http://www.unisim.edu.sg/about-unisim/Overview/PublishingImages/UniSIM_logo.gif

**ANL303e**

**Fundamentals of Data Mining**

**Assignment 2 – Group Based Assignment**

**January 2017 Presentation**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Li Xiaoling | B1571363 |
| Lim Hui Qin | Z1610579 |

Contents

[Assumption 3](#_Toc475921607)

[Question (a) 3](#_Toc475921608)

[Question (b) 4](#_Toc475921609)

[Question (c) 4](#_Toc475921610)

[Question (d) 5](#_Toc475921611)

[Question (e) 5](#_Toc475921612)

[Question (f) 5](#_Toc475921613)

[Question (g) 5](#_Toc475921614)

[Question (h) 6](#_Toc475921615)

[Question (i) 7](#_Toc475921616)

[Bibliography 10](#_Toc475921617)

# Assumption

1. It is assumed that customers’ demographics have no impact on the survey results.
2. It is assumed that the objective of the analysis is to determine what affects customers’ satisfaction with the food.
3. It is assumed that there are no errors in the data collection process and the survey results are complete and accurate.

# Question (a)

The Pre-modelling stage of the data mining process involves four steps.

**Step 1: Identification of business problem.**

The business problem could be to increase customers’ food satisfaction with Sam’s based on various contributing factors. This ensures that the pizza restaurant remains competitive despite of the aggressive marketing by restaurants in the newly opened shopping mall.

**Step 2: Translation of business problem into data mining application.**

The objective of the data mining application is to model the relationship between customers’ satisfaction score for each factors and their overall satisfaction with Sam’s so as to determine the areas of improvement for the pizza restaurant.

**Step 3: Assessment of data needed for the data mining application.**

The data required can be generated internally by conducting a customer satisfaction survey. The survey result can be consolidated into a datasheet for use in the analysis. Additionally, customers’ transaction records or demographic information can be leveraged on to enrich the model.

**Step 4: Preparation of data for data mining.**

The preparation of data requires the data collected to be complete and accurate. Any discrepancies or noise in the data would impact the model’s accuracy. If the data come from various sources, efforts would be needed to ensure that they are stored in a standard format.

# Question (b)

A potential business problem for Sam’s would be to increase customers’ food satisfaction based on various contributing factors. In view of the aggressive marketing by restaurants in the newly opened shopping mall, it is natural that Sam’s existing customers may flock to the competitors. To prevent natural attrition, the customer satisfaction survey conducted by Sam’s can be used to better understand the pizza restaurant’s standing from customers’ perspective. In general, customers who are more satisfied are more likely to maintain their patronage with the restaurant. By analyzing the satisfaction score for each contributing factors, the customers’ overall satisfaction can be increased and thereby, aiding Sam’s in customer retention.

A data mining solution to this problem would be to construct a model with the objective of analyzing the relationship between customers’ satisfaction score for each factors and their food satisfaction with Sam’s. This allows the pizza restaurant to better understand their customers’ point of view, what is good about the restaurant and what can be further improved on. The model result can be leveraged on to develop marketing strategy or increase operation efficiency of Sam’s.

# Question (c)

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Name** | **Measurement** | **Role** | **Justification** |
| CustID | Nominal | RecordID | Nominal is used as there are multiple distinct values. Attribute is an identifier, hence the role of RecordID. |
| Ambience | Ordinal | Input | Ordinal is used as there is an inherent order in the distinct values. Attribute is regarded as a factor to determine Food Satisfaction, hence the role of Input. |
| Service | Ordinal | Input | Ordinal is used as there is an inherent order in the distinct values. Attribute is regarded as a factor to determine Food Satisfaction, hence the role of Input. |
| Food | Ordinal | Target | Ordinal is used as there is an inherent order in the distinct values. Attribute is to be predicted based on satisfaction of other factors, hence the role of Target. |
| Parking | Ordinal | Input | Ordinal is used as there is an inherent order in the distinct values. Attribute is regarded as a factor to determine Food Satisfaction, hence the role of Input. |
| Price | Ordinal | Input | Ordinal is used as there is an inherent order in the distinct values. Attribute is regarded as a factor to determine Food Satisfaction, hence the role of Input. |

# Question (d)

**Mean and Standard Deviation (Std. Dev.), Min and Max**

These are used to determine if the data needs to be normalized. Normalization is necessary if the variables have different ranges.

**% Complete and Valid Records**

These are used to determine the quality of the dataset. A dataset with a high %Complete and Valid Records indicates data of good quality or a complete dataset.

**Outliers and Extremes**

These are used to explain for skewness observed in the variables. Dataset with many outliers and extremes may impact the model’s accuracy.

# Question (e)

CustID contains values from 1 to 500. If treated as an input, it may prompt for normalization to be carried out. However the large values do not represent an order or scale. Hence the model accuracy will be impacted as a result of misinterpreting the data.

# Question (f)

Text

# Question (g)

Text

# Question (h)

Text

# Question (i)

**Organization Background**

The organization discussed in this proposal is a financial institution with global footprints; it will be referred as SBM in this report due to privacy constraints. It has a large customer base of over a million clients and operates branches in various countries. This in turn generates a huge amount of data from the transactions and processes taking place in the day to day operations. The primary business of SBM is to provide financial services such as loans, securities, and futures and options to its clients. While the front office proactively engages the clients and brings in business for the firm, the back office provides a strong and comprehensive support to the firm.

The extensiveness of SBM’s business has become a challenge for the Credit Risk department to effectively mange the clientele portfolios. Situated in the back office, Credit Risk is responsible for setting credit limits, stress testing and reporting of clients’ risk profile and SBM’s credit health. To carry out its functions, Credit Risk need to transform the large amount of data associated with SBM’s clients and transaction into useful information where the team can perform analysis on. This is essential as the analysis will be depended on during decision making. To manage the data explosion, a data mining tool can be employed for the data analysis that the team needs to do. A data mining tool has strong analytical capabilities and is able to handle large amount of data. IBM SPSS Modeler, SAS Enterprise Miner and Statisca are examples of commercial data mining software available in the market.

**About IBM SPSS Modeler**

IBM SPSS Modeler is a powerful and versatile analytic workbench that allows user to build predictive models spontaneously without the use of programming language. It comes with a graphical interface for ease of visualization. The data mining process is represented as a workflow in the software; this promotes users’ interaction with the information where user can easily zoom into the details at any point of the model instead of spending time on programming task. It adopts the Cross Industry Standard Process for Data Mining (CRISP – DM), a structured and well proven methodology for solving business problems with data mining.

The pros of IBM SPSS Modeler lie in its excellent graphical user interface and its flexibility in navigation. Many have appraised the GUI as friendly and intuitive, making it easy for non-programmers or non-statistician to work on model development. Drop-down menus can be leveraged on for simple analysis while programmers can work with coding to create more advance and complex models. However it does not work well with large datasets. Freezing may occur when users load a large amount of data or complicated datasets into it and attempt to run the model.

**About SAS Enterprise Miner**

SAS Enterprise Miner is an advanced data mining tool that allows user to develop descriptive and predictive models in a streamlined manner. The SEMMA approach in its data mining process involves sampling, exploration, modification, modelling and assessment. This is well represented in the user interface where nodes are being used to build the model, allowing users to trace the diagram in a logical manner. The client server architecture deployable via a thin client web portal promotes the collaboration and sharing of data models especially amongst users in large organizations. Additionally, an algorithm is built in to perform scoring for all stages in the model development.

The pros of IBM SPSS Modeler lie in its ability to support complicated data mining processes and sharing of models and results. It has a broad set of tools and robust data crunching capability that aids in the model development. The innovative algorithms built in enhances the predictions’ accuracy where the results can be verified using the model assessment function. However to fully utilize and explore the functionalities that can be achieved, users need to be trained in SAS language. Some of the complex analysis may require advance configuration which can only be done by scripting.

**About Statistica**

Statisca has a comprehensive collection of analytical solutions for data mining applications. It offers a suite of software products that addresses the need for data mining across various industries. The user interface is friendly, involving selections from pull down menus and it is not necessary to use command prompts. Model results are often displayed as geometrical graphs with coordinates clearly labelled. It has an unique feature that allows user to experiment with different graphical representation for the same dataset. Industry standard scripting and COM-compliant programming languages are embedded in the software, making it highly customizable to meet the organization needs.

The pros of Statisca lie in its comprehensive graphical model results and the ease of customization. Detailed graphs are generated to display the model result. On top of the common graphs such as histograms and scatter plots, 3D graphs and matrix plots are widely used to visualize the model’s performance. For further analysis, users can use common programming language such as Java, VB and C# to develop advance models. However to effectively interpret the model’s result, users may be required to be statically trained. A certain degree of complexity is present in the graphical results which may not be easily understood by others.

**Recommendation**

Post the considerations and comparison of the existing data mining software, it is recommended that Credit Risk implement SAS Enterprise Miner in the team to assist with the intense analysis that is expected of the team.

Credit Risk deals with a huge amount of data in their daily functions; hence it is necessary that the recommended data mining tool is capable of managing large datasets. With regards to this point, IBM SPSS Modeler would be less suitable due to its limitations in data crunching. Users in the team are considered to be well versed in information technology and statistics, as it is in their nature to be able to work with software applications and use algorithms or logical formulas to perform risk scoring for the clients. As such, the constraint of SAS programming language can be overcome through conducting relevant training for the team. With a technical background, it would be relatively easy for users to pick up SAS programming.

Although IBM SPSS Modeler has an excellent user interface, the ability to perform complex analysis would hold greater importance. Data quality is crucial in developing an accurate model; hence it is an advantage if the data mining tool is strong in data preparation. While Statisca leverages on common programming languages for advance configurations which match the current skill set of users in Credit Risk, the graphical representations of model results may not be appealing to managers. Comprehensive graphical findings require audience to be statistically inclined, which may not sync well with managers. Basic graphs can be easily understood but complicated graphs such as matrix is prone to misinterpretation.

# Bibliography

*Comparison of data analysis packages: R, Matlab, SciPy, Excel, SAS, SPSS, Stata*. (2009, February 23). Retrieved from AI and Social Science – Brendan O'Connor: https://brenocon.com/blog/2009/02/comparison-of-data-analysis-packages-r-matlab-scipy-excel-sas-spss-stata/

*SAS Enterprise Miner*. (n.d.). Retrieved from SAS Products & Solutions: https://www.sas.com/en\_us/software/enterprise-miner.html

*SPSS Modeler*. (n.d.). Retrieved from IBM Software: http://www-03.ibm.com/software/products/en/spss-modeler

*Statisca Data Visualisation*. (n.d.). Retrieved from Statsoft Statisca: http://statistica.io/statsoft/statistica-data-visualization/