SCHOOL OF SCIENCE AND TECHNOLOGY

COURSEWORK FOR THE BSC (HONS) INFORMATION SYSTEMS (BUSINESS ANALYTICS); YEAR 2

ACADEMIC SESSION AUGUST 2020; SEMESTER 4,5,6

BIS 2216: DATA MINING and KNOWLEDGE DISCOVERY FUNDAMENTALS

DEADLINE: 22ND NOVEMBER 2020

Name : Ng Wei Xiang
Student ID : 18033167

INSTRUCTIONS TO CANDIDATES

IMPORTANT

The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 40%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Lecturer's Remark (Use additional sheet if required)
We (Names and IDs stated above) received the assignment and read the comments

……..... (Signature/date)

Academic Honesty Acknowledgement

"Me .....Ng Wei Xiang.....(student name). verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

……Xiang..... (Student's signature / Date)

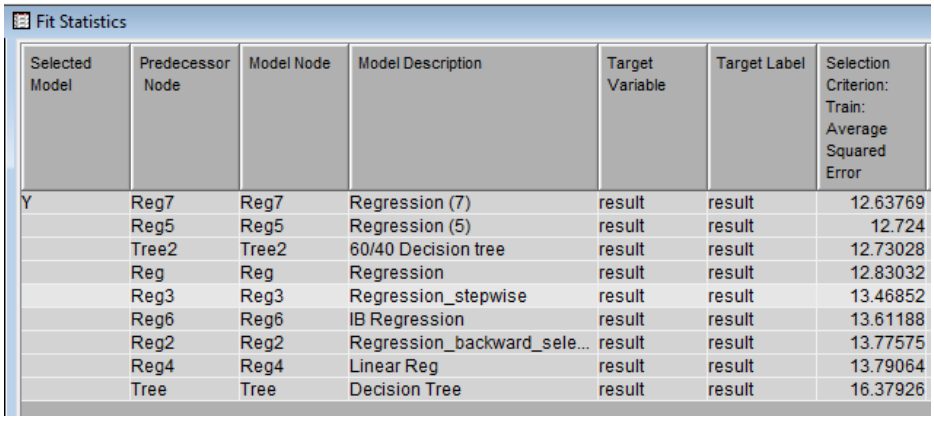
1. **Data preparation**

Attribute Processed	**Rationale for processing**	**Methods for processing**
Absences	Is a numeric datatype ranging from 0 to 93	Change to interval level
Age	Is a numeric datatype ranging from 15 to 22	Change to interval level
Dalc	Is a categorical variable with logical order from 0 to 4	Change to ordinal level
Fedu	Is a categorical variable with logical order from 0 to 4	Change to ordinal level
Fjob	Is a categorical variable with no logical order	Change to nominal level
Gender	Is a categorical variable with no logical order	Change to nominal level
Goout	Is a categorical variable with logical order from 1 to 5	Change to ordinal level
Health	Is a categorical variable with logical order from 1 to 5	Change to ordinal level
Internet	Is a binary datatype with yes or no only	Change to binary level
Medu	Is a categorical variable with logical order from 0 to 4	Change to ordinal level
Mjob	Is a categorical variable with no logical order	Change to nominal level
Result	Is a numeric datatype ranging from 0 to 20	Change to interval level
Studytime	Is a categorical variable with logical order from 1 to 5	Change to ordinal level
Traveltime	Is a categorical variable with logical order from 1 to 5	Change to ordinal level
walc	Is a categorical variable with logical order from 1 to 5	Change to ordinal level

2. **Models and performance** (T for training, V for validation)(Benchmark for model performance is Train Average Squared Error(ASE))

No.	**Modelling Technique**	**Partition Ratio**	**Other preparation Methods**	**Model performance**
1	Linear Regression	40T : 30V	Variable selection to exclude “fedu”, “fjob”, “medu”, “mjob”	13.79
2	Logistic Regression	40T : 30V	None	12.83
3	Logistic Regression	40T : 30V	Dropped “age” and performed backward selection	13.77
4	Decision Tree	40T : 30V	Dropped “age”	16.37
5	Logistic Regression	60T : 40V	Performed stepwise selection	13.46
6	Logistic Regression	60T : 40V	None	12.72
7	Decision Tree	60T : 40V	None	12.73
8	Logistic Regression	60T : 40V	Performed interactive binning	13.61
9	Logistic Regression	70T : 30V	None	12.63

3. **Screenshot of the Model Comparison Node's Fit Statistics result using SAS Miner.**



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Average Squared Error
Y	Reg7	Reg7	Regression (7)	result	result	12.63769
	Reg5	Reg5	Regression (5)	result	result	12.724
	Tree2	Tree2	60/40 Decision tree	result	result	12.73028
	Reg	Reg	Regression	result	result	12.83032
	Reg3	Reg3	Regression_stepwise	result	result	13.46852
	Reg6	Reg6	IB Regression	result	result	13.61188
	Reg2	Reg2	Regression_backward_sele...	result	result	13.77575
	Reg4	Reg4	Linear Reg	result	result	13.79064
	Tree	Tree	Decision Tree	result	result	16.37926

4. **Present the best model.**

The best model selected among all is the model "Regression(7)" which is selected based on the lowest value among all models at 12.63 based on the average squared error of the training set. The formula for this model would be $Y = 1.01(dalc=1)x_1 – 0.08(dalc=2)x_2 + 1.08(dalc=3)x_3 – 1.75(dalc=4)x_4 - 0.56(fjob=at\ home)x_5 – 0.37(fjob=health)x_6 + 0.17(fjob=other)x_7 – 0.52(fjob=services)x_8 - 0.13(goout=1)x_9 +1.2(goout=2)x_{10} + 0.45(goout=3)x_{11} - 0.34(goout=4)x_{12} + 1.02(health=1)x_{13} - 0.49(health=2)x_{14} - 0.44(health=3)x_{15} + 0.29(health = 4)x_{16} - 1.01(studytime=1)x_{17} - 0.32(studytime=2)x_{18} + 0.76(studytime=3)x_{19}.$

5. **Interpret the best mode selected.**

Based on te output of the model, it had a p-value of <0.001 which means it had significant variables affecting the model. The R-square value is at 0.1631 which means the model fits the variables at roughly 16%, which indicates it's a weak model. By evaluating the variables in this model individually, the variables that are below the 0.05 significant figure are : "dalc", "fjob", "goout", "health", and "studytime". That being said, the best model that I am able to generate for this given dataset had a conclusion that says : workday alcohol consumption, father's job, the total time of going out, health status, and time spent on studying are variables that may affect one's result.

6. Present the screenshot of the whole modelling process diagram in SAS Miner.