Singapore Management University
ANLY482 Analytics Practicum

Supervisor Minutes 9 as on 20th March 2017

Time Start:	2.30pm
Time End:	3.30pm
Location:	SIS Meeting Room 4-5
Recorded by:	Heng Kok Chin
Vetted By:	Peh Zhan Hao

Attendees:	
Prof. Kam Tin Seong Heng Kok Chin Peh Zhan Hao	Associate Professor of Information Systems (Practice) Undergraduate, Singapore Management University Undergraduate, Singapore Management University

Agenda

1. **Conference Paper**
2. **R Model**
3. **JMP Pro**

No.	Discussion:	Action by:	Deadline:
1	**Conference Paper**		
	• Structure is okay • Zhan Hao asked about the difference between Feature Construction and Data Modelling - Data Modelling is a term used by database designers - It is a process where we carefully examine the data that we have, in a system form or raw form - After carefully examining those data, we will try to provide the data model. Like personal data, transaction data, come up with the data type, numerical or categorically, whole number or decimal - Then do a physical model, like create four data tables to link them - Then only translate into a UML program or process - Feature Construction/Engineering are terms that are coined by the data scientists, it can be defined as log transforming a collection of variables, deriving a new variable based on existing variable (because its skewed) - Look more into the terms used in Prof’s slides, data preparation, ETL (Extraction, Transformation and Loading) - For our case, our sponsor gave us the data, so there isn’t much extraction - We did more of transformation such as data checking, manipulation and data cleaning. That is where you check for missing value, dirty data		

	- We also did more of what we called data wrangling, reorganizing the data, make it into ‘by year’ - Data was in different files but we combined it into one file • Model Development includes the multivariate, variable selection, method selection (selecting the appropriate method) whether use multiple linear regression or use step-wise, forward or backward • What are the things we considered when selecting the method and the criteria (am I supposed to use P-value or AIC – Akaike Information Criterion) • Then we have our model calibration, then model assessment, usually construct several models then carefully assess it to see which model provide me the best values for predicted error, adjusted R square • Understand the parameter estimates, if we are doing the multiple linear regression, we will also test for the assumption test. Like normality and ensuring the model is not violating all these statistical assumption • All these is part of the Model Development • Model Iteration is part of Model Development • Next phase is recommendations • R Shiny is considered Application Development and Deployment • For the paper, we can choose what to focus on or just do two papers		
2	**R Model**		
	• How big is the sample size (after considering the range) • For considering the range for the subjects, instead of AND operator, perform an OR operator • Add in more subjects into consideration based on JMP Pro analysis of subjects that are significant • We can consider doing “Subject Level Analysis” where we arrange by subjects instead of by CA1, SA1, CA2, SA2 • Can see how consistent a student is across the year (less variation) • Rename the current boxplots for CA1, SA1, CA2, SA2 to “Overall Performance Analysis” • Boxplots a bit ugly • Axis can just include the title of the subject • Student dropdown can be a search + dropdown list • Size and proportion of the screen needs to be considered (should fit standard screen sizes or make the panels dynamically stretch to 100%) • Letter grades (A1, A2, B3, …) might be more in use for the school than numerical scores (possibly bullet charts?)	Kok Chin	27th March 2017 (Monday)
3	**JMP Pro**		
	• For the fit model, can use this model to build our confidence interval • Factor Profiling -> Profiler, a simulator that can perform Monte Carlo Simulation • JMP Pro considers all the subjects		

	• R only considers three subjects now • The simulation model should be based on my predictive model • JMP has this workflow, eliminate the non-significant, then use for Profiler • JMP can also export Javascript but it only consider 3 years data (its not dynamic) • Need to consider dynamic • Should be explanatory -> prescriptive way, workflow is designed, neat in JMP Pro		