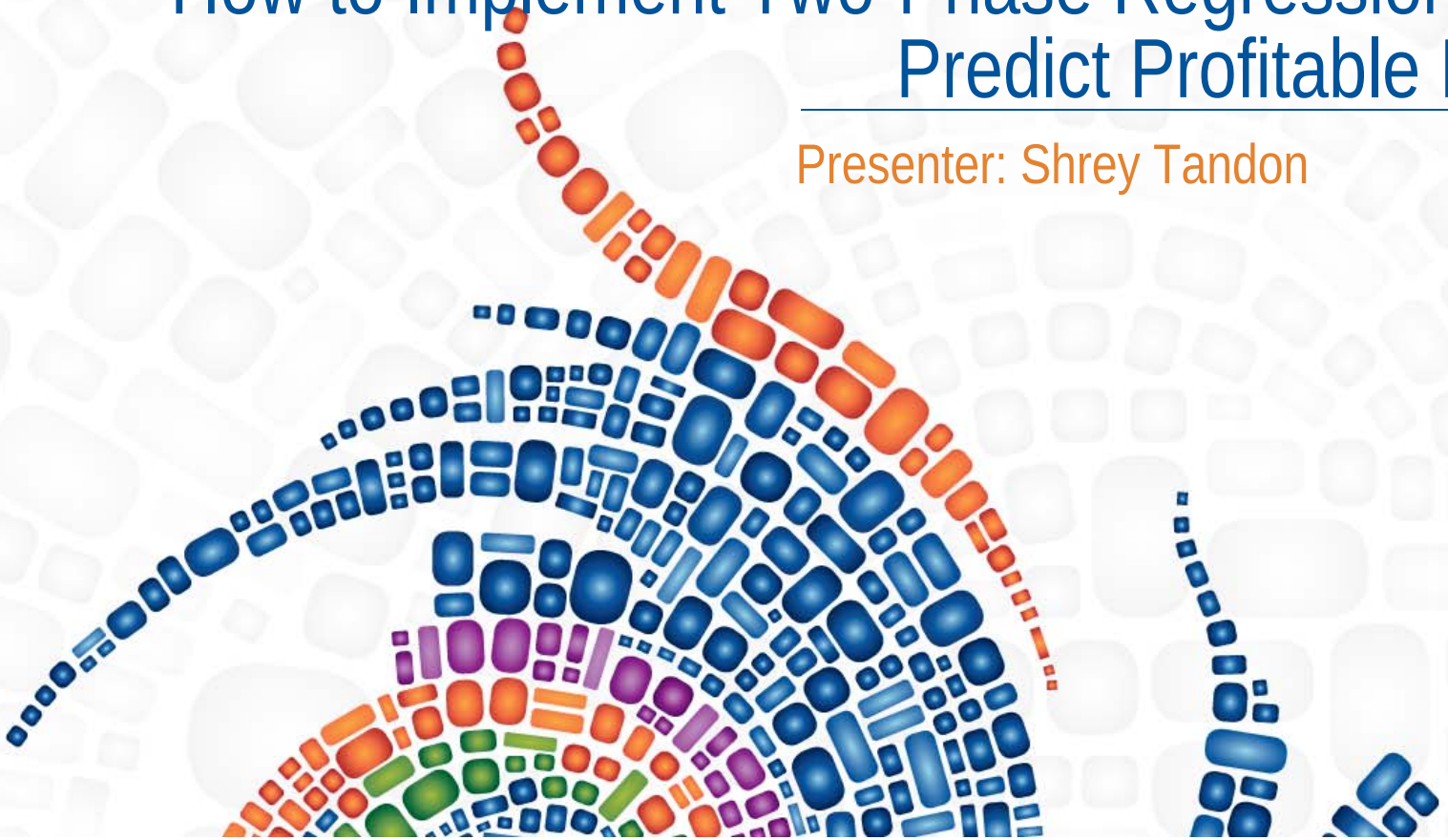
SAS® GLOBAL FORUM 2015

The Journey Is Yours

How to Implement Two-Phase Regression Analysis to Predict Profitable Revenue Units

Presenter: Shrey Tandon



SAS GLOBAL FORUM

How to Implement Two-Phase Regression Analysis to Predict Profitable Revenue Units

Shrey Tandon
Manager, Market and Customer Insights, Sobeys West

Abstract

•A utility company wanted to establish new kiosks and posed the following questions :
- **Which demographic and economic factors are crucial for the success of a payment kiosk?**
- **How well are the current profitable kiosks expected to perform?**

•Predictive models built for a utility company with kiosks in the US
•The company also provided inputs for modeling based on business acumen

Objectives

•To predict the average monthly transactions at each kiosk location

Approach:

1. Study the correlation of independent variables with the target variables.
2. Interpret the relationship between the variables and the target variable.

Methods

•Logistic regression to predict which kiosks would be profitable.
•Linear regression to predict the average monthly revenue at each profitable kiosk.

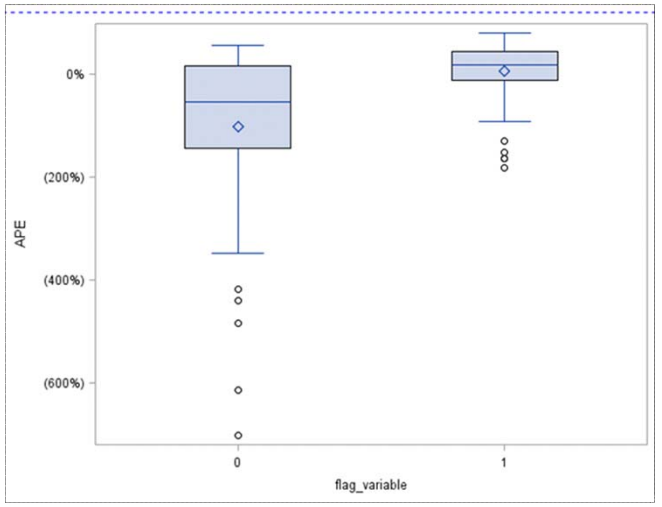
SAS Enterprise Guide was used for modeling.

Assumptions

•The kiosks have been in operation for more than 24 months
•For metro areas, the socio-demographic factors in a 3 mile radius were compared to a 5 mile radius in non-metro areas.

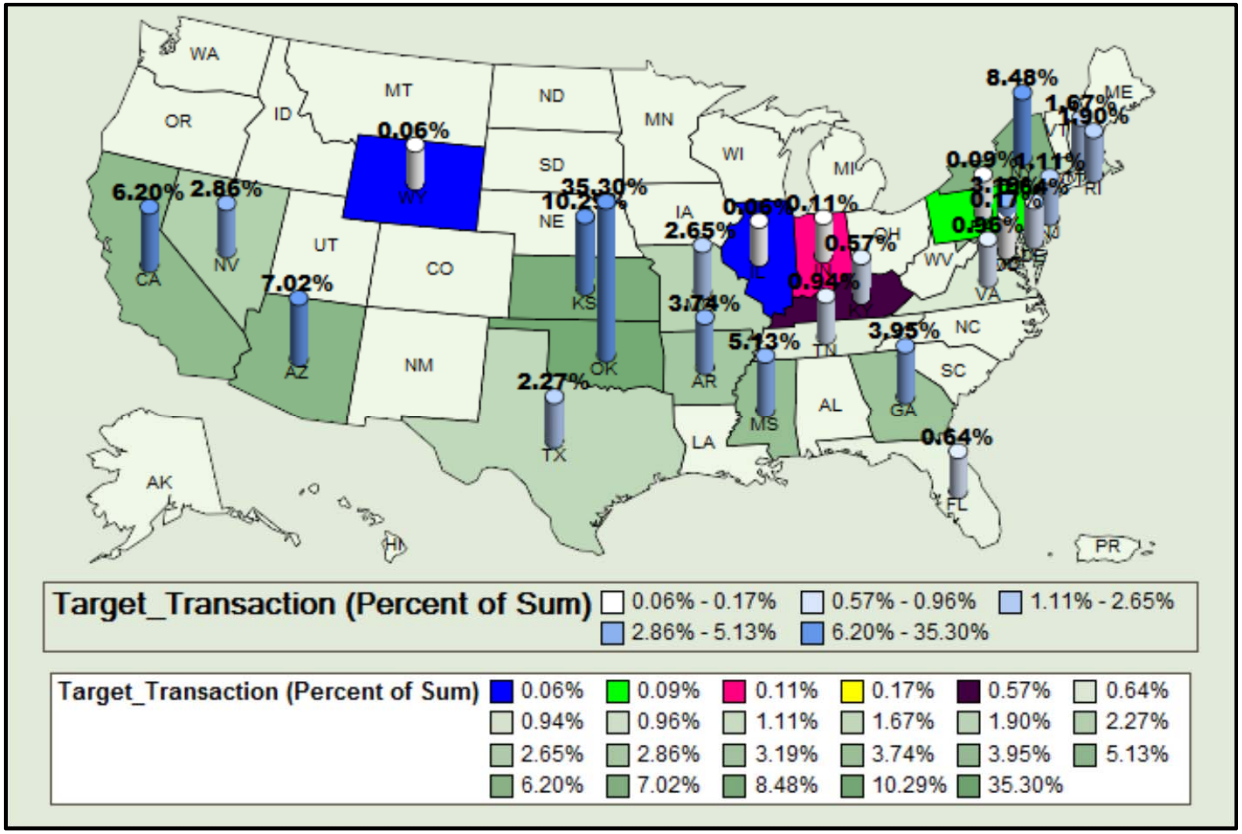
Conclusions

•Dependent and independent variables with highly skewed distributions perform better with two-phase regression model.
- Model predicts the Average Percentage Error (APE) for the target variable when the kiosk is yielding > $350 per month more accurately



Oklahoma leads in monthly average revenue with most of the kiosks located in Oklahoma City

- Major states accounting for 67.29% of the total annual transaction volume in 2011 were:
  - OK (35.30%), KS (10.29%), NY (8.48%), AZ (7.02%) and CA (6.20%)
- Most of the kiosks are located in Oklahoma City (42 kiosks), 17 kiosks in Kansas City and 7 kiosks in Norman City.



Two Phase Regression improved forecast for kiosks earning monthly average revenue greater than $350

Phases: Define parameters and scope → Extract, clean and transform data → Segmentation and correlation analysis → New variables created and data partitioning → Stepwise Logistic Regression to find the most correlated variables → Linear Regression to predict average revenue at profitable kiosks → Conclusions

Process:

- To predict the average monthly transactions at each kiosk
- Study correlation of independent variables with the target variables.
- Predict target variable with independent or transformed variables
- Threshold value for a kiosk to be profitable : $350

- Revenue, demographic and economic data extracted and cleaned
- Assumptions set for rural and urban areas
- Kiosks in operation for 2 or more years to be considered

- Segments were created to capture the important variables based on kiosk location
- Segments created to capture the varied levels of the target variable
- Segmentation analysis helped to study the difference in behaviors for the high revenue kioks vs. Ow revenue kiosks

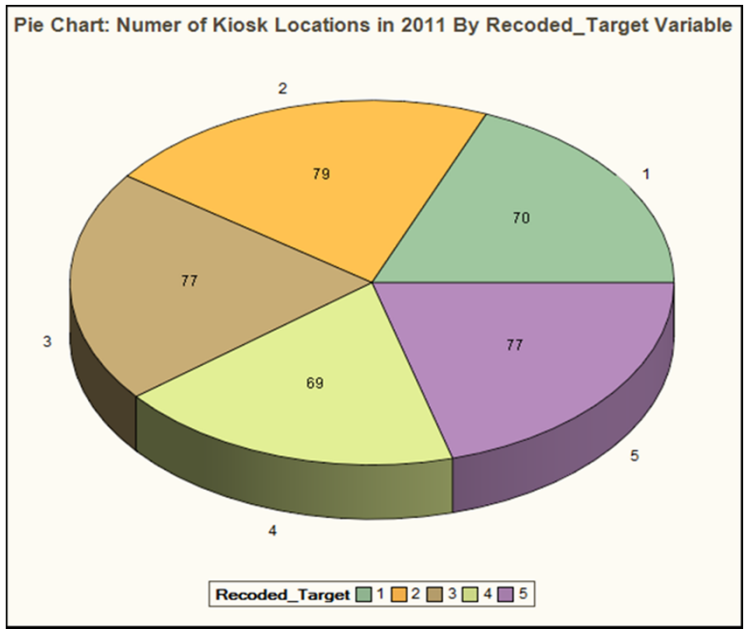
- Flag variable for target variable created in SAS Enterprise Guide using Query Builder
- Profitable kiosks with monthly transactions more than $350 to be assigned 1
- Validation and training datasets built on total data because of small size of the dataset (235 observations)
- Random sampling used

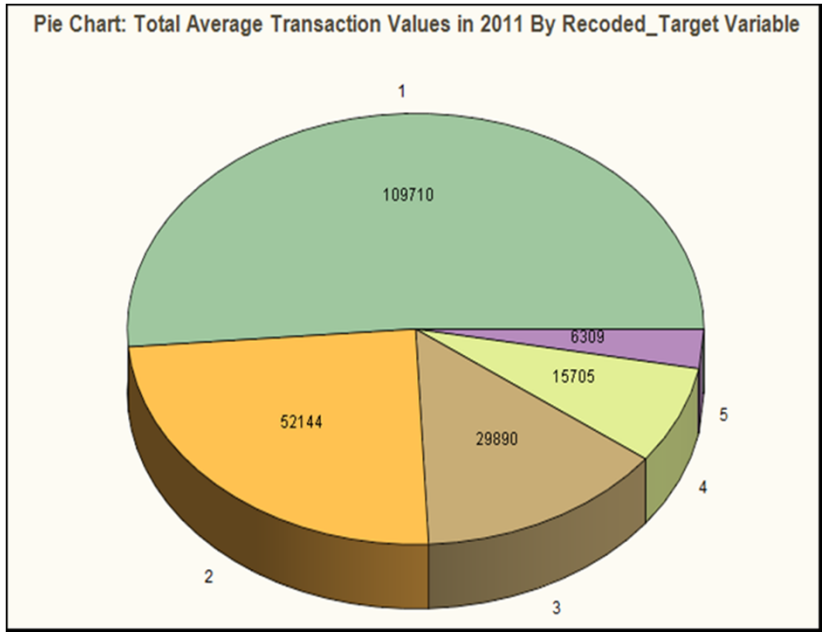
- Stepwise logistic regression performed based on the most correlated variables
- Dependent variable: Flag variable for the target variable

- Stepwise linear regression applied to the logistic model to predict the average revenue at the profitable kiosks
- Dependent variable used in linear regression : the log transformed target variable because
  - –The target variable had a very skewed distribution

- A Regression model also built on most important variables from the client's perspective
- The effectiveness of the model is more for predicting the Average Percentage Error (APE) for the target variable when the kiosk yields > $350 per month.
- If the threshold value for is moved to even $500, the predictive power of the model in terms of APE will substantially increase.

Top 20% of kiosks constituted 52% of average revenues in 2011





The recoded_target variable 1 representing segment 1 has 52% of the total average transaction value in 2011, followed by segment 2 which has 24% of the total transaction value, followed by segments 3, 4 and 5 comprising 14%, 7% and 3% respectively of the total average transaction value in 2011.

Top revenue grossers have lower median ages, higher Hispanic population and have been in operation for more than 29 months

	Caucasian_1mile	**Median_Age_1 mile**	**Caucasian_5mile**	**Caucasian_3mile**	**Hispanic_3mile**	**Hispanic_5mile**	**Hispanic_1mile**	**Median_age_3 mile**	**Hours**	**Months of operation**
1	71% have a concentration between 0.075 and 0.556.	84% between 25 and 36	74% between 0.145 and 0.579	75% between 0.08 and 0.555	65% higher or equal to overall average between 0.198 and 0.869 (Overall: 40%	67% higher than average between 0.141 and 0.866 (Overall: 43%	56% higher than average between 0.172 and 0.865 (Overall: 30%)	29% between 29 and 32 (Overall: 10%) and 16% for 36 (Overall: 13%	24X7 and 9am-6pm and 9am-12 am stores	64% between 29 and 83 months
	Hispanic_3mile	**Divorced_5mile**	**Caucasian_3mile**	**American_Indian_5mile**	**Store**	**American_Indian_3mile**	**Competitors_5mile**	**Hispanic_1mile**	**Caucasian_5mile**	**Medain_income_3 mile**
2	81% between 0.09 and 0.6 (overall: 59%)	22% for 0.145 and 39% for 0.183 (Overall: 18%; 20%)	54% between 0.555 and 0.674 (Overall: 34%)	47% between 0.018 and 0.042 and 6% for 0.122 (Overall: 33%; 2%)	7 Eleven (17), Homeland (9) and Verizon East (7) (around 42% of the transactions; overall: 50%)	57% between 0.018 and 0.193 (Overall: 45%)	2 competitors (lower than average at 53% compared to 71% overall average)	Higher than average at 34% for 0.103 and 13% for 0.24 (overall average: 22%,	Higher than average at 58% between 0.579 and 0.688 (overall average: 38%)	Higher than average at 65% between $4404 and $56515 (overall average: 50%)
	Competitors_5mile	**Store**	**Median_Income_mile**	**Divorced_3mile**	**Months_of_operation**	**Caucasian_5mile**	**American_Indian_1mile**	**Median_Incom e_3mile**	**Caucasian_3mile**	**Competitors_3mile**
3	Mainly 2 competitors (69% compared to 71% overall average)	7 Eleven (13), Verizon East (14) , Buy For Less (7) (44% of transactions; overall: 40%)	25% for $29981 and 19.5% for $47242 (overall average: 21% ; 18%)	Equal to and higher than average at 25% for 0.174 and 23% for 0.196 (overall average: 25%; 18%)	36% for 8 months and 19.5% for 40 months (overall average: 33%; 16%)	57% of values between 0.688 and 0.796 (overall average: 44%)	49% of values have 0.006 and 16% have values of 0.028 (overall average: 55%; 10%)	31% and 29% of values have $56515 and $32293 (overall average: 29%; 21%)	59% of values between 0.674 and 0.792 (overall average:44%)	Mainly 2 competitors (lower than average at 57% compared to 67% overall average)

Low revenue earners have higher Caucasian population, have lesser businesses and households in a 1 to 5 mile radius

	Divorced_5mile	**Caucasian_3mile**	**Average_Income_1mile**	**Caucasian_5mile**	**Average_Income_5mile**	**Caucasian_1mile**	**College_Degree_3mile**	**College_Degree_1mile**	**Median_Income_5mile**	**College_Degree_5mile**
4	*23% for 0.145 and 23% for 0.183 (Overall: 18%; 20%)*	*62% of values between 0.792 and 0.911 (overall average:40%)*	*45% of values between $51695 and $60827 (overall average: 49%)*	*Higher than overall average at 54% between 0.796 and 0.905 (overall average: 34%)*	*45% of values between $51695 and $60827 (overall average: 50%)*	*58% between 0.797 and 0.917 (Overall: 32%)*	*45% of segment members likely to have college degree population of 0.232 to 0.322 (lower than overall average of 57%)*	*46% of segment members likely to have college degree population of 0.191 to 0.291 (lower than overall average of 55%)*	*46% between $38728 and $46787 (overall average: 52%)*	*19%, 23% and 19% of segment members likely to have college degree population between 0.229 to 0.385(overall average of 67%)*
	→	↑	→	↑	→	↑	↓	↓	→	↓
	Pop_18plus5mile	**Pop_5 mile**	**Businesses_5miles**	**Store**	**Households_5mile**	**Households_3mile**	**College_Degree_3mile**	**Pop_18plus3mile**	**Businesses_1mile**	**Pop_5 mile**
5	*56% have lower value 25018 (overall avg: 38%)*	*57% have lower value 33294 (overall average: 38%)*	*49% have lesser businesses at 1179 businesses (overall average:35%)*	*Verizon_East (20) and Homeland(6)*	*49% have lower value 11078 (overall avg: 34%)*	*52% have lower value 4545 households (overall: 34%)*	*54% between 0.232 and 0.322 (low values) (overall avg: 57%)*	*52% have low value of 10475 (overall avg: 32%)*	*53% have few businesses at 116.25 businesses (overall avg: 32%)*	*53% have lesser population at 13842 (overall avg: 32%)*
	↓	↓	↓	→	↓	↓	→	↓	↓	↓

LEGEND

↑ Higher than overall average

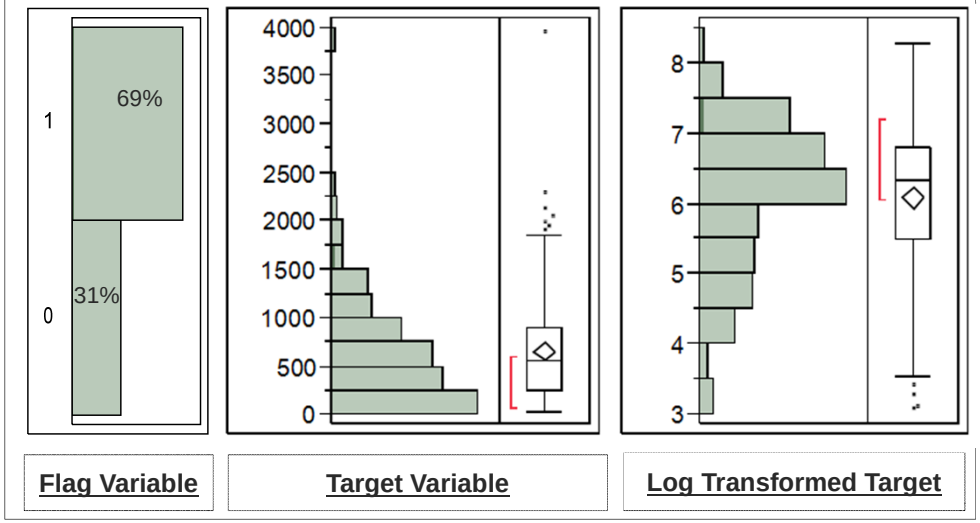
→ Equal to overall average

↓ Lower than overall average

The distribution of the target variables, the means and the standard deviations for the partitioned datasets are similar

Distribution of Flag Variables, Target Variables and Log Transformed Target Variables

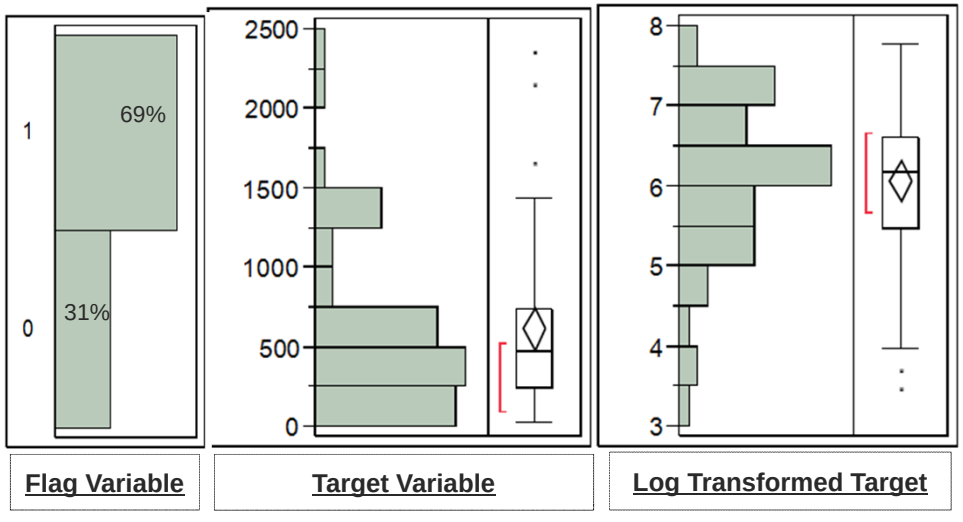
Training Dataset



Flag Variable | Target Variable | Log Transformed Target

Mean	662.48674
Std Dev	555.68424
Std Err Mean	41.886276
Upper 95% Mean	745.15402
Lower 95% Mean	579.81947
N	176

Validation Dataset



Flag Variable | Target Variable | Log Transformed Target

Mean	620.41954
Std Dev	513.6729
Std Err Mean	67.448556
Upper 95% Mean	755.48294
Lower 95% Mean	485.35614
N	58

Stepwise Logistic Regression revealed number of households to be the most important variable for predicting the flag variable

The flag variable was more accurately predicted with lower misclassification rate in the model with all variables vs. select variables

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.4097	0.6355	0.4156	0.5192	
_college_degree	1	-9.4718	2.2523	17.6851	<.0001	-0.6713
households	1	0.000102	0.000025	16.8865	<.0001	0.6879
others	1	45.0443	12.5672	12.8471	0.0003	0.5376

To predict the Flag variable, stepwise regression with p-values of 0.1 for enter and stay p-values and the following variables were significant for modeling in the logistic regression model:

- Households: ***If the number of households is more in the retail kiosk location, there is a higher possibility that the average monthly transaction is higher at that location.***
- College_Degree***: If the percentage of college degree holders Around the kiosk is higher, the average monthly transaction is likely to be lower at that location.***
- Others: ***If the percentage of other population is more in the kiosk location, the average monthly revenue is likely to be higher at that location.***

Logistic Regression model seems to predict which kiosks are profitable reasonably accurately

The FREQ Procedure

Selection Indicator=1

Table of _INTO_ by _FROM_

INTO(Formatted Value of the Predicted Response)		_FROM_(Formatted Value of the Observed Response) 0	1	Total
0	Frequency	37	11	48
	Row Pct	77.08	22.92	
	Col Pct	63.79	9.32	
	Cumulative Col%	63.79	9.32	27.27
1	Frequency	21	107	128
	Row Pct	16.41	83.59	
	Col Pct	36.21	90.68	
	Cumulative Col%	100.00	100.00	100.00
Total	Frequency	58	118	176

Table of _INTO_ by _FROM_

INTO(Formatted Value of the Predicted Response)		_FROM_(Formatted Value of the Observed Response) 0	1	Total
0	Frequency	44	12	56
	Row Pct	78.57	21.43	
	Col Pct	61.11	7.41	
	Cumulative Col%	61.11	7.41	23.93
1	Frequency	28	150	178
	Row Pct	15.73	84.27	
	Col Pct	38.89	92.59	
	Cumulative Col%	100.00	100.00	100.00
Total	Frequency	72	162	234

Sensitivity, specificity and overall accuracy

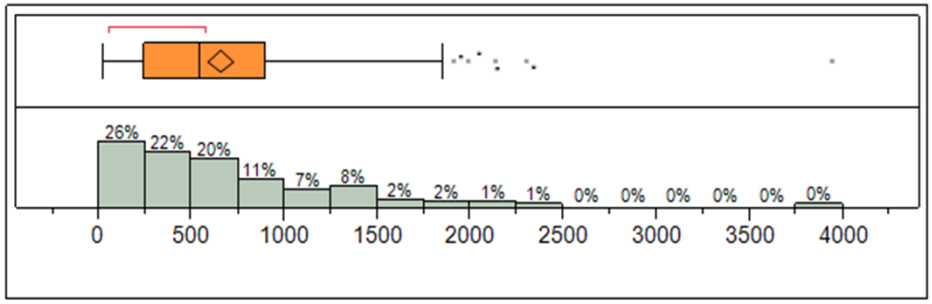
- ***The logistic regression model has an overall accuracy of 82.9%, sensitivity of 92.6% and specificity of 61.1% with comparable figures for the training data set at 81.8%, 90.7% and 63.8% respectively.***
- ***This indicates that the regression model seems to be consistently predicting the profitable kiosks at a reasonably good level.***

Linear Regression performed on log transformed target variable due to its better distribution compared to target variable

Distribution of target and log-transformed target variable

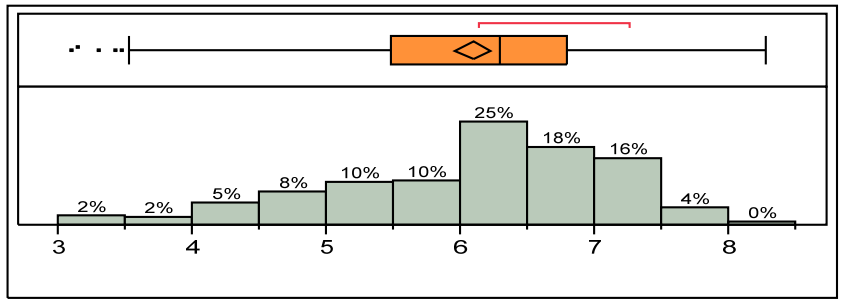
Skewed Distribution of Target variable

- ***The distribution for the target variable has a long right tail, i.e. it is right skewed***
- ***Building a model on such a dependent variable would not be advisable as the target variable is very skewed***



Symmetrical Distribution of Log transformed Target variable

- ***The distribution for the log transformed target variable seems to be closer to a normal distribution.***
- ***Building a model on such a dependent variable would yield better and more consistent results***

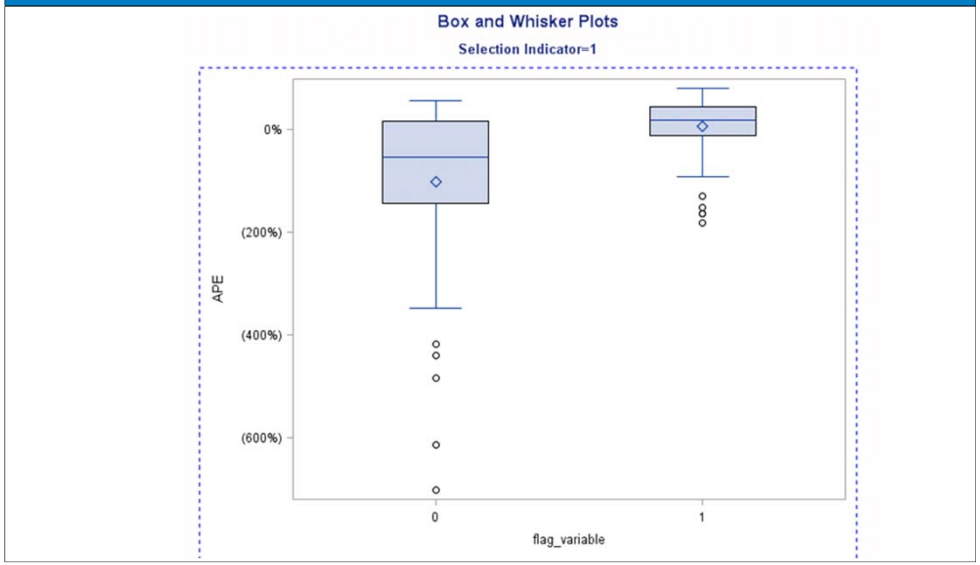


11

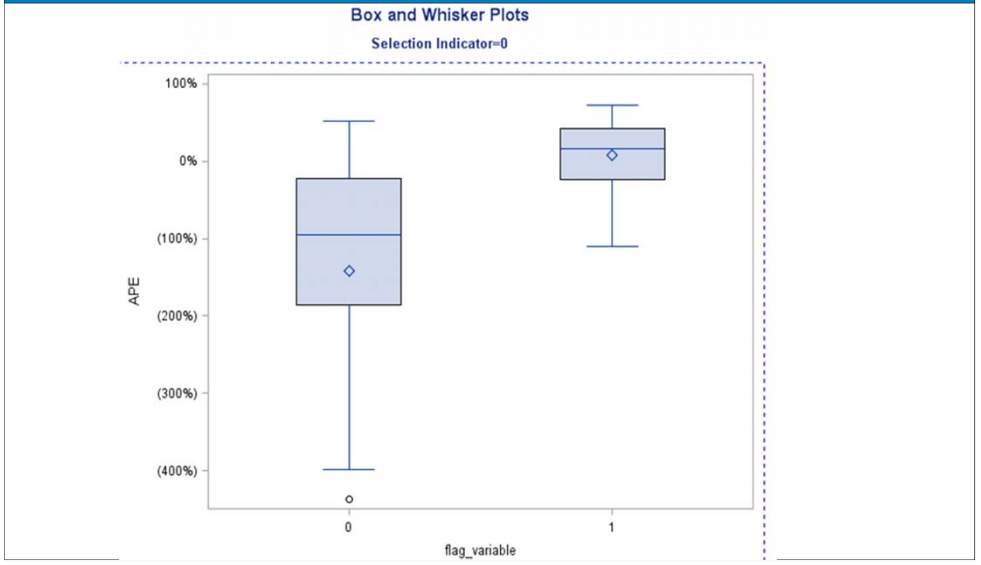
Average Percentage Error is higher and varies more for kiosks earning on an average over $350 monthly

Box and Whisker Plots for Average Percentage Error for the flag variable

Training Dataset



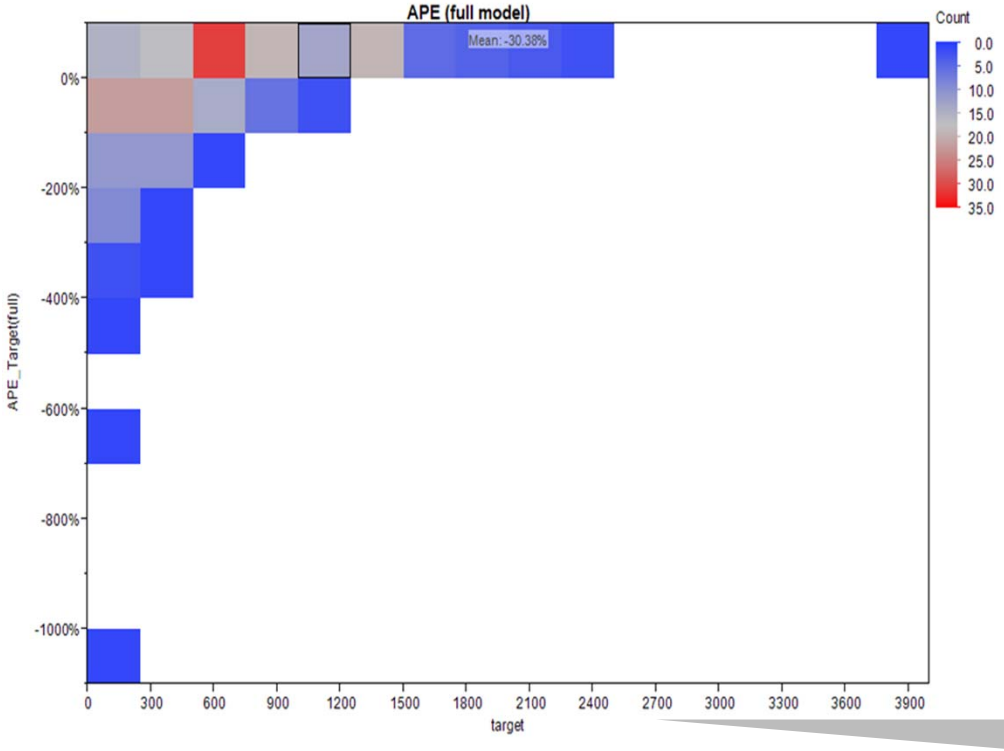
Validation Dataset

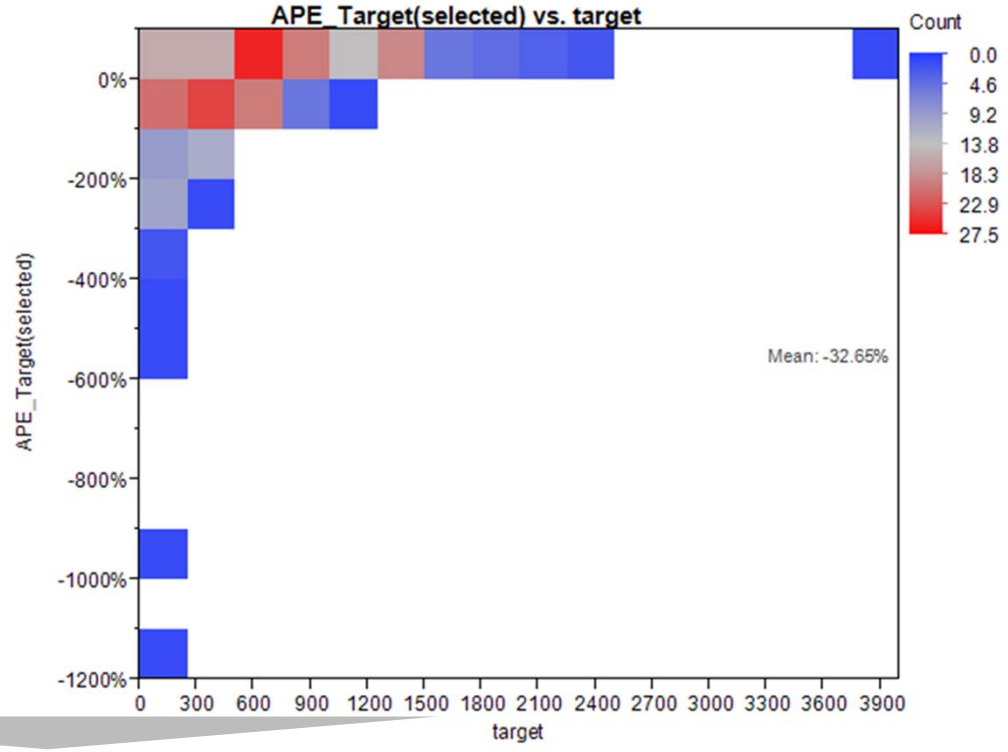


•The tight boxplot for flag variable 1 for APE in training and validation datasets suggests that the kiosks earning more than $350 may be predicted reasonably accurately.
•The broader boxplot for flag variable 0 for APE in training and validation datasets suggests that the kiosks earning less than $350 may be over predicted by the model.
•There are a high number of outliers too for flag variable 0, suggesting that the APEs are significantly high (in negative direction) indicating under forecasting for kiosks earning less than $350

Mean Average Percentage Error is lesser for the regression model based on all the variables as compared to the select variables

Mean Average Percentage Error (MAPE) for the Validation Dataset based on the Regression model





- Model built on all the variables performs slightly better at -30.38% compared to -32.65% for the selected variables.
- Models predicts the Average Percentage Error (APE) for the target variable when the kiosk is yielding > $350 per month more accurately
- ***If the threshold value for the kiosks is moved to $500, the APE will substantially decrease***

The model built with all variables predicted the target variable more accurately for profitable kiosks than underperformers

The MEANS Procedure

Selection Indicator=0

Analysis Variable : APE						
flag_variable	N Obs	Mean	Std Dev	Minimum	Maximum	N
0	14	-1.4191905	1.5946329	-4.3706253	0.5223169	14
1	44	0.0760882	0.4663690	-1.0948886	0.7257622	44

Selection Indicator=1

Analysis Variable : APE						
flag_variable	N Obs	Mean	Std Dev	Minimum	Maximum	N
0	58	-1.0208700	1.6446389	-7.0184005	0.5586894	58
1	118	0.0723178	0.5220123	-1.8243815	0.8055689	118

For the predicted values (not log-transformed) of the target vs. the actual value of the target:

- ***MAPE is 7.2% for the kiosks earning more than $350 in the training dataset***
- ***MAPE for kiosks earning less than $350 is -102% in the training dataset***
- ***MAPE is 7.6% for the kiosks earning more than $350 in the validation dataset***
- ***MAPE for kiosks earning less than $350 is -142% in the validation dataset***
- ***Similar MAPEs for validation and training datasets suggest that the linear regression model seems to have lesser MAPEs for profitable kiosks.***

Conclusion: Plugging in the demographic input variables for a kiosk can help to predict the monthly average revenues at that location provided all the assumptions are met . This model can be used to find the log-transformed target variable which can be converted into the target variable by applying the antilogarithm to the predicted result.

SAS® GLOBAL FORUM

April 26-29
Dallas, TX