**AP Supervisor Meeting Minutes – 24 February 2016:**

|  |  |
| --- | --- |
| Date: | 24 February 2016 |
| Time: | 10:30 a.m. |
| Venue: | SMU SIS Meeting Room 4-1 |
| Attendees: | Denise Quek Si Ying |
|  | Tan Wei Song |
|  | Prof. Kam Tin Seong |
| Agenda: | 1. Discuss project progress |
|  |  |
|  |  |

1. **Review of meeting’s discussion**

|  |  |  |  |
| --- | --- | --- | --- |
| **S/N** | **Issues Discussed** | **Action By** | **Due Date** |
| 1. | **NDA requirements**   * Prof. Kam vetted through and confirmed that it is OK for him to sign the NDA for the purpose of our project. | All to note | - |
| 2. | **Recursive Partition**   * To evaluate our models, it is important to look at 2 things: Misclassification rate, and Confusion Matrix. Misclassification rate (e.g. model is able to achieve ~71% accuracy most of the time) indicates the global accuracy of the model, while the Confusion Matrix helps us to evaluate the portion that we want to focus on – achieving high true positive (wouldn’t mind having high false positive based on nature of industry/based on focus of analysis) * After which, look at the variables that have high log-worth, these are the predictor variables for our model. We need to see if they are significant or not for our model. * Compare our model split proportions with the original proportion of personal care customers and non-customers. * The true positive and true negative are quite close for both models using decision tree and bootstrap forest. * Note that R^2 is only useful for continuous data, hence it is not applicable for our model. * To view final results of the model: Save column > Save prediction formula > look at the probability value from the formula * Set the value ordering of the response variable, G1, by shifting 1 on top of 0, so that the resulting probability will evaluate the probability of customers over the probability of non-customers instead of the other way round. | All to note | - |
| 3. | **Data Cleaning:**   * For section D, it is OK to have missing values in some of our columns, since the decision tree will take care of those missing values. * For variable options that need to be combined to increase the value counts, there are 2 options to go about doing so:   + Make sure that any grouping is not going to affect the initial sequence of the distribution   + Group options that tend towards a higher proportion of “yes” result together, and vice versa. E.g. >50% yes, then group these options together | Wei Song | 26 February 2016 |
| 4. | **Further Research to be done**   * Explore the advantages and limitations of JMP for decision trees | Wei Song | 27 February 2016 |
| 5. | **Miscellaneous:**   * Come up with a flow diagram of our analysis process – so that we will know where we are doing revisions and where we are doing iterations * At the end of every model assessment, list down all the possible factors affecting our results, so that we are able to go back and modify/transform/prepare a new column, etc. and then subsequently recalibrate our model again * After each iteration, examine our results and compare with the results from previous iterations to see whether our model is improving or not. Also record down what we have observed through these changes. * Ideally, we would want to achieve a low misclassification rate and a low number of variables at the end of it | Denise | 27 February 2016 |

1. **Next meeting’s agenda**

|  |  |  |
| --- | --- | --- |
| **S/N** | **Issues Discussed** | **Action By** |
| 1. | Next supervisor meeting will be held next week. Date to be confirmed once prof’s schedule has been made available to us. By then, prepare for next week’s meeting based on the agenda stated below.  Agenda:   * Revised partition model * Discuss objective 3 | Denise, Wei Song |

The meeting was adjourned at 11:30pm. These minutes will be circulated and adopted if there are no amendments reported within the next three days.