Predictive Analytics using SAS
Homework Assignment 5 – Group 8

Guru Prasad Kumar

Jeyenth Kumar

Avinash Panigrahi

Mehraj Shaik

Rahul Kumar

Tej Kashiparekh

1. **What are the top 6 brands in the category in terms of dollar sales? What are the market shares of the 6 brands (assuming there are only 6 brands in the market).**

L2	L4	L5	Percent
GROUND COFFEE	PRIVATE LABEL	PRIVATE LABEL	0.55
GROUND COFFEE	PROCTER & GAMBLE	FOLGERS	0.19
GROUND COFFEE	KRAFT FOODS INC.	MAXWELL HOUSE	0.12
GROUND COFFEE	PROCTER & GAMBLE	FOLGERS COFFEE HOUSE	0.07
GROUND COFFEE	KRAFT FOODS INC.	STARBUCKS	0.04
GROUND COFFEE	KRAFT FOODS INC.	YUBAN	0.04

L2	L4	L5	Percent
WHOLE COFFEE BEANS	PRIVATE LABEL	PRIVATE LABEL	0.68
WHOLE COFFEE BEANS	KRAFT FOODS INC.	STARBUCKS	0.18
WHOLE COFFEE BEANS	PROCTER & GAMBLE	MILLSTONE	0.04
WHOLE COFFEE BEANS	PROCTER & GAMBLE	FOLGERS	0.04
WHOLE COFFEE BEANS	EIGHT O'CLOCK COFFEE COMPANY	EIGHT O CLOCK ROYALE	0.03
WHOLE COFFEE BEANS	EIGHT O'CLOCK COFFEE COMPANY	EIGHT O CLOCK	0.03

2. **Which companies are the major players in the category? Which company owns which brands?**

L4	L5	tot_sell	mkt_share
PROCTER & GAMBLE	FOLGERS	5938257	32.74417
KRAFT FOODS INC.	STARBUCKS	3578066	19.72983
KRAFT FOODS INC.	MAXWELL HOUSE	3515862	19.38683
PRIVATE LABEL	PRIVATE LABEL	2406279	13.26847
BERARDI'S FRESH ROAST INC.	BERNARDS FRESH ROAST	1907219	10.5166
PROCTER & GAMBLE	FOLGERS COFFEE SINGLES	789628.9	4.354096

[Category: Ground decaffeinated Coffee]

[Mkt_share has been multiplied by 100 to indicate percent]

3. **Create a 7th brand called "Other" that has all other brands that are not in the top 6.**
   Please refer to the 4th question for the same.
4. **Find average prices, display, features of each of the 7 brands.**

Average of price, features and display

Obs	L5	avg_pr	avg_f	avg_d
1	STARBUCKS	7.71394	6.80950	1.59151
2	FOLGERS COFFEE SINGLES	4.66233	0.05368	0.10259
3	FOLGERS	4.24499	5.74832	7.36717
4	MAXWELL HOUSE	3.84517	7.69904	6.87145
5	Other	3.24500	5.46341	4.62410
6	PRIVATE LABEL	3.04788	5.80704	6.16709
7	BERNARDS FRESH ROAST	2.38748	5.85999	6.73130

[Average feature and Average display have been multiplied by 100 to show percents]

5. **What are the top 5 regions in terms of dollar sales?**

Top 5 Region with sales

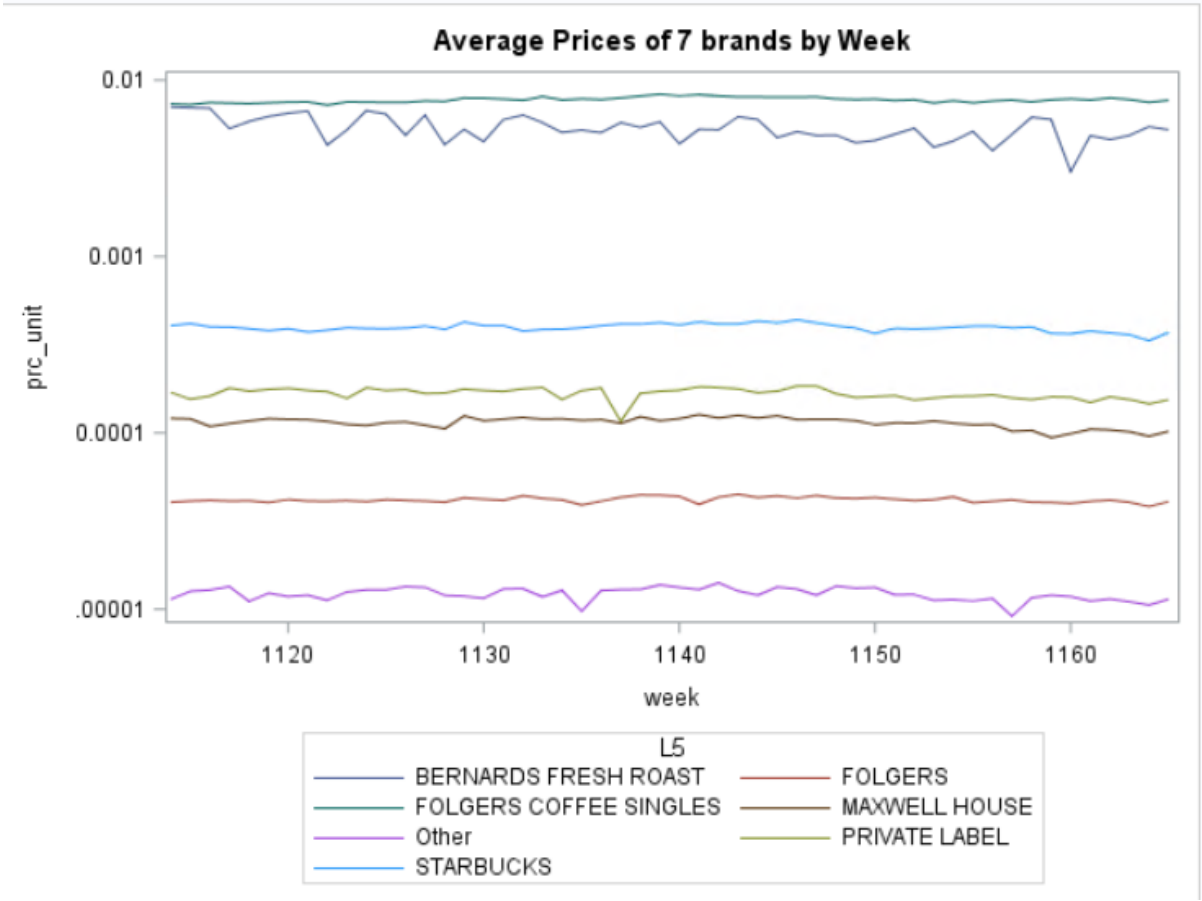
region	tot_sale
NEW YORK	11618012
LOS ANGELES	8971968
NEW ENGLAND	6295677
BOSTON	5923812
CHICAGO	5848896
PHILADELPHIA	4748225
ST. LOUIS	4208519

6. **What are the top 10 store chains that sell a lot of your category in terms of dollar sales?**

Top 10 stores with highest sales in Decaf coffee

chain	sale
Chain10	3484361
Chain11	3460209
Chain13	3064137
Chain12	1978788
Chain89	1103078
Chain94	858707.1
Chain30	777200.2
Chain75	658010
Chain50	595028.1
Chain6	550398

7. **What is the average price per unit of 7 brands by week? Plot the average price by week (I wish to see a line plot of price by week). Comment on your findings.**



As seen in the graph, overall prices remain almost similar through out the year however, Our brand Private Label sees a significant dip in price between weeks 1130 and 1140.

8. **Assume you are manager of a specific brand (out of the top 6). Write a short paragraph stating what you learned from this descriptive analysis (steps 1-7).**
The analysis showed us that with 13% market share, our brand has a significant market share trailing the other 4 top shareholders. Our brand has the lowest average prices while offering a wide variety of features other top brands are offering for a higher price. The brand has very high sales in NEW YORK, LOS ANGELES, NEW ENGLAND, BOSTON, CHICAGO, PHILADELPHIA and ST. LOUIS. We can learn more about these markets to understand what is working in our favor to extend those factors to other markets to improve sales. Chain10, Chain11, Chain13, Chain12 and Chain89 are the retailers selling the bulk of our brand.

9. **Do large stores (top 3 stores) have higher average price per unit than small stores (stores ranked 8-10) for brand 1 (the top brand in Q1). Test and report your results and comments.**

Testing with Ttest, with 95% confidence we have enough evidence to suggest that big stores have higher average price per unit for the decaf private label coffee.

10. **Develop three additional hypotheses linking useful variables to dollar sales, test them and report your findings.**

    Hypothesis 1> Do different flavours affect sales of brand in different region

    We tested using anova test for the hypothesis, with 95% confidence we conclude that flavor is affecting the sales by region

    Hypothesis 2> Is type of package affecting sales in decaff segment brand

    Using anova test, with 95% confidence we conclude that the type of package is indeed affecting the sales in category.

    Hypothesis 3> What is the correlation between unit volume and sales? Is it significant?

    Volume and Sales are significantly correlated with a factor of .430. Volume will be a good predictor of sales of a brand packing.

11. **For your brand: Run a regression model with weekly dollar sales as dependent variable. Use average weekly price per unit, average display, average feature, and other useful variables in your regression and answer the following questions:**

    a. **What is the R-sq and adj R-sq of the model?**
    With total sales as dependent variable, R^2 and adj. R^2 are 0.9454 and 0.9452. So, the model seems very well fit

    b. **Which coefficients are significant?**
    With total sales as dependent variable, only volume and total quantity of sales are significant. With logarithmic of total sales as dependent variable almost all variables except packaging with Foil bag are significant at 95% confidence.

    c. **Which variables are most important in explaining sales?**
    - Standardized estimates suggest that total quantity,volume of sales and package with box and can are the most important variables in determining sales.

    d. **Interpret the meaning of the price coefficient? What is the price elasticity?**
    - standardized coeff of log(prc_unit) is -0.82471 with log(tot_sal) implies that 1% change in unit price reduces total sales each week by 0.82%

    e. **Interpret the meaning of the display coefficient?**
    standardized coeff of avg display is 0.01937 with log(tot_sal) implies that 1-unit change in avg display leads to 5.44 % change in total sales each week

    f. **Test whether there is an interaction between display, feature and price. Comment on your findings.**
    There is very significant interaction between price per unit & avg.feature, price per unit & avg.display, price per unit & avg.feature & avg.display

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

There are good standardized estimates and significant p-values for higher exponents for weekly unit price. So, the effect of Weekly unit price is non-linear

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

- There are tolerance levels of less than 0.1 and VIF greater than 10 and collin's table's conditional index values suggest that there is collinearity
- Between total sales and price per unit which is also a common phenomenon in market.

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

Since the residuals are hanging around 0 and there is no pattern, there is no presence of multicollinearity

---

Question 2:

2.1 Include a clean table of coefficients, t-values, and odds ratio only Interpret the logistic output explaining AIC/BIC, meaning of coefficients, significance of betas, prediction accuracy (percent concordance), odds-ratios etc.

S.NO	Parameters	DF	Estimate	Standard Error	Wald Chi Square	Pr >= Chi Sq	T-Values
1	Intercept	1	-0.104	0.0092	127.7628	<.0001	11.30322078
2	change_rev	1	0.000364	0.000197	3.4038	0.065	1.844939023
3	hnd_webcap_NA	1	0.4722	0.0257	338.8242	<.0001	18.40717795
4	hnd_webcap_UNKW	1	-0.3059	0.1637	3.4899	0.0617	1.868127405
5	blck_dat_Mean	1	-0.00768	0.0087	0.7787	0.3775	0.8824398
6	hnd_webcap_WC	1	0.3491	0.0222	247.2634	<.0001	15.72461128
7	drop_dat_Mean	1	0.00407	0.0155	0.0693	0.7924	0.263248932
8	roam_Mean	1	0.00454	0.0011	17.1347	<.0001	4.13940817
9	mou_opkd_Mean	1	-0.00097	0.000597	2.6449	0.1039	1.626314853
10	threeway_Mean	1	-0.0417	0.00766	29.6596	<.0001	5.446062798
11	callfwdv_Mean	1	-0.00928	0.0164	0.3218	0.5705	0.567274184

Akaike's Information Criterion, AIC of the model = $-2 \text{ Log L} + 2((k\text{-}1) + s)$= 97036.899

Bayesian Information Criterion, BIC of the model = $-2 \text{ Log L} + ((k\text{-}1) + s)*\log(\Sigma f_i)$= 97046.055

The lower these values, the better the model.

Meaning of coefficients, significance and odds-ratios:

Coefficient of *change_rev (% change of revenue)* = 0.000364

Interpretation: For 1% increase in change of revenue, the log of odds of customer churning will increase by 0.000432

Significance:

t-value = 2.20 This value is greater than the critical value = 1.96 at 95% confidence. Hence, this coefficient is statistically different from zero with more than 95% confidence level.

Odds Ratio: 1, We can say, for a one percentage change of revenue, we expect to see about 0.3% increase in the odds of customer churn, keeping other factors the same.

Coefficient of *hnd_webcap_NA(Handset Web Capable)* = 0.4841

- *Interpretation*: For 1% increase in the change of NA dummy variable of hand set but web capability, the log of odds of customer churning will increase by 0.4722

  Significance:
- t-value = 18.407. This value is greater than the critical value = 1.96 at 95% confidence. Hence, this coefficient is statistically different from zero with more than 95% confidence level.
- *Odds Ratio*: 1.604, This means that when there is an handset which is not Webcapable (NA) Limit, the odds of the customer churning over odds of customer not churning increase by 0.623 times compared to when there are handset which has webcam (WCMB), keeping other factors the same.

Coefficient of *hnd_webcap_UNKW(Handset Web Capable Unknown)* = -0.3059

- *Significance*:

  t-value = 1.86. This value is lesser than the critical value = 1.96 at 95% confidence. Hence, this coefficient is not statistically different from zero with more than 95% confidence level.

Coefficient of blck_dat_Mean(No of blocked calls data) = -0.00768

- *Interpretation*: For 1% increase in the change of the number of blocked calls, the log of odds of customer churning will decrease by 0.768 %
- *Odds ratio:* From odds ratio table, the odds ratio estimates of blck_dat_mean is given by 0.992. With every unit increase in block call data, the odds decrease by 0.8%.

Coefficient of hnd_webcap_WC (Handset web capable) = +0.3491

- *Interpretation*: For 1% increase in the change of the handset web capable units, the log of odds of customer churning will increase by 34.91%
- *Significance*: Since p value <0.0001, it is highly significant.
- *Odds ratio:* From odds ratio table, the variable has a point estimate of 1.418, with every unit increase in handset web capable unit, the odds increases by 41.8%

Coefficient of drop_dat_mean (no of dropped calls data) = +0.0047

- *Significance:* Since p value is 0.7 (>0.0001), it is not significant
- *Odds Ratio:* From the odds ratio table, the variable has a point estimate of 1.004.

Coefficient of mou_opkd_mean (nbr_unrnd_mou_off_peak_data_calls) = -0.00097

- *Significance:* Since p-value is 0.139 (>0.0001), it is not significant
- *Odds Ratio:* From the odds ratio table, the variable has a point estimate of 0.999.

Coefficient of threeway_Mean (Three way calls) = -0.0417

- *Interpretation:* For 1% increase in threeway_mean, the log of odds of customer churning will decreases by 4.17%
- *Significance:* Since p value <0.0001, it is significant.
- *Odds Ratio:* From the odds ratio table, the variable has a point estimate of 0.959. With every unit increase in threeway_mean (threeway calls), the odds decrease by 4.1%.

Coefficient of roam_mean *(*Roaming mean*)* = 0.00454

- *Interpretation*: For 1% increase in the change of the of roaming mean, the log of odds of customer churning will increase by 0.454%
- *Significance*: t-value = 3.54. This value is greater than the critical value = 1.96 at 95% confidence. Hence, this coefficient is statistically different from zero with more than 95% confidence level.
- *Odds Ratio*: From the odds ratio table, the variable has a point estimate of 1.005. With every unit increase in roam_mean (roaming mean), the odds increase by 0.5%.

Coefficient of callfwdv_mean (call forward calls) = -0.00928

- *Significance:* Since p value is 0.5705 (>0.0001), it is not significant
- *Odds Ratio:* From the odds ratio table, the variable has a point estimate of 0.991.

Prediction accuracy (percent concordance):

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	54.3	Somers' D	0.091
Percent Discordant	45.2	Gamma	0.092
Percent Tied	0.4	Tau-a	0.046
Pairs	1224900144	c	0.546

Concordance is one of the measures for telling how well the model is predicting. A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value (honcomp = 0) has a lower predicted mean score than the observation with the higher ordered response value (honcomp = 1). Percent Concordance is the percentage of number of cases where the model has thrown a probability for a 1 > probability for a 0. Higher the concordance percent, better the model.

For our model, the percent concordance = 54.3%

2.2 The top three factors that affect churn in our model are:

change_rev - % change of revenue

hnd_webcap_NA - Handset Web Capable

roam_mean – Roaming mean

2.3 Variables (that if collected) would help to improve the fit of the model:

Apart from the variables that are present in the dataset, there can be several other key factors which if included could enhance the model and help in more accurate prediction of churn. Few of these are:

a) *Plan Type*: Most telecom companies usually give out plans as part of individual and family categories. Understanding which plan suits which kind of customer will in turn increase the likelihood of churn.
b) *Population Type*: We can group the customers in 3 different broad categories like Students, Professionals, Business People. Professional and Businesspeople tend to go for more Calling plan when compared to Students which has better internet facilities.
c) *Discount package*: A list of discounts offered to people based on the plan they have can give us an idea about churn conversion. Eg. A person who gets most out of the discount provided and has a long term plan will likely to churn than the others.

2.4 (We need to calculate for our model)

Frequency Percent Row Pct Col Pct	Table of churn by outcome		
	outcome		
churn	0	1	Total
0	12105 40.35 79.43 53.42	3134 10.45 20.57 42.70	15239 50.80
1	10555 35.18 71.51 46.58	4206 14.02 28.49 57.30	14761 49.20
Total	22660 75.53	7340 24.47	30000 100.00

The hit ratio (% events correctly classified) of our model with the top 10 variables is (16311/30000) 0.5437. It implies that our model successfully predicts churn with an accuracy of 54.37%. There are several features which have not taken into consideration while running the models with top 10 variables only. With those factors, the hit ratio is likely to improve with respect to the current performance.

2.5 Using the model parameters predict the churn for the holdout sample as well and compute the hit ratio.

The SAS System

The LOGISTIC Procedure

Fit Statistics for SCORE Data											
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.TEST	30000	-20641.2	0.4554	41304.45	41304.46	41395.85	41395.85	0.010045	0.013394	0.550762	0.247414

The hit ratio for the hold out set.

Frequency Percent Row Pct Col Pct	Table of churn by outcome		
	outcome		
churn	0	1	Total
0	12105 40.35 79.43 53.42	3134 10.45 20.57 42.70	15239 50.80
1	10555 35.18 71.51 46.58	4206 14.02 28.49 57.30	14761 49.20
Total	22660 75.53	7340 24.47	30000 100.00