Spring 2020 | BUAN6337.002 | HW 5

Predictive Analytics using SAS

Group2

Adam Butcher

Dong In Kim

Maria Phetteplace

Mahmoud Ali

Q1- Store Level Scanner Data

Summarize data:

1. Brands and Market Share

Top Brands (According to Total Revenue)	Total Revenue	Top Brands (According to Market Share)	Market Share
1st - Tide	$75,923,991.67	1st - Tide	56.08
2nd - All	$15,116,464.29	2nd - All	11.17
3rd - Purex	$12,333,546.56	3rd - Purex	9.11
4th- Wisk	$12,253,480.95	4th- Wisk	9.05
5th - Gain	$10,343,370.10	5th - Gain	7.64
6th - Cheer	$9,412,878.91	6th - Cheer	6.95

2. Companies of top brands

Company	Brand
Procter & Gamble	Tide, Gain, Cheer
Lever Brothers Co	All, Wisk
The Dial Cooperation	Purex
Church & Dwight Co Inc	Xtra

3. Make "Other" Category

Top Brands (With OTHER)	Total Revenue
1st - Tide	$75,923,991.67
2nd - OTHER	$65,890,448.80
3rd - All	$15,116,464.29
4th- Purex	$12,333,546.56
5th - Wisk	$12,253,480.95
6th - Gain	$10,343,370.10
7th - Xtra	$6,532,363.26

4. Average prices, Display, Features of each of the 7 brands

Brand	**Average Price**	**Features**	**Display**
Tide	10.1293	0.4925	0.605
All	6.620871	0.1375	0.13
Purex	4.813258	0.14	0.16
Wisk	7.447229	0.1825	0.245
Gain	7.057398	0.085	0.105
Cheer	8.428371	0.085	0.075
Other	5.518814	0.655	0.915

5. Top 5 regions in terms of dollar sales

Top 5 Regions	**Dollar Sales**
1st – New York	$19,308,422.26
2nd - Los Angeles	$14,287,702.04
3rd – Chicago	$8,698,014.21
4th – Philadelphia	$7,506,907.97
5th – Boston	$7,440,975.91

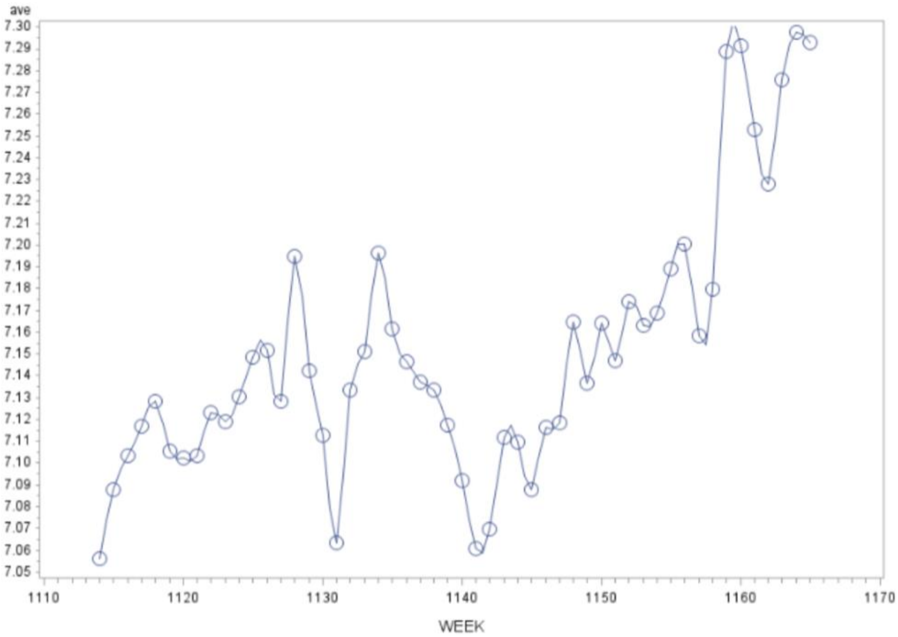
6. Top 10 store chains in terms of dollar sales

Store ID	Dollar Sales
Chain89	$6,277,335.89
Chain134	$6,181,434.34
Chain94	$5,046,937.94
Chain124	$4,983,893.19
Chain55	$3,699,386.19
Chain31	$3,444,069.01
Chain75	$3,072,811.08
Chain117	$2,777,188.78
Chain44	$2,777,188.78
Chain10	$2,732,687.07

7. Average price/unit by Week

Week	Average Price Per Unit
1 (1114)	7.0560
2 (1115)	7.0879
3 (1116)	7.1035
4 (1117)	7.1166
5 (1118)	7.1282
6 (1119)	7.1055
7 (1120)	7.1022
8 (1121)	7.1035
9 (1122)	7.1233

10 (1123)	7.1301



8. We are the manager of the brand, **Purex.** There are two ways thinking about this. One would be the locations that we can focus on, and from the analysis, we can see that our brand is pioneering in three markets (NY, LA, CH) which are the same top markets for the top selling brand “Tide.” We can consider that our brand is the low-priced competitor for Tide. We should start to focus on the same markets in which Tide is pioneering but Purex is not, such as SF and WS.

   Another option for Purex would be to focus on a competitor in the same price range. Gain has an average price of almost $2 above ours and they are pioneering 3 markets that Purex is getting almost less than 1% of our sales from those markets. These markets are Houston, Dallas, and Raleigh/Durham.

   We can see also from the analysis of the chain’s sales performance that the top 5 chains selling detergents in the US market are not within our full focus as we focus only on the top two chains lagging in the other three (94, 124, 55).

Statistical Analysis:

9. Stores (Top 3 vs Bottom 7) in average price per unit [T-test]

H_0	Large stores (top 3 stores) <= average price per unit than small stores (bottom 3 stores).	
H_A	Large stores (top 3 stores) > average price per unit than small stores (bottom 3 stores).	
Result	Reject	Conclusion

t-value: -30.01 p-value: <0.0001	Reject the H_0	Large stores (top 3 stores) have higher average price per unit than small stores.

10. Hypothesis Testing to Dollar Sales
    a. Hypothesis #1

H_0	Minor display dollar sales <= Major display dollar sales	
H_A	Minor display dollar sales > Major display dollar sales	
Result	Reject	Conclusion
t-value: -2.16 p-value: 0.1629	Cannot reject the H_0	Minor display dollar sales <= Major display dollar sales

    b. Hypothesis #2 – ANOVA analyzing variable, 'F' (A-large, B-medium, C-small, or No Feature) on 'Dollars'

H_0	There is not a significant difference in dollar sales between feature types.	
H_A	There is a significant difference in dollar sales between at least one feature type.	
Result	Reject	Conclusion
F-value: 128290 p-value: <0.0001	Reject the H_0	There is a significant difference in dollar sales between at least one feature type.

    c. Hypothesis #3 – T Test

H_0	There is not a significant difference in dollar sales between when there is a price reduction and when there is not a price reduction.	
H_A	There is a significant difference in dollar sales between when there is a price reduction and when there is not a price reduction.	
Result	Reject	Conclusion
t-value: -215.42 p-value: <0.0001	Reject the H_0	There is a significant difference in dollar sales between when there is a price reduction and when there is not a price reduction.

11. POWERED GAIN: Regression Model

a. R-sq

R-sq	Adjusted R-sq
0. 0.6560 0.6560 implies that all the 3 explanatory variables explain around 66% of the variance in the dependent variable, weekly sales.	**0. 0.6560** When factoring in a penalty on any variable added to the model that has a very small explanatory power, it did not go down at all. This means that's all three of the chosen variables do a good job explaining the dependent variable.

b. Significant Variables

Significant Variables	p-value
Average price	<.0001
average display	<.0001
average feature	<.0001

c. Which variables are most important in explaining sales?

Variable with Most Importance	STB Value
Average Display	0.44144

d. Price coefficient

Price Coefficient	Interpretation
-14.29293	For every $1 increase in average price, there is a $14.29 decrease in average weekly sales per unit.

price per ounce elasticity = -14.26614 * 0.015182 = -0.21658853748
price per unit elsticity = -2.40423 * 0.076017 = -0.18276235191

Price Elasticity	
Approach	To compute average estimate of price elasticity, multiply the price coefficient with the average price and divide by the average units.
Computation	price per ounce elasticity = -14.26614 * 0.015182 = **-0.21658853748** price per unit elasticity = -2.40423 * 0.076017 = **-0.18276235191**

e. Display coefficient

Display Coefficient	Interpretation
72128	If powered Gain had a display, average weekly sales per unit increased by $72128 compared to when there was not a display.

f. Test whether there is an interaction between display, feature and price. Comment on your findings.

New model	**Run a new model adding interaction terms**	
Variable	**Coefficient**	**p-value**
avg_Price	-6.74692	<.0001
avg_Display	30167	<.0001
avg_Feature	339027	<.0001
'avg_Price* avg_Feature'	-185.38917	<.0001
'avg_Display* avg_Feature'	857454	<.0001
Conclusion		
When adding 'avg_price | avg_display', 'avg_display' changed to a negative coefficient. Using our own logic, this did not seem appropriate since having a display should increase average weekly sales per unit. We removed that term and kept 'avg_price | avg_feature' and 'avg_display | avg_feature'.		

g. Test whether the effect of price is non-linear. Comment on your findings.

New model	**Run a new model with the same variables but add a new variable 'avg_price^2'**
p-value for Price^2	<.0001
Conclusion	
When checking whether price has a non-linear effect on weekly sales, it is a significantly different from zero at the 99% confidence level. The avg_price coefficient changed drastically when adding the new variable, avg_price_sq. There is a slight U-shaped curve. We can conclude this since the coefficient for avg_price is -254.91581which brings the U-shaped curve down. The avg_price_sq coefficient is 0.06698 which brings the U-shaped curve up only slightly.	

h. Test using VIF and COLLIN whether there is multicollinearity in the model? Comment on your findings.

Highest VIF	1.93696
Lowest COLLIN	0.00022624
Conclusion	
When looking at the model with only avg_price, avg_feature, and avg_display, there is no evidence of multicollinearity. When we look at the models with interaction terms and non-linear term, there evidence of multicollinearity which is to be expected since the variables are dependent on each other.	

i. Test for presence of heteroscedasticity using White test. Do A WLS if needed. Comment on your findings.

H_0	There is no heteroscedasticity.
H_A	There is a heteroscedasticity.

White's Test F-test	71392 (p-value = <.0001)
Conclusion	
We should reject the null hypothesis of no heteroscedasticity. We should progress with a WLS model.	

Q2- Churn Data

1a.) Table of coefficients, t-values, and odds ratio

Variable	Coefficient	P-value	odds ratio
Blck_dat_mean	-0.00792	0.4813	0.992
Callfwdv_mean	-0.00630	0.6308	0.994
Callwait_mean	-0.00413	0.0084	0.996
Change_mou	-0.00029	<0.0001	1.000
Comp_dat_mean	0.000569	0.5447	1.001
Custcar_mean	-0.00248	0.1332	0.998
eqpdays	0.000885	<0.0001	1.001
Roam_mean	0.00239	0.0242	1.002
Threeway_mean	-0.0275	0.0009	0.973
Asl_flag	-0.2870	<0.0001	0.750

creditcd	-0.1668	<0.0001	0.846
forgntvl	-0.0769	0.0216	0.926
Refurb_new	0.3058	<0.0001	1.358

1b.) Report of findings

Keeping all other variables stay the same, with each additional mean number of **call-waiting** calls, the odds of customer churning decreases 0.4%. This variable is significant at even the 99% level.

Keeping all other variables stay the same, with each additional percent change in **monthly minutes** of use vs previous three-month average, the odds of customer churning decreases 0.0%. This variable is significant at even the 99% level.

Keeping all other variables stay the same, with each additional day of **equipment age**, the odds of customer churning increases 0.1%. This variable is significant at even the 99% level.

Keeping all other variables stay the same, with each additional number of mean **roaming calls**, the odds of customer churning increases 0.2%. This variable is significant at the 95% confidence level but not at the 99% level.

Keeping all other variables stay the same, with each additional number of **three-way calls**, the odds of customer churning decreases 2.7%. This variable is significant at even the 99% level.

Keeping all other variables stay the same, if the customer's account's **spending limit** has a flag, the odds of customer churning decreases 25.0% compared to when the customer's account's spending limit has a flag. This variable is significant at even the 99% level.

Keeping all other variables stay the same, if the customer has a **credit card**, the odds of customer churning decreases 15.4% compared to when the customer doesn't have a credit card. This variable is significant at even the 99% level.

Keeping all other variables stay the same, if the customer has had **foreign travel**, the odds of customer churning decreases 7.4% compared to when the customer has not had foreign travel. This variable is significant at the 95% confidence level but not at the 99% level.

Keeping all other variables stay the same, if the device is **refurbed**, the odds of customer churning increases 35.8% compared to when the device is not refurbed. This variable is significant at even the 99% confidence level.

The **AIC** which has a lower absolute value is a better model. The model with intercepts and covariates has an AIC value of 95,596.538 compared to the model with intercepts only which has a value of 97,032.999. That is a difference of 1,436.461. The model that was used does a better job of predicting the churn rate of a customer better than just including intercepts.

Similar to the adjusted R^2, **SC** penalizes for additional variables in a model. The SC with intercepts and covariates is 95,724.725 and the SC for intercepts only is 97,042.156. That is a 1,317.431 difference which leads to the same conclusion as the AIC.

A **concordant pair** is defined as that pair formed by an *event* with a PHAT higher than that of the *no-event*. Since the ratio of churn=1:churn=0 is 49.91:50.09, we want our concordant percentage to be higher than that. The concordant percentage is 58.7% which means our model predicts better than assigning at random.

2.) To find the top three factors that affect churn, we ran STB to get the standardized betas. According to that output, our top three factors were:

Variable	**STB value**	**Meaning of variable**
Eqpdays	0.1251	Number of days (age) of current equipment
Refurb_new	0.0595	Handset: refurbished or new
Asl_flag	-0.0549	Account spending limit

3.) This data set was extensive having 173 variables, but we could get more insight if we collected a rating score of customer service calls. If they collected data having a rating score of 1-5 for examples, we could factor that into predicted churning of a customer.

4.)

34560 / (34560 + 91) = 0.9974

18068 / (18068 + 14064) = 0.5623

3049 / (3049 + 31602) = 0.0880

130 / (130 + 34521) = 0.0038

7 / (7 + 34644) = 0.0002

Hit Ratio = (0.9974 + 0.5623 + 0.0880 + 0.0038 + 0.0002) / 5 = 0.3304

5.)

TN = 1

FN = 15088

TP = 14908

FP = 3

Precision = 14908 / (14908 + 3) = 0.9998
Accuracy = (14908 + 1) / 30000 = 0.4970

Hit Ratio = 14908 / (14908 + 15088) = 0.4970