ASSESSMENT 3: CAPSTONE PROJECT

Overview

Name: Capstone Project

Type: Written report and SAS Visual Analytics workbook submission

Due: 11.59 PM AEST Wednesday of Week 7

Weight: 40%

This capstone project assessment involves writing a report that summarises a data science related investigation that you have selected from one of the datasets listed in this document and available in the SAS Visual Analytics (VA) learning environment on Teradata University Network (TUN). This is to demonstrate that you have grasped the important concepts and techniques associated with statistical modelling, model selection and model evaluation using a coherent written analysis of a clearly stated hypothesis.

The investigation must involve:

1. An initial data exploration and explanation of the SAS datasets listed in this document. These are available online in the Teradata University Network and have available metadata in the SAS Data Dictionaries https://www.teradatauniversitynetwork.com/Software/Online/SAS/Student/SAS-Visual-Analytics-Data-Dictionaries/ . Please note that any dataset used in the previous weeks for tutorials or assignments are not to be used in this assignment (e.g., BIGPVA, PVA_DATA, INSIGHTTOY_SALES, CARS, HEART and BIGORGANICS).
2. A hypothesis that you wish to test based on your initial data selection from point 1 above and the modelling techniques to be deployed.
3. Modelling and model comparison for optimisation with **SAS Visual Analytics** completed
4. A final written submission in **Word** and evidence of the steps taken using **SAS Visual Analytics** to reach your conclusions (e.g., model visualisations, summary stats tables and so on) and insights generated.

The written summary should cover the pre-processing and exploratory steps carried out, the statistical models evaluated and finally the comparisons you have made. These are independent choices and decisions that require full explanation of your analysis process and the hypothesis that you have tested. Please include relevant screenshots of all the charts or model diagnostics that you created and any calculated fields, parameters used during model selection and evaluation.

The report should not exceed 10 pages. It should meet the graduate standard of academic writing (structure, reference, grammar and so on) and include the following:

- Effective use of hypothesis, analytical approach and a conclusion
- Relevant screenshots showing the visualisations used in SAS Visual Analytic

While marks are not awarded for neatness, students may be penalised for poorly written or extremely untidy work.

Report format

The **main body** of the report (containing title, abstract, introduction, data, methods, results and discussion, and conclusions) must be presented in 12pt Roman style font on no more than **five (5) A4 pages**, using single line spacing. Either a single column or double column format may be used.

References and appendices can be listed on no more than **five (5) additional A4 pages**.

In total, the report must not exceed **ten (10) A4 pages**.

WARNING: Only the main body and the references will be formally assessed and graded.

Further details about the report structure are provided in the following section.

Report structure

The capstone project report should include the following sections marked clearly:

- **Title:** In today's busy world, it is very important to make the most of your title. Make the title eye-catching, informative and succinct, it should be aligned with your aim and an accurate representation of the contents of the report.
- **Abstract:** The abstract provides a short sharp overview of the contents in the report and will be around 200 – 300 words. The abstract has five (5) parts:
    i. Introductory statement: Background to the study, important issues addressed in the report (approximately 1-2 sentences)
    ii. Purpose: State the objectives of the report (1-2 sentences)
    iii. Methodological approach: Overview of the data and methods (2-3 sentences)
    iv. Findings or achievements: List one or two of the main findings or achievements from your investigation (1-2 sentences)
    v. Conclusions and implications: What conclusions can be drawn from your investigation? How can the findings and/or achievements listed in your report deliver a benefit to people, things, systems or processes? (1-2 sentences)
- **Introduction:** The introduction sets the scene for the investigative efforts. It provides motivation for the work and relevant background information and references that will enable the reader to put into context the key objectives and achievements in your report. Address the important issues that have motivated your investigation. At the end of the introduction clearly state the objectives of the report that state your null and alternative hypotheses.

  Note: Do not put any results from your investigation in the introduction. Do not discuss details about the data and methods in this section. Do not discuss your conclusions or key findings in the introduction.

- **Data:** This section should provide details about the data and what the data represents. You should include information such as:
    i. Identify the dataset selected from the *Assessment datasets* section and the data you derived from the data
    ii. Details of the sample size, number of variables and types of variables
    iii. Describe how you quantified the reliability of the data
    iv. Any known interventions or pre-processing that were necessary to improve the data or derive further meaning from the data
    v. Any other information that is relevant to the understanding and assessment of your work and report.
- **Methods:** This section should summarise the data science methods, including the evaluation framework, operational variables used, any custom categories used, any custom heirachies used and implementation that was used to process and to statistically analyse the data using SAS Visual Analytics. The methods should be appropriate to ensure that the stated objectives are met. It is important to provide a sufficient level of details so that your methodology could be repeated by an independent person, while also being clearly and objectively presented so that it can be understood. Describe your proposed model that will answer your hypothesis and any proposed features that will be used for model comparison, tuning or parameter selection.
- **Results and discussion:** This section presents and discusses the results. The discussion centres on the outputs from the pre-processing and evaluation framework implemented (modelling and visualisations). You should include information such as:

i. Summarise results, model reliability and comparison to any benchmark analyses or models. Provide visualisations or statistical outputs as screenshots from SAS Visual Analytics to support your written summary.
ii. Critically analyse results, for example, limitations of data, setup or approach, characteristic errors, subtractive feature analysis and possible improvements
iii. Discuss what are the main outcomes, why are they useful and what for, how are they interesting and why
iv. Discuss how the results support or reject your initial hypothesis stated in the introduction? What are the main achievements and their implications?
v. Conclude with what you have learned from this study and what would you recommend for further follow up analysis? Provide reasons.

- **Conclusions:** This section includes final remarks about the key achievements of the investigations and what makes them interesting or useful – right now or for future work. Achievements or findings should be contrasted with the original objectives or hypotheses of the project. Make sure that you mention any limitations of your work here. Limit the conclusions to no more than two or three paragraphs.
- **References:** List the sources your investigation has drawn from. You should use at least four (4) references. Note that all references should be referred to in the text.
- **Appendices (optional):** Add any supporting materials that might be useful to help assess your work.

Important notes

1. The **entire project** must be accomplished using **SAS Visual Analytics**. The report can be written using a text editor of your choice (e.g., Microsoft Word or similar).
2. You must submit your SAS Visual Analytics workbook as a PDF appended to the report. Refusal to comply with this requirement may incur in your work being considered as not delivered.

A word on plagiarism and self-plagiarism:

Plagiarism is the act of using another's words, works or ideas from any source as one's own. Plagiarism has no place in a University. Student work containing plagiarised material will be subject to formal university processes. In case significant portions of your own previous work (e.g. a report for a related subject you did in this or any other university) is recycled in a way that it could be fully or partially graded twice ('double-dipping'), this is considered **self-plagiarism** and will not be tolerated.

Assessment submission

Your submission for the capstone project assessment should be uploaded to LearnJCU as a single ZIP file, containing the following three (3) files:

1. **Written report:** A single written document (DOCX or similar format)
2. **SAS Visual Analytics workbook:** A single workbook exported from SAS VA as a workbook (PDF format)
3. **Task cover sheet:** Completed.

Be sure to use the file naming convention, including your first name and last name, for all individual file names and the ZIP file name. Upload all submission files for this assessment in one ZIP file. You can upload as many times as you want, but only the last submission is graded.

Assessment instructions

Return to SAS Visual Analytics on Teradata University Network and select your dataset from the data dictionaries. If you have any problems with your project then follow these steps

1. Sign in to Teradata University Network and select Software > Cloud > SAS Visual Analytics.
2. Scroll down the page to find the link for SAS Data Dictionaries and click on the link to open the metadata for the available datasets
3. Select a dataset for analysis and then open the dataset in SAS Visual Analytics.

Assessment datasets

For your capstone project assignment, you will need to pick **one** of the following datasets from the SAS Data Dictionaries .

ACME Bank June 2013 Data Dictionary for ACME Bank June 2013
Autism Adult Data Dictionary for Autism Adult
Bank Direct Marketing Data Dictionary for Bank Direct Marketing
BirdStrikes Data Dictionary for BirdStrikes
Hollywood Movie Dataset Data Dictionary for Hollywood Movie Dataset
Ozone Data Dictionary for Ozone
PATSAT Data Dictionary for PATSAT
Post Operative Discharge Data Dictionary for Post Operative Discharge
Primary Billing Cirrhosis Data Dictionary for Primary Billing Cirrhosis
Readmissions Scored Data Dictionary for Readmissions Scored
Readmit Historical Data Dictionary for Readmit Historical
Retail Demo 2 Data Dictionary for Retail Demo 2
RUGBY Data Dictionary for RUGBY
Support Data Dictionary for Support
Telco 03 Data Dictionary for Telco 03
Very Low Birth Weight Data Dictionary for Very Low Birth Weight
Workshop Educ Final Data Dictionary for Workshop Educ

JAMES COOK UNIVERSITY

Faculty of Science Engineering and Information Technology

School of Maths, Physics and IT

Marking scheme

Please adhere to the strict formatting requirements. The report will not be assessed if it is not formatted appropriately. Total marks possible 120.
The overall mark will be based on clarity of question, method (approach and implementation), discussion of results and quality of analysis which are to be structured according to the sections defined in the report structure.

Dimension	**Sophisticated [100% marks]**	**Competent [50% marks]**	**Needs Work [0% marks]**
Title [2 marks]	The title is a concise (less than 20 words) and accurate reflection of the contents of the report. Author is listed below the title.	The title is concise (less than 20 words) and moderately reflects the contents of the report. Author is listed.	The title is not informative or exceeds the word length or Author not listed.
Abstract [6 marks]	Clearly addresses the five parts of the abstract so that the reader has a clear overview of the reports.	Partially addresses the five parts of the abstract and or addresses all five parts but the writing is not clear in places.	Unclear, does not overview the report, or the writing is poor overall and mostly unclear
Introduction [16 marks]	Position and exceptions, if any, are clearly stated. Organisation of the argument is completely and clearly outlined and implemented.	Position is clearly stated. Organisation of argument is clear in parts or only partially described and mostly implemented.	Position is vague. Organisation of argument is missing, vague, or not consistently maintained.

Data [20 marks]	Data are suitable, the report explains how the data were obtained, and all of the following information items (whenever applicable) are clearly explained: i. What the source of the data is. ii. The sample size, the number of variables and types of variables. iii. How the reliability of the data was quantified. iv. Any known interventions or pre-processing that precede the ones described in the report. v. Any other information that is relevant to the understanding and assessment of the work/report.	Data are suitable, the report explains how the data were obtained, and most of the applicable data information items are addressed and reasonably explained. i. What the source of the data is. ii. The sample size, the number of variables and types of variables. iii. How the reliability of the data was quantified. iv. Any known interventions or pre-processing that precede the ones described in the report. v. Any other information that is relevant to the understanding and assessment of the work/report.	Little information/explanation about the data is provided and/or the grammar structure is difficult to follow and/or the data do not meet the minimum requirements. i. What the source of the data is. ii. The sample size, the number of variables and types of variables. iii. How the reliability of the data was quantified. iv. Any known interventions or pre-processing that precede the ones described in the report. v. Any other information that is relevant to the understanding and assessment of the work/report.

Methods [28 marks]	Lists all the steps in the order in which they were performed to pre-process and/or explore the data. These steps, if executed appropriately and interpreted appropriately, will ensure that the objectives of the report are clearly met. At least 6 of the following targeted key topics from the subject have been explored and explained in depth: 1. Data representation 2. Exploratory visualisation using SAS VA 3. Unstructured to Structured data 4. Data cleaning 5. Type conversion 6. Missing value imputation (informative missingness) 7. Assessment criteria 8. Data subset selection, grouping and/or subsampling 9. Group-based data summarisation 10. Variable selection and/or transformation 11. Modelling and comparison in SAS VA	Most of the steps are listed and explained, but some details are vague or questionable. At least 4 of the targeted key topics from the subject have been reasonably explored and explained. 1. Data representation 2. Exploratory visualisation using SAS VA 3. Unstructured to Structured data 4. Data cleaning 5. Type conversion 6. Missing value imputation (informative missingness) 7. Assessment criteria 8. Data subset selection, grouping and/or subsampling 9. Group-based data summarisation 10. Variable selection and/or transformation 11. Modelling and comparison in SAS VA	The methods clearly will not allow the objectives of the report to be met and/or the details of methodological steps and procedures are very difficult to follow and/or the listed key topics from the subject have been poorly or not appropriately explored. 1. Data representation 2. Exploratory visualisation using SAS VA 3. Unstructured to Structured data 4. Data cleaning 5. Type conversion 6. Missing value imputation (informative missingness) 7. Assessment criteria 8. Data subset selection, grouping and/or subsampling 9. Group-based data summarisation 10. Variable selection and/or transformation 11. Modelling and comparison in SAS VA

Results and discussion [22 marks]	The results and discussion are explained correctly, clearly, and in sufficient detail. The results and discussion clearly follow from the data collection and the methods.	The results and discussion are explained correctly, clearly and in sufficient detail most of the time. There exists a connection of some type between the results and discussion, and the data collection and methods.	One or more of the items are are not explained correctly, unclearly and lack detail most of the time. There is no connection between the results and discussion, and the data collection and methods.
Conclusion [10 marks]	The original objectives and hypotheses are restated and contrasted against the obtained achievements and/or findings. The conclusion summarises and draws a clear, effective conclusion of the investigation and enhances the impact of the report – e.g., it provides a recommendation or action that should be undertaken in the future. It may also highlight unavoidable limitations of the investigation.	Conclusion is clearly stated and connections to the original objectives and hypotheses are mostly clear.	Conclusion may not be clear and/or the connections to the work reported are incorrect or unclear or just a repetition of the findings without a suitable summarisation and interpretation and/or the underlying logic has major flaws.

Writing [16 marks]	Report is coherently organised and the logic is easy to follow. There are no spelling or grammatical errors and terminology is clearly defined. Writing is clear, concise and persuasive. Each figure/table will be numbered, followed by a caption, and referred to in the body of the text, most noticeably in the results and/or discussion section. The figures/tables provided reinforce the most relevant achievements of the work. All references have been listed and referred to in the appropriate places in the body of the text and listed at the end of the report. At least 4 references have been provided.	Report is generally well organised and most of the argument is easy to follow. There are only a few minor spelling or grammatical errors, or terms are not clearly defined. Writing is mostly clear but may lack conciseness. Each figure/table will be numbered, followed by a caption, and referred to in the body of the text, most noticeably in the results and/or discussion section. Most references have been listed and referred to in the appropriate places in the body of the text and listed at the end of the report. At least 4 references have been provided.	Report is poorly organised and difficult to read – does not flow logically from one part to another. There are several spelling and/or grammatical errors, technical terms may not be defined or are poorly defined. Writing lacks clarity and conciseness. Figures/tables and/or references are sloppy or missing.