missing values from each variable, the data left with 88,005 observations.

The second step in the data cleaning process was checking the correlation relationship among the variables. Although the dataset contained multiple variables, including all of them may not be beneficial, as some were highly correlated, as shown in Appendix A. For example, there was a high correlation between Stockage and Days Since Last Drop, since they had a correlation of 0.94. This can be explained by the fact that the metric calculating Stockage in fashion is partially based on Days Since Last Drop. Additionally, the number of High Value Customers was highly correlated with Total orders, Total Coupons, suggesting that high valued customers seem to be more attracted by coupons. To improve the efficiency and accuracy of our analysis and align with the main goal of the thesis, some highly correlated variables were removed. It can help to reduce the noise of the model, which improves the model to become more accurate.

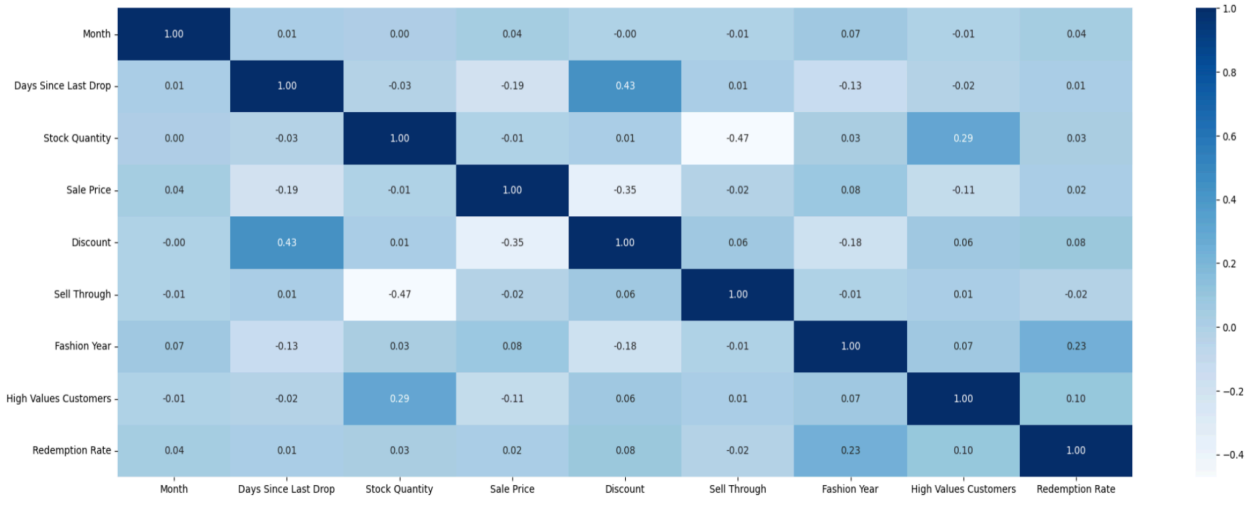

Figure 4.2.Correlation Matrix After Cleaning Data

After selecting essential variables between highly correlated variables, the overall correlation between dependent variables was relatively low, from 0 to 0.36, which was an acceptable number for the model. The initial analysis showed that all independent

variables have a significant correlation with the redemption rate. However, none of these correlations were too high to dominate the prediction of the dependent variable, suggesting a balanced influence from each variable on redemption rate. This indicated that multiple factors actually contribute meaningful insights to the model performance, enhancing its robustness and accuracy.

The next step for data exploration was detecting outliers. Those can significantly decrease the accuracy and reliability of the model. Even though tree-based models are generally robust to outliers, Sinwar (2015) points out that outliers can still have a considerable effect on model performance. They can potentially cause the incorrect or misleading results by distorting the true pattern within the data. For instance, in the context of decision trees, outliers can disproportionately affect the splits and decision rules created by the model, leading to overfitting where the model becomes too tailored to these unusual data points rather than generalizing well to new data. This can reduce the model's overall predictive performance and robustness.

Additionally, the majority of the variables have extreme values that can be considered as outliers. To address this issue, a Quantile method is applied in order to clean the data. This method creates lower and upper boundaries based on 25th and 75th percentiles, respectively. This approach effectively removes only extreme values, thereby retaining the majority of the data while reducing the influence of outliers.

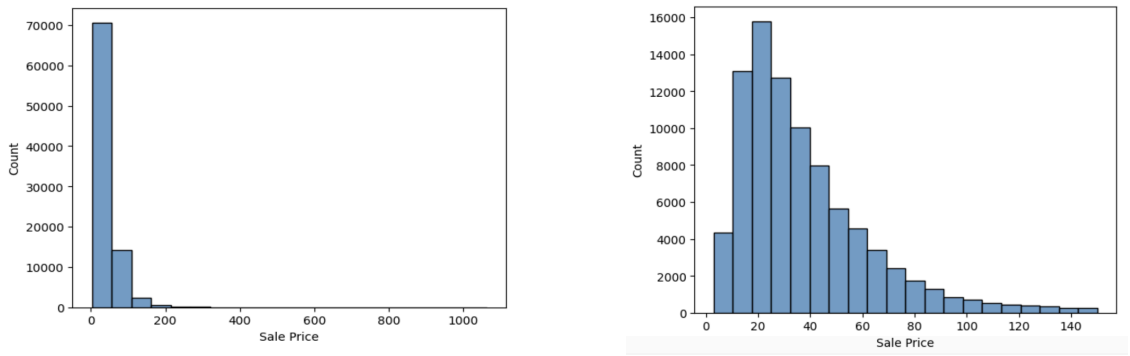

Figure 4.3. Sale Prices Distribution Before and After Cleaning Outliers

One exceptional case that this method can not be applied was the Sale Price variable because this variable was extremely left-skewed (Figure 4.3). Applying the Quantile method could result in the removal of important low value observations that were characteristics of skewed distribution. After thorough exploration, it was found that only 1.41% observations had the sale price higher than 150 EUR, therefore, I decided to only keep those products lower than 150 EUR. After cleaning the outliers, the distribution of all variables are presented in Figure 4.4 below.

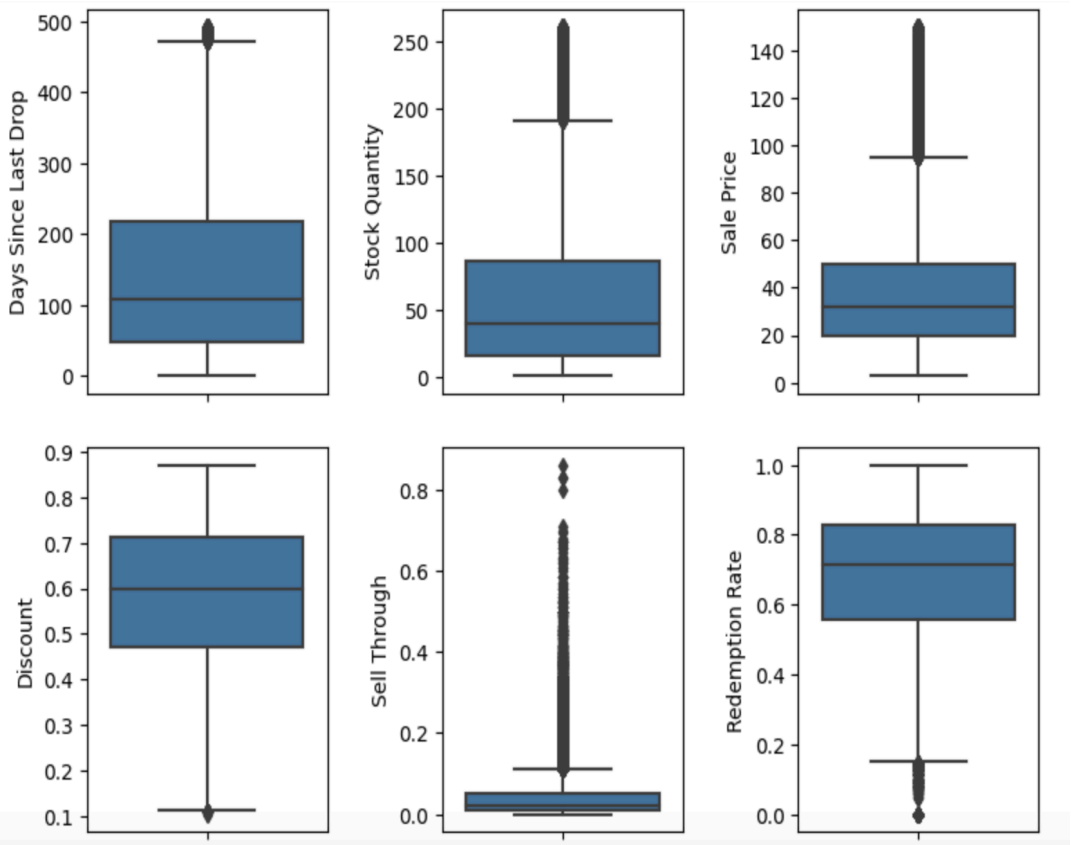

Figure 4.4. Box Plots of Numerical Variables

In conclusion, after the cleaning process, the data remain 73,299 observations, with 5 numerical variables and 5 categorical variables. The summary statistics for numerical are shown below (Table 4.2).

Table 4.2. Summary Statistics

	count	mean	50%	std	min	max
Month	73299	6.3	6	3.75	1	12
Days Since Last Drop	73299	144.61	108.68	116.72	0	493
Stock Quantity	73299	59.58	39.48	57.92	1	261.84
Sale Price	73299	38.44	31.67	25.45	2.98	150
Discount	73299	0.58	0.6	0.16	0.11	0.87
Sale Through	73299	0.04	0.02	0.06	0	0.86
Redemption Rate	73299	0.67	0.71	0.21	0	1

Additionally, it is essential to check the distribution of the target variable - the Redemption Rate (Figure 4.5). Overall, the dataset shows a right skewed distribution. The average of the data is 0,67 and the median is 0.71, suggesting a significant proportion of the products has a high redemption rate. This means that there is a significant amount of sales driven by coupons.

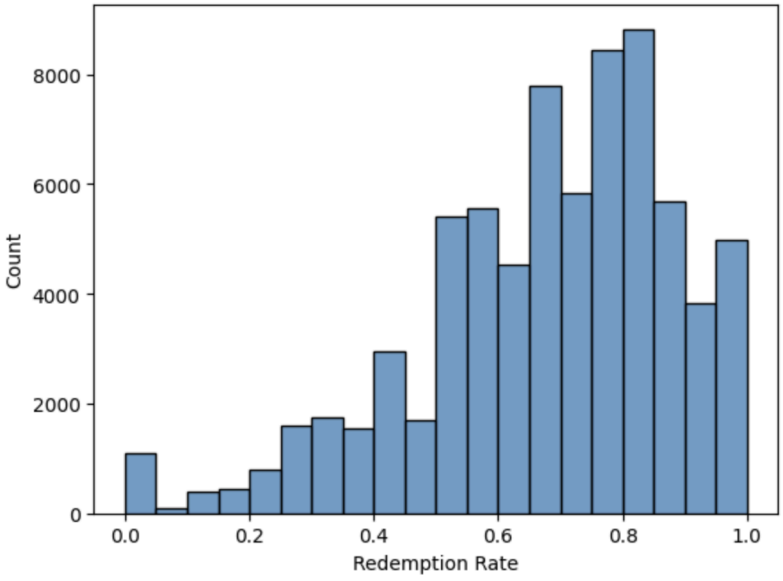

Figure 4.5. Distribution of Redemption Rate

Addition to the numerical variables, the distribution of categorical variables also has some informative findings. Overall, while the dataset does not show any extreme

imbalance, there are some noticeable disparities in the distribution of certain categories (Figure 4.6). For instance, Size Range Types are not evenly distributed across three size categories. Majority of products fall under Full Size Range value, around 50,000 observations, accounting for 67.6% of observations while Broken Size Range and Last Size Range only account for 25.4% and 6.8% respectively. However, this aligns with the real fashion business context, where maintaining a full size range is crucial for maximizing customer satisfaction and sales. The same case with the Main Season variable, when Fall Winter Collection and Spring Summer collection contribute 47.7% and 40.9% while Never out of Season Products (NOS) only has 11.4%. This is due to the common feature of the fashion industry, when items are typically released to suit a specific season rather than being flexible for year-round use.

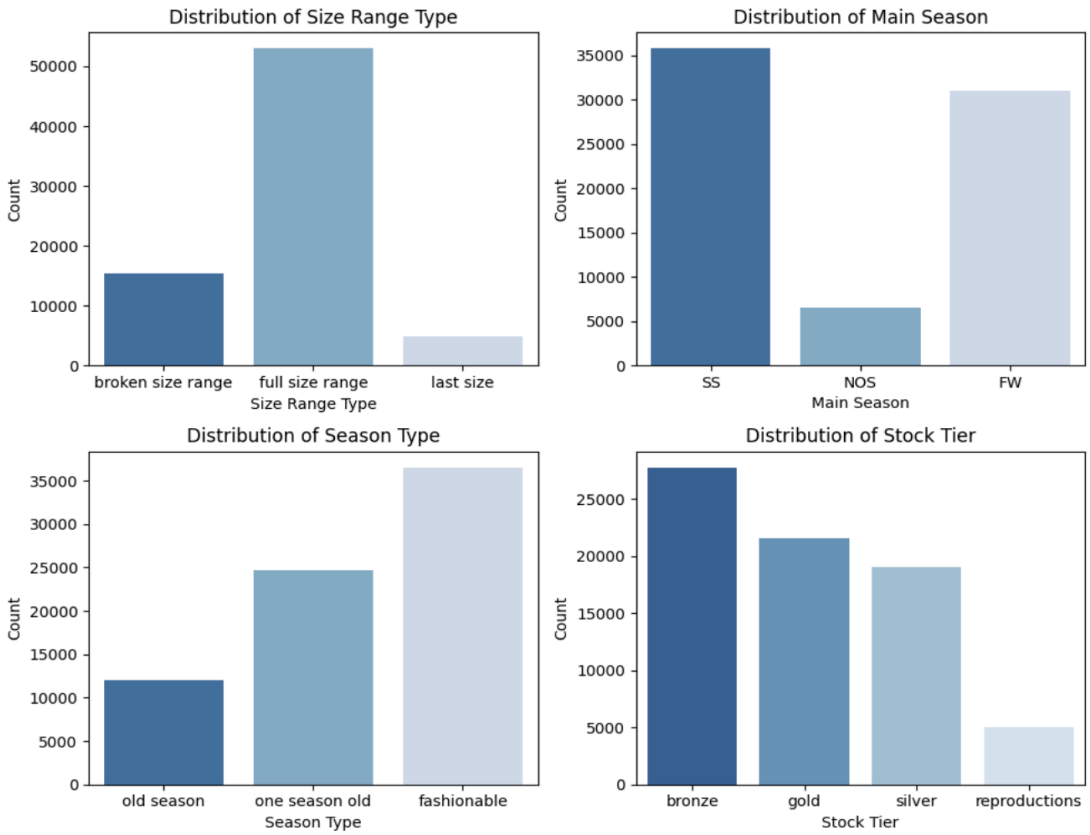

Figure 4.6. Distribution of Categorical Variables

However, there is one categorical variable that has an extreme imbalance, which is Product Category (Figure 4.7). It is common sense that clothing are the main products that stores sold, which account for approximately 78.2%. This reflects the main core focus of the store. Sportswear and Shoes follow, ranking second and third respectively, highlighting their significant but secondary role. Products for kids, both boys and girls, also constitute a notable portion of 5.68%. Lastly is the Accessories, which only account for 2.32% of the products.

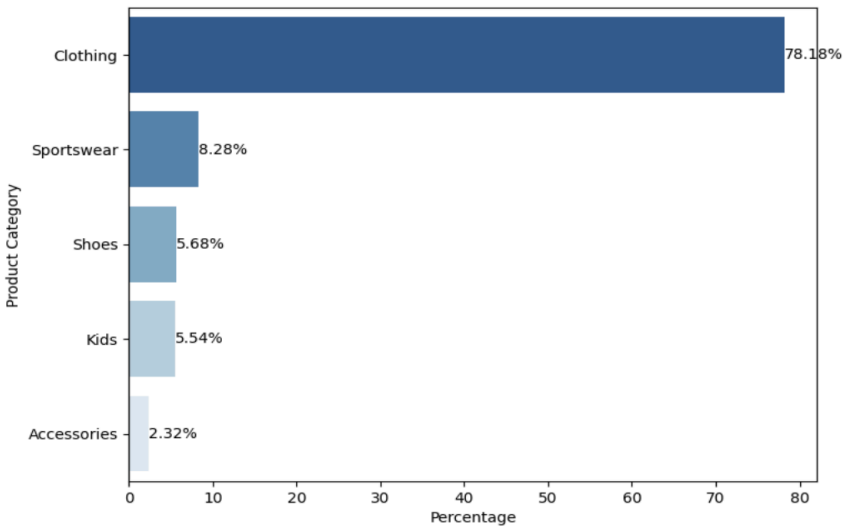

Figure 4.7. Product Category Distribution

5. Results

5.1. Regression Model

5.1.1. Model Performance

In order to predict redemption rate for each product, the linear regression method was applied for two models below. The first model is a simple linear regression without any interaction terms, resulting in an root mean squared error (RMSE) of approximately 0.274 for the training data and 0.288 for testing data. This model showed no sign of overfitting, however, the RMSE were relatively high. One possible reason could be that the model did not account for interaction between variables, which may increase the

error. Additionally, the R-squared value of this model was low, which was only around 0.46.

The second model had a slightly higher R-squared of 0.48, however, it was still a relatively small number. This model included all possible interaction between variables which helps to slightly decrease the RMSE to 0.232 for training data and 0.238 for testing data. Even so, the RMSE is still relatively high. One explanation for this is that even calculating the interaction terms, it still can not capture the nonlinear relationship between the dependent variable and independent variable.

To avoid multicollinearity, the baseline for this regression model includes: Products had Last Size Range, in Reproduction Stock Tier and in Shoes category. They were Never Out of Season Products and were sold in January.

As mentioned above, the regression model having interactions gets better accuracy, therefore, this interpretation part mainly focuses on this model. The coefficient table of two models is added in Appendix B.

5.1.2. Model Interpretation

For the sake of comprehensiveness, the interpretation will follow the conceptual framework and categorize the variables into three groups: Product Quantity, Product Characteristics, Product Price.

a. Product Quantity

In general, the coefficient for this variable aligns with common expectations. For instance, as shown in Table 5.1, the coefficient of Stock Quantity is 0.093, with p-value less than 0.01, indicating a significant positive relationship between Stock Quantity and Redemption Rate. This is rational because if the quantity of a product is high may cause

the lack of perceived scarcity. Customers are less inclined to rush their purchase and are not looking for additional incentives.

Table 5.1. Variables Coefficients

Feature	Model with interaction
Intercept	0.126***
Stock_Quantity	0.093**
Size_Range_Type_broken_size_range	-0.002*
Size_Range_Type_full_size_range	-0.009*
Season_Type_fashionable	-0.030**
Season_Type_one_season_old	-0.013***
Stock_Tier_bronze	-0.004
Stock_Tier_gold	0.017***
Stock_Tier_silver	0.007
Sell_Through	-0.091***
Days_Since_Last_Drop	0.121***

*** p - value < 0.001, ** p-value < 0.01 , * p- value < 0.05

b. Product Characteristics

In order to make the interpretation more comprehensive, the analysis is based on a theoretical framework, which focuses on 2 groups of products. The first group of variables, including Size Range Type, Season Type, Stock-tier, Sell Through and Days Since Last Drop, has no interaction effect with the control variable Months of the Year on the Redemption Rate. The second group includes those variables having interaction effects with the control variable on Redemption Rate such as: Product Category, Main Season. The full coefficient table is added in Appendix B.

Group 1: Product Characteristics Variables without Interaction Effects by Month

In addition to the Quantity Group, other variables in Product Characteristics such as Size Range Type, Season Type and Stock Tier also influence the Redemption Rate (Table 5.1). The logic behind these effects is logical and aligns with customer behaviors.

Table 5.1 shows that, compared to the baseline "One Size Left" products, "Full Size Range" and "Broken Size Range" have lower Redemption Rates (-0.009 and -0.002, respectively, $p < 0.05$). This suggests that items with more size options or choices require fewer coupons, causing lower Redemption Rate. With "One Size Left" items, the rate is highest, which means those are sold most with coupons. For Season Type, “Fashionable” items have the lowest Redemption Rate (-0.03, $p < 0.001$), indicating customers are less likely to use coupons due to high demand of the product themself. In contrast, "One Season Old" and "Old for Seasons" items, which are less popular, have higher Redemption Rates as customers are more inclined to use coupons for these less hot items.

However, it is worth mentioning that while the effect of Size Range Type and Fashion Type are significant, as they have p-value < 0.05, they are relatively minor compared to other variables in the group such as Sell-Through rate or Day Since Last Drop.

For instance, the effect of Sell Through rate on Redemption Rate is notable, with a coefficient of -0.091 (p-value <0.01). This is logical, since the Sell-Through rate is used to monitor how fast moving items are, therefore, the higher this rate gets, the lower the need for customers using coupons is.

In addition, Days Since Last Drop also proves the same thing, as the coefficient of 0.021 (p-value < 0.001). This finding suggests that as the time since a product online, the Redemption Rate tends to rise. This likely happens because these products may become

less visible or appealing overtime, leading to a decrease in customer demand and requiring coupons to attract more customers.

Group 2: Product Characteristics Variables with Interaction Effects by Month

There are 2 main variables which also have interaction with Month of the Years in Product Characteristics Group, which are Product Category, Main Season. For the sake of comprehensiveness, this part will analyze each variable along with the interactions of them.

Table 5.2. Variables Coefficients

Feature	Model with interaction
Product_Category_Accessories	0.058**
Product_Category_Clothing	-0.055**
Product_Category_Kids	0.080*
Product_Category_Sportswear	0.072***

*** p - value < 0.001, ** p-value < 0.01 , * p- value < 0.05

For Products Category, those variables have a significant effect on Redemption Rate, as shown in Table 5.2. As the baseline is Shoes, the effect of other categories on Redemption Rate are positive. Kids and Sportswear have the highest coefficients (0.08 and 0.072, p-value < 0.05), this shows that for those specific products, they actually are more likely to be bought by customers if coupons are available, in comparison to baseline Shoes. In addition, Clothing has a negative coefficient (-0.055, p-value < 0.001). This means the Redemption Rate for coupons in Clothing are actually lower than for Shoes. Since customers are more focused on clothing than shoes, Clothes may not need to have as much need of using coupons as Shoes.

Moreover, when diving into the interaction between Product Category and Months of the Years, there are some interesting relationships. As illustrated in Figure 5.1, using January

as the baseline, if months have a value of 0, they indicate no significant effect between these variables and those months.

The figure highlights that for Clothing, the highest interaction effects occur in June (0.068) and July (0.066), showing a higher Redemption Rate compared to January. Conversely, in March, the interaction effect is negative, and it is also negative in November and December. This suggests that in these months, this rate is lower as customers are already inclined to purchase clothing without additional incentives. This finding is quite counterintuitive, because November and December is the biggest sale time of the year, therefore the usage of coupons may increase since customers are more actively looking for them, especially for Clothing Items. This phenomenon requires a deeper investigation and validation with the XGBoost model.

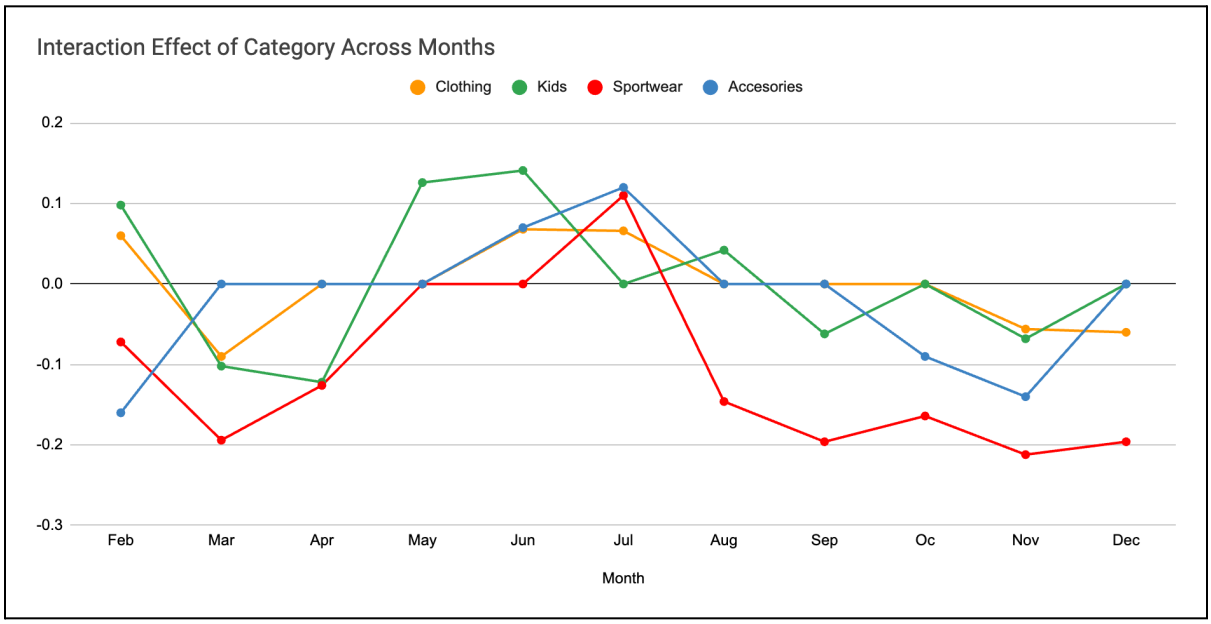

Figure 5.1. Interaction Effect of Product Category Across Months

The interaction effect of Clothing Category across Month is less pronounced, compared to Kids products and Sportswear products. For instance, Kids products see a significant

boost in Redemption Rates in May and June (0.126 and 0.141) and a large decrease in March and April (-0.102, -0.122). This seasonal variation highlights the importance of timing in coupons strategy, particularly for Kids products.

Sportswear and Accessories have quite the same pattern. It is worth mentioning that compared to January, the additional effects of those products in each month are all negative, except June and July. This suggests that the Redemption Rate in July and January for Sportswear and Accessories is highest, while in the end of the year period, from September to December, the Redemption Rate is lower than other times of the year.

The second variable interacting with the control variable in Product Characteristics is the Main Season. There are two main collections of the year: Spring Summer and Fall Winter. The figure 5.2 illustrates the additional effect of these collections in each month. The baseline is set to "Never out of Season" products in January. If a month does not have any value, it indicates that the interaction for that month is insignificant.

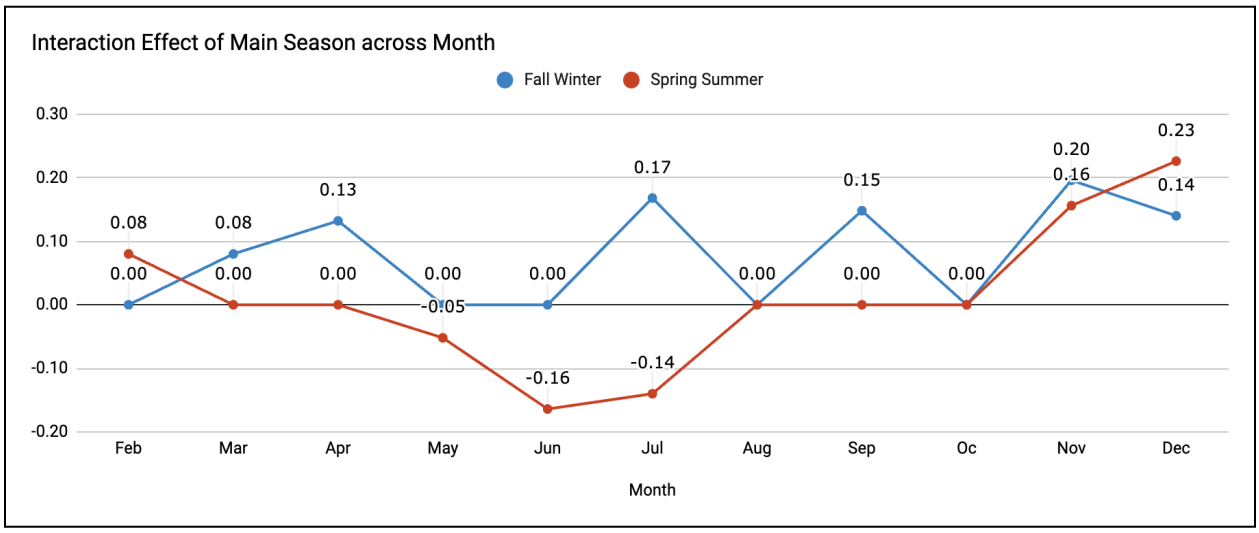

Figure 5.2. Interaction Effect of Fall Winter and Spring Summer Across Months

For the Spring Summer collection, a notable pattern emerges. From May to July, the additional effect on Redemption Rate from Spring Summer products is lower compared to other months. This is expected, as these products are in season and customers are more likely to purchase them without coupons. In contrast, from November to February, the effect increases, which suggests during the colder months, when Spring Summer is out of season, coupons become essential to drive customers' interest and encourage purchases.

For Fall Winter, the interaction effect on the Redemption Rate remains positive for nearly the whole year. This suggests that the need for coupons is consistently higher compared to both the Spring Summer collection and the baseline “Never Out of Season” products, even during its peak season. This relationship is unusual and needs further investigation.

c. Pricing Group

Lastly, in the Pricing Group, the variables show significant effects in the model when the interactions are not considered. However, once the interactions with months are included, the effect of those variables becomes extremely small, as shown in Table 5.3. This indicates that the influence of Pricing Group variables on Redemption rate is not equal across different months.

Table 5.3. Variables Coefficients

Feature	Model without interaction	Model with interaction
Sale_Price	0.090***	0.001**
Discount	0.138***	0.022**
Sale_Price:Month_9		0.157**
Sale_Price:Month_10		0.112*
Sale_Price:Month_11		0.241**
Sale_Price:Month_12		0.179***
Discount:Month_10		-0.12**
Discount:Month_11		-0.069*

*** p - value < 0.001, ** p-value < 0.01 , * p- value < 0.05

Additionally, the effect of Sale Price and Discount are particularly significant towards the end of the year (Table 5.3). For Sale Price, from September to December, the interaction between Sale Price and Months shows a positive relationship, this suggests that during these months, a one unit increase in Sale Price corresponds to a proportional increase in the Redemption rate. This seems logical because of the price increases, customers are more likely to purchase if they receive coupons.

For Discount, there are 2 significant interaction effects of this variable with October (-0.12, p-value < 0.01) and November (-0.069, p-value <0.05). This indicates that higher discounts during these months reduce the necessity for additional coupons incentives by 0.12 in October and 0.068 in November.

5.2. XGBoost Model

5.2.1. Model Performance

In the tuning process for model optimization, three key parameters were focused: the Regularization Term Gamma, the Learning Rate Eta, and the Maximum Tree Depth. As mentioned in the methodology part, Gamma controls the minimum loss reduction to make a further partition on a leaf node while the learning rate adjusts the impact of each boosting step to enhance the model strength against overfitting. Additionally, the Maximum Depth of the trees determines how deeply the model can capture patterns in the data, with deeper trees capable of capturing more complexity at the risk of overfitting.

To identify the optimal combination of these parameters, cross validation was used to evaluate the model performance. The average RMSE and the standard error for each group of parameters were calculated across 5-folds. This method informed a robust assessment of model reliability. Through extensive testing, Max Depth of 20 showed the best result in terms of accuracy.

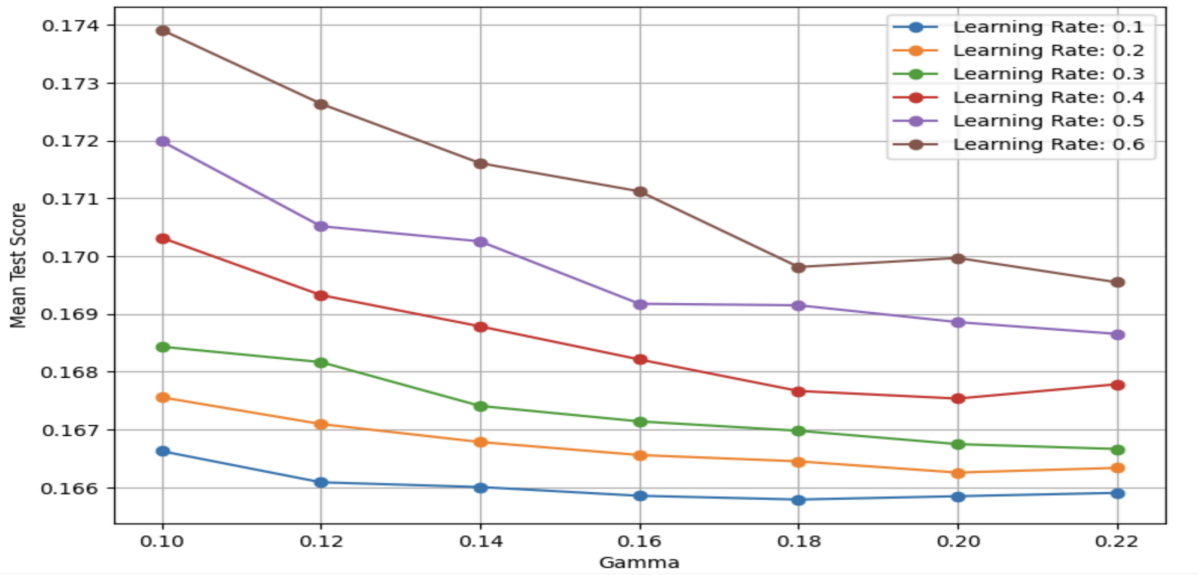

Figure 5.3. Model Performance in Max Depth of 20

Figure 5.3 illustrates the model performance in Max Depth of 20. The mode accuracy is highest when the learning rate is 0.1 (represented by the blue line), at this value, learning rate is the most effective regularization term for this model, striking the best balance in terms of accuracy. Furthermore, when examining the gamma conjunction with learning rate of 0.1, the RMSE initially decreased when gamma increased from 0.1 to 0.18. Beyond this point, the RMSE began to rise. Therefore, the combination of gamma of 0.18 and learning rate of 0.1 can be considered as the optimal group in achieving the lowest RMSE and increasing model performance.

Additionally, I also examined the pattern of other combination actors at different Maximum Tree Depths (from 10 to 40). The pattern remained consistent, with accuracy variations being minimal, differing by only about 0.00001. For detailed information, please refer to the Appendix C.

Finally, when evaluating the RMSE for the resulting model, the RMSE was 0.166 for the training set and 0.174 for the testing set. This minimal difference indicated that the model does not exhibit signs of overfitting, demonstrating its robustness and generalizability.

5.2.2. Model Interpretation

In order to check the most influential variables in the XGBoost model's decision-making process, the Feature Importance Method is applied. As shown in Figure 5.4, the order of effect among all variables are slightly different compared to the Regression coefficients. This can be explained by the fact that while Regression only can capture the linear relationship between the dependent variable and independent variables, XGBoost can capture a more complex relationship between them. In order to visualize those relationships, the Accumulated Local Effect (ALE) method is applied in order to interpret those relationships in the XGBoost model.

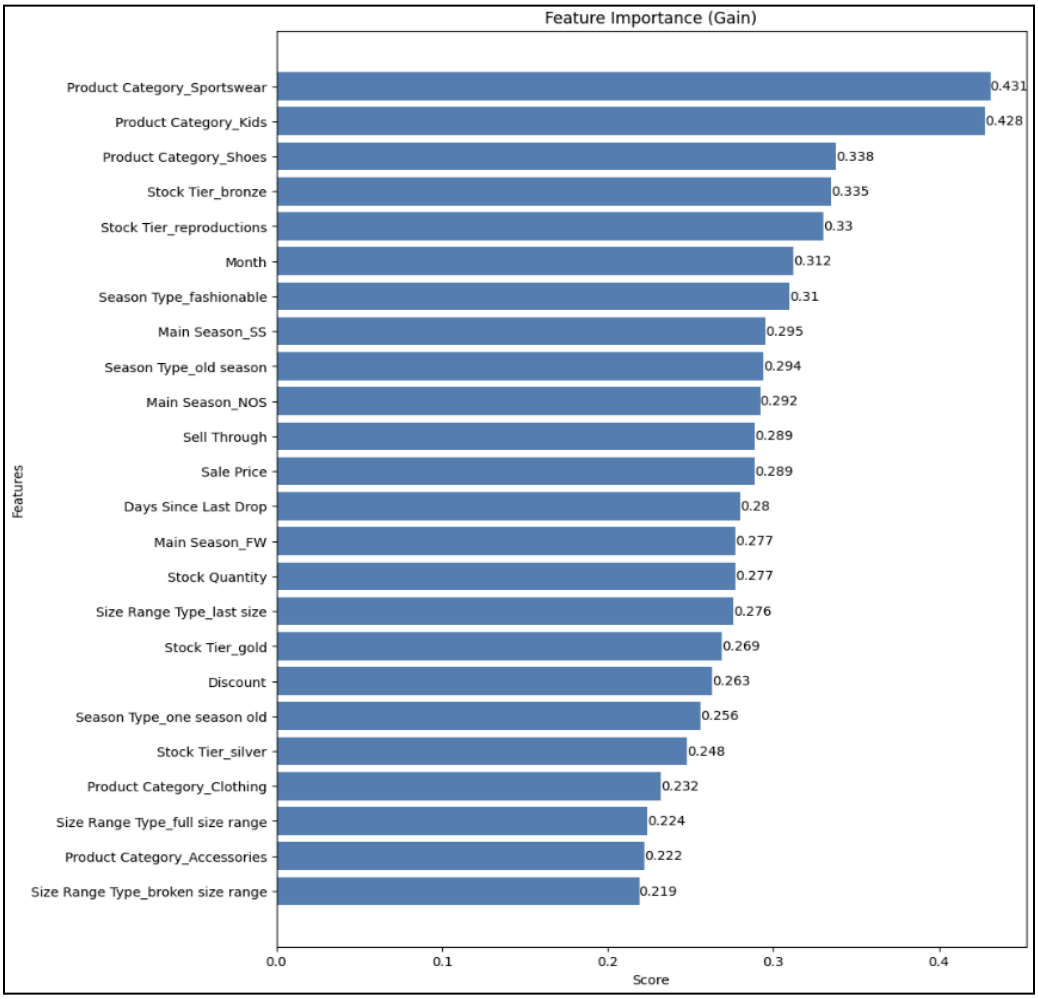

Figure 5.4. Feature Importance Plot

In addition, for the sake of comprehensiveness, the interpretation will follow the theoretical framework and categorize the variables into three groups: Product Quantity, Product Characteristics and Product Price.

a. Product Quantity Group

For this group, both the Regression Model and XGBoost Model show a positive relationship between Stock Quantity and Redemption Rate (Figure 5.5). This indicates that as the Quantity of products increases, the Redemption Rate also grows, showing a greater number of coupons to capture customer interest.

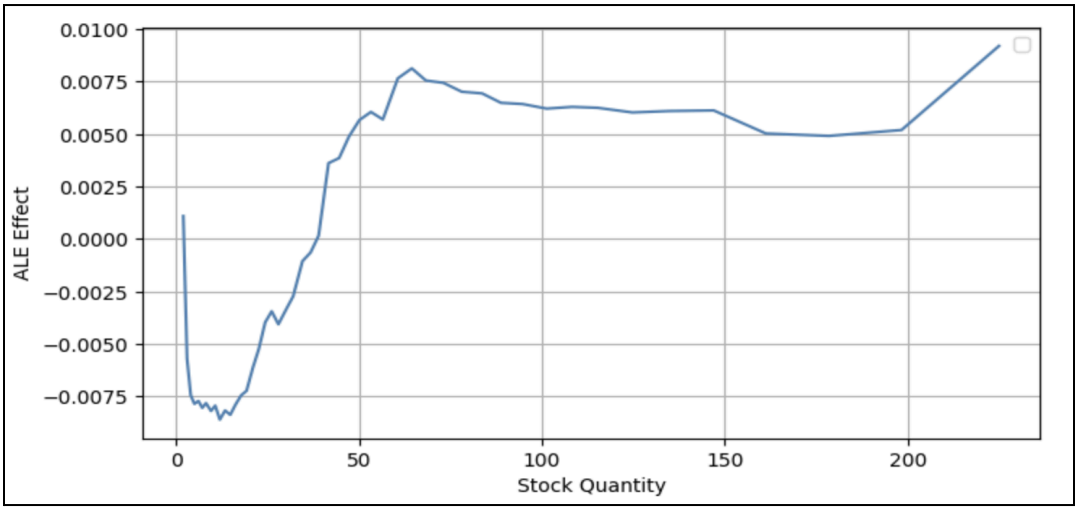

Figure 5.5. ALE Plot for Stock Quantity

However, in the XGBoost model, the Figure 5.5 reveals a big drop in the Redemption Rate for Stock Quantity ranges from 1 to 20 items. This suggests that within this specific range, the Redemption Rate for coupons is actually lower than the average and only increases when the quantity surpasses 20 items. This can be explained as customers might be inclined to purchase these items before they run out, decreasing the effectiveness of coupons in driving sales.

b. Product Characteristics Group

For this specified group, as mentioned in the conceptual frameworks, only Main Season and Product Category may interact differently across the Months of the Year. To better understand these relationships, we will evaluate their performance on a monthly basis. This approach can help in identifying seasonal patterns or variations specific to each product category and collection season.

Group 1: Product Characteristics Variables with Interaction Effects

For Product Category Group, the impact of it varies throughout the year, as shown in Figure 5.6. The Clothing Category has the variation in Redemption Rate ranging from -0.1 to 0.1, reflecting a relatively small and stable trend. Noticeably, there are notable increases in Redemption Rate during two periods: from March to May and from October to December. Based on the business context, the rate may increase during the October - December period mainly because of the holiday season. Consumers are more inclined to purchase clothing as gifts. There are also many competitive discount strategies to attract shoppers, therefore, the increasing use of coupons is plausible.

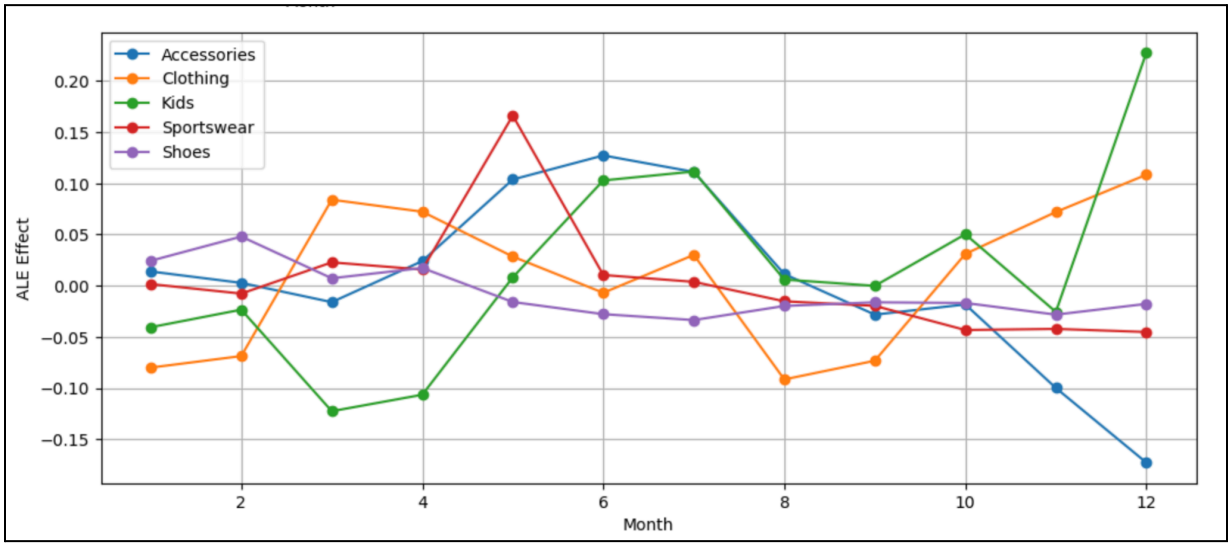

Figure 5.6. ALE Plot for Product Category across Months

For Kids products, the Redemption Rate peaks in December to over 0.2, showing a heightened utility for coupons during this period. December is a peak for the shopping period, due to holidays such as Christmas, which increase the demand of Kids products. However, parents are likely to seek out deals to minimize their spending during this time, leading to a higher ratio of using the coupons. In other words, it can be said that coupon availability encourages them to convert into orders.

The second peak of Kid's products are from June to July, which reaches 0.1, showing an increased need for coupons during this time. This may be because this is the beginning of back-to-school shopping, when customers start looking for some items, but only prompted to make a purchase when they receive a coupon incentive. Conversely, in March and April, the Redemption Rate for Kids products is notably lower to -0.1. This ratio during this period of time decreases since parents may be buying for specific needs or occasions for the incoming Summer season, making them less reliant on coupons. This trend is also aligned with the Regression model, when there is an increase in Redemption Rate in May and June and decrease in March and April.

Sportswear and Accessories have quite the same pattern in the beginning of the year with the Redemption Rate for those products being highest between April and June. This period is the transaction from Spring to Summer. Sportswear and Accessories such as activewear, swimwear, and outdoor gear, become more relevant as people engage in more outdoor activities and exercise. While demand for these products rises, it is less urgent compared to the holiday season. During this time, customers are often encouraged to make purchases through coupon offers, leading to a high Redemption Rate in this period. Additionally, even though Sportswear and Accessories has a close pattern, there is a significant drop of Redemption Rate only happens in Accessories items to around -0.15.

Shoes effect on redemption rate remains stable over month of the year. This suggests that the use of coupons in this category is not heavily influenced by seasonal variations or specific promotional periods.

Beside Product Category, Main Season is also another variable which has a special relationship with Month of the year. In general, the effect of the Main Season on Redemption Rate varies across different months and collection categories, as shown in Figure 5.7.

For the Spring Summer collection, The effect on Redemption Rates are relatively low in June to August, ranging from 0 to -0.05. This decline is expected, as these months align with the peak of season for Spring Summer products, when customers are willing to purchase these items without the need for coupons. However, from October to February, the effect on Redemption Rate increases up to 0.05, showing that coupons are highly used. This is because Spring Summer items are less relevant during the colder months. As a result, customers are more inclined to purchase these products if they have incentives.

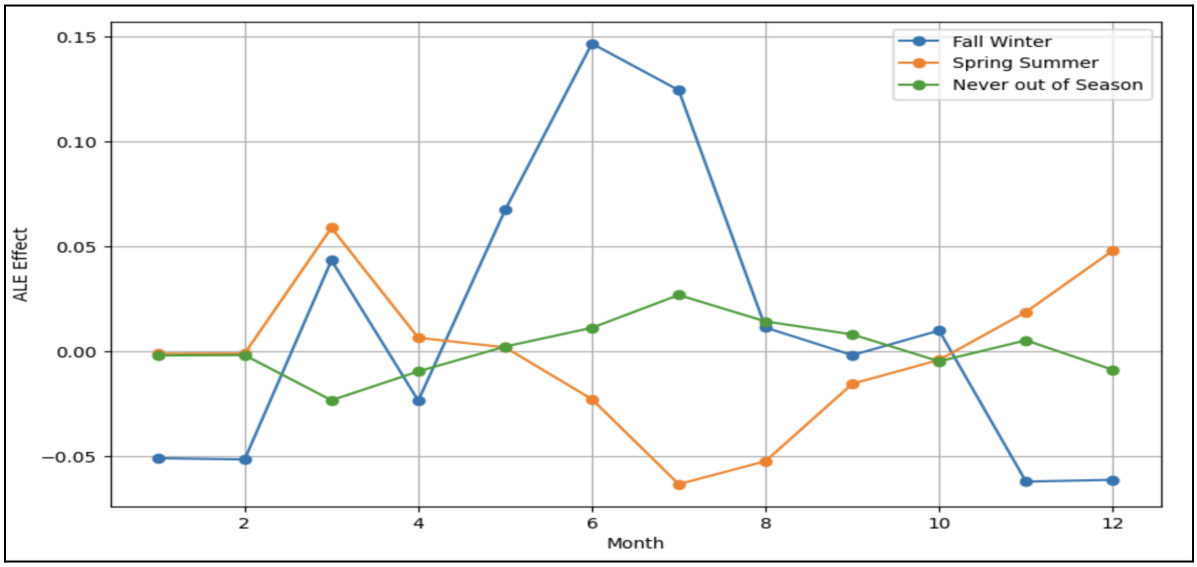

Figure 5.7. ALE Plot for Main Season across Months

For Never Out of Season products, the effect of them on Redemption Rate across months stays almost the same, showing a steady need for coupons throughout the year. Since those can be considered as essential items that customers may need to buy throughout the year, the use of coupons does not change much across each month.

Lastly, for the Fall Winter collection, the effect on Redemption Rate peaks from May to July, indicating a high Redemption Rate for coupons during this time for the Fall Winter collection. This makes sense since this is the time when Fall Winter products are out of season, as customers are more focused on other products. This means that in order to have customer attention, additional incentives are required. In contrast, from November until next February, the effect on Redemption Rate turns out to be lower, indicating a lower rate of used coupons.

Group 2: Product Characteristics Variables without Interaction Effects

For other variables in the Product Characteristics Group, their effect on Redemption Rate in XGBoost suggest the same pattern in the Regression Model. For instance, the Sell Through Rate exhibits a similar negative linear relationship with Redemption Rate (Figure 5.8, left figure). This suggests that as this rate increases, the use of coupons decreases. This relationship seems logical as higher Sell-Through rates proves strong product demand, reducing the need for additional incentives.

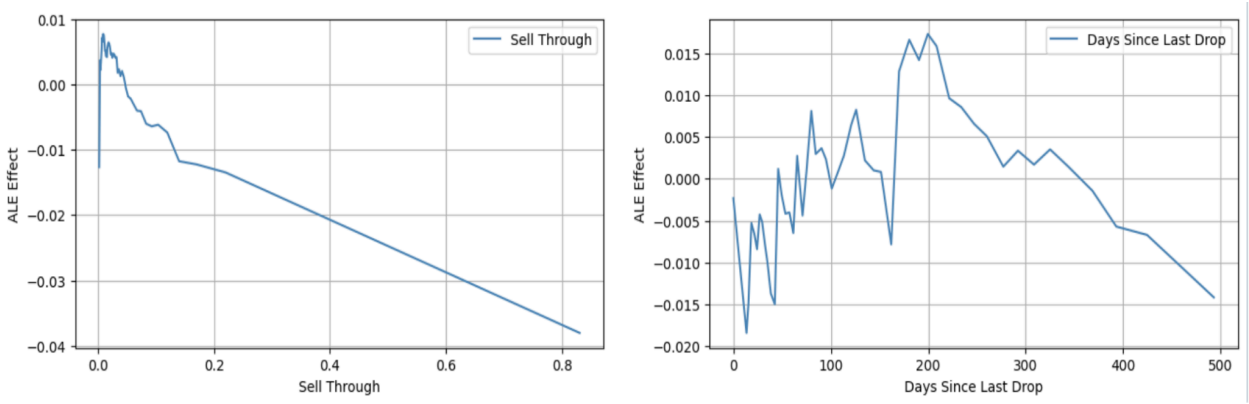

Figure 5.8. ALE Plot for Sell Through Rate and Days Since Last Drop

Despite a generally linear relationship between the majority of variables and Redemption Rate, the influence of Days Since Last Drop in XGBoost shows a more complex pattern than the simple Regression Model, as shown in Figure 5.8, the right figure.

The figure shows that for the first 200 days after product online, the Redemption Rate fluctuates swiftly but tends to increase over time. This suggests that products are sold more with coupons, by attracting customer attention as they become old gradually. However, after 200 days, the pattern shifts. The Redemption Rate starts to decrease when products remain on the website for a longer time. This reversal implies that products which have been available for over 200 days might experience reduced effectiveness of coupons. The explanation for this phenomenon is that after 200 days, the focus of customers may change and coupons alone may no longer be enough to convert buying attention to actual purchase.

c. Product Price

In the Product Price Group, Discount is one of the key variables significantly contributing to the model's accuracy. As illustrated in Figure 5.9, on the right, the overall trend for discount is negatively related to the Redemption Rate. When the Discount increases, the Redemption Rate for coupons decreases. This relationship makes sense, because higher discounts directly reduce product price, reducing the urgency of consumers looking for coupons. Consumers are likely to perceive the discounted price as a good deal on its own, reducing their need to seek out additional promotions.

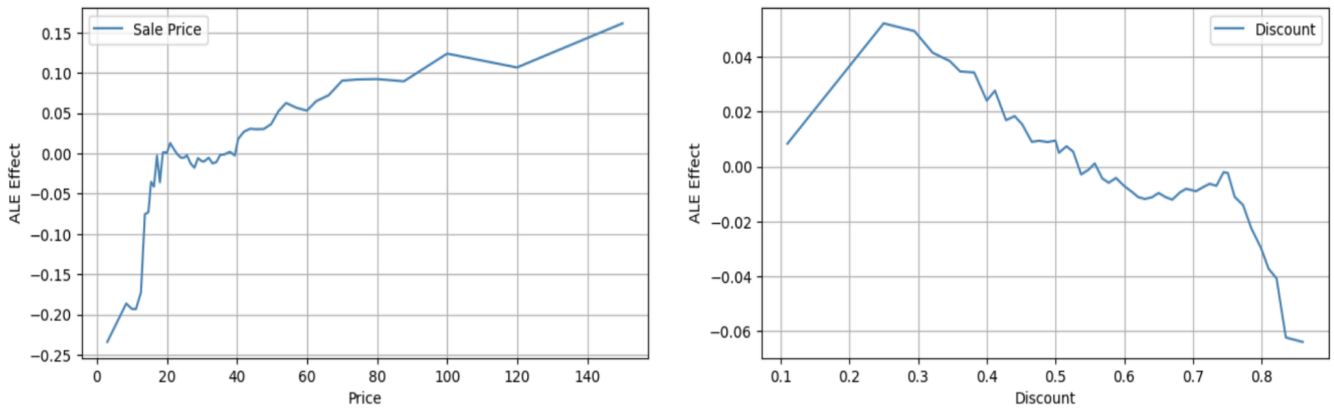

Figure 5.9 ALE Plot for Price and Discount

Additionally, the Discount reveals that the Redemption Rate for coupons peaks when discounts are in the range of 20% to 30%. This insight is valuable for marketers, suggesting that releasing coupons for products within this discount range is used most, since after this discount, the effect for coupons starts to be lower.

Furthermore, when discounts exceed 30%, the utility of additional coupons begins to decrease, but after exceeding 80%, the drops are even stronger. This trend suggests that with a very high discount, customers are less motivated to seek out extra coupon incentives, as the perceived value of the product is already sufficiently enhanced.

The analysis of Price also reveals a positive relationship with the Redemption Rate. As the price of a product increases, the need of using coupons also rises, leading customers to actively search for discounts and increasing the Redemption Rate. Conversely, when prices fall below 20€, the use of coupons diminishes significantly compared to other price ranges. This suggests that customers are less inclined to seek out coupons for lower-priced items, indicating a lower redemption rate in these scenarios.

5.3. Model Comparison

This thesis used 2 main models, Regression and XBBoost to achieve its objective of predicting coupon necessity. When comparing the results between them, it is evident that the XGBoost model outperforms the Regression model in terms of accuracy, as illustrated in Table 5.4.

Table 5.4. Model Performance

Model	RMSE of training data	RMSE of testing data	Running Time
Linear Regression without Interaction	0.274	0.288	1m30s
Linear Regression with Interaction	0.232	0.238	1m30s
XGBoost	0.166	0.174	212m:40s

Both model Regression and XGBoost have advantages and disadvantages. Regression models are generally simpler and more interpretable compared to more complex models like XGBoost. This makes it easier to understand the relationship between the predictors and the response variable. Additionally, the Regression models typically require less computational power and time to train. This makes them suitable for large datasets or real-time applications where speed is crucial. As can be seen in Table 5.4, the running time for the Regression model is much shorter than for XGBoost when it only took almost 2 minutes to run a Linear Regression model while it took over 3 hours to complete the XGBoost.

While all three models did not show overfitting, the RMSE of the Regression model showed a higher RMSE compared to XGBoost. This is because Regression attempts to fit

a straight line through all the data points, which may not be suitable in this case when the relationship between 2 variables is complicated. In contrast, XGBoost, an ensemble learning method using gradient boosting, excels at capturing these non-linear relationships, resulting in better performance and lower RMSE. This difference is highlighted most in models interpretation which is mentioned below.

In terms of interpretation between two models, the majority of the relationships discovered via the XGBoost model align with those found in the Regression model. However, there are some noticeable disagreements between 2 models, regarding the effect of Clothing Category and the effect of Fall Winter Collection on Redemption Rate across different months as well as the fluctuated effect of Days Since Last Drop on Redemption Rate.

Frisly, there is the difference between the relationship of Clothing Category and Redemption Rate in two models. XGBoost suggests that there are notable increases in Redemption Rate from October to December while Regression suggests a decrease in this duration. However, based on the business context, the Redemption Rate may increase during the October - December period mainly because of the holiday season. Consumers are more inclined to purchase clothing as gifts. There are also many competitive discount strategies to attract shoppers, therefore, it increases the use of coupons or Redemption Rate in this period.

Secondly, when comparing the Fall-Winter collection effect at the end of the year, the Regression model suggests that the impact on Redemption Rate remains relatively high, whereas the XGBoost model indicates a lower effect, suggesting the reduced utility of coupons. This difference is because seasonal demand often exhibits complex, non-linear effects that linear regression, even with interaction terms, may not be able capture effectively. The impact of coupons may show sudden spikes or drops across different

seasons, which are difficult for linear models to represent accurately. Additionally, real-world data typically involve higher-order interactions among variables, not only between Main Season and Month but can also include other factors such as Product Category or Price. These patterns and interactions are better captured by models like XGBoost which builds decision trees where each split can partition the data based on different criteria, capturing non-linear relationships and interactions. Moreover, in the business context, Fall-Winter collections are naturally in high demand at the end of the year. Customers purchase them out of coupon necessity, not because of promotions, which aligns with the XGBoost model's findings.

Lastly, the XGBoost model reveals a more complex pattern regarding the influence of Days Since Last Drop compared to the simpler Regression model, as illustrated in Figure 5.8 (on the right). The figure demonstrates that the relationship between Days Since Last Drop and Redemption Rate is not linear but rather exhibits fluctuations, with varying effects at different values of Days Since Last Drop. The XGBoost model captures a more intricate relationship because it accounts for non-linear and varying effects across different values of Days Since Last Drop. Unlike the Linear Regression model, which assumes a constant effect, XGBoost reveals that the impact on Redemption Rate fluctuates, showing both increases and decreases depending on the time since the last drop.

In conclusion, in order to validate these opposite findings, it is essential to applying the understanding of business context and market dynamics. In this case, the XGBoost model provides complex and logical insights that align more closely with practical business knowledge than Regression. In addition, XGBoost is capable of capturing both non-linear and linear relationships, whereas the Regression model can only identify linear relationships. This makes insights from XGBoost become more trustworthy. However, the differences highlight the dynamic nature of these two models, and mark the need for further investigation to fully understand their differing insights.

6. Implications

The aim of this thesis is to evaluate the effects of product features on Redemption Rate. It also takes into account the effect of them across different times of the year. By understanding these factors and their timing, marketers can develop effective coupon strategies.

According to the model, it is suggested to implement a dynamic coupons strategy that is adjusted based on every month of the year. For Kid's products and Clothes, stores should focus targeted coupon campaigns on the Holiday Season, particularly at the end of the year, when Redemption Rates reach highest. Companies can develop holiday marketing with gift-oriented promotion and festival coupon offers. This helps the products stand out during the holidays season, maintain competitive features in the markets. It also can help to increase demand and ensure that promotional efforts align with peak shopping time.

Moreover, in March and April, when parents are less coupon dependent because of incoming season preparation for their kids, marketers should not release too many coupons, instead, offering bundles that combine multiple items parents are likely to need, such as a summer outfit or travel essential pack for kids.

For Sportswear and Accessories, concentrate coupon effects from April to June to align with peak outdoor activity period since the need for coupons in this period is high. Additionally, for Shoes Category, since the effect of it on Redemption Rate remains stable throughout the year, stores should maintain consistent, moderate discount offers (e.g., 10-20%) to keep customers engaged without needing to ramp up discounts during specific periods.

Additionally, the strategy for each Main Season collection should be tailored to different months. For Spring Summer collection, limiting coupon usage from June to August

which is the peak season can help in preventing profit erosion. In the off season (October to February), stores should implement significant discount and coupons campaigns to clear remaining inventory with end of season and clearance events. Similarly, for Fall Winter collection, it should focus on clearance sales by high discount during off season from May to July. During the main season from October to next February, when the need for coupons is low, instead of increasing more coupons, marketers should focus marketing efforts via content to highlight how these products meet seasonal needs and fashion trends. Additionally, bundling Spring-Summer items with Fall-Winter items can help to manage inventory. Lastly, for “Never out of season” products, since the need for coupons remain low and table all year around, marketers should offer the periodic promotions such as "Monthly Essentials" sales, where select Never Out of Season items are offered at a special discount. This can create a sense of urgency without relying on seasonal demand.

For Price Group, the Discount reveals that the Redemption Rate for coupons peaks when discounts are in the range of 20% to 30%. This range has been identified as the sweet spot where coupons are used most. If the product itself has a higher discount than 30%, marketers should not focus too much on those products since the natural purchase of them is already high. For products priced below 20 EUR, reduce the emphasis on coupon distribution. Instead, focus on highlighting the affordability and value of these items without additional discounts, as customers are less likely to seek coupons for lower-priced goods. For higher-priced products, ensure that coupon campaigns are prominently featured. Since the Redemption Rate increases with price, offering attractive coupons for these items can significantly boost sales.

7. Limitations and Further Research

This thesis has several limitations. Firstly, the scope of the research is confined to a single main outlet store, representing the outlet fashion industry. Therefore, this can apply in this specific industry, however, to generalize findings to the larger fashion industry, it needs more validation and additional data. The model was built and tailored for practical needs of one company, with many features adjusted to meet its particular demands. Consequently, some insights derived from the model may not be applied across all industries.

Secondly, using the Redemption Rate to estimate the need for coupons can be risky sometimes. The Redemption Rate indicates the percentage of sales driven by coupons, if the percentage is high for specific products attributes at a certain time of the year, it suggests a high demand for coupons for those products during that period. However, there are cases when customers may become too familiar with using coupons. This means while customers are willing to make purchases, they still actively seek out coupons. Although this study found no coupon addiction from customers through the relationship between the Sell Through rate (indicating how quickly items are purchased) and the Redemption Rate, the possibility of customer addiction to coupons cannot be entirely ruled out.

Lastly, there is significant potential for innovation within this model. For example, this thesis can be developed more by taking into account Customer Behaviors. Multiple researches in this topic suggest approaches to customer segmentation. For example, Marcus (1998) applied Customer Value Matrix for small retail and service businesses and Malhotra (2022) applied clustering algorithms such as K-means. Another study combined K-means with LRFM (Length, Recency, Frequency, Monetary) feature selection to

categorize customers into four loyalty levels: Premium, Inertia, Latent, and No Loyalty (Nikmah et al., 2023).

In conclusion, this thesis presents a predictive model designed to estimate the Redemption Rate based on specific product attributes. The model has yielded several valuable insights and demonstrated a relatively high level of accuracy without indications of overfitting. Despite these successes, there remains potential for future improvement. Expanding the research to by adding more observations from various segments of the fashion industry would enhance the model's generalizability. Additionally, refining customer segmentation methods, as discussed, could help to capture customer behavior better and further improve model’s effectiveness. Overall, this work provides a solid foundation for predicting Redemption Rate of coupons and ample opportunities for continued enhancement and application.

Reference

Allenby, G.M., Jen, L., & Leone, R.P. (1996). Economic Trends and Being Trendy: The Influence of Consumer Confidence on Retail Fashion Sales. *Journal of Business & Economic Statistics, 14*, 103-111.

Barat, S. (2007). An empirical investigation of how perceived devaluation and income effects influence consumers' intended utilization of savings from coupon redemption.

Cachon, G.P., & Swinney, R. (2011). The Value of Fast Fashion: Quick Response, Enhanced Design, and Strategic Consumer Behavior. *Manage. Sci., 57*, 778-795.

Chiang, J. (1995). Competing Coupon Promotions and Category Sales. *Marketing Science, 14*, 105-122.

Choi, T. (2007). Pre-season stocking and pricing decisions for fashion retailers with multiple information updating. *International Journal of Production Economics, 106*, 146-170.

Duan, G., & Ma, X. (2018). A Coupon Usage Prediction Algorithm Based On XGBoost. *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 178-183.

Guo, T., Zhong, S., Wang, X., & Li, G. (2021). Does product display quantity increase purchase intention? The mediation of diminished pain of payment. *Journal of Research in Interactive Marketing*.

Heazlewood, E. (2013). Tis' the season to be selling! *PS Post Script*, 59.

Hsu, H., & Chen, J. (2011). Applying Choice Theory to Promotion Asymmetry Effect.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (1st ed.) [PDF]. *Springer.*

Ignacio Osuna, Jorge González, Mario Capizzani (2016) Which Categories and Brands to Promote with Targeted Coupons to Reward and to Develop Customers in Supermarkets. *Journal of Retailing*

Kawakatsu, H. (2010). An Optimal Replenishment Policy for Seasonal Items in Retailing.

Kim, N., Lee, M., & Kim, H. (2008). The Effect of Service Coupons on the Consumer Trade-Offs Between Price and Perceived Quality. *Journal of Promotion Management, 14*, 59 - 76.

Malhotra, S., Agarwal, V., & Ticku, A. (2022). Customer Segmentation - A Boon for Business. *SSRN Electronic Journal*.

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of Consumer Marketing, 15*, 494-504.

McKinsey. (2023, November 29). The State of Fashion 2024: Finding pockets of growth as uncertainty reigns.

Nayal, P., & Pandey, N. (2020). Redemption Intention of Coupons: A Meta-Analytical Review and Future Directions. *Journal of Promotion Management, 26*, 372 - 395.

Ngwe, D. (2017). Why Outlet Stores Exist: Averting Cannibalization in Product Line Extensions. *Mark. Sci., 36*, 523-541.

Nikmah, T.L., Harahap, N.H., Utami, G.C., & Razzaq, M.M. (2023). Customer Segmentation Based on Loyalty Level Using K-Means and LRFM Feature Selection in Retail Online Store. *Jurnal ELTIKOM*.

Nudell, N. (2023). Fashion Time: The Fashion Calendar and Scheduling an Industry. *The Journal of American Culture*.

Oliver, R.L., & Shor, M. (2003). Digital redemption of coupons: satisfying and dissatisfying effects of promotion codes. *Journal of Product & Brand Management, 12*, 121-134.

Osuna, I., González, J., & Capizzani, M. (2016). Which Categories and Brands to Promote with Targeted Coupons to Reward and to Develop Customers in Supermarkets. *Journal of Retailing, 92*, 236-251.

Thomas, M.T., Bawa, K., & Menon, G. (2003). Spending More to Save More : The Impact of Coupons on Premium Priced Products.

Prajapati, D.A., Pandya, D.D., Patel, R.K., & Jadeja, A. (2022). A Perception of Promotional code Technique in E-commerce that uses Data Analytics and Data Mining for Consumer Response. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*.

Reibstein, D.J., & Traver, P.A. (1982). Factors Affecting Coupon Redemption Rates. *Journal of Marketing, 46*, 102 - 113.

Ren, Y., Fu, P., & Yu, W. (2021). Prediction of Coupon Usage Behavior Based on Customer Segmentation and XGBoost algorithm. *2021 2nd International Conference on Big Data Economy and Information Management (BDEIM)*, 42-47.

Senter, H.F. (2008). Applied Linear Statistical Models. *Journal of the American Statistical Association, 103*, 880 - 880.

Shamout, M.D. (2016). The Impact of Promotional Tools on Consumer Buying Behavior in the Retail Market.

Sinwar, D., & Dhaka, D.V. (2015). Discovering Outliers from Real-World Data using Decision Trees based Classifiers.

Song, L., & Yang, W. (2018). A Novel Digital Coupon Use Prediction Model Based on XGBoost.

Subhojit Banerjee (2009). Effect of product category on promotional choice: comparative study of discounts and freebies. *Institute of Business Management, VBS Purvanchal University, Jaunpur, India*

Triantafillidou, A., Siomkos, G.J., & Papafilippaki, E. (2017). The effects of retail store characteristics on in-store leisure shopping experience. *International Journal of Retail & Distribution Management, 45*, 1034-1060.

Vakeel, S., & Kaushik, R. (2020). Sustainability in the Fashion Industry. *Circular Economy and Re-Commerce in the Fashion Industry*.

Von Mutius, B., & Huchzermeier, A. (2021). Customized Targeting Strategies for Category Coupons to Maximize CLV and Minimize Cost. *Journal of Retailing*.

Wheatley, J.J., & Chiu, J.S. (1977). The Effects of Price, Store Image, and Product and Respondent Characteristics on Perceptions of Quality. *Journal of Marketing Research, 14*, 181 - 186.

Zhang, Y., & Hu, X. (2022). Digital Coupon Promotion and Inventory Strategies of Omnichannel Brands. *Axioms, 12*, 29.

Zhang, Z., Ma, M., Leszczyc, P.T., & Zhuang, H. (2020). The influence of coupon duration on consumers' redemption behavior and brand profitability. *Eur. J. Oper. Res., 281*, 114-128.

Zheng, J., & Mu, Y. (2010). Joint Decision of Pricing and Package Coupon Value. *2010 International Conference on Management and Service Science*, 1-4

Appendix

Appendix A. Correlation Matrix before Data Cleaning Process

	Days Since Last Drop	Stock age	Stock Quantity	Retail Price	Sale Price	Discount Live	Total Sold 7 Days	Total Sold 30 Days	Sell Through Rate	Fashion Year	Low Value Cus	High Value Cus	Total Orders	Total Coupons	Redemption Rate
Days Since Last Drop	1	0.94	-0.09	0.04	-0.08	0.33	-0.08	-0.08	-0.03	-0.33	-0.05	-0.08	-0.08	-0.07	-0.12
Stockage	0.94	1	-0.07	0.04	-0.09	0.35	-0.07	-0.06	-0.03	-0.36	-0.04	-0.07	-0.07	-0.06	-0.11
Stock Quantity	-0.09	-0.07	1	-0.05	-0.06	-0.03	0.33	0.31	-0.09	0.09	0.23	0.36	0.35	0.33	0.12
Retail Price	0.04	0.04	-0.05	1	0.82	0.11	-0.05	-0.05	-0.02	-0.07	-0.04	-0.05	-0.05	-0.05	-0.06
Sale Price	-0.08	-0.09	-0.06	0.82	1	-0.19	-0.06	-0.06	-0.05	0.01	-0.05	-0.07	-0.07	-0.06	-0.09
Discount Live	0.33	0.35	-0.03	0.11	-0.19	1	0.02	0.01	0.09	-0.29	0.01	0.02	0.02	0.03	0.05
Total Sold 7 Days	-0.08	-0.07	0.33	-0.05	-0.06	0.02	1	0.91	0.25	0.08	0.62	0.87	0.87	0.82	0.23
Total Sold 30 Days	-0.08	-0.06	0.31	-0.05	-0.06	0.01	0.91	1	0.23	0.08	0.54	0.77	0.77	0.71	0.22
Sell Through Rate	-0.03	-0.03	-0.09	-0.02	-0.05	0.09	0.25	0.23	1	0.01	0.11	0.18	0.18	0.16	0.25
Fashion Year	-0.33	-0.36	0.09	-0.07	0.01	-0.29	0.08	0.08	0.01	1	0.05	0.09	0.09	0.1	0.16
Low Value Cus	-0.05	-0.04	0.23	-0.04	-0.05	0.01	0.62	0.54	0.11	0.05	1	0.64	0.76	0.74	0.14
High Value Cus	-0.08	-0.07	0.36	-0.05	-0.07	0.02	0.87	0.77	0.18	0.09	0.64	1	0.99	0.95	0.26
Total Orders	-0.08	-0.07	0.35	-0.05	-0.07	0.02	0.87	0.77	0.18	0.09	0.76	0.99	1	0.97	0.25
Total Coupons	-0.07	-0.06	0.33	-0.05	-0.06	0.03	0.82	0.71	0.16	0.1	0.74	0.95	0.97	1	0.3
Redemption Rate	-0.12	-0.11	0.12	-0.06	-0.09	0.05	0.23	0.22	0.25	0.16	0.14	0.26	0.25	0.3	1

Appendix B. Regression Coefficients Table

Feature	Model without interaction	Model with interaction
Intercept	0.156***	0.126***
Stock_Quantity	0.017***	0.093**
Size_Range_Type_broken_size_range	-0.004	-0.002*
Size_Range_Type_full_size_range	-0.018***	-0.009*
Season_Type_fashionable	-0.018***	-0.030**
Season_Type_one_season_old	-0.012***	-0.013***
Stock_Tier_bronze	-0.019***	-0.004
Stock_Tier_gold	0.016***	0.017***
Stock_Tier_silver	0.000	0.007*
Sell_Through	-0.080***	-0.091***
Days_Since_Last_Drop	0.012*	0.121***
Product_Category_Accessories	0.073***	0.058***
Product_Category_Clothing	0.038***	-0.055**
Product_Category_Kids	0.088***	0.080*
Product_Category_Sportswear	0.011*	0.072***
Main_Season_FW	-0.020***	-0.053***
Main_Season_SS	-0.036***	-0.031**
Sale_Price	0.090***	0.001
Discount	0.138***	0.022
Month_2		-0.060**
Month_3		-0.062**
Month_4		-0.087***
Month_5		-0.162***
Month_6		-0.265***
Month_7		-0.232***
Month_8		-0.129***
Month_9		-0.189***
Month_10		-0.087***
Month_11		0.100***
Month_12		-0.060**

Product_Category_Clothing:Month_2	0.06*
Product_Category_Clothing:Month_3	-0.09***
Product_Category_Clothing:Month_4	-0.013
Product_Category_Clothing:Month_5	-0.016
Product_Category_Clothing:Month_6	0.068*
Product_Category_Clothing:Month_7	0.066*
Product_Category_Clothing:Month_8	0.003
Product_Category_Clothing:Month_9	0.001
Product_Category_Clothing:Month_10	-0.023
Product_Category_Clothing:Month_11	-0.056*
Product_Category_Clothing:Month_12	-0.06*
Product_Category_Kids:Month_2	0.098**
Product_Category_Kids:Month_3	-0.102**
Product_Category_Kids:Month_4	-0.122***
Product_Category_Kids:Month_5	0.126**
Product_Category_Kids:Month_6	0.141***
Product_Category_Kids:Month_7	0.029
Product_Category_Kids:Month_8	0.042*
Product_Category_Kids:Month_9	-0.062*
Product_Category_Kids:Month_10	-0.006
Product_Category_Kids:Month_11	-0.068*
Product_Category_Kids:Month_12	-0.004
Product_Category_Sportswear:Month_2	-0.072*
Product_Category_Sportswear:Month_3	-0.194***
Product_Category_Sportswear:Month_4	-0.126***
Product_Category_Sportswear:Month_5	-0.031
Product_Category_Sportswear:Month_6	0.010
Product_Category_Sportswear:Month_7	0.11*
Product_Category_Sportswear:Month_8	-0.146***
Product_Category_Sportswear:Month_9	-0.196***
Product_Category_Sportswear:Month_10	-0.164***
Product_Category_Sportswear:Month_11	-0.212***
Product_Category_Sportswear:Month_12	-0.196***
Product_Category_Accesories:Month_2	-0.16***

Product_Category_Accesories:Month_3	0.056
Product_Category_Accesories:Month_4	0.041
Product_Category_Accesories:Month_5	0.089
Product_Category_Accesories:Month_6	0.07***
Product_Category_Accesories:Month_7	0.12***
Product_Category_Accesories:Month_8	0.021
Product_Category_Accesories:Month_9	0.034
Product_Category_Accesories:Month_10	-0.09***
Product_Category_Accesories:Month_11	-0.14***
Product_Category_Accesories:Month_12	0.056
Main_Season_FW:Month_2	0.019
Main_Season_FW:Month_3	0.08*
Main_Season_FW:Month_4	0.133*
Main_Season_FW:Month_5	-0.007
Main_Season_FW:Month_6	-0.017
Main_Season_FW:Month_7	0.171*
Main_Season_FW:Month_8	0.102
Main_Season_FW:Month_9	0.151*
Main_Season_FW:Month_10	0.015
Main_Season_FW:Month_11	0.159***
Main_Season_FW:Month_12	0.140**
Main_Season_SS:Month_2	0.08*
Main_Season_SS:Month_3	-0.006
Main_Season_SS:Month_4	0.018
Main_Season_SS:Month_5	-0.054*
Main_Season_SS:Month_6	-0.163*
Main_Season_SS:Month_7	-0.140*
Main_Season_SS:Month_8	-0.027
Main_Season_SS:Month_9	0.038
Main_Season_SS:Month_10	0.020
Main_Season_SS:Month_11	0.164**
Main_Season_SS:Month_12	0.232***
Sale_Price:Month_9	0.157***
Sale_Price:Month_10	0.112***

Sale_Price:Month_11	0.241***
Sale_Price:Month_12	0.179***
Discount:Month_10	-0.12**
Discount:Month_11	-0.069*

*** p - value < 0.001, ** p-value < 0.01 , * p- value < 0.05

Appendix C. XGBoost Model Tuning Results Table

Max Depth	Gamma	Learning Rate	RSME Average	RSME SD	Rank
10	0.1	0.1	0.16597	0.00164	130
20	0.1	0.1	0.16663	0.00222	23
30	0.1	0.1	0.16662	0.00216	29
40	0.1	0.1	0.16662	0.00216	29
10	0.1	0.2	0.16652	0.00153	146
20	0.1	0.2	0.16757	0.00207	41
30	0.1	0.2	0.16756	0.00224	17
40	0.1	0.2	0.16756	0.00224	17
10	0.1	0.3	0.16752	0.00170	116
20	0.1	0.3	0.16855	0.00217	25
30	0.1	0.3	0.16843	0.00199	64
40	0.1	0.3	0.16843	0.00199	64
10	0.1	0.4	0.16841	0.00150	139
20	0.1	0.4	0.17059	0.00237	1
30	0.1	0.4	0.17032	0.00226	12
40	0.1	0.4	0.17032	0.00226	12
10	0.1	0.5	0.16991	0.00182	94
20	0.1	0.5	0.17212	0.00219	21
30	0.1	0.5	0.17199	0.00183	91
40	0.1	0.5	0.17199	0.00183	91
10	0.1	0.6	0.17053	0.00128	147
20	0.1	0.6	0.17387	0.00225	12
30	0.1	0.6	0.17391	0.00183	87
40	0.1	0.6	0.17391	0.00183	87
10	0.12	0.1	0.16581	0.00146	133
20	0.12	0.1	0.16605	0.00198	59
30	0.12	0.1	0.16609	0.00189	79
40	0.12	0.1	0.16609	0.00189	79
10	0.12	0.2	0.16673	0.00173	98
20	0.12	0.2	0.16699	0.00237	4
30	0.12	0.2	0.16709	0.00206	33

40	0.12	0.2	0.16709	0.00206	33
10	0.12	0.3	0.16706	0.00119	136
20	0.12	0.3	0.16793	0.00206	35
30	0.12	0.3	0.16817	0.00206	37
40	0.12	0.3	0.16817	0.00206	37
10	0.12	0.4	0.16869	0.00164	100
20	0.12	0.4	0.16922	0.00180	83
30	0.12	0.4	0.16932	0.00185	76
40	0.12	0.4	0.16932	0.00185	76
10	0.12	0.5	0.16920	0.00149	115
20	0.12	0.5	0.17007	0.00229	6
30	0.12	0.5	0.17052	0.00210	25
40	0.12	0.5	0.17052	0.00210	25
10	0.12	0.6	0.16995	0.00134	121
20	0.12	0.6	0.17266	0.00216	18
30	0.12	0.6	0.17264	0.00237	2
40	0.12	0.6	0.17264	0.00237	2
10	0.14	0.1	0.16609	0.00160	96
20	0.14	0.1	0.16608	0.00200	46
30	0.14	0.1	0.16600	0.00203	34
40	0.14	0.1	0.16600	0.00203	34
10	0.14	0.2	0.16650	0.00172	80
20	0.14	0.2	0.16682	0.00223	9
30	0.14	0.2	0.16678	0.00227	4
40	0.14	0.2	0.16678	0.00227	4
10	0.14	0.3	0.16717	0.00155	98
20	0.14	0.3	0.16760	0.00237	1
30	0.14	0.3	0.16741	0.00227	3
40	0.14	0.3	0.16741	0.00227	3
10	0.14	0.4	0.16836	0.00183	61
20	0.14	0.4	0.16882	0.00211	15
30	0.14	0.4	0.16878	0.00206	20
40	0.14	0.4	0.16878	0.00206	20
10	0.14	0.5	0.16901	0.00210	17

20	0.14	0.5	0.16998	0.00211	14
30	0.14	0.5	0.17025	0.00223	5
40	0.14	0.5	0.17025	0.00223	5
10	0.14	0.6	0.16956	0.00135	97
20	0.14	0.6	0.17114	0.00176	63
30	0.14	0.6	0.17161	0.00193	43
40	0.14	0.6	0.17161	0.00193	43
10	0.16	0.1	0.16603	0.00165	67
20	0.16	0.1	0.16588	0.00223	4
30	0.16	0.1	0.16585	0.00206	15
40	0.16	0.1	0.16585	0.00206	15
10	0.16	0.2	0.16650	0.00156	79
20	0.16	0.2	0.16642	0.00193	40
30	0.16	0.2	0.16656	0.00202	22
40	0.16	0.2	0.16656	0.00202	22
10	0.16	0.3	0.16723	0.00194	33
20	0.16	0.3	0.16730	0.00203	18
30	0.16	0.3	0.16714	0.00197	25
40	0.16	0.3	0.16714	0.00197	25
10	0.16	0.4	0.16801	0.00187	41
20	0.16	0.4	0.16807	0.00195	29
30	0.16	0.4	0.16821	0.00200	23
40	0.16	0.4	0.16821	0.00200	23
10	0.16	0.5	0.16860	0.00158	64
20	0.16	0.5	0.16910	0.00182	38
30	0.16	0.5	0.16917	0.00196	23
40	0.16	0.5	0.16917	0.00196	23
10	0.16	0.6	0.16914	0.00140	72
20	0.16	0.6	0.17127	0.00141	71
30	0.16	0.6	0.17112	0.00129	73
40	0.16	0.6	0.17112	0.00129	73
10	0.18	0.1	0.16614	0.00164	51
20	0.18	0.1	0.16589	0.00202	20
30	0.18	0.1	0.16579	0.00195	22

40	0.18	0.1	0.16579	0.00195	22
10	0.18	0.2	0.16643	0.00163	48
20	0.18	0.2	0.16637	0.00194	22
30	0.18	0.2	0.16645	0.00210	11
40	0.18	0.2	0.16645	0.00210	11
10	0.18	0.3	0.16684	0.00165	43
20	0.18	0.3	0.16695	0.00208	11
30	0.18	0.3	0.16698	0.00214	9
40	0.18	0.3	0.16698	0.00214	9
10	0.18	0.4	0.16793	0.00170	37
20	0.18	0.4	0.16802	0.00207	9
30	0.18	0.4	0.16767	0.00193	19
40	0.18	0.4	0.16767	0.00193	19
10	0.18	0.5	0.16855	0.00142	54
20	0.18	0.5	0.16908	0.00225	3
30	0.18	0.5	0.16915	0.00231	1
40	0.18	0.5	0.16915	0.00231	1
10	0.18	0.6	0.16917	0.00137	51
20	0.18	0.6	0.16980	0.00162	37
30	0.18	0.6	0.16981	0.00158	39
40	0.18	0.6	0.16981	0.00158	39
10	0.2	0.1	0.16624	0.00156	41
20	0.2	0.1	0.16588	0.00190	18
30	0.2	0.1	0.16584	0.00190	16
40	0.2	0.1	0.16584	0.00190	16
10	0.2	0.2	0.16633	0.00168	28
20	0.2	0.2	0.16626	0.00179	21
30	0.2	0.2	0.16625	0.00177	23
40	0.2	0.2	0.16625	0.00177	23
10	0.2	0.3	0.16668	0.00132	40
20	0.2	0.3	0.16664	0.00204	6
30	0.2	0.3	0.16675	0.00215	4
40	0.2	0.3	0.16675	0.00215	4
10	0.2	0.4	0.16761	0.00154	32

20	0.2	0.4	0.16756	0.00203	6
30	0.2	0.4	0.16753	0.00201	7
40	0.2	0.4	0.16753	0.00201	7
10	0.2	0.5	0.16822	0.00156	27
20	0.2	0.5	0.16877	0.00159	24
30	0.2	0.5	0.16886	0.00173	17
40	0.2	0.5	0.16886	0.00173	17
10	0.2	0.6	0.16884	0.00143	28
20	0.2	0.6	0.16988	0.00178	16
30	0.2	0.6	0.16997	0.00188	10
40	0.2	0.6	0.16997	0.00188	10
10	0.22	0.1	0.16616	0.00154	21
20	0.22	0.1	0.16590	0.00194	9
30	0.22	0.1	0.16590	0.00194	7
40	0.22	0.1	0.16590	0.00194	7
10	0.22	0.2	0.16643	0.00157	17
20	0.22	0.2	0.16634	0.00181	7
30	0.22	0.2	0.16634	0.00181	7
40	0.22	0.2	0.16634	0.00181	7
10	0.22	0.3	0.16657	0.00159	13
20	0.22	0.3	0.16665	0.00202	6
30	0.22	0.3	0.16666	0.00203	4
40	0.22	0.3	0.16666	0.00203	4
10	0.22	0.4	0.16772	0.00165	5
20	0.22	0.4	0.16778	0.00219	1
30	0.22	0.4	0.16778	0.00219	1
40	0.22	0.4	0.16778	0.00219	1
10	0.22	0.5	0.16843	0.00165	2
20	0.22	0.5	0.16865	0.00149	5
30	0.22	0.5	0.16865	0.00149	5
40	0.22	0.5	0.16865	0.00149	5
10	0.22	0.6	0.16853	0.00178	1
20	0.22	0.6	0.16955	0.00162	1
30	0.22	0.6	0.16955	0.00162	1

40	0.22	0.6	0.16955	0.00162	1