11/20/2023

Boosting Juice Category Performance: Analyzing and predicting customer's purchasing behavior

Dolly Sah

MBA & MSBA'24, UNIVERSITY OF UTAH

Contents

Problem statement

The business goal of the grocery store is to increase the store profit (Profit = Revenue – Cost). There are two ways to increase store profit without changing cost, by increasing sale volume or by increasing price (Profit = ∑(no of item sold*price) - Cost). However, without changing the sale volume and cost, profit can also be increased by simply selling the high profitable value item.

Working towards the same goal, the brand manager and sales manager of the store chain have distinct challenges and have specific objectives. Both wants to increase the bottom line by influencing the sale of high profitable item in juice brand, in this case is Minute Maid (MM), by understanding what factor influence the purchase and if customer's probability of purchasing the product can be predicted.

Brand Manager's Challenge

Problem: The brand manager is interested in understanding the variables that influence a customer's probability of purchasing MM. This would be helpful to orchestrate a marketing strategy to increase the sale of MM.

Objective: Identify key factors influencing MM purchases to increase the probability of customers choosing MM.

Sales Manager's Challenge

Problem: The sales manager aims to build a predictive model that can accurately predict the probability of a customer purchasing MM.

Objective: Develop a robust predictive model to estimate the likelihood of MM purchases.

Expected outcome:

- To craft a comprehensive report about the methods to be utilized to address specific business concerns.
- Providing clear explanations on why one method is superior to the other, and present specific recommendations.
- The step-by-step execution of the recommendation is currently not part of the scope.

Exploratory Data Analysis

Dataset Description

The dataset contains 1070 purchases and 14 variables in which the customer either purchased Citrus Hill (CH) or Minute Maid (MM) Orange Juice.

The dataset source: http://data.mishra.us/files/project/OJ_data.csv

Dataset information:

Total Purchases	1070
Minute Maid Purchases	417
Citrus Hill Purchases	653
Target Variable	Purchase
No of Binary Variables	3
No of Int Variables	3
No of Float Variables	11

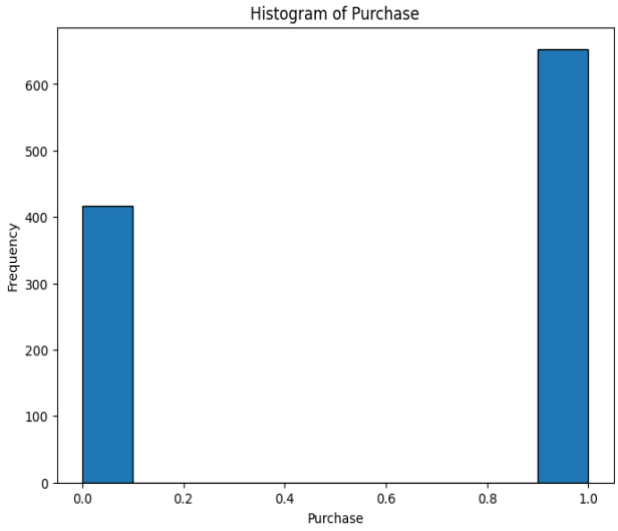
The dataset contains the following variables:

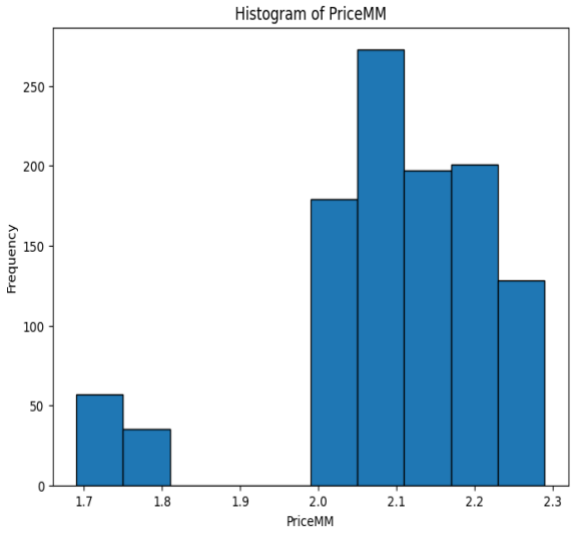
Variables	**Description**	**Minimum Value**	**Maximum Value**
Purchase	A factor with levels 0 and 1 indicating the purchase of Citrus Hill (1) or Minute Maid Orange Juice (0).	0	1
PriceCH	Price charged for CH. Also called List Price for CH.	1.69	2.09
PriceMM	Price charged for MM. Also called List Price for MM.	1.69	2.29
DiscCH	Discount offered for CH.	0	0.5
DiscMM	Discount offered for MM.	0	0.8
SpecialCH	Indicator of special on CH. Special can be a free gift, loyalty points, etc.	0	1
SpecialMM	Indicator of special on MM. Special can be a free gift, loyalty points, etc.	0	1
LoyalCH	Customer brand loyalty for CH. Probability to buy CH (over MM) based on prior purchase behavior.	0.000011	0.999947
SalePriceMM	Sale price for MM. This is the difference between the list price and discount.	1.19	2.29
SalePriceCH	Sale price for CH. This is the difference between the list price and discount.	1.39	2.09
PriceDiff	Sale price of MM less sale price of CH.	-0.67	0.64
PctDiscMM	Percentage discount for MM.	0	0.40201
PctDiscCH	Percentage discount for CH.	0	0.252688
ListPriceDiff	List price of MM less list price of CH.	0	0.44

Variables Histogram

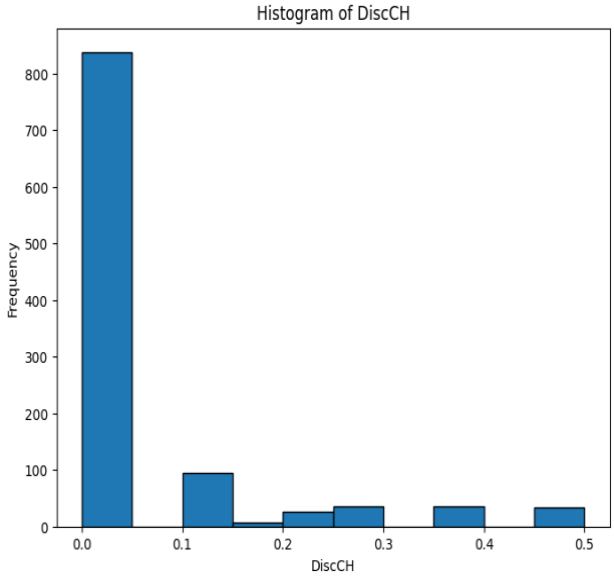
The histograms of all 14 variables were examined visually, and it is evident that no outliers are present in the data. There are no missing values, and the ranges of the variables aligns with the expectations.

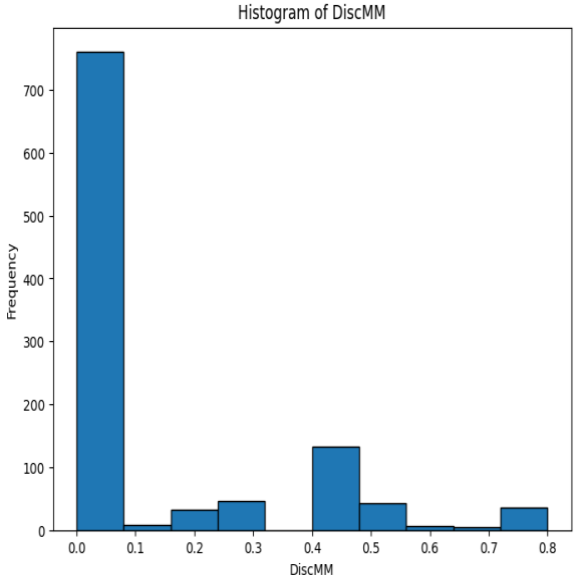
As observed, the binary variable 'Purchase' has values of either 0 or 1, with no missing values. The MM price ranges from 1.69 to 2.29, and there are no missing values. Additionally, there are no values less than 0 for MM price.



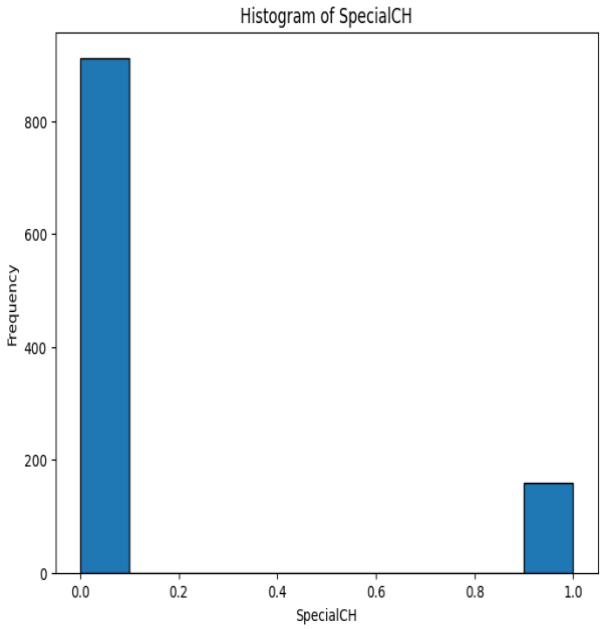


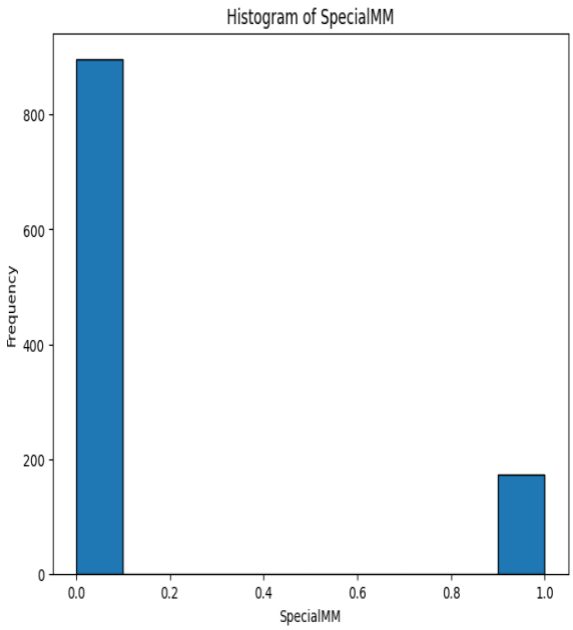
The discount for CH varies from 0 to 0.5, with no missing values and no values less than 0. Similarly, the discount for MM varies from 0 to 0.8, with no missing values and no values less than 0.



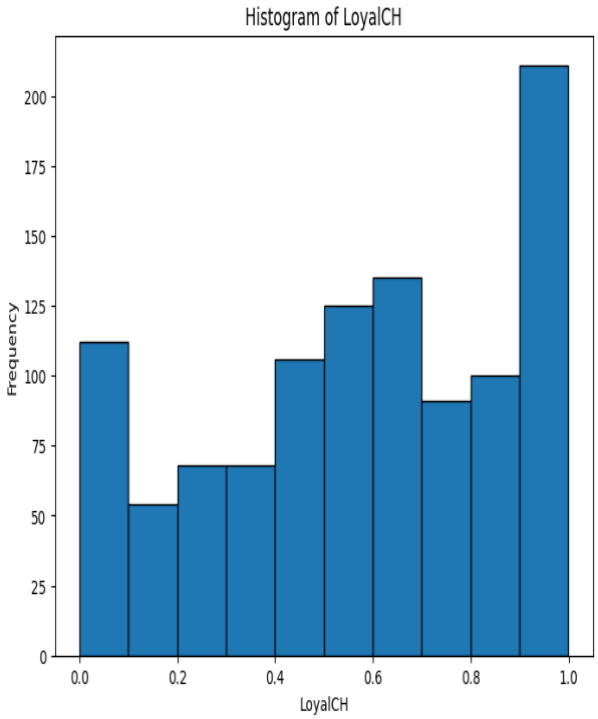


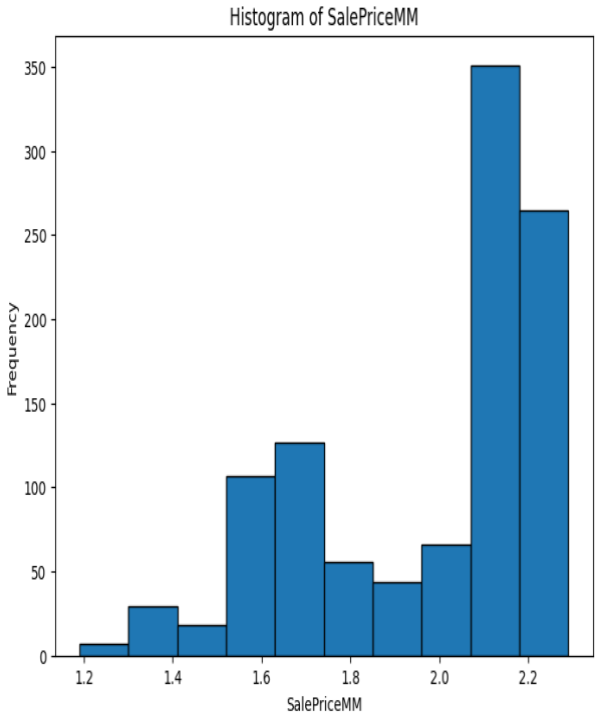
SpecialCH is a binary variable with values of either 0 or 1, and it has no missing values. There are no values other than 0 or 1. Similarly, SpecialMM is also a binary variable with values of either 0 or 1, and it has no missing values. There are no values other than 0 or 1.



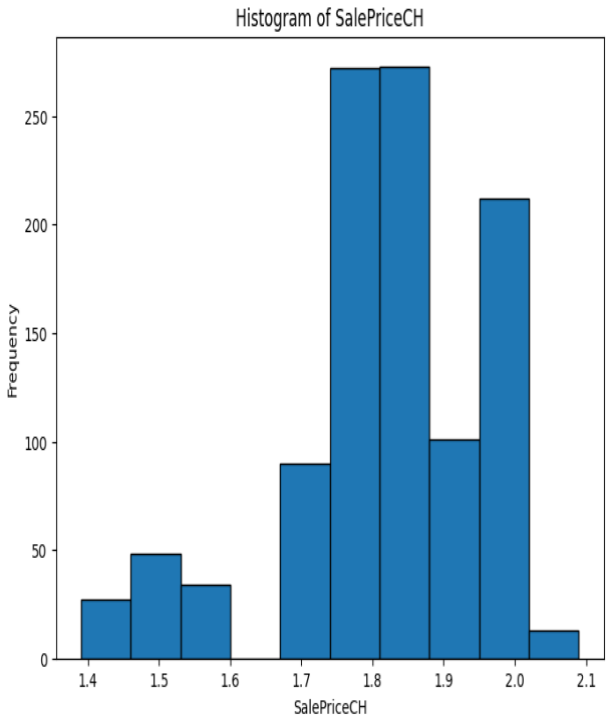


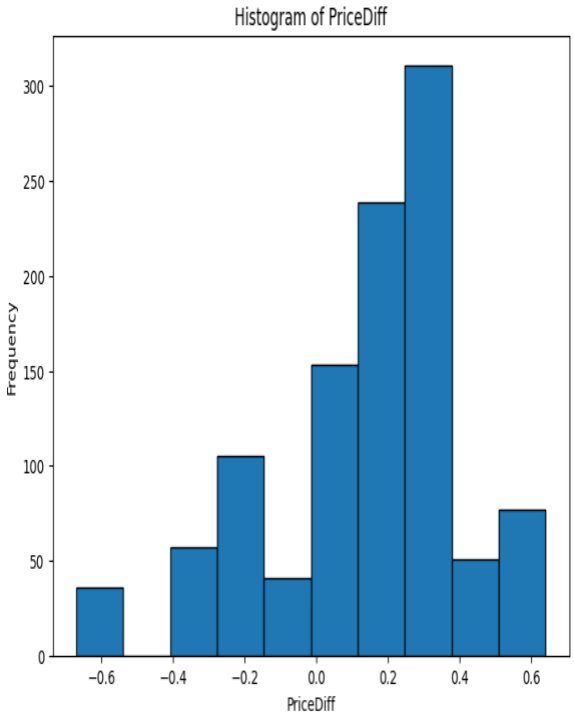
LoyalCH varies from 0.000011 to 0.999947, and there are no missing values. SalepriceMM ranges from 1.19 to 2.29, with no missing values and no values less than 0.



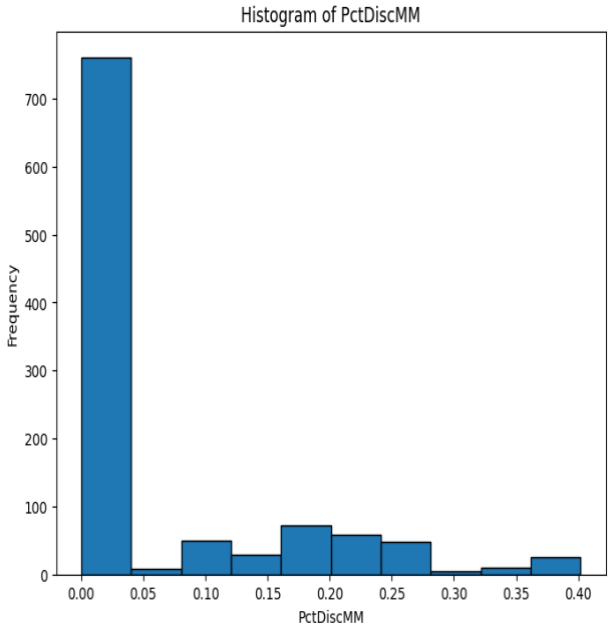


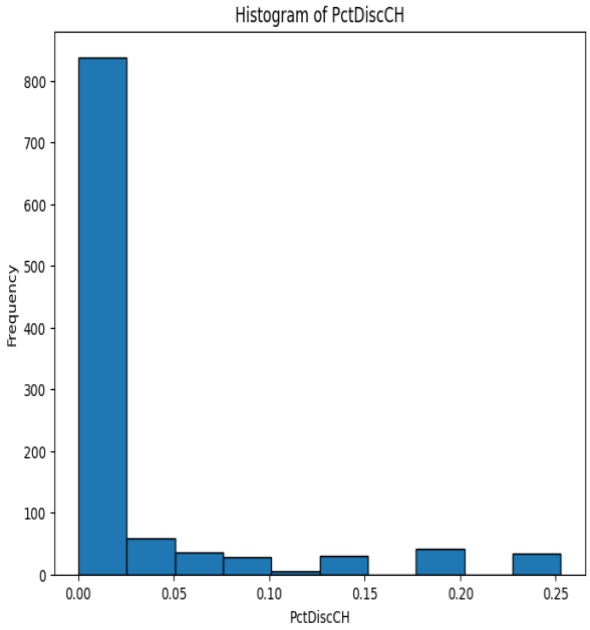
SalePriceCH varies from 1.39 to 2.09, with no missing values and no values less than 0. The price difference varies from -0.67 to 0.64, with no missing values and no outliers detected.



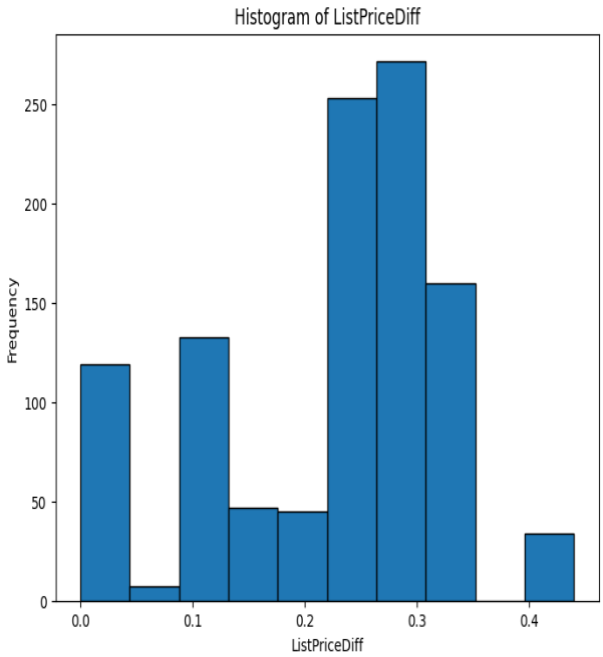


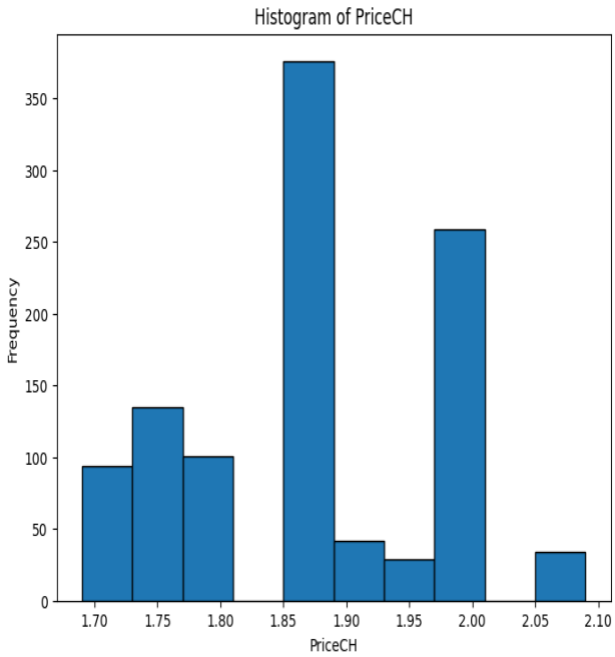
PctDiscMM varies from 0 to 0.40201 indicating maximum discount of about 40% discount, with no missing values and no values greater than 1 or less than 0. PctDiscCH varies from 0 to 0.252688 indicating max discount of about 25%, with no missing values and no values greater than 1 or less than 0.





List price difference varies from 0 to 0.44, with no missing values. Citrus Hill price varies from 1.69 to 2.09, with no missing values and no values less than 0.

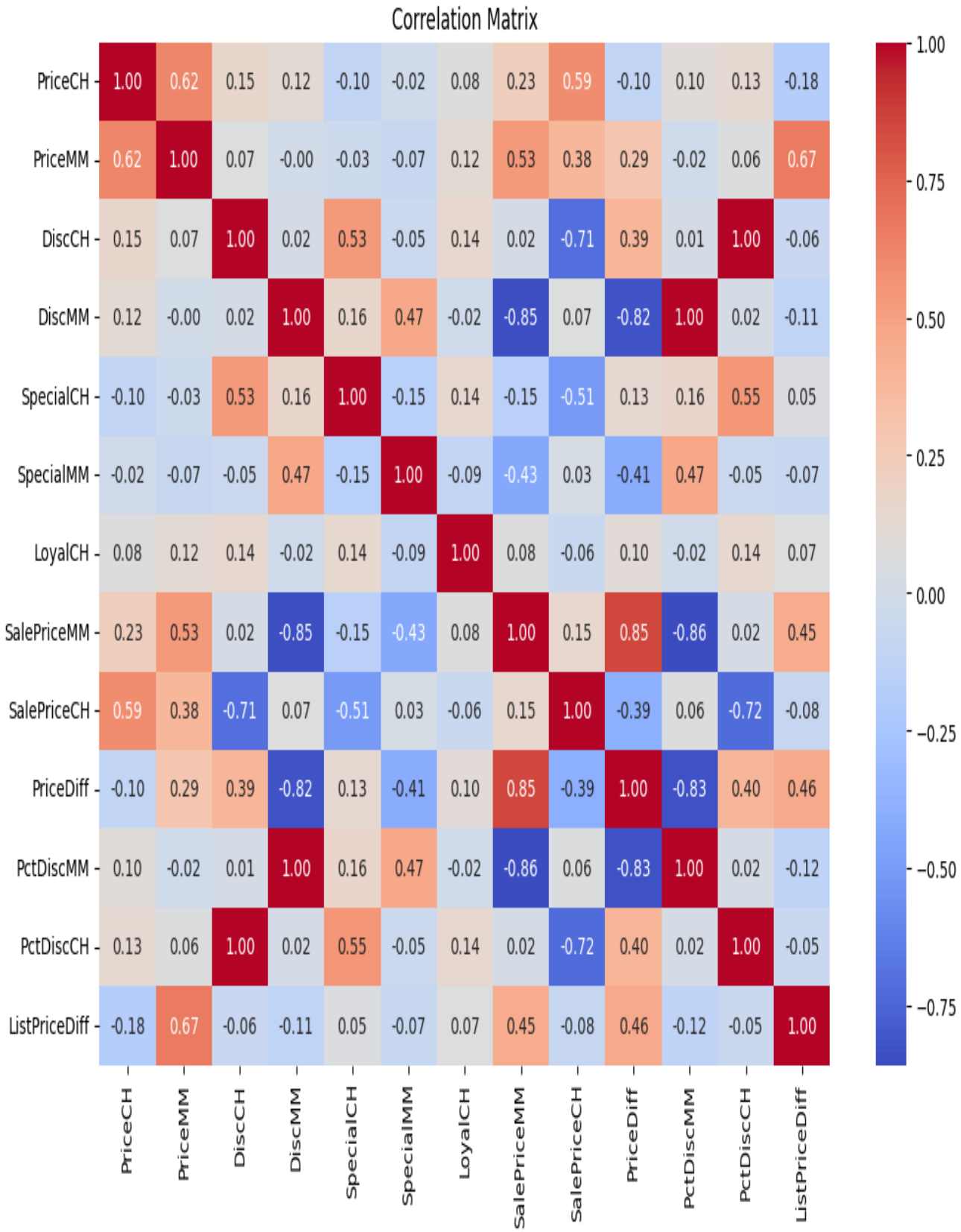




Correlation

It is crucial to check correlation among variables as it helps uncover relationships and dependencies within the dataset, guiding variable selection and model building. Understanding the degree of correlation ensures that the chosen variables contribute meaningful and diverse information to the analysis, preventing issues such as multicollinearity and improving the overall robustness of predictive models.

The correlation graph below (excluding the target variable 'purchase') indicates that each variable is highly correlated with at least one other variable, except for 'Loyal CH'. which is not correlated with any other variables. The non-collinearity of 'Loyal CH' variable indicates it is providing new information not captured by any other variables.


Correlation Matrix
PriceCH 1.00 0.62 0.15 0.12 -0.10 -0.02 0.08 0.23 0.59 -0.10 0.10 0.13 -0.18
PriceMM 0.62 1.00 0.07 -0.00 -0.03 -0.07 0.12 0.53 0.38 0.29 -0.02 0.06 0.67
DiscCH 0.15 0.07 1.00 0.02 0.53 -0.05 0.14 0.02 -0.71 0.39 0.01 1.00 -0.06
DiscMM 0.12 -0.00 0.02 1.00 0.16 0.47 -0.02 -0.85 0.07 -0.82 1.00 0.02 -0.11
SpecialCH -0.10 -0.03 0.53 0.16 1.00 -0.15 0.14 -0.15 -0.51 0.13 0.16 0.55 0.05
SpecialMM -0.02 -0.07 -0.05 0.47 -0.15 1.00 -0.09 -0.43 0.03 -0.41 0.47 -0.05 -0.07
LoyalCH 0.08 0.12 0.14 -0.02 0.14 -0.09 1.00 0.08 -0.06 0.10 -0.02 0.14 0.07
SalePriceMM 0.23 0.53 0.02 -0.85 -0.15 -0.43 0.08 1.00 0.15 0.85 -0.86 0.02 0.45
SalePriceCH 0.59 0.38 -0.71 0.07 -0.51 0.03 -0.06 0.15 1.00 -0.39 0.06 -0.72 -0.08
PriceDiff -0.10 0.29 0.39 -0.82 0.13 -0.41 0.10 0.85 -0.39 1.00 -0.83 0.40 0.46
PctDiscMM 0.10 -0.02 0.01 1.00 0.16 0.47 -0.02 -0.86 0.06 -0.83 1.00 0.02 -0.12
PctDiscCH 0.13 0.06 1.00 0.02 0.55 -0.05 0.14 0.02 -0.72 0.40 0.02 1.00 -0.05
ListPriceDiff -0.18 0.67 -0.06 -0.11 0.05 -0.07 0.07 0.45 -0.08 0.46 -0.12 -0.05 1.00
PriceCH
PriceMM
DiscCH
DiscMM
SpecialCH
SpecialMM
LoyalCH
SalePriceMM
SalePriceCH
PriceDiff
PctDiscMM
PctDiscCH
ListPriceDiff
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Data Pre-processing

Imputation

It is evident from the exploratory data analysis that the quality of the data is good and there are no missing values or outliers, eliminating the need for any imputation or removal of anomalies.

Transformation purchase variable

The values in the 'purchase' variable has been transformed from 0 to 1 and 1 to 0. This transformation was performed so that 1 now represents Minute Maid purchase, while 0 represents a Citrus Hill purchase. This adjustment ensures that a higher probability of purchasing Minute Maid, aligning with the target variable (Minute Maid purchase).

Standardization

As evident from the summary, the variables have different range. To ensure that all variables are on a comparable scale, preventing certain features from dominating others in predictive models, standardization is crucial. Although this won't impact the predictive model used further in the analysis, normalization of all the variables has been performed for ease in interpretating the magnitude of impact while comparing coefficient.

Overfitting

Dividing the dataset into training and testing sets is crucial to evaluate the model's performance on unseen data, providing an unbiased assessment of its generalization ability. This practice helps prevent overfitting and ensures that the model's effectiveness can be reliably measured in real-world scenarios beyond the training data. A data split of 80:20 has been performed to generate training and testing set.

Variable Selection

For the first part of the problem statement, to identify the variables that influence the purchase of MM, one approach is to use Principal Component Analysis (PCA). PCA will reduce the dimensionality of a dataset by transforming the original variables into a new set of uncorrelated variables. This would also handle the multi-collinearity as evident from the EDA above. However, PCA being a black box model, won't be much helpful in this case because the result won't be easily interpretable.

Other approach is to employ variable selection techniques. Since, for the second part of the problem, to build predictive model that informs about the probability of customers buying MM, there are two prescribed approaches, logistic regression and eXtreme Gradient Boosting (XG Boost), the same can be deployed for variable selection.

Logistic regression is a statistical method used for modeling the probability of a binary outcome, expressing the relationship between one or more independent variables and the probability of the event occurring.

Assumptions of logistic regression include:

1. **Linearity:** The log-odds of the dependent variable are assumed to have a linear relationship with the independent variables.
2. **Independence of Errors:** Errors in the model's predictions should be independent of each other.
3. **No Multicollinearity:** Independent variables should not be highly correlated with each other.
4. **Large Sample Size:** Logistic regression performs well with a reasonably large sample size for robust parameter estimates.
5. **No Outliers:** The absence of influential outliers that could unduly affect the model's performance.
6. **No Perfect Separation:** The independent variables should not perfectly predict the dependent variable, as it can lead to estimation issues.

EXtreme Gradient Boosting is an ensemble machine learning technique that combines the predictive power of multiple decision trees, sequentially building each tree to correct the errors of the previous ones, ultimately producing a robust and accurate predictive model.

Assumptions of eXtreme Gradient Boosting include:

1. **Additive Model Assumption:** The model assumes that the relationship between the predictors and the target variable can be expressed as an additive combination of the individual trees.
2. **Weak Learner Assumption:** The individual trees, often referred to as weak learners, should perform slightly better than random chance, allowing the boosting process to iteratively improve the model.
3. **No Collinearity:** The input features are assumed to be not highly correlated, as collinearity may affect the stability and interpretability of the model.

4. **No Outliers:** Like many machine learning models, Gradient Boosted Trees can be sensitive to outliers, potentially impacting the performance and generalization of the model.

Variable Selection Approach

There are multiple variable selection techniques such as forward selection, backward elimination, and so on. However, confounding variables can affect these techniques by leading to an overemphasis on confounder variables. This may cause the prioritization of variables associated with the confounders, even if those variables are not directly related to the outcome of interest.

For this project, five techniques in a waterfall approach has been used to arrive at the final set of variables. First, removal of variables with high correlation and with high Variance Inflation Factor (VIF). Following that, excluding variables deemed less important, i.e., those with a low coefficient absolute magnitude. Subsequently, eliminating variables based on p-values. Finally, the use of parsimony concept to further refine the selection of variables.

Treating Multicollinearity

All the assumptions for using logistic regression or gradient boosted tree are met except for multicollinearity.

By examining the correlation map, it is evident that some variables are highly correlated with each other. Another way to assess this correlation is by using the Variance Inflation Factor (VIF), which is displayed below.

Variables	Variance Inflation Factor
PriceCH	Inf
PriceMM	Inf
DiscCH	Inf
DiscMM	inf
SpecialCH	1.906880
SpecialMM	1.422967
LoyalCH	1.045105
SalePriceMM	Inf
SalePriceCH	Inf
PriceDiff	Inf
PctDiscMM	520.342425
PctDiscCH	714.050117
ListPriceDiff	Inf

As observed, the VIF is approaching infinity for most of the variables. Results from the correlation plot indicates that 'DiscMM' is perfectly collinear with 'PctDiscMM'; 'DiscCH' is perfectly collinear with 'PctDiscCH.' Additionally, 'SalePriceMM' has a correlation of 0.85 with 'PriceDiff'. 'ListPriceDiff' and 'PriceDiff' are displaying same information i.e. price difference before and after discount between CH & MM.

For the first iteration, removing 'DiscMM', 'DiscCH', ‘‘SalePriceMM’, and ‘ListPriceDiff’ variables. The resulting VIF are:

Variables	Variance Inflation Factor
PriceCH	508.157273
PriceMM	197.041538
SpecialCH	1.906880
SpecialMM	1.422967
LoyalCH	1.045105
SalePriceCH	1372.866858
PriceDiff	838.554729
PctDiscMM	520.342425
PctDiscCH	714.050117

The industry standard recommends that the VIF greater than 5 indicates high correlation. After removing the above four variables, although VIF is not infinite for any variables, it still exceeds 5 for most of the variables.

Since variables are still highly corelated, further processing and removal of variables are processed through recursive feature elimination using logistic regression and feature importance from gradient boosted trees.

Variable Selection Using P-value

The below table shows the remaining 9 variables in logistic regression with p-values of each variable, along with the absolute magnitude of their coefficients, sorted in descending order.

Feature	Coefficient	Standard Error	Z-Score	P-value
LoyalCH	-1.935326	0.034822	-55.5778	0
PriceDiff	-0.90208	0.969006	-0.93093	0.351888
PctDiscMM	-0.311497	0.761309	-0.40916	0.682422
PriceMM	-0.128333	0.469359	-0.27342	0.784529
SpecialMM	0.126558	0.04069	3.110311	0.001869
SalePriceCH	0.099561	1.266084	0.078637	0.937321
PriceCH	0.066476	0.784371	0.084751	0.93246
PctDiscCH	-0.047821	0.928746	-0.05149	0.958935
SpecialCH	-0.037449	0.046148	-0.8115	0.417076

Eliminating variables SpecialCH, PctDiscCH, PriceCH, SalePriceCH based on their low coefficient value and re-running model with the remaining variables.

Feature	Coefficient	Standard Error	Z-Score	P-value

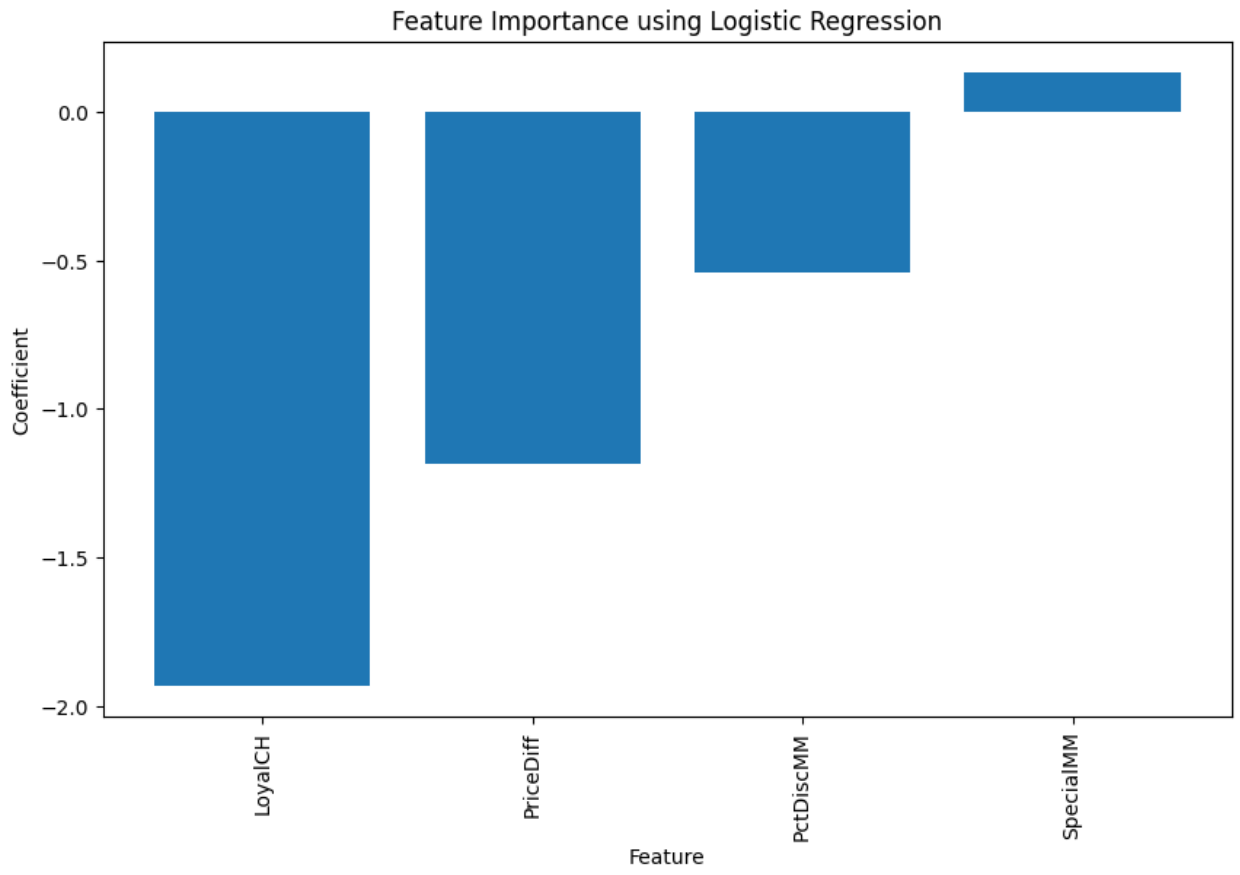
LoyalCH	-1.93493	0.034614	-55.8997	0
SpecialMM	0.134981	0.038749	3.483519	0.000495
PriceDiff	-1.22528	0.070362	-17.4139	0
PriceMM	0.03948	0.039149	1.008458	0.313235
PctDiscMM	-0.57319	0.067932	-8.43763	0

Eliminating PriceMM based on p-value >0.05 and re-running model with the remaining variables.

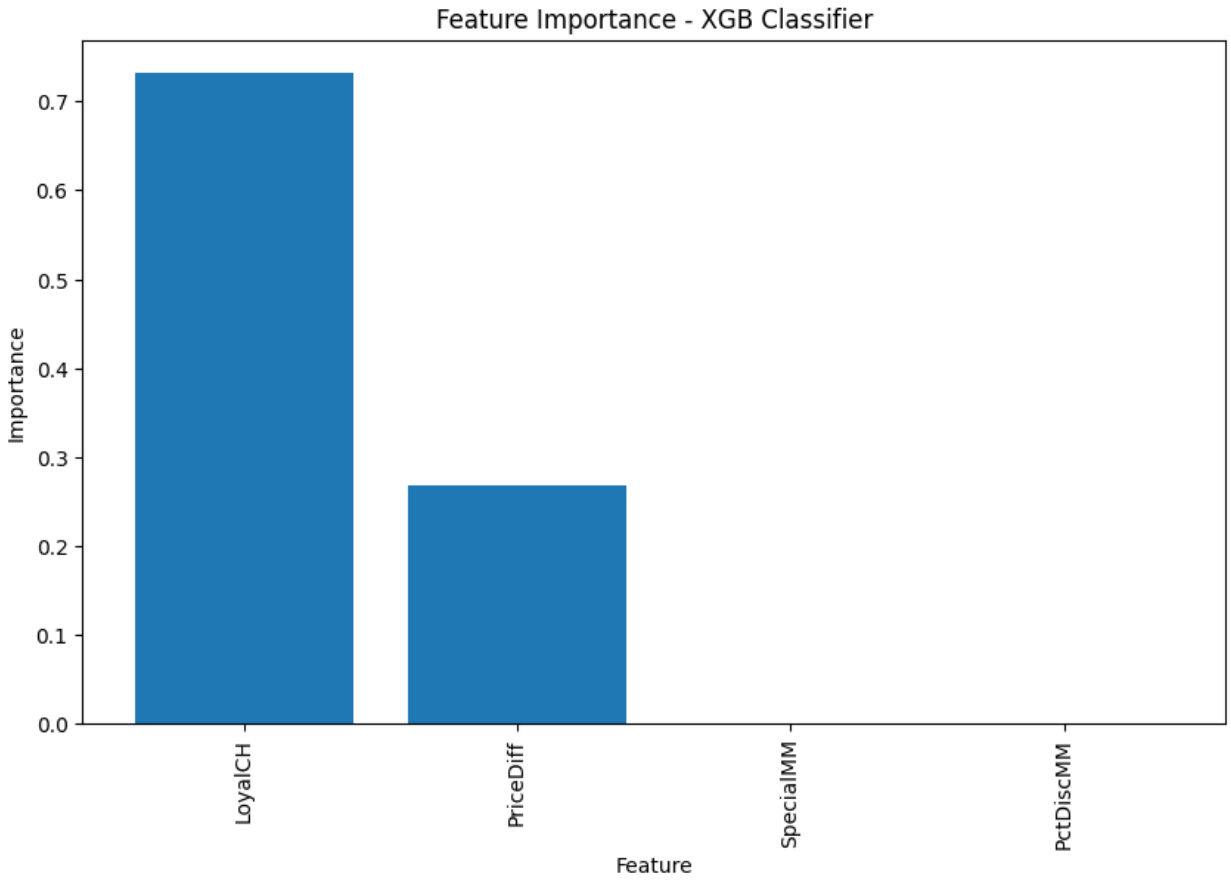
Feature	Coefficient	Standard Error	Z-Score	P-value
LoyalCH	-1.93171	0.034569	-55.879	0
SpecialMM	0.134699	0.038734	3.477546	0.000506
PriceDiff	-1.18249	0.061901	-19.1031	0
PctDiscMM	-0.53992	0.062626	-8.62136	0

Feature Importance using Logistic Regression

Below is the feature importance graph, and all variables have p-values less than 0.05.



Feature Importance using XGB Classifier



As observed above, both in logistic regression and gradient boosting techniques, 'LoyalCH' is the most important variable, followed by 'PriceDiff'. However, in logistic regression, PctDiscMM is the third most important, followed by SpecialMM. In the gradient boosting approach, SpecialMM and PctDiscMM are not at all important.

Building Predictive Models

AUC-ROC (Area Under the Receiver Operating Characteristic Curve), precision, and recall are used to evaluate the performance of logistic regression and gradient boosting model. These metrics collectively provide a comprehensive assessment of the model's predictive capabilities, balancing different aspects of classification performance.

Area Under the Receiver Operating Characteristic Curve

The Area Under the Receiver Operating Characteristic (AUC-ROC) is a metric that quantifies the performance of a binary classification model by measuring the area under its ROC curve, which plots the true positive rate against the false positive rate across different classification thresholds. A higher AUC-

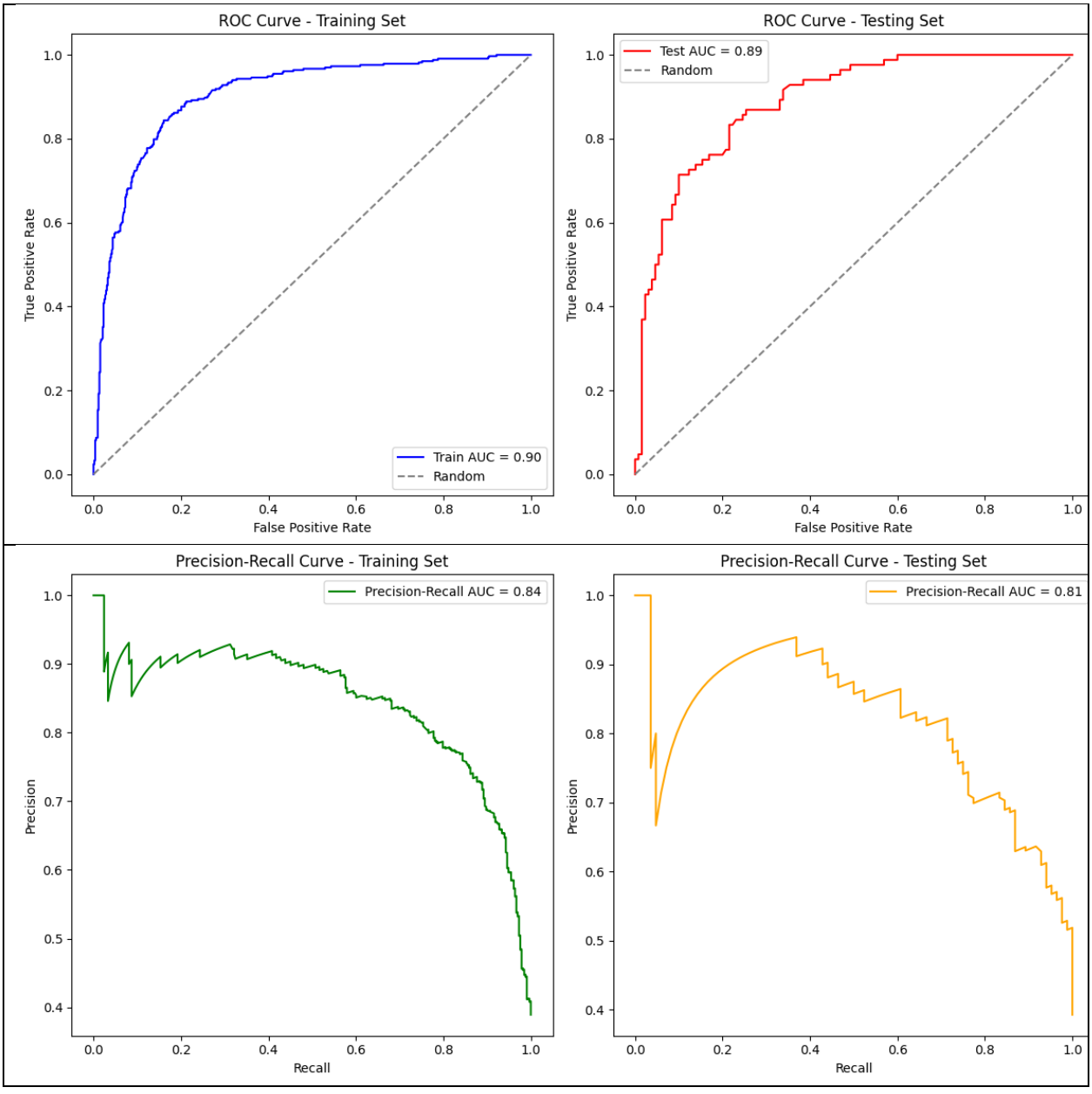
ROC value (closer to 1) indicates better discrimination ability, making it a widely used metric to assess and compare the overall performance of binary classifiers.

Precision, Recall

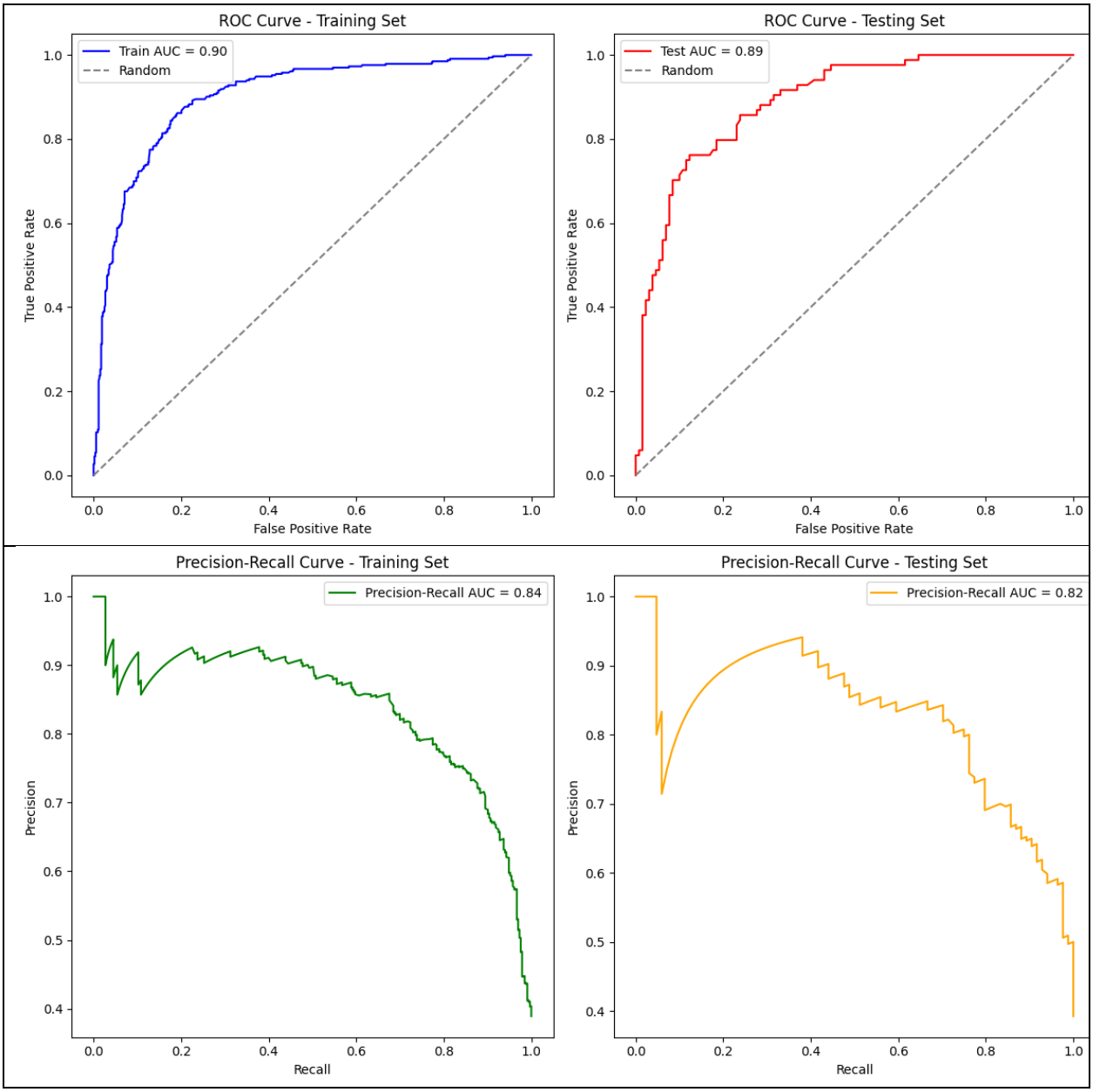
Precision measures the accuracy of positive predictions, recall assesses the model's ability to capture all relevant instances of the positive class.

Performance of Predictive Model using Logistic regression

Utilizing the earlier selected four variables 'LoyalCH', 'SpecialMM', 'PriceDiff', and 'PctDiscMM', the training AUC-ROC is 0.90, and the testing AUC-ROC is 0.89. Additionally, the training AUC for the precision-recall curve is 0.84, while the testing AUC is 0.81.



Now, the goal is to achieve parsimony, selecting fewer variables while maintaining the same performance as the aforementioned four variables. It was discovered that by using only 'LoyalCH' and 'PriceDiff' variables, the model achieves comparable performance, as evidenced by the following graphs. The training AUC-ROC is 0.90, and the testing AUC-ROC is 0.89. Additionally, the AUC for the precision-recall curve is 0.84 for training and 0.82 for testing



The following are the variable coefficients obtained from logistic regression:

Features	Coefficient

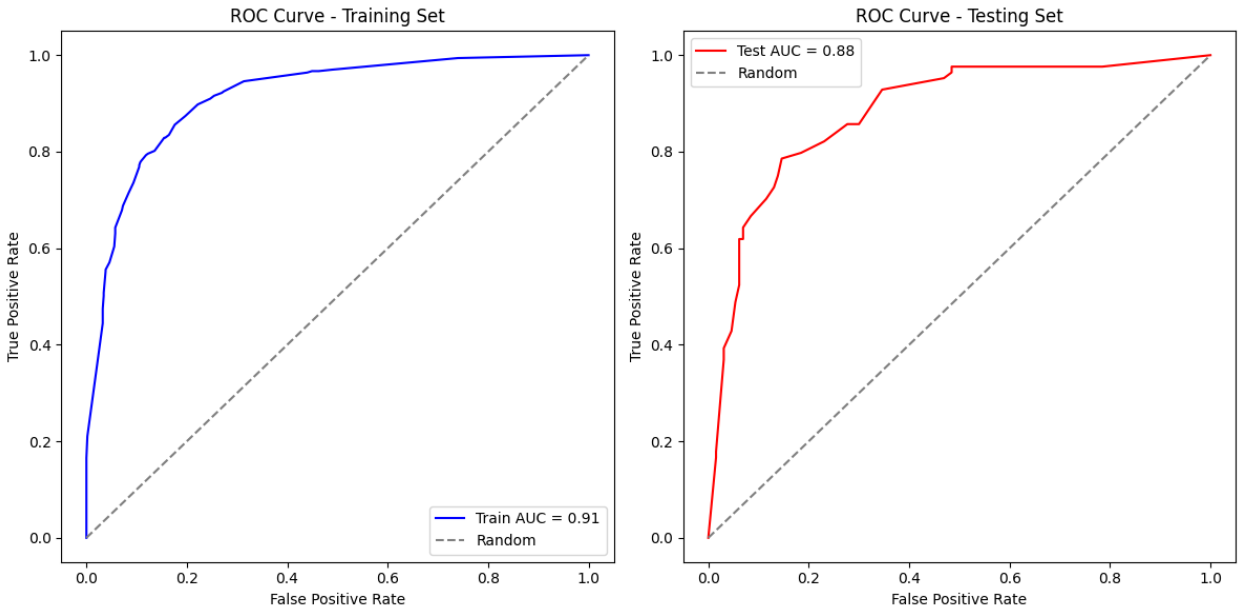
LoyalCH	-1.96
PriceDiff	-0.75
Intercept	-0.80

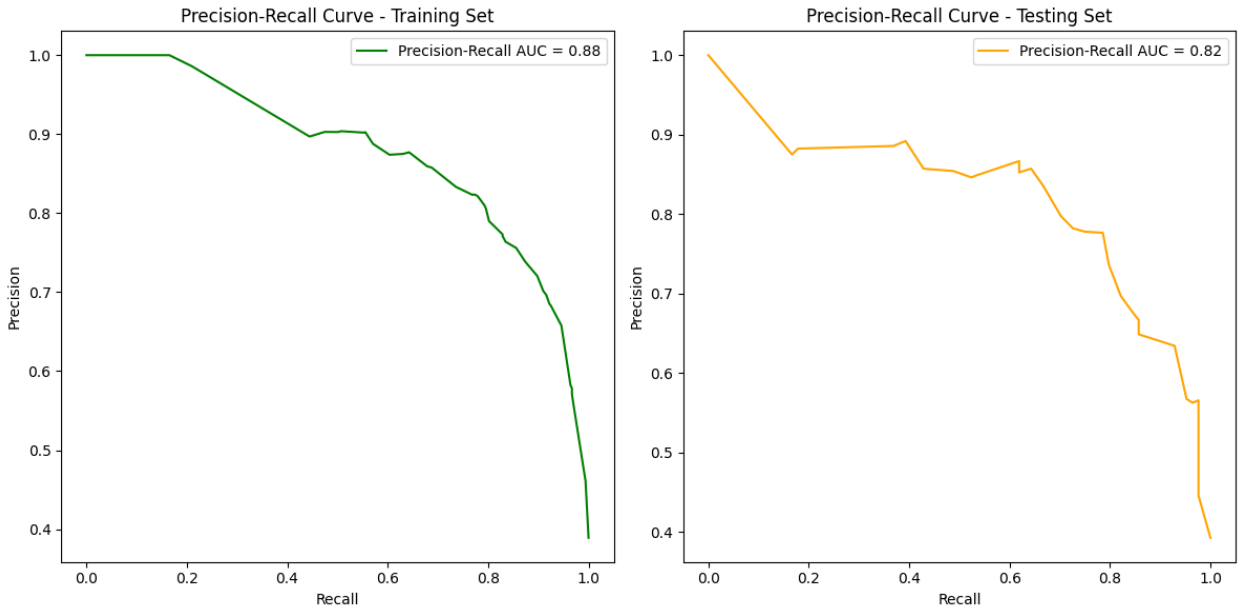
The VIF for the selected two variables is less than 5:

Features	VIF
LoyalCH	1.01099
PriceDiff	1.01099

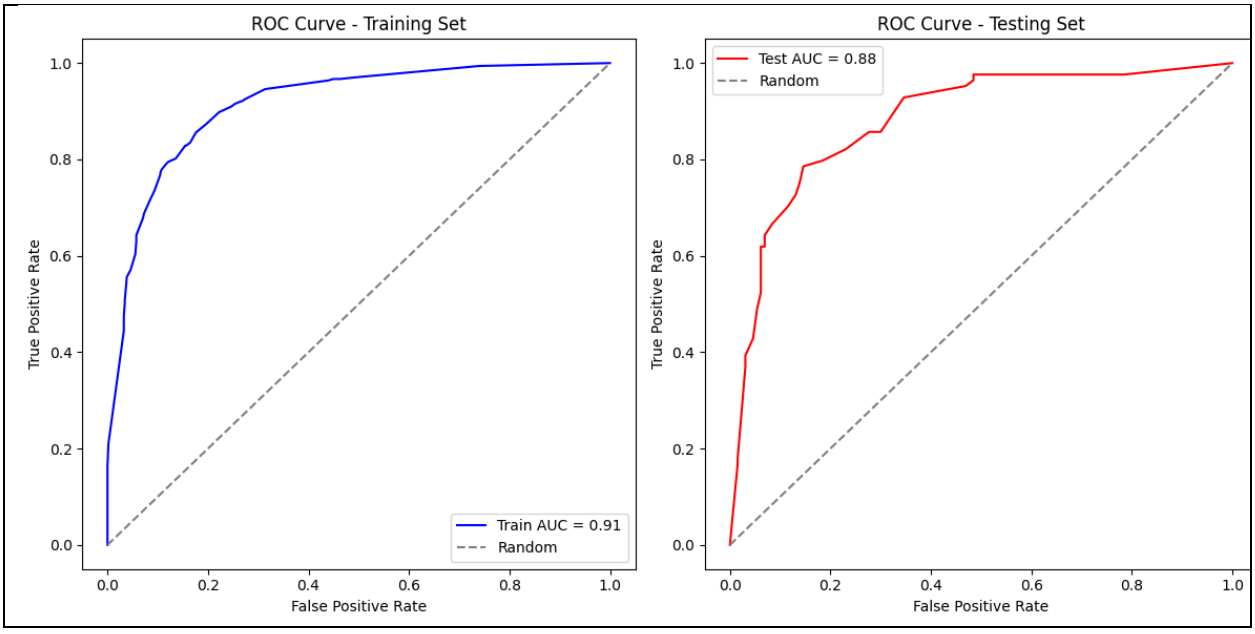
Performance of Predictive Model using Gradient Boosted Trees

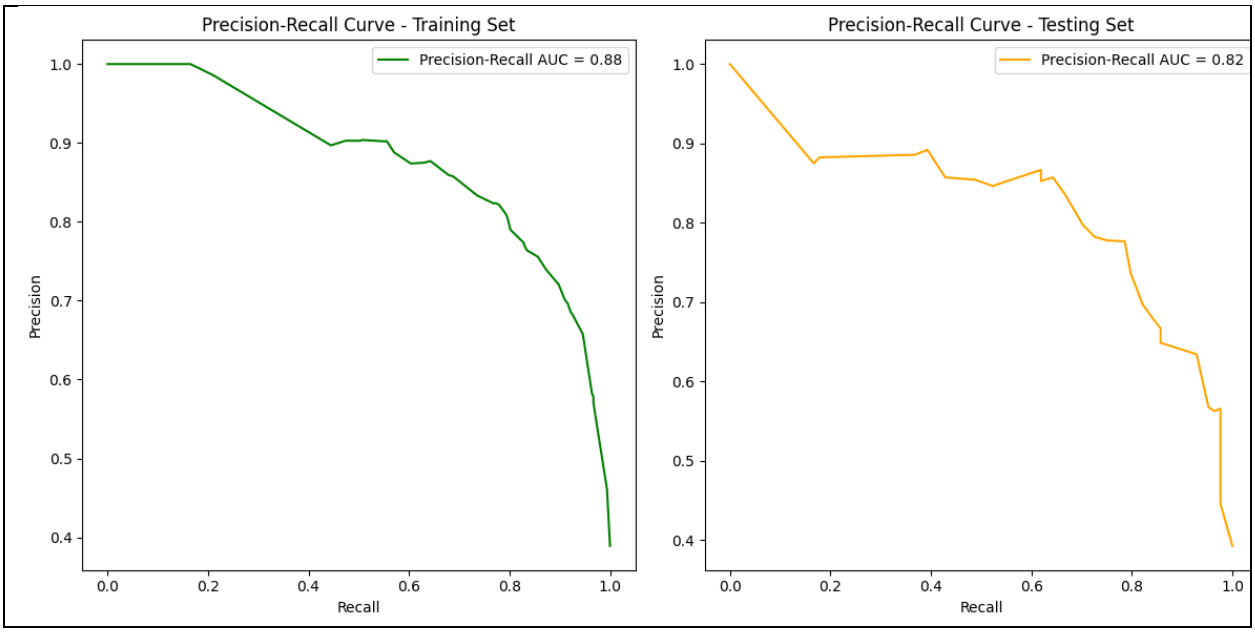
Using the same four variables 'LoyalCH', 'SpecialMM', 'PriceDiff' , and 'PctDiscMM', the training AUC-ROC is 0.91, and the testing AUC-ROC is 0.88. Additionally, the training AUC for the precision-recall curve is 0.88, while the testing AUC is 0.82.





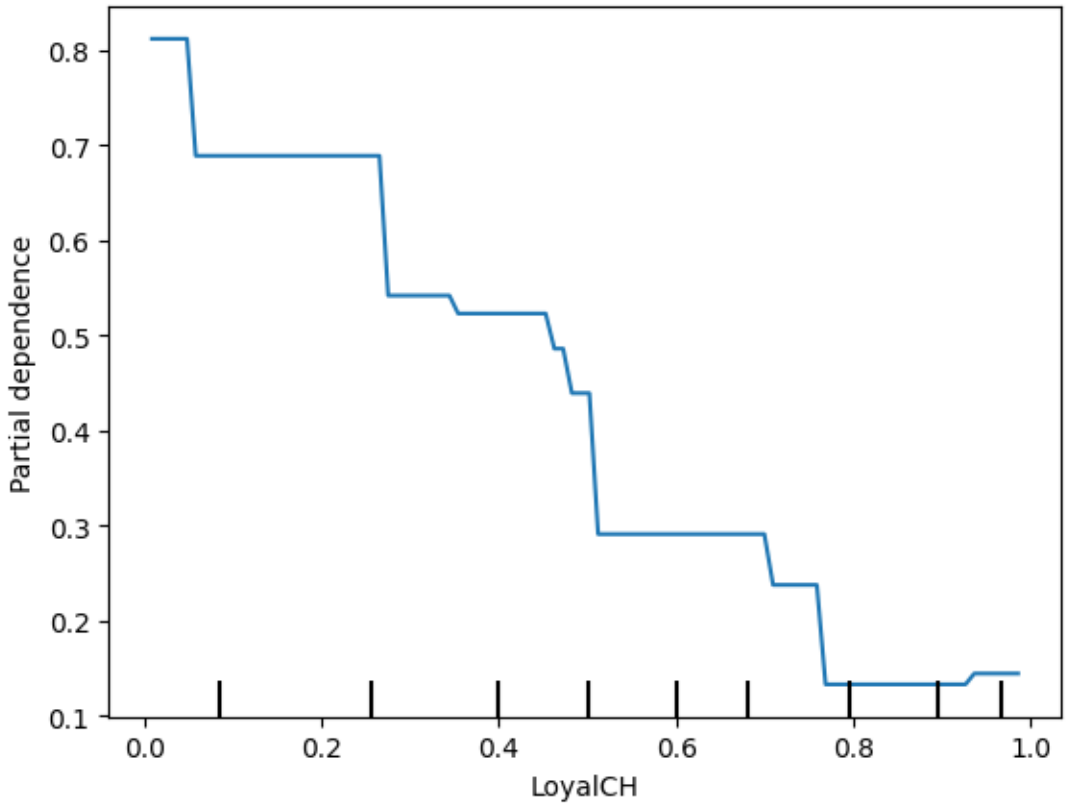
The training AUC-ROC is 0.91, and the testing AUC-ROC is 0.88 for parsimonious model with only two variable. Additionally, the training AUC for the precision-recall curve is 0.88, while the testing AUC is 0.82.



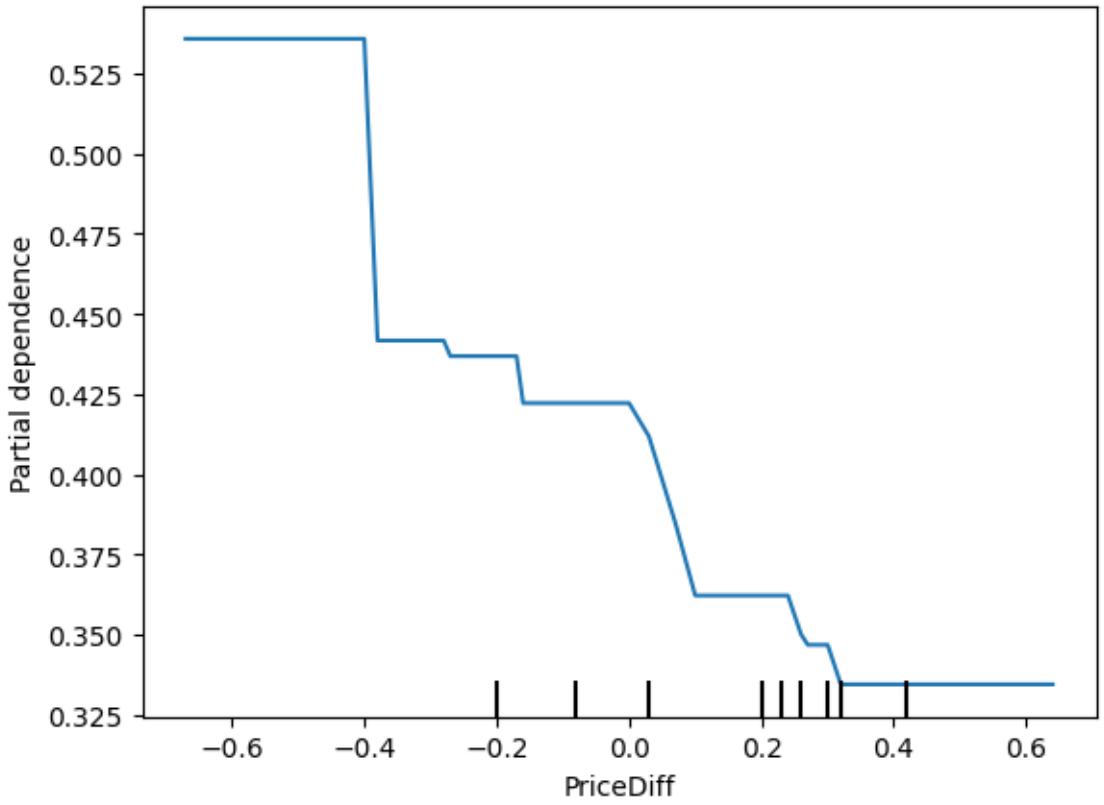


Partial Dependence Plot using XG Boost

By plotting the partial dependence plot for two selected variables, it can be observed that as the value of 'LoyalCH' increases, the probability of a customer purchasing MM decreases, keeping other factors constant.



By examining the partial dependence plot for PriceDiff, it can be concluded that as the value of the price difference increases, i.e., when the sale price of MM is greater than the sale price of Citrus Hill, the probability of a customer purchasing MM decreases, and vice versa, keeping other factors constant.



In conclusion, for LoyalCH, if the value is greater than 0.5, the probability of a customer purchasing MM is very low. On the other hand, for values less than 0.3, the probability of a customer purchasing MM is high. Additionally, for LoyalCH values between 0.3 to 0.5, a smaller price difference (i.e., the sale price of MM being less than the sale price of CH) will increase the probability of a customer purchasing MM.

Results and Conclusion

For Brand Manager

1. What predictor variables influence the purchase of MM?

As per to logistic regression model, four important variables influencing the purchase of MM are LoyalCH, SpecialMM, PriceDiff, and PctDiscMM. On the other hand, the gradient boosting algorithm identifies two important variables: LoyalCH, and PriceDiff. Remarkably, both methods converge on two variables—LoyalCH, and PriceDiff. In a pursuit of parsimony, LoyalCH and PriceDiff alone suffice and perform comparably to the original set of five variables in influencing MM purchases.

2. Are all the variables in the dataset effective or are some more effective than others?

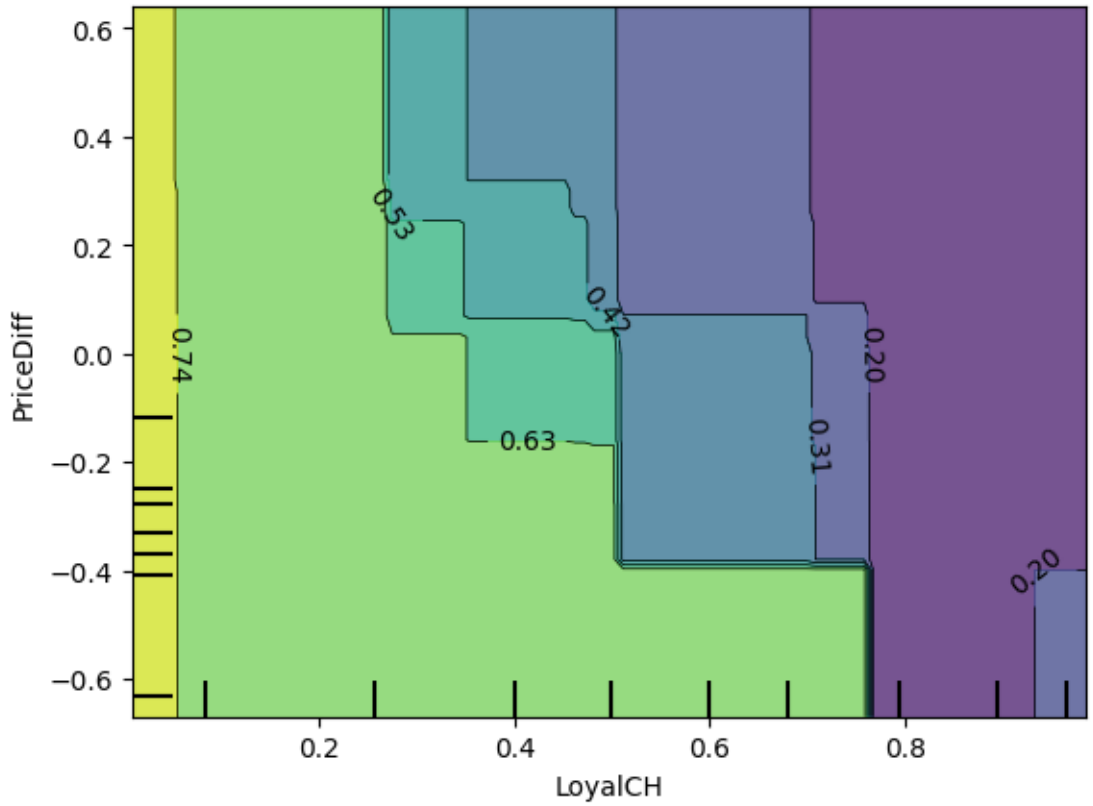
All variables in the dataset have an impact on influencing the purchase of MM, with LoyalCH being the most influential variable, followed by PriceDiff.

3. How confident are you in your recommendations?

No model is perfect, but based on the model result, it is safe to conclude 95% confidence (p-value <5%) in saying that LoyalCH is the most important factor affecting MM purchases. Following that, PriceDiff is the next significant factor in influencing the MM purchases.

4. Based on your analysis what are the specific recommendations you have for the brand manager?

Focusing on LoyalCH and PriceDiff variables would increase the likelihood of customers purchasing MM. If the LoyalCH value is greater than 0.5, there is a lower chance that the customer will buy MM, while a value less than 0.3 indicates a higher likelihood of MM purchase. For values between 0.3 and 0.5, price difference becomes critical. In this range, trying to keep the sale price of MM lower than CH, increases the chances of customers purchasing MM. This is evident by the partial dependence plot below:



For Sales Manager

1. Can you build a predictive model that can inform him the probability of customers buying MM?

Yes. For this task, the logistic regression model is good model to predict the probability of customers purchasing MM. This decision was influenced by the lower variability observed in both training and testing performance for AUC-ROC and AUC for the precision-recall curve. Additionally, logistic regression offers higher interpretability.

2. How good is the model in its predictions?

The model's performance, as indicated by the AUC-ROC and AUC for the precision-recall curve in the sections above, suggests that it is quite effective. An AUC-ROC of 0.9 for the training set and 0.89 for the

testing set indicates a strong ability to discriminate between classes. Similarly, AUC values of 0.84 for training and 0.82 for testing in the precision-recall curve suggest good precision and recall trade-offs.

3. How confident are you in your recommendations?

The high AUC-ROC scores (0.9 for training and 0.89 for testing) indicate a strong discriminatory power of the model in correctly classifying instances. The AUC values of 0.84 for training and 0.82 for testing in the precision-recall curve signify good precision and recall trade-offs, emphasizing the model's ability to balance correctly identified positive instances with minimizing false positives.

The statistical metrics, combined with the consistency between training and testing performance, provide a robust indication of the model's reliability. The high AUC scores and precision-recall curve values contribute to a statistically significant level of confidence in the model's predictive accuracy and effectiveness.

Based on the p-values associated with the logistic regression coefficients, where all p-values are less than 0.05, the model have a confidence rate of 95%. The statistical significance of these coefficients supports the reliability of the model, indicating that the included predictor variables, namely LoyalCH and PriceDiff, have a significant impact on the predicted outcome.

Recommendations

LoyalCH and PriceDiff emerge as the two most influential variables affecting customers' choices when purchasing MM. It is recommended to focus on the two variables to increase the likelihood of customers purchasing MM. If the LoyalCH value is greater than 0.5, there is a lower chance that the customer will buy MM, while a value less than 0.3 indicates a higher likelihood of MM purchase. For values between 0.3 and 0.5, price difference becomes critical. In this range, trying to keep the sale price of MM lower than CH increases the chances of customers purchasing MM. This recommendation is based on logistic regression to calculate the probability of a customer buying MM. The model exhibits a high discriminatory power, and indicates 95% confidence in prediction, given the statistical significance of the variables.

Code Link

https://drive.google.com/file/d/1I3LCfeX3G6mttMv4W8jBGWZfigncbU8f/view?usp=sharing