**Faculty of Technology – Course work Specification 2020/21**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Module name:** | | **Data Mining Techniques and Applications** | | | | | | |
| **Module code:** | | **IMAT 5238** | | | | | | |
| **Title of the Assignment:** | | **Framingham Data Set Assignment** | | | | | | |
| **This coursework item is:** (delete as appropriate) | | | | Summative | | |  | |
| **This summative coursework will be marked anonymously** | | | | | | Yes | | ~~No~~ |
| **The learning outcomes that are assessed by this coursework are:**   1. To discuss what is meant by the term data mining, to be able to express a business problem within a data mining framework and identify an appropriate target variable. 2. To be able to identify necessary prequisites for the application of a data mining framework and discuss the limitations of data mining methods with regard to data and methods in the context of the business problem. 3. To describe and analyse the organisational structure of large data sets to facilitate effective data mining and to be able to correctly interpret and critically evaluate the results to make informed decisions within a data mining framework 4. To be express the data mining framework for a particular business problem through the correct application of data mining tools. 5. To interpret the results produced from data mining tools and evaluates the effectiveness of data mining methods and where necessary making appropriate recommendations for use in the virtuous cycle of data mining. | | | | | | | | |
| This coursework is: **(delete as appropriate)** | | | **Individual** | | | | **~~Group~~** | |
| If other or a mixed ... explain here: | | | | | | | | |
| **This coursework constitutes** 70 % **to the overall module mark.** | | | | | | | | |
| **Date Set:** | **1/02/2021** | | | | | | | |
| **Date & Time Due:** | **21/05/2021 @ noon** | | | | | | | |
| **Your marked coursework and feedback will be available to you on:**  If for any reason this is not forthcoming by the due date your module leader will let you know why and when it can be expected. The Associate Professor Student Experience ([cemstudentexperience@dmu.ac.uk](mailto:cemstudentexperience@dmu.ac.uk) ) should be informed of any issues relating to the return of marked coursework and feedback.  Note that you should normally receive feedback on your coursework by **no later than 20 University working days after the formal hand-in date,** provided that you have met the submission deadline. | | | | | **19/06/2021** | | | |
| **When completed you are required to submit your coursework to:**   1. Turnitin | | | | | | | | |
| **Late submission of coursework** **policy:** Late submissions will be processed in accordance with current University regulations which state:  *“the time period during which a student may submit a piece of work late without authorisation and have the work capped at 40% [50% at PG level] if passed is* ***14 calendar days****. Work submitted unauthorised more than 14 calendar days after the original submission date will receive a mark of 0%. These regulations apply to a student’s first attempt at coursework. Work submitted late without authorisation which constitutes reassessment of a previously failed piece of coursework will always receive a mark of 0%.”* | | | | | | | | |
| **Academic Offences and Bad Academic Practices:**  **These include plagiarism, cheating, collusion, copying work and reuse of your own work, poor referencing or the passing off of somebody else's ideas as your own. If you are in any doubt about what constitutes an academic offence or bad academic practice you must check with your tutor. Further information and details of how DSU can support you, if needed, is available at:**  <http://www.dmu.ac.uk/dmu-students/the-student-gateway/academic-support-office/academic-offences.aspx> and  <http://www.dmu.ac.uk/dmu-students/the-student-gateway/academic-support-office/bad-academic-practice.aspx> | | | | | | | | |
| **Tasks to be undertaken: Data Mining Case Study See Assignment Brief** | | | | | | | | |
| **Deliverables to be submitted for assessment: Word processed report 2000 words max** | | | | | | | | |
| **How the work will be marked: See Marking Scheme** | | | | | | | | |
| **Module leader/tutor name:** | | **Anthony Williams** | | | | | | |
| **Contact details:** | | **anthony.williams@dmu.ac.uk** | | | | | | |

MSc Business Intelligence & Data Mining /MSc Intelligent Systems:

**IMAT5238 Data Mining Assignment**

*Anthony Williams*

*GH6.58, School of Computer Science and Informatics*

e-mail: [anthony.williams@dmu.ac.uk](mailto:anthony.williams@dmu.ac.uk)

**Task**

The assignment is an individual assessment, requiring you to analyse the FRAMINGHAM heart study dataset within SAS Enterprise Miner, using the data mining techniques covered in the IMAT5238 module, and detailing your results, interpretations, conclusions and recommendations in a well-structured technical report.

* A sample of data from the FRAMINGHAM dataset is shown in Appendix A.
* The FRAMINGHAM dataset contains 5,209 observations and 13 variables. The variables in the dataset are shown in Appendix B.
* The coursework will be assessed according to the marking grid in Appendix C.

***Each of you will be working on your own random sample of data generated by typically inserting the last 5 figures of your DMU student id number into the random seed generator within the Data Partition node. (Note: If spurious output for any of the models should occur, insert the last 4 figures (or the last 3 figures) of your DMU student id number into the random seed generator – to enable you to generate sensible output that you can interpret.)***

**Your practical tasks are outlined below:**

* Conduct a thorough exploratory data analysis including investigation of outliers;
* Develop Regression, Decision Tree and Neural Network models, using data replacement, data transformation and variable selection methods, where appropriate;
* If appropriate you may choose to add one of the nodes from the research activity published in the general discussion board (DB Activity 3).
* Use your data mining skills and appropriate lift and diagnostic charts to ensure you get the best model from the above directed data mining techniques;
* Detail your results, interpretation, conclusions, recommendations and evaluation (including a critical appraisal of the data analyses conducted) of the first flow of the data mining cycle in a report. Assume the report is for the Chief Executive Officer (CEO) of The National Health Service.

**Deliverable**

You are required to submit a well-structured, detailed, technical report of 20 pages maximum (not including appendices) that gives details of:

* Description of the business problem and appropriate data mining problem and use of an appropriate data mining framework
* the exploratory data analysis;
* the technical interpretation of the models considered;
* a comprehensive assessment of the models;
* a full justiﬁcation of the choice of ﬁnal directed data mining model;
* if you have added a node from the research activity provide a justification for its use what you anticipate to find you should also assess it performance and limitations. You may wish to include one of the presentations in the appendix for reference, cite the original author.
* conclusions and recommendations resulting from your work and a critique of the analyses conducted.
* In the Appendix you must provide a workflow diagram of your process, and a short reflection identifying any lessons learned in doing the assignment, indicate any difficulties and how you overcame them.

**Assessment**

This assignment contributes **70% of your ﬁnal module grade.** You will need to submit a copy of your report using the TurnitinUK link in the assessments section of the Data Mining module shell on Blackboard (to be made available prior to the coursework deadline).

It is not necessary to submit a hardcopy to the CEM advice centre, however you might find it useful to print a hardcopy for proof reading of your work. Note research has demonstrated that humans are more critical with paper materials than electronic screens.

Note **The filename of your report should consist of your p-number and your name.**

**Assessment deadline: Week 33 Fri May 21th 2021 at noon**

**This coursework adheres to the Faculty of CEM Late Work Guidelines.**

**Background information: The FRAMINGHAM dataset**

Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study - under the direction of the National Heart Institute embarked on an ambitious project. At that time, little was known about the general causes of heart disease, but the death rates for CVD had been increasing steadily since the beginning of the century and had become an American epidemic.

The objective of the Framingham Heart Study was to identify the risk factors that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed symptoms of CVD or had not suffered a heart attack or stroke. The researchers recruited 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts, and began the first round of extensive physical examinations and lifestyle interviews that they would later analyse for common patterns related to CVD development.

Since 1948, the subjects have continued to return to the study every two years for a detailed medical history, physical examination, and laboratory tests, and in 1971, the Study enrolled a second generation of the original participants' adult children and their spouses - to participate in similar examinations. Then in 2002 the study entered a new phase, the enrolment of a third generation of participants, the grandchildren of the original cohort. Today, the Framingham Heart Study remains a world-class epicentre for cutting-edge heart research and has been acclaimed as some of the most important to medicine. The concept of CVD risk factors has led to the development of effective treatment and preventive strategies in clinical practice.

You are a data miner and have been commissioned by The National Health Service to analyse the FRAMINGHAM dataset and to submit a technical report detailing your findings, which should include the best model that should be used to identify the risk factors that contribute to Cardiovascular Disease (CVD) and the characteristics of the participants after cluster analysis.

Note The analysis being conducted represents the first flow of the virtuous cycle of data mining.

**Notes**

**Report Guidance:**

Your report should be no longer than 20 A4 sides and minimum font size 12. Marks are awarded for technical correctness, descriptions of models, appropriate justification of node and parameter choices, appropriate actions to guard against overfitting, indications of model robustness, model limitations, data insights, analysis supported by appropriate charts.

Reports are expected to be written to a professional standard, clear concise, good use of English. Text supported by relevant choice of diagrams and use of tables to summarise data, avoidance of repetition and redundancy, appropriate use of appendices, table of contents and use of page numbers, table numbering and figure numbering, presence of an informative abstract or executive summary.

All diagrams must be legible and appropriately labelled if short of space use appendices. If you use coloured diagrams to illustrate or contrast points then a hardcopy of your report must be printed in colour. Assume that the audience of the report is senior management with no knowledge of the technical details of data mining.

**Time Management Guidance:**

You should budget half your time on using the SAS enterprise miner software to generate informative models, explore nodes and appropriate non-default options. The other half of your time should be budgeted on production of a clear, well structure report which describes your work and addresses the assignment brief.

The SAS software is a professional data mining product full of features and options. You are free to explore these options, however if you use something not covered in the course you must justify its use to receive credit. You will be penalized for inappropriate use of features you cannot adequately justify or explain. There will come a point of diminishing returns where no matter how much effort you put into the software you cannot improve on model performance. This will be the point you should focus on writing the report and analyse your results

You should start work on the assignment immediately and make allowance for any difficulties you might face. Leaving the work till the last minute will result in poor quality report. You should **not underestimate** the time it takes to become fluent in the use of the software. If you have followed all the labs attentively and with understanding you should not face too many hurdles. The completion of the assignment represents a cap-stone moment which will integrate everything you have learnt on the course.

*A suggestion plan is that after completing a weeks lab activities on the HMEQ data set you apply what you have learnt to the Framingham data set. That way you do not leave it all the to the end. This would be Tim Urban’s approach not a good idea… you know how that end up for him.*

*Note you will have three weeks over the Easter vacation weeks 28-30) where you should be able to do the bulk of your work and you will have the knowledge to complete the later tasks. It is suggested that you split the work into 7 tasks.*

1. *Description of the business problem and appropriate data mining problem and use of an appropriate data mining framework the exploratory data analysis;*
2. *Develop Regression Model*
3. *Develop Decision tree*
4. *Develop Neural network*
5. *Research node Discussion board activity 3*
6. *Model comparison and recommendations*
7. *Proof reading and editing*

*Use the report template !*