ANLY482 Analytics Practicum
Meeting Minutes

Date:	03/02/2015
Time:	1800h
Venue:	SIS Meeting Room 4.1
Attendees:	Prof KAM Tin Seong
	CHENG Fu Mei
	LEONG Wai Sum
	Lynette SEOW Hui Xin
Absent with Apologies:	–
Minutes Taker:	LEONG Wai Sum

Agenda:

Topic	Customer Analytics
Content	- Descriptive Analytics - RFM: Recency, Frequency, Monetary - Is there a need to adapt to the data we have at hand? E.g. geo-demographics, routes - Customisation of RFM: Refer to "Reading 2" - Using RFM to cluster: Refer to "Reading 3" - Other possible derived variables: (Frequency of) Transaction behaviour: Many small orders, or a few large orders throughout the year? Retention rate / Length of account? - Customer profiling/segmentation - Behaviour - Nature/Characteristics - Use clustering? K-means / Hierarchical - Recommend targeted solutions - Who are the high value customers? - Is there potential for growth in certain segments? - Drives marketing strategies

Topic	Future Directions (Next steps)
Content	1. Data collection – How to improve? - Reduce missing data (by making fields mandatory) - Collect new fields/variables (e.g. industry indicators) 2. What is important to customers? - Price? Speed? Quality (no damage)? 3. Benchmark performance against industry 4. Cost-side benefits - Further work can be to include cost factor to get a more complete picture - E.g. Route optimisation (may want to show the concentration of the transport nodes), economies of scale

	• Structure/Process: EDA → Clustering using RFM

Topic	Data Preparation

Content	• Assumption: Different data sources, e.g. CRM and Sales
	• **E**xtraction: Merge data
	• **T**ransformation: Remove unnecessary fields; Derive variables
	• [**L**oading: (Linking the databases) Write data into target database]
	• Aggregation of data:
	    o Quantity, shipment, monetary value
	    o Monthly, quarterly, etc.
	• Number of active customers
	• "Sales & Marketing (S&M)" perspective

Topic	Data Cleansing

Content	• Missing data/gaps – First try to recover from data custodian
	• E.g. >20% missing
	    o Exclude (do not use) field
	    o Check to see whether remaining 80% is biased (tested by looking at distributions before and after) → If not biased, maybe can still be included and used
	• Do not use imputation (e.g. average, random number, remove missing instances) to rectify; recommend how to improve the system instead
	• Recommendations:
	    o How to improve current system to reduce dirty data
	    o Improving current system to prepare for future analysis and analytics work
	    o Identifying which is the crucial data

Topic	Exploratory Data Analysis (EDA)

Content	• Univariate Analysis
	• Distribution of revenue, quantity of shipments, weight, etc.
	• To get a feel of the distribution of the customer base, like who the major customers are, mass market vs. niche market, what industries the customers are from, are they non-homogeneous
	• Have the objectives in mind when running the tests like in distribution analysis
	• Pareto Rule: 80/20
	• Distribution of customers by revenue, volume, and frequency of sales
	• Then slice according to industry, sales channel
	• Indicators to prepare for "Customer Profiling" which is the next stage
	    o To understand the data first in order to know how to slice or profile your customers appropriately
	    o Output of EDA: Which variable to use for customer profiling; which are the ones that are indicative

Topic	Classification
Content	- Not so much of predictive, but more of profiling the customer base, industry, niche, any specific sales channel - To help with planning of S&M strategy - Leave S&M to the domain experts - Cluster analysis - Each is homogeneous - Characteristics of each cluster (i.e. what type of customers) - Questions to think about: - How to bin RFM? - What form of RFM to use? - Apart from clustering, are there any other techniques that we can use? - How to decide on the number of clusters to be used? - Churn analysis – Neural network (Backpropagation multi-perception); tree decision → c.f. Reading 4 - Check out: Customer value classification model and application based on analytic network process and k-means clustering (LUO, YAN, WAN, 2013) - Deciding what k is in clustering → Reading 3 - Implementing more sophisticated RFM analysis → Reading 6 - How to compare efficacy of different methods

Topic	AOB
Content	- More research required (tools): - JMP software – visualisation tools - D3.js - CHAID - Tie in with the deliverables (e.g. Mid-term Client Presentation)

Follow-up Actions:

S/N	Task	Person Assigned	Due Date
1	Data preparation and cleansing	All	08/02/2015, 1159h
2	Basic data exploration	All	09/02/2015, 1159h