Project Report

PREDICTIVE AND INFLUENTIAL BIG BRANDS MILK PRODUCTS SOLD ACROSS VARIOUS GROCERY STORES.

Contents

Project Description

The purpose of this project is to examine the provided grocery and milk datasets; to design business models that will help us make better and informed decisions. The data comprises information on milk categories, its brands and other factors that contribute towards milk sales. With this analysis we will provide valuable insights that can help improve sales and understand top market brands.

The project will be analyzed by studying the below two segments:

Consumer Behavior Analysis

This involves predicting the forecast of dollar sales depending on various factors like fat-content, packaging, flavor, etc.

Brand Influence

Here, we'll focus on the top-selling brands and analyze if units of brands purchased are influenced based on various factors that are captured in the data like feature, price reduction, flavor, packaging, etc.

Objective

To study and find insights from the data provided, and to analyse dollar sales along with the factors that affect them. It also involves finding the brands that are most preferred based on different features.

Overview of Data

We are provided with four datasets, namely: Milk_groc, Prod_Milk, IRI Week and Delivery Stores.

The first file Prod_Milk contains transactional data for different purchases in various stores. It has 8,831 observations and consists of the product type, package, flavour/scent, additives, etc.

The second file Milk_groc consists of transactional data for different purchases for various stores (Keys: IRI_KEY, WEEK, SY+GE+VEND+ITEM)

The third file IRI Week has the time frames, and the fourth file Delivery Stores consists of the details like name, and location of the stores.

Data Preparation

The data consists of 4 million+ records from scanners, so for further processing, we have randomly sampled the data to approximately 1 million records.
As milk packages across brands are sold in different measuring units, to devise a uniform measurement by standardizing the measuring quantities (volume) to ounces and calculating the quantities per package sold per unit along with price per ounce.

To get better understanding of the data provided, we have performed initial exploration on data by studying its properties further.

Creating Dummies

Since we have multiple categories in our data for most of the variables, we have created dummy variables based on their frequencies in the dataset. We have considered the top categories and brands whilst considering minor categories as "Others".

Correlation

Summarizing our metrics from SAS correlation (PROC CORR) output, we can say:

- Packaged Volume (in ounce) has a high positive correlation with packages that are plastic jugs, in order words plastic jugs can contain more volume as compared to other containers like carton. Also, it has a P-value less than the significant value of 95% confidence interval, hence we can say that volume is highly significant with plastic jug packages. This can be inferred
- There is a positive correlation between the price of milk products and units sold along with the volume of milk sold. This makes sense because more units and volume indicate the higher price of milk. From p-value, it's clear that price is statistically significant with both these variables at 95% confidence interval.
- Milk and milk substitutes like yogurt have a negative correlation between them, but they have p-value less than 0.05 indicating they are highly significant. The negative correlation could be an indicator that customers prefer milk more than its substitutes.
- Regular flavored milk has a positive correlation with buttermilk and is statistically significant at 95% intervals due to its p-value<0.05. This is an indicator that customers prefer flavored milk over buttermilk.
- Skimmed milk is negatively correlated to milk substitutes like yogurt and is statistically significant at 95% interval due to its p-value. Again, substitutes are not preferred over skimmed milk.
- Due to high p-value, volumes of milk sold are not significant with low-fat milk.

Multicollinearity

Variable	DF	Parameter Estimate	Standard Error	t value	Pr > |t|	Variance Inflation
Intercept	1	28.6909	1.81248	15.83	<.0001	0
F_flag	1	145.3647	1.87038	77.72	<.0001	1.40428
DISPLAY	1	159.9517	5.6031	28.55	<.0001	1.1675
Feature_display	1	515.8255	15.20818	33.92	<.0001	1.16876
PR	1	36.2919	1.2402	29.26	<.0001	1.39477
NESTLE_NESQUIK	1	120.2801	1.7591	68.38	<.0001	1.58867
LACTAID100	1	74.82596	1.9409	38.55	<.0001	2.25316
SILK	1	81.67301	1.92891	42.34	<.0001	1.56761
PRIVATE_LABEL	1	319.0555	1.00143	318.6	<.0001	1.31329
DEANS	1	11.20567	2.08599	5.37	<.0001	1.38768
FLAVOR_SCENT_CHOCOLATE	1	-27.4653	1.35809	-20.22	<.0001	2.23061
FLAVOR_SCENT_WHITE	1	107.6278	1.57164	68.48	<.0001	4.52E+00
PACKAGE_CARTON	1	-87.744	1.43537	-61.13	<.0001	3.82717

PACKAGE_PLASTIC_JUG	1	198.8107	1.55273	128.04	<.0001	3.23681
PACKAGE_PLASTIC_BOTTLE	1	-74.6642	1.65572	-45.09	<.0001	2.67407
PROCESS_PASTEURIZED_HOMOGENIZED	1	64.32296	0.97084	66.25	<.0001	1.66193
PRODUCT_TYPE_BUTTERMILK	1	-119.947	2.30133	-52.12	<.0001	1.89925
PRODUCT_TYPE_MILK	1	-100.355	1.93471	-51.87	<.0001	4.80479
TYPE_LOWFAT	1	-13.3483	1.78513	-7.48	<.0001	2.7714
TYPE_REDUCED_FAT	1	112.6085	1.72358	65.33	<.0001	3.17959
TYPE_Regular	1	23.87684	1.61843	14.75	<.0001	4.02826
TYPE_SKIM	1	-6.1083	1.81536	-3.36	0.0008	2.64114

The Variance Inflation Factor (VIF) for all the explanatory variables is less than 10. Therefore, there is no multicollinearity present in the model.

Heteroscedasticity

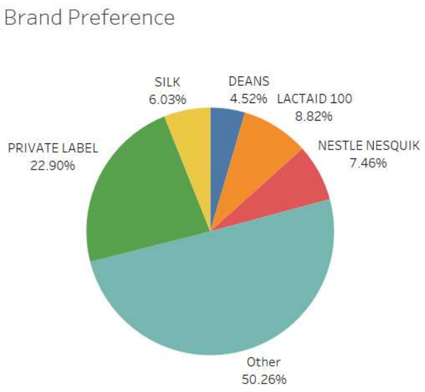
Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
DOLLARS	White's Test	121E3	88	<.0001	Cross of all vars

The p-value is less than 0.05, therefore, we reject the null hypothesis that there is no heteroscedasticity. Hence, the model has Heteroscedasticity.

Brand Preference & Insights

Top brand focus

To study the most preferred brand by customers, we analyze the effect of sales and other variants like flavor, display, etc. All brand level data is grouped into 5 categories – Deans, Silk, Nestle Nesquik, Lactaid 100, Private Label.



A PROC LOGISTIC was run with choice of brands as dependent variable along with sales and its factors as independent variables. The regression equation used is

Brand = $\beta 0$ + $\beta 1$Dollars + $\beta 2$Feature + $\beta 3$Display + $\beta 4$FeatureDisplay + $\beta 5$PR + $\beta 6$ChocolateFlavor + $\beta 7$WhiteFlavor + $\beta 8$Carton + $\beta 9$PlasticJug + $\beta 10$PlasticBottle + $\beta 11$Pasteurized_Homogenized + $\beta 12$Buttermilk + $\beta 13$Milk + $\beta 14$LowFat + $\beta 15$Fat + $\beta 16$Regular + $\beta 17$skim

Breaking the results further,

Based on AIC, SC and -2Log L values, intercepts and covariates model is the best fit model for the data:

Model Information	
Data Set	WORK.LOGIT_DATA
Response Variable	L5_flag
Number of Response Levels	6
Model	cumulative logit
Optimization Technique	Fisher's scoring

Model Fit Statistics		1055061
Criterion	Intercept Only	
AIC	2953502.7	
SC	2953562.0	
-2 Log L	2953492.7	
Number of Observations Read		
Number of Observations Used		1055061

Analysis of Maximum Likelihood Estimates						
				Standar d	Wald	

Parameter		DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	5	1	-2.4878	0.00873	81198.4035	<.0001
Intercept	4	1	-1.5388	0.00804	36604.4742	<.0001
Intercept	3	1	-0.8745	0.00784	12436.5045	<.0001
Intercept	2	1	-0.2824	0.00777	1319.8057	<.0001
Intercept	1	1	0.8581	0.00782	12035.8174	<.0001
DOLLARS		1	0.000704	4.621E-6	23180.4621	<.0001
F_flag		1	0.4110	0.00935	1932.4433	<.0001
DISPLAY		1	0.3551	0.0280	160.3915	<.0001
Feature_display		1	-0.4511	0.0754	35.7771	<.0001
PR		1	-0.2355	0.00641	1351.1166	<.0001
FLAVOR_SCENT_CHOCOLA		1	-0.2408	0.00657	1343.0209	<.0001
FLAVOR_SCENT_WHITE		1	-1.6782	0.00748	50331.9088	<.0001
PACKAGE_CARTON		1	-0.2285	0.00628	1322.7025	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
PACKAGE_PLASTIC_JUG		1	-0.8267	0.00727	12927.1191	<.0001
PACKAGE_PLASTIC_BOTT		1	-0.8918	0.00754	13999.4041	<.0001
PROCESS_PASTEURIZED_		1	0.3610	0.00481	5634.0192	<.0001

PRODUCT_TYPE_BUTTERM		1	0.5012	0.0110	2073.7310	<.0001
PRODUCT_TYPE_MILK		1	1.8062	0.00817	48881.2165	<.0001
TYPE_LOWFAT		1	-1.5141	0.00785	37246.1509	<.0001
TYPE_REDUCED_FAT		1	-1.6635	0.00720	53348.8746	<.0001
TYPE_Regular		1	-1.4101	0.00653	46632.6837	<.0001
TYPE_SKIM		1	-1.0503	0.00741	20077.8715	<.0001

Interpretation:

Given other variables are held constant, the likelihood of choice selection between brands are listed out below

- Customers are less likely to choose to buy any of the brands - Deans, Silk, Nestle Nesquik, Lactaid 100 over Private Label when private label is on display
- Customers are less likely to choose flavored or packaged product when private label is on display
- Customers are more likely to buy pasteurized when private label is on display
- Customers are less likely to buy buttermilk over milk type when private label is on display

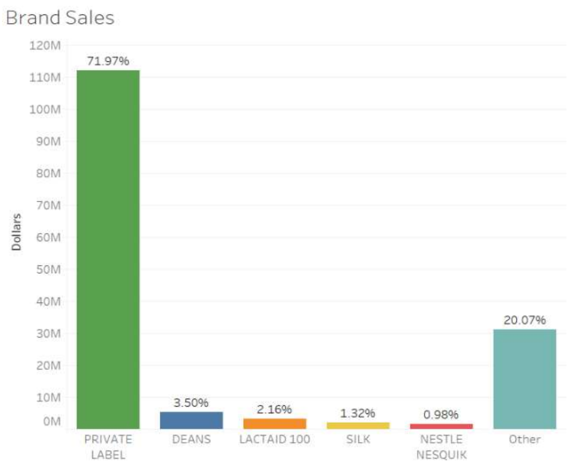
Regression Model:

The below table shows the hypothesis for each variable in the model. The dependent variable is Dollars. From the chi-square test statistic and the associated p-values shown in the table, we can see that the variables mentioned below are significant at 95% Confidence Interval (i.e., p-value < 0.05)

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
B0	145.955	1.8702	78.04	<.0001
B1	161.9973	5.6023	28.92	<.0001
B2	515.1347	15.2099	33.87	<.0001
B3	38.25483	1.2341	31	<.0001
B4	117.696	1.7517	67.19	<.0001
B5	80.55785	1.9071	42.24	<.0001
B6	92.33784	1.8076	51.08	<.0001
B7	317.7037	0.9979	318.37	<.0001
B8	20.62153	1.9996	10.31	<.0001
B9	-21.7288	1.309	-16.6	<.0001
B10	112.0279	1.547	72.41	<.0001
B11	-71.8396	1.0252	-70.07	<.0001

B12	212.192	1.3026	162.9	<.0001
B13	-58.3283	1.2949	-45.04	<.0001
B14	64.98139	0.9701	66.99	<.0001
B15	-111.833	2.2438	-49.84	<.0001
B16	-95.3091	1.9085	-49.94	<.0001
B17	-6.76125	1.7362	-3.89	<.0001
B18	118.0099	1.6897	69.84	<.0001
B19	30.21007	1.5684	19.26	<.0001
B20	-1.40684	1.7911	-0.79	0.4322

B0(F_flag), B1(Display), B2(Feature_display), B3(PR), B4(Nestle_Nesquik), B5(Lactaid100), B6(Silk), B7(Private_label), B8(Deans), B9(Flavor_scent_chocolate), B10(Flavor_scent_white), B11(Package_carton), B12(Package_plastic_jug), B13(Package_plastic_bottle),B14(Process_pasteurized_homogenized), B15(Product_type_buttermilk),B16(Product_type_milk), B17(Type_lowfat), B18(Type_reduced_fat), B19(Type_regular)



Interpretation:

- If there is a **feature** present, then the sales will increase by $145.95
- If the product is on **display**, then the sales will increase by $ 515.13
- If there was a **price reduction**, then the sales $38.25
- Brand:
    - If the brand is Nestle Nesquik, there’s a $117.69 increase in sales as compared to others.
    - If the brand is Lactaid100, then there’s an increase of $80.55 in the sales as compared to others.
    - If the brand is Silk, there’s an increase of $92.34 in sales as compared to others.
    - If the brand is Private Label, then there is an increase of $317.70 in sales as compared to others.
    - If the brand is Deans, then there is an increase of $20.62 in sales as compared to others.
    - Therefore, we can say that Private Label is the **costliest** brand whereas Deans is the **least costly** from our top 6 brands.
- Flavor:
    - If the flavored scent is Chocolate, then the sales decrease by $21.73
    - If the flavored scent is White, then the sales increases by $112.07
- Packaging:
    - If the package type is “carton”, then the sales decreases by $71.84

    - If the package type is “plastic jug”, then there is an increase of $212.19 in sales
    - If the package type is “plastic bottle”, then there is a decrease of $58.33 in sales.
- If the process used is Pasteurized Homogenized, then there is an increase of $64.94 in sales
- Product Type:
    - If the product type is buttermilk, then the sales decrease by $111.83
    - If the product type is milk, then the sales decrease by $95.31
- Fat Percentage:
    - If the milk type is Low-fat, then the sales decreases by 6.76
    - If the milk type is Reduced Fat, then the sales increases by $118
    - If the Milk Type is Regular, then the sales increases by $30.21