COM0152M

Department of Computer Science

Data Mining and Text Analysis

SUMMATIVE ASSESSMENT BRIEF



|  |  |
| --- | --- |
| **Author** | Dr Phoebe Barraclough (checked by Dr Joseph Akinyemi) |
| **Assessment type** | Summative assignment |
| **Weighting** | 100% |
| **Release** | Week 3 |
| **Deadline** | Monday following Week 8, 13:00 (UK time) \* |

\* If this date falls on a UK public holiday or a University of York closure day, the submission date will change. Please check the submission point in the ‘Assignments’ area of the module in Canvas for the exact submission deadline.

# Module Learning Outcomes

The module learning outcomes for this module are as follows:

**MLO 1.** Analyse different data mining and text processing tasks and the algorithms most appropriate for addressing them.

**MLO 2.** Critically evaluate and select the appropriate open-source or commercial data mining and text processing toolkits and implement the algorithms.

**MLO 3.** Critically evaluate the algorithms with respect to the accuracy of their results.

**MLO 4.** Develop and communicate a data mining and text processing solution to a real-world problem.

**MLO 5.** Identify and discuss the challenging research issues in the area of data mining and text processing.

This assessment addresses **all** the module learning outcomes listed above.

# Assessment Background/Scenario

The manager of a multinational bank is concerned about the future of the business and would like to identify the rate at which customers are leaving (also referred to as exiting) despite the services offered by the bank. They also wish to explore the factors that influence an individual's eligibility for a credit card. You have been assigned to investigate customers’ exit patterns, the factors influencing a person’s eligibility for a credit card and profile different customer groups.

Customer leaving, also known as customer attrition, exit or churn, refers to the loss of a customer (i.e. to a rival business), or a period of inactivity, meaning the customer is no longer making transactions. For example, some customers will eventually discontinue their usage, either because they are unhappy with their user experience, or they can no longer afford to keep using the service. For brevity, we will henceforth use the term exit, exiting as appropriate.

The manager has highlighted five problem areas they would like to identify and test potential solutions for, using the data mining and/or text analysis techniques that you have learned about in this module.

You are required to use the data set provided by the manager to explore and present possible solutions for any **two** of the following problem areas:

1. **Profiles**: Examine the data and identify two distinct profiles (differing sets of personal attributes; there may be some overlap, and each profile should contain more than one attribute) that are linked to high levels of attrition actions. You will need to define and clearly state what you have identified as ‘high level’ as part of your assumptions for this problem.
2. **Credit Card Eligibility:** Scrutinise the data and identify two distinct profiles that are linked to the factors that influence a person’s eligibility for a credit card. You will need to define and clearly state what you have identified as ‘factors’ as part of your assumptions for this problem.
3. **Age:** Determine whether ‘age’ is a good predictor of the likelihood of credit card eligibility and clearly demonstrate this against two of the other attributes in the data set. Make sure you clearly state which other attributes you selected as comparators and justify why you selected them for the role.
4. **Churning Scrutiny**: The bank is looking to provide an extra level of security for high-value targets. Ascertain if there is a strong relationship between the attribute “balance” and the “tenure”. Consider what other attributes (within the data set) could be included to help protect high-value targets.
5. **Recommendation:** Consider how you could use the data to promote better services to the bank’s **exiting** customers. As part of this problem, also consider how this information could be best communicated to **exiting** customers visually and put forward or demonstrate one option. You may need to consider the data sparsity problem here and research possible solutions.

The bank has provided you with a simulated sample set of data as one CSV file. How you use this data will depend on the approach you choose to take.

This is a hybrid data sample given under a Usability 10.00 License [CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/) and available from the Kaggle data repository at the time of writing.

# Assessment Tasks

Given the scenario above, your task is to write a report detailing possible solutions to your **two** selected options. You should clearly draw on the current literature and can use examples from your work throughout the module, including your formative assessment as supporting evidence for your approach. There is **NO** requirement for you to cite yourself where you have reused work from your formative. However, please be aware that the context for this assessment is different, and you may find that success in one does not necessarily translate to success in the other.

Your report should provide an initial executive summary (**see the list of three tasks below**). Your response in one section will **not** contribute to grades in another, so you should consider this assessment. Further formatting details are given below.

The maximum word count for the three tasks combined is 3,000 words, and your executive summary should be a maximum of 400 words.

## Executive summary (MLOs 1, 2, 3, 4 and 5, 10%)

(maximum 400 words, and no less than 390)

Write an overview/summary of the report which meets the word count restrictions above and contains at least the following points:

* Which options you have chosen to present solutions for;
* What was achieved/undertaken;
* What processes were applied;
* What the results demonstrated;
* What should be reconsidered in the future.

## Task 1: Discussion of techniques used in your two solutions (MLOs 1, 3 and 4, 40%)

(Suggested word count for this section: 1300, i.e. 650 words per solution)

Given the scenario above, design and discuss the potential solution(s) to the problem(s) you have selected. You **will** need to write small programs and/or use tools to run simulations as supporting evidence on the given test data. In this task you should make it clear **which** problem you are presenting a solution for.

Your report should clearly cover the following:

* Any assumptions you are making about the scenario or selected problem;
* Any pre-processing you would undertake to make the data fit for purpose;
* Which data mining/text analysis techniques you have employed in your solutions;
* Justification for the selection of those techniques, given the nature of the data and the requirements of the problem you are attempting to solve;
* An evaluation of the techniques you have applied in terms of the accuracy of their results. You will need to clearly define and state the measures/methods by which you are evaluating the techniques. It is perfectly acceptable for your techniques to have been unsuccessful. Whether successful or not, it should be clear how your evaluation has informed your conclusion.

All code examples and results (output) should be presented in the appendices as screenshots only, **not** as handwritten (typed) code. All supporting evidence in the appendices **must** be referred to and discussed in the body of the report. To attain a pass, you need to present evidence of your prototype programs and tests to pass this assessment.To attain a higher grade, your discussion should be supported by reference to relevant literature in this section.

## Task 2: Evaluation of the tools/languages (MLO2, 20%)

(Suggested word count for this section: 400 words)

Given the languages/tools you have selected and used, provide a critical evaluation of their effectiveness in the context of the given scenario. You should clearly make comparisons to other options available and draw on the specific requirements of the scenario when presenting your argument. To attain a higher grade, your discussion should be supported by reference to relevant literature in this section.

Consider the question: If you undertook this assignment a second time, would you use the same languages/tools, and why?

## Task 3: Discussion of the current literature (MLO5, 30%)

(Suggested word count for this section: 900 words)

Given the scenario above and the nature of the problems you have selected, research and identify the main areas of investigation the research community is currently tackling. Consider the following questions:

* What are the current ‘problem’ areas?
* What solutions have been put forward and how are they being evaluated?
* Given your experience, would you consider these potentially successful solutions?
* Justify why you consider them successful or not.

Present a discussion around these questions and consider how current research could potentially change or improve your solutions to the given scenario. To attain **a pass,** your discussion **must** be supported by reference to relevant literature in this section.

# Deliverables

Your assignment should be laid out following the formatting guidelines that are specified in the ‘Submission Formatting’ page in Canvas. This includes restrictions on the length of the appendices, expectations on how your work should be presented and any penalties when these expectations are not met.

Your report should not exceed 3,000 words in total and consist of three clear sections – one for each task. Your response to one section/task will not contribute to grades in another. Further formatting details and essential points are given below.

You are to produce and submit a **REPORT** that presents your response to the three tasks for **two** of the problems presented in the scenario. You are **NOT** required to submit your code files or adapted data sets as separate files.

You may choose to redistribute the given indicative word counts between the three tasks as you see fit, providing your total response to these does not exceed 3,000 words. However, the **executive summary must be 400 words and you cannot redistribute that word count to support other sections.**

## Referencing

You are required to use the [IEEE referencing style](https://subjectguides.york.ac.uk/referencing-style-guides/ieee) for citing books, articles, and all other sources (such as websites) used in your assignment.

Good referencing is essential in order to meet the standards of academic integrity set by the University. All your sources must be acknowledged, regardless of whether you have included direct quotes or not. Visit your **Academic Integrity Tutorial** module in Canvas for additional guidance on effective referencing.

# Marking Criteria

The following table identifies the marking criteria and the marks available for each, given the tasks you’ve been set.

Please note that academic best practice is being evaluated in Task 3 directly. While it is not being evaluated directly in the rest of this assessment, poor academic practice can, and will, affect the quality of your communications and justifications, and the attainment of higher grades at Master’s level.

|  |  |  |  |
| --- | --- | --- | --- |
| **Learning Outcome** | **Section/Task** |  | **Available marks** |
| **All** | **Executive summary**  All the indicated areas have been covered and the summary provides an effective overview of the report and its outcome. | | **10** |
|  | **Task 1: Design and Discussion of Solution**  The solutions to the two selected problems have been presented effectively, supported by relevant evidence of prototype programs and/or tests using appropriate applications. Where appropriate, decisions and justifications are further supported by the literature. | | |
| **1** | * **Pre-processing techniques**   Appropriate and effective pre-processing techniques have been considered for both solutions. | | **10** |
| **4** | * **Selected data mining/text analysis techniques**   Appropriate and effective data mining/text analysis techniques have been selected for both solutions. | | **10** |
| * **Justification**   The selected data mining/text analysis techniques have been effectively justified in the context of the scenario and the selected problems. | | **10** |
| **3** | * **Evaluation**   The effectiveness of each solution has been evaluated using appropriate measures/methods which support the presented outcome/conclusion. | | **10** |
| **2** | **Task 2: Evaluation of tools**  There is a clearly identifiable evaluation process/criteria which has been applied. The evaluation is objective and is supported by presented evidence (results and/or measures). There is a clear conclusion which follows logically from the evaluation process. Where appropriate, the evaluation approach and the conclusions are further supported by the literature. | | |
| * **Effective evaluation approach has been applied**   The approach to evaluate each tool/language is clear and has been used objectively. | | **10** |
| * **The outcome(s) is/are clearly supported by the evidence**   There is a clear and rational link in the discussion between the outcome of the evaluation process and conclusions/overall outcome of the discussion. | | **10** |
| **5** | **Task 3: Review of Current Literature**  There has been a clear and comprehensive investigation into research areas that relate generally to the scenario area, and specifically to the problems selected for investigation. The discussion provides relevant insight into the selected problems and demonstrates an objective academic approach to the appraisal of the current literature and research activities in the subject area. | | |
| * **Presentation of current literature**   There is a clear discussion around relevant areas of the current literature that informs the reader of the status quo. | | **10** |
| * **Critical evaluation of the literature**   The various approaches are reviewed critically, drawing out rational and supported arguments in the evaluation and/or comparison of those approaches. | | **10** |
| * **Academic best practice**   Academic best practices have been effectively adhered to, demonstrating effective citations and referencing, a high level of critical thinking, and an objective and rational approach to the presentation of ideas and arguments. | | **10** |
|  | **TOTAL**: | | **100** |

|  |
| --- |
| NOTE: To attain an overall **Pass** in this assessment you **MUST** present and discuss evidence of your prototype programs and the tests you have undertaken as figures/screenshots in your appendices, and provide relevant citations and references to support your arguments/discussion. |

# Marking Criteria: Grade breakdown

## Executive Summary (10%)

|  |  |  |
| --- | --- | --- |
| 0-39% | Fail | * The executive summary is either missing, or incomplete and poorly structured. * There is no indication of the options chosen. * Few or none of the indicative points are covered. * There is little or no cohesion or relevance to the content of the report or its outcomes |
| 40-49% | Marginal fail | * The executive summary is incomplete and/or poorly structured. * There is no indication of the options chosen. * Few of the indicative points are covered. * There is limited cohesion or relevance to the content of the report or its outcomes |
| 50%-59% | Pass | * The executive summary is complete and reasonably structured. * There is a clear indication of the options chosen. * All of the indicative points are covered. * There is a cohesive and relevant summary representative of the contents of the report and its outcomes |
| 60%-69% | Merit | * The executive summary is complete and well structured. * There is a clear indication of the options chosen. * All of the indicative points are covered, with one or two additional points, which are mostly relevant. * There is a cohesive and relevant summary which very clearly presents the contents of the report and its outcomes. |
| 70%-100% | Distinction | * The executive summary is complete and very well structured. * There is a clear indication of the options chosen. * All of the indicative points are covered, with one or two additional points, which are relevant and well thought out. * There is a cohesive and relevant summary which very clearly and logically presents the contents of the report and its outcomes. |

## Task 1: Design and Discussion of Solution (40%)

|  |  |  |
| --- | --- | --- |
| 0-39% | Fail | * Discussion of the design and solution(s) for one or two of the problems have been presented. * There is little or no effective supporting evidence of prototype programmes or tests. * There is little or no relevant supporting evidence from the literature. * **Pre-processing:** There is little or no evidence of pre-processing considerations. * **Selected Techniques:** There is little or no evidence of which techniques have been selected, or the selected techniques do not relate to the problem, solution or stated selection criteria, if present. * **Justification of Techniques:** There is little or no relevant justification of the techniques selected, given the context of scenario and selected problems. * **Evaluation of Solution(s):** There is little or no effective evaluation of any presented solutions. The measures or methods selected are incompatible or incomplete and do not support the presented outcomes. |
| 40-49% | Marginal fail | * Discussion of the design and solutions for two of the problems have been presented. * There is limited effective supporting evidence of prototype programmes or tests. * There is limited relevant supporting evidence from the literature. * **Pre-processing:** There is some evidence of pre-processing considerations; however, they are not correctly aligned with the requirements of the selected problem and/or of the proposed solution. * **Selected Techniques:** There is some evidence of which techniques have been selected; however, they do not relate to the problem, solution or stated selection criteria (if present). * **Justification of Techniques:** There is some justification of the techniques selected; however, it is not completely relevant given the context of the scenario and selected problems. * **Evaluation of Solutions:** There is some evaluation of any presented solutions. However, the measures or methods selected are incomplete or poorly applied and do not support the presented outcomes. |
| 50%-59% | Pass | * Discussion of the design and solutions for two of the problems have been presented. * There is supporting evidence of prototype programmes or tests, which is somewhat effective. * There is supporting evidence from the literature, which is somewhat relevant. * **Pre-processing:** There is evidence of pre-processing considerations, which align to some degree with the requirements of the problems selected and the proposed solutions. * **Selected Techniques**: There is evidence of which techniques have been selected; these align to some degree with the selected problems, solutions and/or stated selection criteria. * **Justification of Techniques:** There is a justification of the techniques selected, which is somewhat relevant and effective, given the context of the scenario and selected problems. This considers a limited range of potential options. * **Evaluation of Solutions:** There is clear and somewhat effective evaluation of the presented solutions. The measures/methods selected are fairly compatible, have been applied partially effectively, and support the presented outcomes to some extent. |
| 60%-69% | Merit | * Discussion of the design and solutions for two of the problems have been presented. * There is supporting evidence of prototype programmes or tests, which is mostly effective. * There is supporting evidence from the literature, which is mostly relevant. * **Pre-processing:** There is clear evidence of pre-processing considerations, which mostly aligns with the requirements of the problems selected and the proposed solutions. * **Selected Techniques:** There is clear evidence of which techniques have been selected, which mostly aligns with the selected problems, solutions and/or stated selection criteria. * **Justification of Techniques:** There is a clear justification of the techniques selected, which is mostly relevant and effective, given the context of the scenario and selected problems. This considers a reasonable range of potential options. * **Evaluation of Solutions:** There is a clear and mostly effective evaluation of the presented solutions. The measures/methods selected are generally compatible, have been applied in an approach that is largely effective, and support the presented outcomes to a considerable extent. |
| 70%-100% | Distinction | * Discussion of the design and solutions for two of the problems have been presented. * There is effective supporting evidence of prototype programmes or tests. * There is relevant supporting evidence from the literature. * **Pre-processing:** There is clear evidence of pre-processing considerations, which aligns fully with the requirements of the problems selected and the proposed solutions. * **Selected Techniques**: There is clear evidence of which techniques have been selected, which align fully with the selected problems, solutions and/or stated selection criteria. * **Justification of Techniques**: There is a clear and effective justification of the techniques selected, given the context of the scenario and selected problems. This considers a good range of potential options. * **Evaluation of Solutions:** There is clear and effective evaluation of the presented solutions. The measures/methods selected are compatible and have been applied effectively, supporting the presented outcomes. |

## Task 2: Evaluation of tools (20%)

|  |  |  |
| --- | --- | --- |
| 0-39% | Fail | * The evaluation of tools for one or two of the problems has been presented. * There is little or no relevant supporting evidence from the literature. * **Effective evaluation approach**: There is little or no evaluation approach presented, which has not been applied objectively to the key/main tools/languages used in the presented solution. * **The outcome(s) is/are clearly supported by the evidence:** There is little or no discussion, presenting the outcome(s) of the evaluation, there is no clear statement of effectiveness for the key tools/languages used. The outcome is little or no evidence presented that supports the conclusion. |
| 40-49% | Marginal fail | * The evaluation of tools for two of the problems has been presented. * There is limited relevant supporting evidence from the literature. * **Effective evaluation approach:** A limited, or incomplete evaluation approach has been presented, which has not been applied objectively to the key/main tools/languages used in the presented solution. * **The outcome(s) is/are clearly supported by the evidence:** There is limited discussion, presenting the outcome(s) of the evaluation, there is no clear statement of effectiveness for the key tools/languages used. The outcome is poorly supported by the evidence presented, and the conclusion is limited or unclear. |
| 50%-59% | Pass | * The evaluation of tools for two of the problems has been presented. * There is supporting evidence from the literature, which is somewhat relevant. * **Effective evaluation approach**: A somewhat effective evaluation approach has been presented, which has been applied objectively to the key/main tools/languages used in the presented solution. * **The outcome(s) is/are clearly supported by the evidence**: A somewhat rational discussion, presents the outcome(s) of the evaluation, which clearly states effectiveness of the tools/languages used. The outcome is supported to some degree by the evidence presented, and the conclusion is fairly rational. |
| 60%-69% | Merit | * The evaluation of tools for two of the problems has been presented. * There is supporting evidence from the literature, which is mostly relevant. * **Effective evaluation approach:** A mostly effective evaluation approach has been presented, which has been applied objectively to the key/main tools/languages used in the presented solution. * **The outcome(s) is/are clearly supported by the evidence:** A mostly rational discussion presents the outcome(s) of the evaluation, which clearly states the effectiveness of the tools/languages used. The outcome is mostly supported by the evidence presented, and there is a mostly rational conclusion. |
| 70%-100% | Distinction | * The evaluation of tools for two of the problems has been presented. * There is relevant supporting evidence from the literature. * **Effective evaluation approach:** A clear and effective evaluation approach has been presented, which has been applied objectively to the key/main tools/languages used in the presented solution. * **The outcome(s) is/are clearly supported by the evidence:** A clear and rational discussion, presents the outcome(s) of the evaluation, which clearly states the effectiveness of the tools/languages used. The outcome is clearly supported by the evidence presented, and there is a solid and rational conclusion. |

## Task 3: Review of Current Literature (30%)

|  |  |  |
| --- | --- | --- |
| 0-39% | Fail | * IEEE referencing may not have been applied. * **Presentation of current literature:** There is little or no relevant discussion of the current literature with little relation to the given scenario, and/or the selected problems. This does not provide the reader with an understanding of the current state of research in these areas. * **Critical evaluation of the literature:** Few or no approaches have been presented. There is little or no evidence of these having been reviewed, with little or no evidence of effective evaluation and/or comparison. * **Academic best practice:** Academic best practice has not been adhered to in all aspects, there is little or no correct use of citations and referencing. There is little or no effective presentation of ideas and arguments, which are not well supported by given evidence. |
| 40-49% | Marginal fail | * IEEE referencing may not have been applied. * **Presentation of current literature:** There is limited discussion of the current literature which is not always relevant generally to the given scenario, and/or specifically to the selected problems. This provides the reader with a limited understanding of the current state of research in these areas. * **Critical evaluation of the literature:** Few approaches have been presented. These have been reviewed in a limited way, drawing out little rational argument in the evaluation and/or comparison of those approaches. * **Academic best practice:** Academic best practice has not been adhered to in all aspects, and there is little correct use of citations and referencing. The presentation of ideas and arguments is limited, and not well supported by given evidence. |
| 50%-59% | Pass | * Correct IEEE referencing has been applied. * **Presentation of current literature:** There is a somewhat relevant discussion of the current literature that relates generally to the given scenario, and specifically to the selected problems. This provides the reader with some understanding of the current state of research in these areas. * **Critical evaluation of the literature:** The various approaches presented have been critically reviewed to some degree, drawing out some rational arguments in the evaluation and/or comparison of those approaches. * **Academic best practice:** Academic best practice has been partially adhered to with regard to demonstrating citations and referencing, displaying a fairly high level of critical thinking in the presentation of ideas and arguments. |
| 60%-69% | Merit | * Correct IEEE referencing has been applied. * **Presentation of current literature:** There is a mostly relevant discussion of the current literature that relates generally to the given scenario, and specifically to the selected problems. This provides the reader with a reasonable understanding of the current state of research in these areas. * **Critical evaluation of the literature:** The various approaches presented have been critically reviewed in a mostly effective way, drawing out mostly rational arguments in the evaluation and/or comparison of those approaches. * **Academic best practice:** Academic best practice has been mostly adhered to with regard to demonstrating citations and referencing, displaying a generally high level of critical thinking, and an objective and rational approach to the presentation of ideas and arguments. |
| 70%-100% | Distinction | * Correct IEEE referencing has been applied. * **Presentation of current literature:** There is a clear and relevant discussion of the current literature that relates generally to the given scenario, and specifically to the selected problems. This provides the reader with a good understanding of the current state of research in these areas. * **Critical evaluation of the literature:** The various approaches presented have been critically reviewed in an effective way, drawing out rational and supported arguments in the evaluation and/or comparison of those approaches. * **Academic best practice:** Academic best practice has been effectively adhered to, demonstrating effective citations and referencing, a high level of critical thinking, and an objective and rational approach to the presentation of ideas and arguments. |

# Assessment Submission

You will submit your assessment in the ‘Assignments’ area of the module in Canvas. Please check your Canvas module for the specific submission date for this assignment.

This assessment requires you to anonymously upload your submission to Canvas. If you are submitting multiple files, you must upload all files simultaneously to ensure that they are marked as a single submission. If you want to resubmit one component of your work, you need to re-upload all other files at the same time: every submission must include **all** files required by the assessment brief.

We recommend that you allow at least 30 minutes before the deadline to upload your submission, as failure to upload your assessment file within the allotted time is not admissible as an exceptional circumstance.

The webpage [How do I submit an online assignment?](https://community.canvaslms.com/t5/Student-Guide/How-do-I-submit-an-online-assignment/ta-p/503) provides further technical information on how to upload an assessment. The advice given here comes directly from Canvas. We do not recommend uploading assignments by mobile. We recommend you view the submission immediately after uploading your work to ensure the correct file has been submitted and no technical errors have occurred.

If you face any technical difficulties whilst trying to submit this assessment, then contact Canvas support on [support@instructure.com](mailto:support@instructure.com) or +44 80 0060 8442 (available 24 hours) in advance of the deadline. You should also email [york-online-assessment@york.ac.uk](mailto:york-online-assessment@york.ac.uk) as a matter of urgency to report the issue and receive further instruction.

# Assessment Policies

This assessment is subject to the policies stated on the ‘Summative Assessment Policies’ page in Canvas. These policies include (but are not limited to):

* Academic Integrity and submission of student work to Turnitin
* Advice on anonymising your assessment
* Penalties for late submission
* Marking policy for multiple submissions
* The Fit to Sit / Submit policy
* Passing mark and module reassessment

Please ensure that you have read and understood these policies before starting the assessment.