Guidelines for Data Science Case Studies

Prepared by Intern at AnalytixLabs

ANALYTIXLABS
www.analytixlabs.co.in

First Floor, SCO 394, Sector 29, Gurgaon
info@analytixlabs.co.in
(+91)75308-18107

Contents

Credit Card Segmentation

Issue faced:

Uneven clustering solution obtained, with majority of observations restricted to one cluster and multiple clusters having very low amount of observations.

25% above overall		Cluster5					Cluster6					
25% below overall	Segment Size	0%	16%	17%	2%	64%	62%	0%	0%	16%	16%	5%
	All	1	2	3	4	5	1	2	3	4	5	6
ONEOFF_PURCHASES	604.9	23497	293.32	1743.65	3545.19	177.22	160.02	6473.03	26084.12	1404.21	270.82	3442.39
PAYMENTS	1700.04	13621.78	3224.6	2448.66	7441.58	855.59	849.17	12094.79	13609.39	1952.02	3152.99	5843.24
ONEOFF_PURCHASES_FREQUENCY	0.21	0.85	0.12	0.67	0.56	0.08	0.07	0.53	0.84	0.66	0.12	0.62
CASH_ADVANCE	994.18	1062.5	4197.95	207.75	1173.48	394.43	394.21	1190.93	1456.23	172.65	4174.73	801.48
CASH_ADVANCE_FREQUENCY	0.14	0.04	0.44	0.04	0.12	0.09	0.09	0.06	0.06	0.04	0.44	0.09

Fig: Uneven spread of data points among clusters

Possible causes:

1. The K- means algorithm is a ***distance based algorithm*** for clustering and its solutions can be adversely affected in case of:
    i. Improper outlier treatment leading to clusters being formed with uneven amount of observations among them.
    ii. Inappropriate missing value imputation such that the imputed values behave as outliers.

Clustering based on overshadowing variables can result in highly skewed clusters, since the population can behave in a homogeneous manner with respect to the variable.

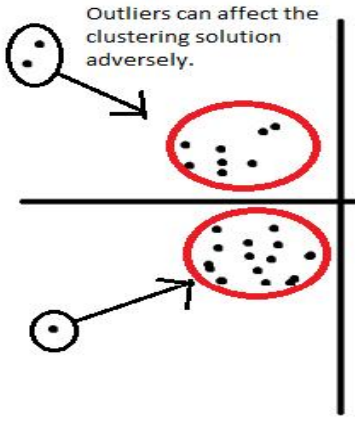

Fig : Outliers can cause formation of extraneous clusters

Possible Solutions:

1. Make certain that all variables that are being used for clustering have been ***standardized***.
2. Ensure that all missing values have been dealt with and ***missing value imputation*** does not result in outliers.
3. Ensure that all variables have been ***properly treated for outliers***.
4. If outliers have been capped at high percentile values (99 and above), retry the procedure by ***capping the variables at either the 95th percentile or mean+3 standard deviation level***. This will reduce the observation's distance effect and has high probability of leading to better, well defined clusters.
5. Verify ***appropriate selection of variables***, taking into account maximum variability and reduced multi-co linearity. Best practices for variable selection include :
    i. Factor analysis (through **PROC FACTOR**)
    ii. Variable clustering (through **PROC VARCLUS**)

Proactive Attrition Management: Logistic Regression

Issue Faced:

How to perform variable reduction and choose the appropriate variables from a large set of possible predictors?

Possible Solutions:

For eliminating categorical variables:

Perform ***chi-square test*** between the categorical variable and the response variable. Eliminate variables with Chi-square below a certain threshold and phi-coefficients tending to zero.

For eliminating continuous variables:

Calculate a ***correlation matrix*** (using PROC CORR), and drop variables with high correlations among themselves.

For selecting continuous variables:

Continuous variables can be selected based on their ***comparative histograms on the response variable*** (churn in this case).

The more significant the difference in distribution of independent variable across the different response categories, the more likely it is to be a significant predictor. The figure below shows the histogram of variable X4 for observations related to response variable categories 0 and 1. The comparative histogram clearly shows a marked difference in distributions and hence, can be kept as a possible candidate for the regression model.

The comparative histogram can be obtained with the help of PROC UNIVARIATE procedure in SAS.

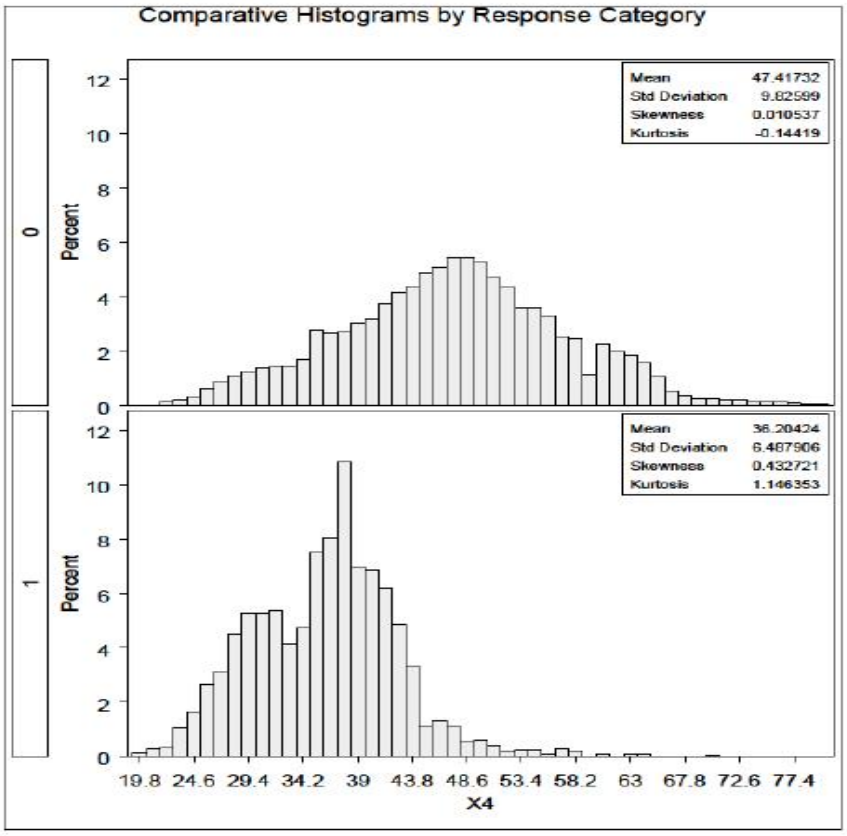

Fig: Variable X4 shows significantly different distributions in different response categories

Utilizing Factor Analysis:

The final step for variable reduction can be ***factor analysis***. Based on the rotated factor patterns obtained, an apt number of variables that loads highly on to the factor can be chosen and these variables will serve as the input set to PROC LOGISTIC.

The final set of predictor variables will be output by the PROC LOGISTIC method through stepwise selection.

Issue Faced:

Low value of concordance leading to poor predictive capability of the logistic model.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	58.8	Somers' D	0.176
Percent Discordant	41.2	Gamma	0.176
Percent Tied	0.0	Tau-a	0.088
Pairs	392802680	c	0.588

Fig: Low Concordance Value for Logistic Regression Model

Possible Causes:

1. ***Improper variable selection***: If key variables that explain maximum variability in the data are left out, it will lead to models with poor predictive power.
2. ***Incorrect outlier treatment***: If outlier treatment has not been performed it will lead to biased variable estimates, which severely impacts model capability. The variable coefficient estimates are based on ***maximum likelihood***, which tries to find an estimate that best explains the variation in the data. Hence, if there are outliers it will cause biased estimates to be calculated.
3. ***Non-linear relation of independent variable with the logit function***: The assumption of logistic regression is that the independent variable has a linear relationship with the link function (logit). Violation of this assumption will lead to reduced prediction accuracy.
4. ***Data itself might not be amenable to prediction with high accuracy***: High degree of missing values in important variables, categorization of too many continuous variables, data compilation errors etc. will lead to subpar performance of regression models on such data.

Possible Solutions:

1. ***Proper variables need to be selected***, in accordance with methods discussed in the previous section.

2. ***Perform outlier treatment*** on all variables of interest by :
   i. Capping the values at an appropriate percentile level (95^{th}, 99^{th} etc.), if the number of outliers are too many.
   ii. Removing the observations with outliers, if the number of such records is a very small proportion of the data available.

3. ***Check linearity*** between logit function and continuous independent variable :

   i. Deciling :
   a) Divide the data into ten groups based on the independent variable.
   b) Compute the probability of event (***p***) occurring by dividing the number of 1's with total observation within each decile.
   c) Calculate $\log(\frac{p}{1-p})$ for each decile and observe the trend of increase or decrease in the logit value with independent variable decile to determine if they have a linear relationship.

   ii. Using PROC LOESS :

   The PROC LOESS procedure helps identify the relationship between the categorical variable and continuous variable, by plotting a graph of the continuous variable versus the probability value for the binomial variable.

```
proc loess data= linearity_test;
model churn = months;
run;
```

Fig: Sample code for PROC LOESS

Since we assume linearity between logit function and the continuous covariate, we expect to see a sigmoid like graph between the variable and the response probability (as shown in the figure below).

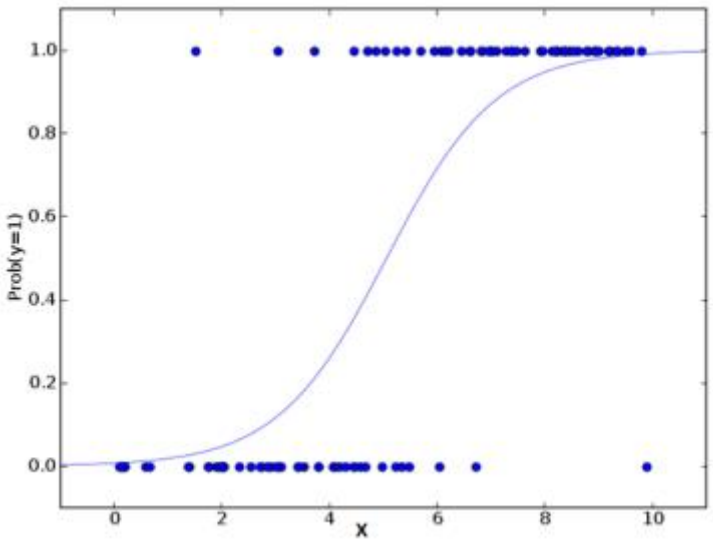

Fig: Sample Output of PROC LOESS given logit of y is linearly dependent on x

iii. If the linearity assumption fails while employing any of the above two procedures ((i) and (ii)), transform the independent variable in a suitable manner and re-run the tests.

iv. The model will yield improved results if the assumption of linearity is fulfilled.

Predicting Credit Card Spend: Linear Regression

Issue Faced:

The linear regression model has a low R-Square value and poor predictive power.

Root MSE	0.55621	**R-Square**	0.3092
Dependent Mean	5.99388	**Adj R-Sq**	0.3074
Coeff Var	9.27954		

Fig: Regression model with low R-Square

Possible Causes:

1. ***Presence of outliers and too many missing values***: The presence of outliers can result in biased coefficient estimates, since the regression is based on ***ordinary least square*** method, which is a distance based procedure. Too many missing values will result in the omission of those records in the PROC REG procedure, which leads to lesser data for the regression to run on and a subsequent output of a poor model.
2. ***Improper variable selection***: The key to a good regression model is the identification and selection of variables that can account for maximum variance and are key drivers for the response variable. A regression run on the basis of an incorrect set of variables will also lead to an underperforming model.
3. ***Violation of linearity and normality assumption***: If the assumptions of linearity and normality between dependent and independent variables is violated, the regression results will be unreliable.

Presence of influential observations: An influential observation is a data record that when deleted results in a noticeable change in the parameter estimates for the regression. Such observations can be detected with the help of Cook's Distance formula.

Possible Solution:

1. ***Outlier treatment*** on variables needs to be done properly. In case of lesser number of outliers they can be removed, else outliers need to be capped at either 95^{th} /99^{th} percentile or mean + 3 STD.
2. ***Proper variable selection*** can be done through the following methods :
    i. ***Correlation matrix***: A correlation matrix between the continuous variables can be calculated through the PROC CORR procedure and variables with high correlation amongst themselves can be dropped. Also, variables that have correlation tending to zero with the response variable can be dropped.
    ii. ***Variables with a high percentage of missing values can also be dropped***.
    iii. Continuous variables with ***variance tending to zero*** can also be dropped since they cannot explain the variance in the response variable.
    iv. ***Comparative histograms*** can be used to select possible predictors (through the PROC UNIVARIATE method) by analyzing whether target variable distributions across the different levels of the independent categorical variable show any significant change.
    v. ***Factor analysis*** can then be employed to select the desired number of variables from each factor and form a variable set to be used in the PROC REG model.

3. ***Check linearity assumptions*** between the dependent and independent continuous variables with the help of scatter plots and transform the variables such that they are linear with respect to the dependent variable.
4. ***Check normality assumptions*** for the variables and transform them if necessary.
5. ***Appropriate transformations required*** can be arrived at with the help of the ***box-cox*** transformations, which is provided through the PROC TRANSREG procedure.

The box-cox transformation identifies the transformation which is most likely to yield a linear relationship between the dependent and independent variables in the regression. The result of the procedure is a value of lambda, from which the appropriate transformation can be applied on the dependent variable.

```
proc transreg data=s.linear_reg;
model boxcox(total_spent) = identity(income employ creddebt total_items);
run;
```

Fig: Sample code for box-cox transformation

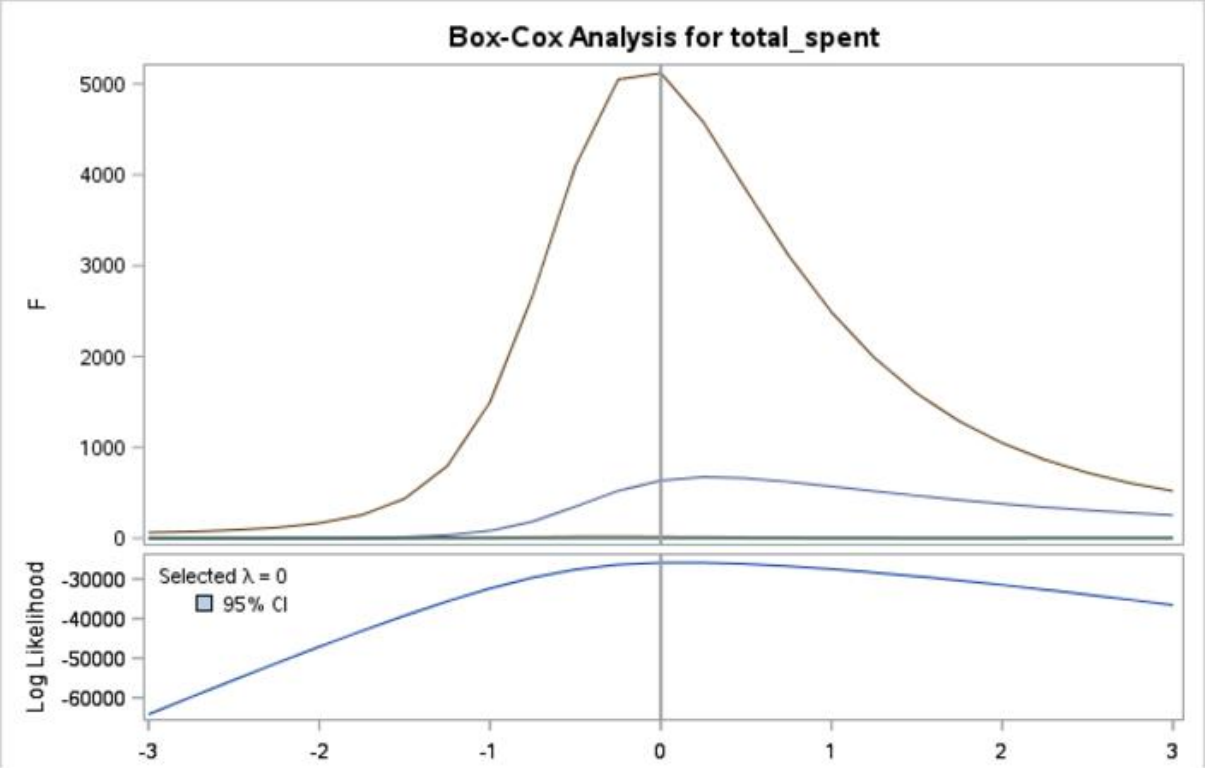

Fig: Results of the box-cox transformation

The above figure shows that a linear relationship is most likely for a value of lambda = 0. From the table below, we see that lambda = 0 corresponds to the log transformation. Therefore, with the help of this procedure an optimum transformation can be settled upon to fulfill the assumption of linearity.

This method can also be used for a single variable ***to identify the transformation required to best approximate normality.***

λ	**Y'**
-2	$Y^{-2}=1/Y^2$
-1	$Y^{-1}=1/Y$
-0.5	$Y^{-0.5}=1/(Sqrt(Y))$
0	Log(Y)
0.5	$Y^{0.5}= Sqrt(Y)$
1	$Y^1=Y$
2	Y^2

Fig: Table of lambda values and corresponding transformations

6. ***Removal of influential observations*** will positively impact the parameter estimates and help remove the bias within the model. The presence of influential observations can be identified through the model diagnostics charts that are output with PROC REG (as shown in figure below).

   The cook's D values can be output into a dataset with the option :

```
/selection = stepwise slentry= 0.05 slstay=0.1 vif stb;
output out = tmp cookd=cd;
```

   Remove the observations with cook's D value greater than ***4/N.*** Where N is the number of observations in the dataset.

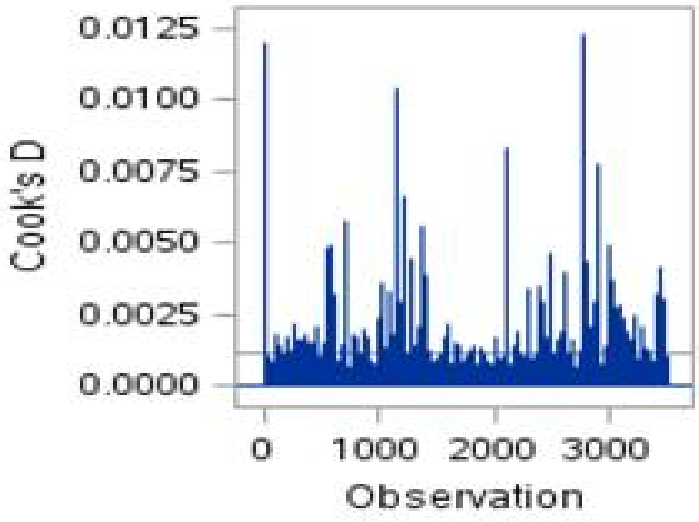

Fig: The Cook's D plot reveals the presence of influential observations