

PREDICTIVE ANALYTICS USING SAS

Project Report

Formulating Marketing Strategies for Diaper Industry



By:

Akanksha Guruprasad - Axx180053

Ashwin Rathore - Axx180015

Yu Zhang - Yxx170831

A marketing strategy generally gives a long term plan to reach the top 1 or top 2 position in the market. On the other hand, there are several short-term marketing strategies which focus on how to beat the competition at present. Both type of marketing strategies are important as one needs to make sure they survive in the current business environment, at the same time targeting the top 3 positions in product market for the long term.

There are several factors to be considered before the formulation of a marketing strategy. But some factors stand apart and we can build marketing strategy just from these crucial factors. We have outlined 3 such factors for brand Huggies:

1) Developing Customer Intimacy

A company can gain a lot by investing time in its customers, getting to know them, their likes, dislikes, preferences. To develop customer intimacy, a company should have an answer to the following questions

- How much is spent on market research with focus on customer satisfaction or customer retention?
- Do we have the proper customer insight at any point of time?
- Do we take notice of consumer trends to fine tune our marketing strategy?

We have customer-level data available with us. Customer-level data includes data about the customer who bought products from brand HUGGIES, which week they bought and how many units they bought. Thus, we can segregate these customers into 3 clusters based on Recency-Frequency-

The SAS System

The CORR Procedure

3 Variables: Recency Frequency Monetary

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Recency	346	1147	14.65523	396706	1114	1165
Frequency	346	3.12717	4.05627	1082	1.00000	31.00000
Monetary	346	42.36965	54.97768	14660	5.99000	378.21000

Pearson Correlation Coefficients, N = 346 Prob > r under H0: Rho=0			
	Recency	Frequency	Monetary
Recency	1.00000	0.34121 <.0001	0.37425 <.0001
Frequency	0.34121 <.0001	1.00000	0.94366 <.0001
Monetary	0.37425 <.0001	0.94366 <.0001	1.00000

Monetary matrix and target these clusters accordingly. This analysis helps us figure out our best value customers. Also, at the time of any kind of promotion we can target only the high value customers rather than targeting all of them and the approach of targeting selected customers is more profitable.

Insights from the Model:

First, we tested the correlation between the matrix - Recency, Frequency and Monetary.

From the results we can say that, Frequency and Monetary are highly correlated. So, by taking anyone of them into account we get

the proper customer segmentations.

We have considered Recency and Monetary as our parameters for further analysis.

Below is the table which shows the ranks of our customers based on Recency and Monetary:-

The SAS System

Obs	PANID	Recency	Monetary	rnkR	rnkM	rank	Customer_Type
1	1101162	1144	25.98	1	2	1-2	Potential
2	1101519	1161	25.98	3	2	3-2	Potential
3	1104935	1120	13.49	0	1	0-1	CustAtRisk
4	1106229	1139	11.79	1	0	1-0	Potential
5	1107367	1119	24.38	0	2	0-2	CustAtRisk
6	1108456	1135	12.99	1	1	1-1	Potential
7	1109595	1146	13.49	2	1	2-1	Potential
8	1114710	1137	14.49	1	1	1-1	Potential
9	1116590	1143	12.99	1	1	1-1	Potential
10	1118786	1161	64.95	3	4	3-4	HighValue

Based on these ranks we have segmented our customers into three buckets:

- (i) Customers at Risk
- (ii) Potential Customers
- (iii) High Value Customers

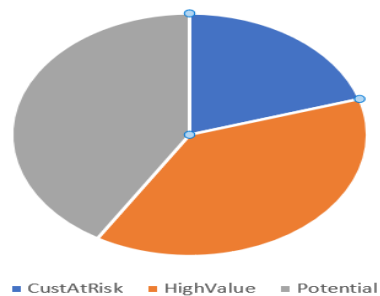
Below table shows the frequency and percentage of type of customers we have for Huggies brand:

The SAS System

The FREQ Procedure

Customer_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
CustAtRisk	70	20.23	70	20.23
HighValue	133	38.44	203	58.67
Potential	143	41.33	346	100.00

Customer Segmentation



Recommendations:

(i) Customers at Risk (CustAtRisk):

These are the customers that are not frequent buyers of the Huggies diaper. Also, the revenue generated from these customers is very less. **20.33%** customers fall in this category.

As these customers are not generating much revenue and are also not recent buyers of Huggies products, we should send them “Miss you” card and emails which contain information of offers and discounts available on the products.

(ii) Potential Customers (Potential):

These customers are of importance because they have bought Huggies diapers recently and, the revenue generated from these customers is better compared to the customers at risk. Also, these customers can potentially convert into high value customers by giving them more or proper type of offer and discounts. **41.33%** customers fall in this category.

To specifically target these customers, we should give them

1. Gift cards
2. Discount coupons
3. Membership cards to become loyal customers which has extra offers and discounts.

(iii) High value customer (HighValue):

These customers have bought Huggies diapers regularly and they are also high revenue generating customers. They are loyal to our brand which is shown by the recency and purchase of these customers. **38.44%** customers fall in this category.

To target these customers, we should give them

1. Special rewards.
2. Thank-you gifts and Thank-you cards when they make purchase.
3. Special offer cards or vouchers for coming week.
4. Specific day offer where loyal customers get extra discounts.

Further based on geographic location and the purchase of these customers we can find out the stores which they buy from and have Huggies diapers on display which will also help in improving the sales.

2) Health of the Brand:

The second most important factor is determining the health of the brand. This takes the external environment into account. From the analysis on customers' brand choice, we get an answer to several questions. Is our competitive strategy strong enough? Where are we positioned in the customers mind?

Even though Customer-Brand loyalty is a good factor on which a brand can bank their dollar sales, there are several factors that affect the brand choice for a customer. Average price of a unit, Display in-store, Feature Ad, Price Reduction are a few factors that affect the sales for brand - Huggies. These factors might also affect the sales of other brands(competitors) in the market. Not only does it affect that brand but a slight change in these factors of the competitors might affect the sales of brand HUGGIES as well.

The top 6 brands for the product diaper in terms of total dollar sales are: HUGGIES, PAMPERS, PRIVATE LABEL, LUVS, DRYPERS AND FITTI. All brands other than the ones mentioned above are considered as "Other" Brand for later analysis

Let us compare the effect of changes in these brand specific variables - Avg_price, Feature, Price_Reduction in brand - HUGGIES and its competitors affecting the customers brand choice.

Insights from the Model:

The SAS System								
Obs	Result	DRYPERS	FITTI	HUGGIES	LUVS	Other	PAMPERS	PRIVATE_LABEL
1	Avg_Price	0.23	0.25	0.31	0.25	0.36	0.31	0.22
2	own_price_elast	-1.61	-2.20	-2.40	-1.70	-3.01	-2.33	-1.77
3	cross_price_elast	0.47	0.05	0.38	0.57	0.16	0.42	0.23

From Avg_Price, we can say that the average unit price of Huggies and Pampers is highest when only considering the top 6 brands. And that of Private_Label is the lowest.

(i) **Own Price Elasticity:** 1% decrease in Avg_Price for brand **Huggies**, keeping all other factors constant, **increases** the probability of customers choosing brand Huggies by **2.40%** which is the highest among all the top 6 competitors.

(ii) **Cross Price Elasticity:** When other competitors **increase their Avg_Price by 1%**, the probability of customers choosing brand **Huggies increases by 0.38%**. The probability of customers choosing

other brands such as **Luvs increases by 0.57%** and that of **Drypers is 0.47%**. As these brands have a higher cross price elasticity, they are more competitive brands compared to others for brand Huggies.

The SAS System

Obs	Result	DRYPERS	FITTI	HUGGIES	LUVS	Other	PAMPERS	PRIVATE_LABEL
1	Feature	0.15	0.00	0.27	0.10	0.04	0.19	0.17
2	own_feat_elast	0.06	0.00	0.14	0.03	0.03	0.10	0.09
3	cross_feat_elast	-0.07	0.00	-0.07	-0.05	0.00	-0.05	-0.04

From Feature, we can say that average no of units having a feature is highest for brand Huggies followed by brand Pampers and Private_Label.

- (i) **Own Feature Elasticity:** For brand **Huggies**, when it puts **Feature** ad on a product, the probability of customers choosing this brand **increases by 0.14%**. This is the highest probability among all its competitors.
- (ii) **Cross Feature Elasticity:** When other competitors put a Feature ad on their product, the probability of customers choosing brand **Huggies decreases by 0.07%**. This decrease in the probability of customers brand choosing due to other competitors putting a Feature ad is very high for brand Huggies as other competitors like **Luvs** and **Drypers** have a less decrease in their probability of customers brand choosing.

The SAS System

Obs	Result	DRYPERS	FITTI	HUGGIES	LUVS	Other	PAMPERS	PRIVATE_LABEL
1	PriceReduction	0.58	0.00	0.38	0.33	0.13	0.27	0.24
2	own_PR_elast	0.11	0.00	0.08	0.06	0.03	0.05	0.05
3	cross_PR_elast	-0.05	0.00	-0.03	-0.03	0.00	-0.02	-0.01

From PriceReduction, we can say that the average no of units having Price Reduction is highest for Drypers followed by Huggies and Luvs.

- (i) **Own PriceReduction Elasticity:** For brand **Huggies**, when it puts a **Price Reduction** of more than 5% on its product, the probability of customers choosing this brand **increases by 0.08%**.
- (ii) **Cross PriceReduction Elasticity:** When other competitors put a Price Reduction on their product, the probability of customers choosing brand **Huggies decreases by 0.03%**. This decrease in probability is even less for brands **Pampers** and **Private_Label**.

	Brand
B1	Huggies
B2	Pampers
B3	Private Label
B4	Luvs
B5	Drypers
B6	Fitti

Let us look at the effect of Non-Brand specific factors such as Age or Income which might affect the brand choice.

We have assumed the base brand to be “Other”.

The dummy brand variables B1-B6 created to capture the non-brand specific estimates are represented by:

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
B1	1	5.4326	1.4302	3.80	0.0001
B2	1	6.1909	1.4179	4.37	<.0001
B3	1	4.1141	1.4628	2.81	0.0049
B4	1	6.2893	1.4085	4.47	<.0001
B5	1	6.3687	1.4121	4.51	<.0001
B6	1	3.4435	1.7581	1.96	0.0502
Age1	1	-0.5399	0.2061	-2.62	0.0088
Age2	1	-0.8335	0.2050	-4.07	<.0001
Age3	1	-0.3994	0.2088	-1.91	0.0558
Age4	1	-0.5722	0.1973	-2.90	0.0037
Age5	1	-0.8861	0.2003	-4.42	<.0001
Age6	1	-0.1578	0.2905	-0.54	0.5870
Income1	1	-0.3785	0.1068	-3.54	0.0004
Income2	1	-0.3213	0.1060	-3.03	0.0024
Income3	1	-0.3736	0.1081	-3.46	0.0005
Income4	1	-0.4387	0.1031	-4.25	<.0001
Income5	1	-0.3983	0.1039	-3.83	0.0001
Income6	1	-0.6302	0.1563	-4.03	<.0001

Effect of Age on Brand Choice:

The impact of age on brand choice is as:

“Other” = “Fitti” > “Private Label” > “Huggies” = “Luvs” > “Pampers” = “Drypers”

From the dummy variable Age1, we can say that as age of the customer increases, probability of customer choosing **Huggies** compared to “Other” brands decreases.

Thus, it can be said that the target customers for brand Huggies could be young customers.

Effect of Income on Brand Choice:

From the dummy variable Income1, we can say that as income increases, the probability of customer choosing **Huggies** compared to “Other” brand decreases.

Thus, it can be said that even though the average unit price for brand Huggies is the highest among its competitors, the probability of a customer with lower income choosing this brand is more.

Recommendations:

For brand HUGGIES, Feature Ad is an effective way of getting more customers.

Also, young customers and customers having lower income could be one of the target groups to increase sales. To increase our market share, we can design some promotions targeting at these customers by giving them membership cards.

3) Marketing Budget

We have weekly store-level data for outlets - Grocery stores and Drug stores from where diapers were bought for the brand **HUGGIES**. Each store is masked with an IRI-KEY and we have the weekly sales data of a store for a year(52 weeks).

We wish to know the factors that affect the total weekly dollar sales of diapers for the brand HUGGIES.

The weekly sales data collected has information related to DOLLARS(Total Dollar Sales of the product), UNITS(Total Unit Sales), VOL_EQ(Count of diapers in each Unit), D(Product on Display(=1) or not), F(Product having any Ad Feature(=1) or not), PR(Products having price reduction > 5% (=1) or not).

All of the above-mentioned explanatory variables are time invariant. Also, we have data from 1826 stores(Cross-sections) for 52 weeks(Time series length).

First, a Durbin-Wu-Hausman test was applied to check autocorrelation and determine which model would be preferred.

H_0 : Autocorrelation doesn't exist. Preferred model is Random effects model.

H_1 : Autocorrelation exists. The model is Fixed Effect model.

As the P-value is less than the significance level, we reject the null hypothesis and we will perform Fixed Effect model.

Hausman Test for Random Effects			
Coefficients	DF	m Value	Pr > m
6	6	51.03	<.0001

The SAS System

The PANEL Procedure

Fixed Two-Way Estimates

Dependent Variable: total_dollar_sale

Model Description	
Estimation Method	FixTwo
Number of Cross Sections	1826
Time Series Length	52

Fit Statistics			
SSE	2089596119	DFE	82611
MSE	25294.4053	Root MSE	159.0421
R-Square	0.9019		

F Test for No Fixed Effects			
Num DF	Den DF	F Value	Pr > F
1876	82611	401.04	<.0001

Parameter Estimates						
Variable	DF	Estimate	Standard Error	t Value	Pr > t	Label
Intercept	1	286.2488	24.2150	11.82	<.0001	Intercept
avg_price	1	-464.629	29.2404	-15.89	<.0001	
avg_D	1	318.5495	107.8	2.96	0.0031	
avg_F	1	7.845099	22.1163	0.35	0.7228	
avg_PR	1	84.7388	3.2110	26.39	<.0001	
avg_price_D	1	-150.28	358.2	-0.42	0.6748	
avg_price_F	1	339.2527	74.3670	4.56	<.0001	

Insights from the Model:

From the Parameter Estimates, we can say that the coefficients of **avg_price**, **avg_D**, **avg_PR**, **avg_price_F** are statistically significant at 95% CI.

- An increase of one dollar in the **Average Unit Price** of a diaper decreases the total weekly sales by 464.62 dollars. This could be because increase in price might lead customers to buy other brands having less price than buying our brand product.
- If the product has **Display in-store**, compared to the product not on display, the total weekly sales increases by 318.54 dollars.
- If the product has a **Price Reduction** of more than 5%, compared to the product having a price reduction of less than 5%, the total weekly sales increases by 84.73 dollars. This could be because customers might tend to buy more units of a product having greater discount.
- For a product having a **Feature**, an increase of one dollar in the **Average Unit Price** of a diaper increases the total weekly sales by 339.25 dollars. So, **having a feature on a product cancels the decreasing effect on total weekly sales due to unit price increase** compared to the one not having a feature. But this comes at a cost of the expenditures done for ads on the product.

Thus, choosing the type of Feature for a product to be displayed at any region is important. There are 4 type of Features possible which are: **C(Small Ad** - Usually one line of text), **B(Medium sized Ad)**, **A(Large sized Ad)** and **A+(Retailer Coupon)**.

Further Analysis on Feature:

For this hypothesis, we will limit our scope to Feature type **A** and **B** and Regions **Los Angeles** and **New York**.

The SAS System

The TTEST Procedure

Variable: DOLLARS

F	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	6343	49.4105	51.8385	0.6509	3.9900	609.4
B	8090	57.9151	61.3740	0.6824	3.9900	1185.8
Diff (1-2)		-8.5046	57.3789	0.9623		

F	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
A		49.4105	48.1345 50.6864	51.8385	50.9519 52.7567
B		57.9151	56.5775 59.2527	61.3740	60.4427 62.3346
Diff (1-2)	Pooled	-8.5046	-Infy -6.9217	57.3789	56.7246 58.0487
Diff (1-2)	Satterthwaite	-8.5046	-Infy -6.9534		

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	14431	-8.84	<.0001
Satterthwaite	Unequal	14352	-9.02	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8089	6342	1.40	<.0001

Region 1 - Los Angeles: The t-test results are:

From the mean and frequency distribution, it, can be said that the sales of dollars are more due to feature B than A.

To check if it is statistically true or not, consider the following hypothesis:

H_0 : Mean Dollar sales due to feature A and B is equal.

H_1 : Mean Dollar sales due to feature B is more than feature A.

From the F-test done on the variances of the mean dollar sales for feature A and B,

it can be said that the variances are unequal. The t-value for this test is -9.02.

The P-Value for the unequal variances is less than significance level 0.05.

Thus, we reject the null and conclude that the dollar sales are more due to feature B than feature A.

Region 2 - New York: The t-test results are:

From the mean and frequency distribution, it, can be said that the sales of dollars are more due to feature A than B.

The SAS System

The TTEST Procedure

Variable: DOLLARS

F	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	3521	73.7966	85.5426	1.4416	6.0000	844.4
B	5632	67.9461	72.7017	0.9688	3.6900	753.4
Diff (1-2)		5.8506	77.8920	1.6734		

F	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
A		73.7966	70.9701 76.6231	85.5426	83.5904 87.5888
B		67.9461	66.0469 69.8452	72.7017	71.3835 74.0698
Diff (1-2)	Pooled	5.8506	3.0977 Infy	77.8920	76.7797 79.0371
Diff (1-2)	Satterthwaite	5.8506	2.9932 Infy		

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	9151	3.50	0.0002
Satterthwaite	Unequal	6578.3	3.37	0.0004

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3520	5631	1.38	<.0001

To check if it is statistically true or not, consider the following hypothesis:

H₀: Mean Dollar sales due to feature A and B is equal.

H₁: Mean Dollar sales due to feature A is more than feature B.

From the F-test done on the variances of the mean dollar sales for feature A and B, it can be said that the variances are unequal. The t-value for this test is 3.37.

The P-Value for the unequal variances is less than significance level 0.05.

Thus, we reject the null and conclude that the dollar sales are more due to feature A than feature B.

Recommendations:

Thus, to have more sales in **Los Angeles**, the company should focus more on **Feature B - Medium sized Ad** and in **New York**, the company should focus more on **Feature A - Large sized Ad**.