The Complete Journey

-

Dunnhumby

CAPSTONE : GROUP - G6

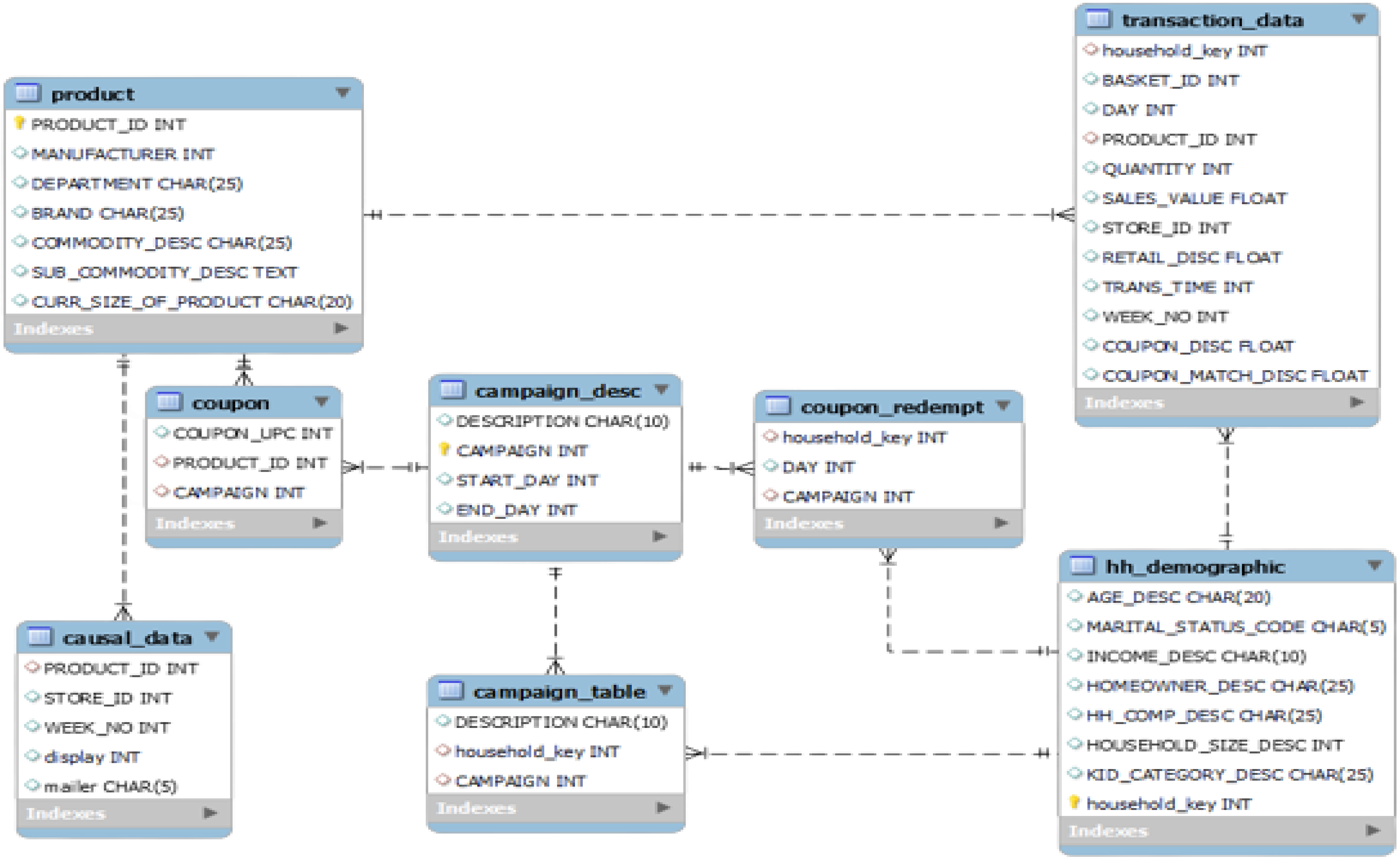
Table of Contents

ABOUT THE DATA

DATA DESCRIPTION

This dataset contains household level transactions over two years from a group of 2,500 households who are frequent shoppers at a retailer. It contains all of each household's purchases, not just those from a limited number of categories. For certain households, demographic information as well as direct marketing contact history are included.

ER DIAGRAM



product
PRODUCT_ID INT
MANUFACTURER INT
DEPARTMENT CHAR(25)
BRAND CHAR(25)
COMMODITY_DESC CHAR(25)
SUB_COMMODITY_DESC TEXT
CURR_SIZE_OF_PRODUCT CHAR(20)
Indexes
transaction_data
household_key INT
BASKET_ID INT
DAY INT
PRODUCT_ID INT
QUANTITY INT
SALES_VALUE FLOAT
STORE_ID INT
RETAIL_DISC FLOAT
TRANS_TIME INT
WEEK_NO INT
COUPON_DISC FLOAT
COUPON_MATCH_DISC FLOAT
Indexes
coupon
COUPON_UPC INT
PRODUCT_ID INT
CAMPAIGN INT
Indexes
campaign_desc
DESCRIPTION CHAR(10)
CAMPAIGN INT
START_DAY INT
END_DAY INT
Indexes
coupon_redempt
household_key INT
DAY INT
CAMPAIGN INT
Indexes
hh_demographic
AGE_DESC CHAR(20)
MARITAL_STATUS_CODE CHAR(5)
INCOME_DESC CHAR(10)
HOMEOWNER_DESC CHAR(25)
HH_COMP_DESC CHAR(25)
HOUSEHOLD_SIZE_DESC INT
KID_CATEGORY_DESC CHAR(25)
household_key INT
Indexes
causal_data
PRODUCT_ID INT
STORE_ID INT
WEEK_NO INT
display INT
mailer CHAR(5)
Indexes
campaign_table
DESCRIPTION CHAR(10)
household_key INT
CAMPAIGN INT
Indexes

DATA CONSIDERED :

Campaign_description	Campaign_table	Causal_data	Product
Coupon_redempt	Coupon	hh_demographic	Transaction_data

TABLES DESCRIPTION

CAMPAIGN TABLE

Gives the detail of the type of campaign and the households/customers recieved a specific campaign

COUPON_REDEMPT

This table provides with the details of the redemption of the coupon households campaign wise providing the day of redemption and _upc.

TRANSACTION_DATA TABLE

This table contains the transaction of past 2 years of 2500 households. But we need to clean the data as we are excluding last 3 campaigns.

COUPON TABLE

This table lists all the coupons sent to customers as part of a campaign, as well as the products for which each coupon is redeemable.

CAMPAIGN DESCRIPTION

The basic and essential information of each campaign thier type and Start and End day.

HH_DEMOGRAPHIC TABLE

This table contains demographic information for a portion of households. Due to nature of the data, the demographic information is not available for all households

PRODUCT TABLE

This table contains information on each product sold such as type of product, national or private label and a brand identifier.

CAUSAL_DATA TABLE

This table signifies whether a given product was featured in the weekly mailer or was part of an in-store display (other than regular product placement)

DATA ANOMALIES

- ✓ Transactions have a "COUPON_DISC" column that shows value of a coupon when it got applied, but some discounts were applied to products that did not receive coupons according to campaign table.
- ✓ There are also some transactions where the quantity and sales value were marked as 0 without any further explanation. Additionally, in some cases, the retail discount was higher than the sales value of the product.
- ✓ CouponUPCS aren't unique with respect to campaigns, couponUPCS don't contain unique products.Products are provided with multiple couponupcs.

Multiple manufacturers provided same couponupcs for some products.We are not sure how many coupons were provided to each household while campaign visited them. Transaction table consists coupondisc column for the transactions for where coupon discount is applied. No elaboration given. For some transactions. Retail discount is greater than the salesvalue for that product

Our Approach

To understand how customers, interact with a business, a comprehensive analysis of sales data is conducted.

This analysis aims to identify trends that may negatively impact the business's growth, as well as trends that may contribute to its growth.

The insights gained from this analysis are data-driven and can assist management in making informed decisions about the future direction of the business.

An in-depth analysis of direct marketing data is carried out to study the effectiveness of promotional activities done by the retailer.

We’ve tried to mine all types of patterns from the engagement of customers in those marketing campaigns. We’ve also attempted to bring out loopholes and ineffective practices that lead to underutilization of resources.

Coupon Redemption
(households)

PROBLEM STATEMENT

Model building for the coupon redemption problem can completely depend on the demographic information of the households . Hh_demographic table contains the demographic information of 801 households . So whole analysis will be done for the 801 households . Aim is to build a model to read the pattern of the households which defines the redemption criteria of the coupons depending on the demographic information of the households.

DATA CONSIDERED :

Coupon

Campaign_table

Transaction_data

Coupon_redempt

hh_demographic

DATA DESCRIPTION

We know from the dataset that 70% of the customers never use the coupons they receive and this would lead to a waste of money and time for the company.

```
Total No. Of Households : 2500
No. Of Households To Which Coupons Were Provided : 1584
No. Of Households Who Redmeed Coupons : 434
Percentage Of Households Where Coupons Were Unused : 72.6 %
```

TARGET VARIABLE CREATION

Following Two tables were used for creating the Target variable:

1. “Coupon_redempt" table contains information about coupons redeemed by unique 434 households.

2. "Hh_demographic" table contains demographic information about 801 households.

- The coupon redempt table and hh_demographic table can be related through the household identifier (or "household key") that is present in both tables. By using the household identifier present in both tables, it is possible to link information about coupon redemptions with the demographic characteristics of the households that made those redemptions. Hence those 311 households redeemed the coupon and having demographic details are marked as 1 and remaining are marked as 0.

```
0    490
1    311
Name: Target
```

FEATURE CREATION

H0: Variable is not significant
HA: Variable is significant

CAMPAIGN_TABLE :

- Number of Campaigns received by unique households

	household_key	cnt_camp_recieved_per_hsld
0	1	8
1	2	1
2	3	3

	F_onewayresult
Test_Statistics	2.112970e+03
pvalue	9.050725e-290

COUPON_REDEMPT :

- Household wise count of campaigns in which the coupon were redempt.

	household_key	distinct_camprdmptn_per_hsld
0	1	2
1	8	1
2	13	7

	F_onewayresult
Test_Statistics	7.039228e+01
pvalue	1.097288e-16

- Coupon redemption rate : It is the ratio of no. of campaigns received and coupon redemption on those campaigns.

	household_key	cnt_camp_recieved_per_hsld	distinct_camprdmptn_per_hsld	camp_rdmptn_rate
2316	2317	17.0	3.0	0.176471
2488	2489	16.0	6.0	0.375000
717	718	15.0	5.0	0.333333

	F_onewayresult of camp_rdmptn_rate
Test_Statistics	1.750850e+02
pvalue	6.600895e-38

TRANSACTION_DATA :

- Total sales value per household : It is the sum of the purchased value made by each households.

	household_key	TOTAL_SALES_VALUE_hsld_wise
0	1	4330.16
1	2	1954.34
2	3	2653.21

	F_onewayresult of TOTAL_SALES_VALUE_hsld_wise
Test_Statistics	1.797826e+03
pvalue	7.944291e-260

- Mean items purchase per transaction : Average numbers of items purchased in a single transaction.

	household_key	mean_items_purch_per_trans
0	1	23.0
1	2	19.0
2	3	182.0

	F_onewayresult of mean_items_purch_per_trans
Test_Statistics	4.597601e+02
pvalue	2.609885e-89

- Number of Total visits : Total number of visits made by unique households over the span of two years.

	household_key	No_of_total_visits
0	1	85
1	2	45
2	3	47

	F_onewayresult of No_of_total_visits
Test_Statistics	1.403534e+03
pvalue	4.536211e-218

ALL THE FEATURES HAVE P VALUE < 0.05 % WHICH MEANS THAT WE REJECT THE H0 , i.e ALL THE FEATURES ARE SIGNIFICANT.

- MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS : Average purchase made by unique households in a single transaction.

	household_key	MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS
0	1	50.94
1	2	43.43
2	3	56.45

	F_onewayresult of MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS
Test_Statistics	2.324380e+03
pvalue	1.967052e-308

- MEDIAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS : Median purchase made by unique households in a single transaction.

	household_key	MEDIAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS
0	1	49.33
1	2	26.94
2	3	36.38

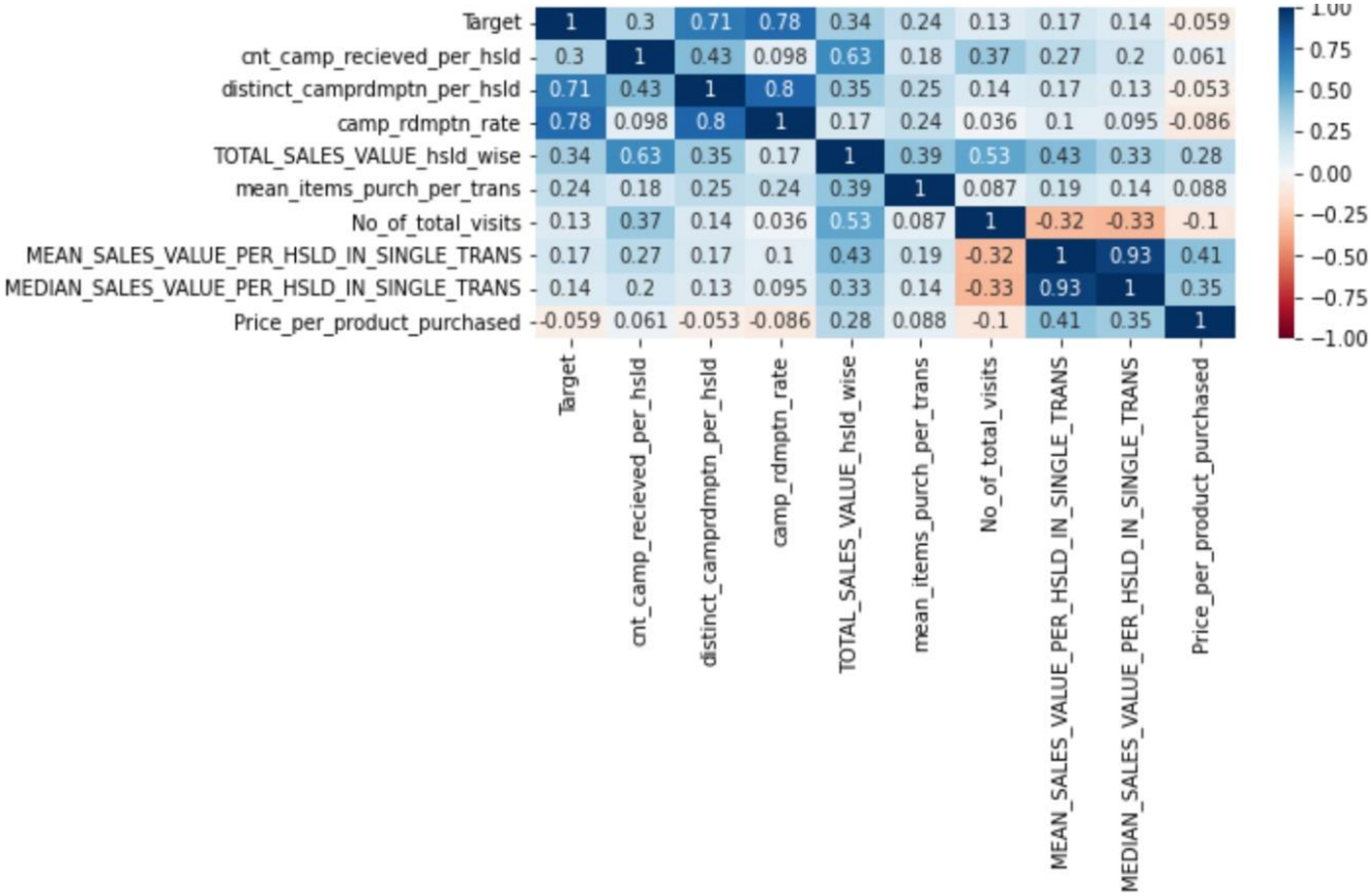
	F_onewayresult
Test_Statistics	1.598726e+03
pvalue	2.108284e-239

- Price per product purchased : It is the average price of product generally purchased by unique household.

	household_key	Price_per_product_purchased
0	1	2.31
1	2	2.53
2	3	1.98

	F_onewayresult of Price_per_product_purchased
Test_Statistics	6913.179884
pvalue	0.000000

RELATIONSHIP BETWEEN VARIABLES



DATA PREPROCESSING

- OUTLIER TREATMENT :

Outlier treatment by IQR (Interquartile Range) is a statistical method used to identify and remove outliers from a dataset.
Replace the outlier with either the upper or lower bound depending on the direction of the outlier.

- SCALING :

Standardization Method was apply to Scale the Data: This method scales the data to have a mean of 0 and a standard deviation of 1. It is calculated by subtracting the mean of the variable from each data point and dividing the result by the standard deviation of the variable.

- ENCODING :

Frequency Encoding: AGE_DESC, HH_COMP_DESC
Label Encoding: MARITAL_STATUS_CODE', HOMEOWNER_DESC
Ordinal Encoding: INCOME_DESC

.

➢ MULTICOLINEARITY TREATMENT :

HOUSEHOLD_SIZE_DESC and MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS are highly correlated with each other so these columns were dropped.
dropped_columns = [HOUSEHOLD_SIZE_DESC, MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS]

	VIF
distinct_camprdmptn_per_hsld	7.294024
camp_rdmptn_rate	6.258057
TOTAL_SALES_VALUE_hsld_wise	5.946472
HH_COMP_DESC	4.951355
AGE_DESC	4.726221
No_of_total_visits	4.598817
INCOME_DESC	4.144141
MEDIAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS	3.335953
HOMEOWNER_DESC	3.037176
cnt_camp_recieved_per_hsld	2.416929
MARITAL_STATUS_CODE	2.234695
KID_CATEGORY_DESC	1.503506
Price_per_product_purchased	1.375146
mean_items_purch_per_trans	1.263182

Conclusion

MODELLING

BASE MODEL : LOGIT

Logit Regression Results

Dep. Variable:	Target	**No. Observations:**	608
Model:	Logit	**Df Residuals:**	592
Method:	MLE	**Df Model:**	15
Date:	Fri, 31 Mar 2023	**Pseudo R-squ.:**	0.9994
Time:	21:04:17	**Log-Likelihood:**	-0.25816
converged:	False	**LL-Null:**	-411.43
Covariance Type:	nonrobust	**LLR p-value:**	1.436e-165

	coef	std err	z	P>|z|	[0.025	0.975]
const	25.3541	323.571	0.078	0.938	-608.834	659.542
AGE_DESC	-20.9215	317.758	-0.066	0.948	-643.716	601.873
MARITAL_STATUS_CODE	-6.6024	119.638	-0.055	0.956	-241.089	227.884
INCOME_DESC	1.1665	14.522	0.080	0.936	-27.297	29.630
HOMEOWNER_DESC	5.4353	165.064	0.033	0.974	-318.084	328.955
HH_COMP_DESC	-49.8100	1008.843	-0.049	0.961	-2027.107	1927.487
KID_CATEGORY_DESC	-5.4639	81.439	-0.067	0.947	-165.081	154.153
cnt_camp_recieved_per_hsld	6.3010	196.527	0.032	0.974	-378.885	391.487
distinct_camprdmptn_per_hsld	27.9580	288.017	0.097	0.923	-536.545	592.461
camp_rdmptn_rate	28.1617	626.675	0.045	0.964	-1200.099	1256.422
TOTAL_SALES_VALUE_hsld_wise	0.3524	136.846	0.003	0.998	-267.861	268.565
mean_items_purch_per_trans	2.3650	45.599	0.052	0.959	-87.007	91.737
No_of_total_visits	-0.7982	174.659	-0.005	0.996	-343.123	341.527
MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS	-2.0221	140.362	-0.014	0.989	-277.126	273.082
MEDIAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS	-1.5761	171.091	-0.009	0.993	-336.908	333.755
Price_per_product_purchased	-0.8270	102.770	-0.008	0.994	-202.253	200.599

ALL MODELS :

	Model	accuracy_score	f1_score	recall_score	precision_score
0	dt	1.000000	1.00000	1.000000	1.000000
1	nb	1.000000	1.00000	1.000000	1.000000
3	rfc	1.000000	1.00000	1.000000	1.000000
4	ada	1.000000	1.00000	1.000000	1.000000
5	gbm	1.000000	1.00000	1.000000	1.000000
6	xgb	1.000000	1.00000	1.000000	1.000000
2	knn	0.967105	0.95935	0.951613	0.967213

Coupon Redemption
(Products)

PROBLEM STATEMENT

Model building for the coupon redemption problem can wholly depend on the product and coupon details. Product table contains the detailed information about the 92,353 products available.

Therefore, whole analysis will be done for the 44,000 unique products. Our aim is to build a model to read the pattern of the products being purchased and the Coupon discount being offered on them which defines the redemption criteria of the coupons depending on the detailed products description.

DATA CONSIDERED :

Coupon

Transaction_data

Hh_demographic

Product

METHODOLOGY

We followed the following steps:-

- Understanding the data and the context of the problem statement.
- Univariate Analysis to understand and mine pattern of each variable.
- Bivariate Analysis to understand the impact of the predictors on the Target Variable.
- Treatment of the Missing Values & Outliers.
- Feature Engineering.
- Scaling of the numerical data.
- Encoding of the Categorical Data
- Modelling

Data Preprocessing:

(Encoding, Scaling & Multicollinearity)

- Encoding 1 => Target Encoding
- Encoding 2 => pd.factorize

Finalizing the encoding method:

- Target encoding is rejected as it contains too many null values.
- Thus selected encoding method is pd.factorize.

TARGET VARIABLE CREATION

- The products table contains data of around 92 thousand products.
- For around 44 thousand products, coupons have been provided for.
- Around 48 thousand were such products for which the coupons were not provided for.
- We are uncertain how the retailer determined the products for which coupons were available.
- To develop a challenge scenario, we recognize that the retailer has selected the products that have coupons, randomly.
- Researching the likely influence of coupons on the product and its sales data is our goal.
- Now, all those products that were randomly selected for the promotions and campaigns will be our training set
- Those products that we did not provide with coupons beforehand => will turn out into `real test set`.

```
-1.0    48220
 0.0    37995
 1.0     6138
Name: Target, dtype: int64
```

Feature Engineering

No. of households that purchased products	unique count of households that has purchased that particular product over the span of 2 years.
No. of stores selling products	unique count of stores that sells that particular product.
Total quantity sold	count of products sold during the span of 2 years.

FEATURES AND THEIR STATISTICAL TESTS

num_hsld_prch_prd	num_stores_has_prd	quantity_sold_total
3.0	3.0	6.0
1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	2.0
...	...	...
1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	1.0

H0: Variable is not significant
HA: Variable is significant

```
num_hsld_prch_prd : 0.0
num_stores_has_prd : 0.0
quantity_sold_total : 0.0
```

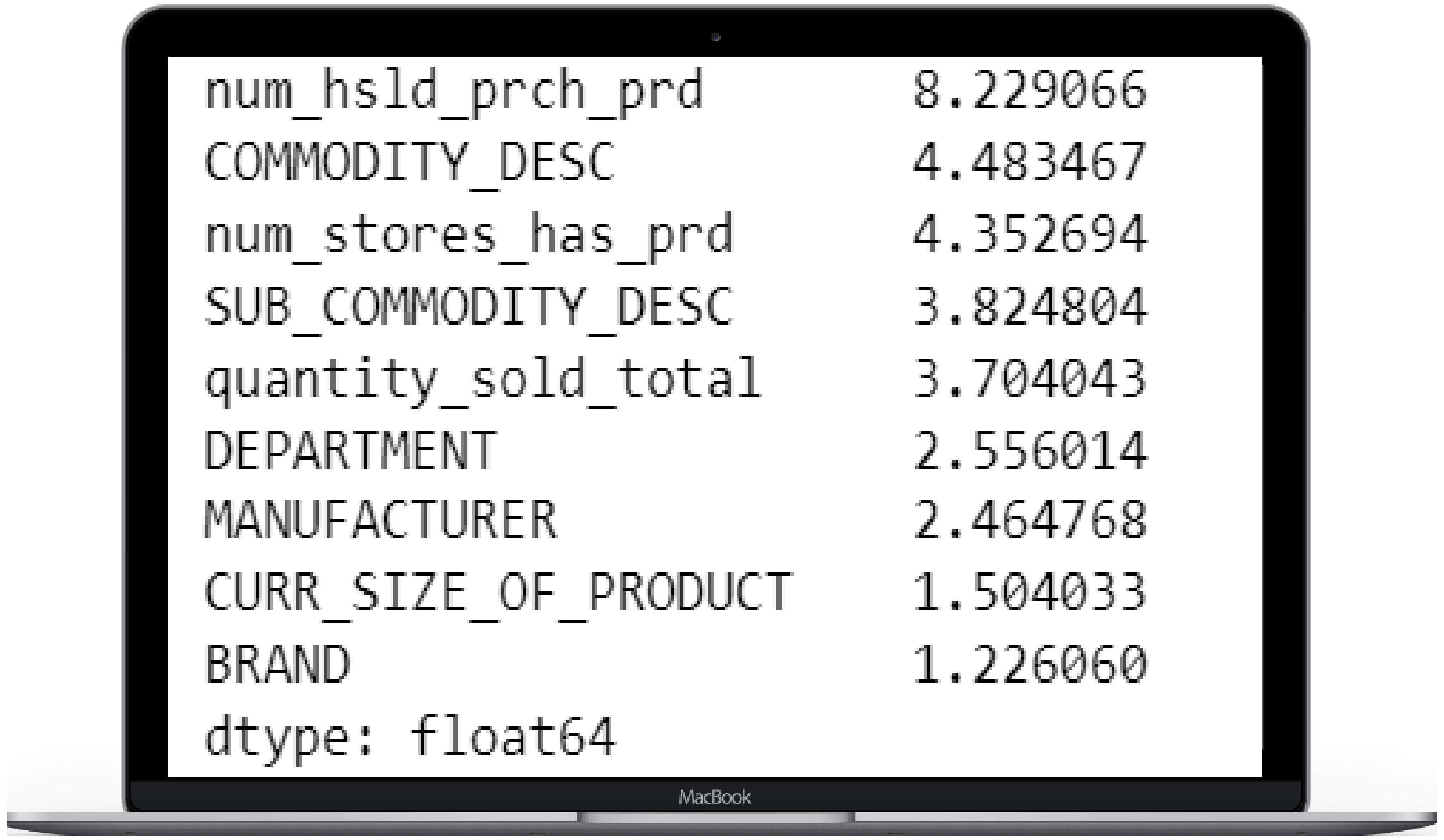
ALL THE FEATURES HAVE P VALUE < 0.05 % WHICH MEANS THAT WE REJECT THE H0 , i.e ALL THE FEATURES ARE SIGNIFICANT.

Correlation of the variable with The Target

As observed , we can see that manufacturer and current_size_of_the_product are highly correlated.

Checking Multicollinearity:-

We see that no multicollinearity is there.



Models tried and tested:

- KNeighborsClassifier()
- GaussianNB()
- DecisionTreeClassifier()
- RandomForestClassifier()
- AdaBoostClassifier()
- GradientBoostingClassifier()

Models performing the best:

1. DECISION TREE
2. RANDOM FOREST

Conclusion

Hyperparameter Tuning on decision tree:

After trying tuning the parameters on Decision tree (best performing model):

It didn’t perform well, rather it was better and the top performer before the hyperparameter tuning.

THANK YOU!