Supervisor Meeting #8

Drafted by: Liam Pang (29/03/2018)
Edited and Vetted by: Ong Geok Ting (29/03/2018)

Date	Time	Venue
29/03/2018	1300Hrs – 1400Hrs	SIS MR 4-06

Participants: Prof. Kam, Liam Pang, Ong Geok Ting, Tan Rui Feng
Agenda:

1. **To clarify issues on data transformation, selection and interpretation**
2. **To understand LCA better**

Meeting Item 1: Data transformation, selection and interpretation				
S/N	**Issue**	**Action**	**By**	**Due**
1	Prof. Kam reviewed the team's project and provided some input. - Learn the differences and similarities between LCA and K-Means more - The team should not be using two different sets of data to conduct the clustering analysis - To separate the observations between those who are active and inactive. The Inactives are influencing the interpretation of the data. Another suggestion would be to separate them and build two different models. Do a statistical test, parametric test, mean test to identify if active and inactive are identical.	Include the performance output as a clustering variable. To conduct data transformation using different techniques. Compare results to find the most optimal.	All	1/4/18

	- Paid or non-paid should not be conveniently left out for K-Means while included for LCA. - To include performance variable of customer enquiries for clustering as it might indicate the relative success of each group. The business justification to include them would probably be that those measures are the customers of the customers of REO. Therefore, it might be useful to investigate further. - The team only have 4 variables, so a multivariate analysis to identify correlation should be done. Currently, total sessions and organic is highly correlated. Therefore, both variables should not be included in the clustering. Decide to drop either sessions or organics but based on the distribution, it might be smarter to drop organic due to the higher proportions of 0. - In cluster analysis, especially K-means, it is very sensitive to different data range. For example, sessions and cobroke have a huge difference in their variance. This is something need to be resolved. This will cause			

	your centroid to be pulled away. Identify if this data range will affect your cluster analysis. One way to do so is to build 2 models, 1 with standardisation and another without. - As K-Means is a method which is sensitive to skewness, a transformation should be done. Identify the pattern of data distribution to determine the best transformation technique. However, JMP has a global transformation method – the Johnson Transformation Method. - Attempt a few transformation methods such as Johnson, Log 10 or others recommended by JMP. Use these sets of data and see which produces the most satisfactory set of clusters. - Use the 2 loglikelihood or AIC to assess. The smaller AICc is the recommended method.			
2	- Use the outlier clean up function to remove them. Other methods to deal with outliers would be normal mixtures and robust normal mixtures. It will identify them and start the seeds without the influence from the outliers before it starts converging	Experiment with various clustering techniques and removal of outlier to determine if there exists a most optimal technique for the data.	All	1/4/18

	towards the optimal answer. - One challenge is to define the K upfront using K-means. JMP allow you to help you define a range, such as CCC – which range give you the optimal configuration. The largest CCC value is the most optimal value. The team had issues previously because the optimal CCC value was in the negative, which means that the clustering was not converging well. CCC can be influenced by outliers, so if it does not converge properly, use other methods such as normal mixtures. If optimal CCC is at the upper/lower limit of the range, re-run the clustering with a greater range as the optimal CCC might sit outside of the input range.			
3	To evaluate the cluster results: - Check if they are evenly distributed - Check for odd membership, such as really low counts. - Look at the cluster mean and the centroids as these are the summary statistics for the clusters. Do note that the centroid values are pre-transformed. Do not use the transformed data to do analysis in means	-	-	-

	and standard deviation as you cannot derive a good interpretation. - Analyse using visualisation such as biplots to identify overlapping clusters and merge them - Save the clustering results and visualise the results using graph viewer with the cluster numbers and distances. The mean will assist with the interpretation of the clusters. - A parallel coordinate plot would be useful to visualise the differences between the clusters.			
Meeting Item 2: To understand LCA better				
S/N	**Issue**	**Action**	**By**	**Due**
1	For latent class clustering, the classes for categorisation must be equal. For example, binning all the variables into quantile. To analyse the results, focus on the frequency of the membership in a sub-class. This is very different as compared to k-means, as you compare the mean and standard deviation.	To implement the suggestions.	Tan Rui Feng	30/3/18