INFS 7203 / 4203 Data Mining Project Proposal

Prepared for: Dr Helen Huang

Andi Nuruljihad 44159069
Thanat Chokwijitkul 44522328
Yong Jin YEO 44063326
Xuanming Fu 43360774
Dongnan Nie 44028053

PROJECT DESCRIPTION

This project will be developed in JavaScript and R studio demonstrating in a form of a web application.

This project will focus on analysing the attributes of the speed dating dataset using data mining techniques. The main aim of using Clustering Analysis and C4.5 Algorithm is to classify and attempt to predict which class the new data belongs to based on the speed dating data.

Objective: The primary goal of this study is to predict whether a participant will find a match based on the answers they provided in the speed dating survey.

ALGORITHM

The C4.5 algorithm is an extension of the ID3 classification algorithm and will be used in this project to generate a decision tree by using a top-down greedy search through the given dataset to test each attribute at every tree node.[Ref]

The Decision Tree generated can then be used for classification of the data. C4.5 builds decision trees from a set of training data using the concept of information entropy.

An initial examination of the 169 questions in the speed dating experiment CFA has been introduced to test the EFA model on neighbourly complaint. The statistical analyses were conducted using AMOS version 24 via structure equation modelling.

The first step was to determine whether a single-factor structure remained the preferred model to best describe neighbourly dispute in current model (which also been termed as test for Common method bias). Thereupon, the method of common latent factor has been adopted to test the common method bias. This test required the comparison between a one-factor structured construct of observed items and a hypothesized multi-factor structure, which is the EFA model in previous phase. The one-factor CFA model forcing all 29 observed items onto a common latent factor did not fit the data well. The RMSEA, CFI, and SRMR were .16, .62, .12, respectively. Since the unitary model is not capable to describe the underlying structure, the priori estimated EFA model was retained as preferred model for neighbourly complaint data.

The next step was to model the relationship between latent factors through EFA model and observed items for original data. In order to verify the performance of mode, Root Mean Squared Error of Approximation (RMSEA), the Standardized Root Mean Squared Residual (SRMR), as well as the Comparative Fit Index (CFI) has been used as the indicator of model fit. In the case of CFI, values above .85 indicate an acceptable fit, RMSEA of .06 or less and for the SRMR values of .09 or less were desired.

At last, in order to achieve a best fitted CFA model, an exploratory step of examining modification indices would enable us to add covariance between observed items and enhance the performance of model. As a first step to identifying the underlying factor structure in the survey, an exploratory factor analysis (EFA) was manipulated on the 169 counted variables.

All analyses were conducted using R Studio. Dating data have been normality test to determine the extraction method for EFA. None of participants were identified as relatively normal distributed, based on both test of Kolmogorov-Smirnov and Shapiro-Wilk reached a p-value of 0. The null hypothesis for both test of normality is that the data are normally distributed. Thus, for the factor extraction method, the referent analysis was Principle Axis Factoring (PAF), as it is commonly implemented, is a recommended best practice approach when data are not multivariate normally distributed. In comparison, Maximum Likelihood Factor Analysis (MLFA) and Generalized Least Squares (GLS) were not appropriate for our non-normally distributed data. Data factorability was assessed against a Kaiser-Meyer-Olkin (KMO) value of 0.92 which indicate great factorability. Analyses were based on Pearson correlations, which are fairly robust for these data. Also, initial communalities were determined by squared multiple correlations. Since this study is grounded on the domain of social science, some correlation between factors are inevitable, so oblique factor method which expected that the factors would be correlated has been adopted. Furthermore, all suburbs with neighbourly complaint in its community has been involved in this study, promax rotation has been chosen for its superiority on dealing with large amount of data.

CFA has been introduced to test the EFA model and achieve dimensional reduction for dataset. The statistical analyses were conducted using AMOS version 24 via structure equation modelling.

The first step was to determine whether a single-factor structure remained the preferred model to best describe neighbourly dispute in current model (which also been termed as test for Common method bias). Thereupon, the method of common latent factor has been adopted to test the common method bias. This test required the comparison between a one-factor structured construct of observed items and a hypothesized multi-factor structure, which is the EFA model in previous phase. The next step was to model the relationship between latent factors through EFA model and observed items for original data. In order to verify the performance of mode, Root Mean Squared Error of Approximation (RMSEA), the Standardized Root Mean Squared Residual (SRMR), as well as the Comparative Fit Index (CFI) has been used as the indicator of model fit. In the case of CFI, values above .85 indicate an acceptable fit, RMSEA of .06 or less and for the SRMR values of .09 or less were desired.

At last, in order to achieve a best fitted CFA model, an exploratory step of examining modification indices would enable us to add covariance between observed items and enhance the performance of model.

DATA

Speed dating data set will be downloaded from kaggle.com. The data set is generated by conducting surveys and interviews with speed dating participants. A total of 8400 interviewee's answers to 160 questions were recorded in the data set.

REFERENCES

http://web.arch.usyd.edu.au/~wpeng/DecisionTree2.pdf