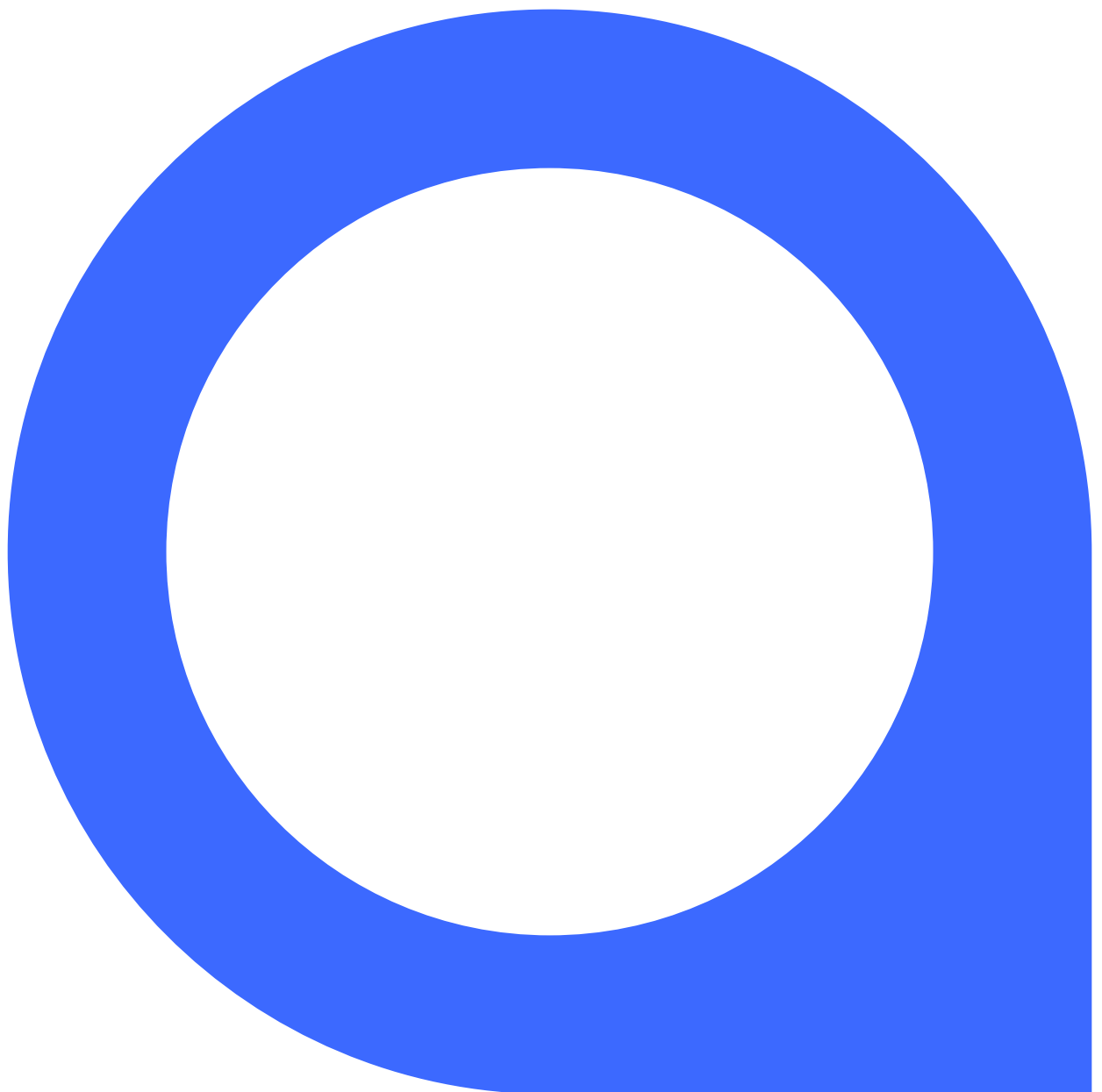


Data Science Applications

Semester 1 2024 Assignment Feedback





Detailed assignment feedback

Question 1 – Prepare the data

In this question, students were required to examine the movie quotes dataset, apply cleaning and vectorisation on the text feature in the dataset, and create a response variable. Sections 1(a) and 1(b) were done well, while questions 1(c) and 1(d) which required vectorisation and creating a response variable were done adequately.

What Students Did Well:

- Data Splitting and Justification:
 - Many students performed data splitting at an appropriate stage and provided good justifications for their choices.
 - Some students showed a clear understanding of the data splitting process, considering issues such as data leakage and the relevance to future modelling.
- Cleaning Steps:
 - Most students performed more than the required five cleaning steps and conducted checks to verify the effectiveness of their cleaning processes.
 - Cleaning steps were generally well explained, and students often provided a rationale for why specific cleaning actions were taken.
 - Multiple checks were performed for each cleaning step, demonstrating thoroughness and attention to detail.
- Exploratory Data Analysis (EDA):
 - Students used a variety of visualisation techniques and analyses to explore the dataset.
 - There was significant attention to identifying data drift and other issues in the dataset.
 - Some students provided extensive analysis covering multiple variables, offering detailed insights and observations.
- Linking to Business Context:
 - Some students effectively linked their analysis and cleaning steps to the problem context, explaining the potential impact on modelling and business decisions.
 - Justifications for the choice of vectorisation methods and the construction of response variables were provided, with appropriate checks.



What Students Did Poorly:

- Clarity and Organisation:
 - Several students had poorly organised notebooks, making it difficult to follow their thought processes and findings.
 - There were instances of duplicated work, lack of explanation for findings, and mechanical responses that seemed copy-pasted.
- Superficial Analysis:
 - A number of students provided high-level, surface observations without delving deeper into the implications or providing detailed analysis.
 - Many responses lacked thorough explanation and interpretation of the data, focusing more on procedural steps rather than insights.
- Linking to Business Context:
 - A common issue was the lack of strong linkage between findings and the business context. Many students did not adequately explain how their observations and cleaning steps would impact the problem or future modelling.
 - Justifications for cleaning steps and vectorisation methods were often weak, with limited reference to the problem context.
- Checks and Validations:
 - While many students performed checks on their cleaning steps, the explanations for these checks were often brief or missing. The checks themselves were sometimes too simplistic and not insightful.
 - There was a lack of thorough validation for constructed response variables, with some students failing to provide sufficient checks or explanations.
- Detailed Observations:
 - Some students failed to adequately explain their observations, often making vague remarks without supporting their statements with detailed analysis or visual evidence.
 - Important variables such as the movie quote field were sometimes not analysed thoroughly, showing a lack of focus on key aspects of the dataset.

Recommendations for Improvement:



- Ensure that notebooks are well-organised and each step is clearly explained.
- Provide detailed explanations and justifications for all cleaning steps, linking them to the problem context and potential impacts on modelling.
- Perform thorough and insightful checks on cleaning steps and constructed response variables, with clear explanations of the outcomes.
- Use a variety of visualisation and analysis techniques to explore the data, ensuring that key variables are thoroughly analysed.
- Make explicit connections between observations and the business context, demonstrating a deep understanding of how data issues affect the overall problem.

Question 2 – Clustering to explore data characteristics

In this question, students were asked to cluster the data, describe insights gained from the clustering, and then suggest practical uses for the clustering insights. All three sections averaged a borderline pass mark. Overall, while there were strengths in justification, description, and communication among some students, a general trend of insufficient justification, overly technical focus, and poor validation practices was observed.

What the Students Did Well:

- Justification of Clustering Methodology:
 - Many students provided good justification for choosing K-means and the number of clusters. They often linked their choices to the context and used internal validation to support their decisions.
 - Some students demonstrated a strong approach by iterating and evaluating different numbers of clusters and using various validation methods.
- Description and Analysis:
 - Detailed descriptions of key characteristics of each cluster were common. Some students were able to provide clear and concise communication suitable for executives.
 - There were instances of students leveraging external datasets to enhance their analysis, and describing the characteristics of clusters with a strong link to the problem context.
- Communication Style:



- Several responses had well-structured and clear summaries that were appropriate for executive-level communication. Bullet points and highlighted key terms were used effectively.
- The students who did well were able to present their findings in a format that was easy to follow and tailored for non-technical stakeholders.

What the Students Did Poorly:

- Lack of Justification and Discussion:
 - A significant number of students failed to justify their choice of clustering algorithm and distance measures adequately. There were instances where internal validation methods were either not used or not well-explained.
 - Some responses were mechanical, lacking in-depth discussion and merely stating the processes without proper rationale.
- Technical Focus Over Contextual Relevance:
 - Several students provided responses that were too technical and not sufficiently tailored to the stakeholders' needs. There was often a lack of connection between the technical analysis and the real-world application.
 - Communication styles varied, and some were not concise enough, making it hard for executives to extract key insights.
- Evaluation and Validation:
 - Many students did not perform or discuss internal validation adequately. In some cases, the internal validation performed did not align with their conclusions or was not properly interpreted.
 - Manual validation outputs were often not generated or saved, leading to incomplete analyses and descriptions.
- Repetitiveness and Hypothetical Points:
 - Feedback indicated that many students had repetitive points in their answers and included suggestions that were hypothetical or not directly derived from their clustering analysis.
 - There were instances where points raised did not follow logically from the clustering outcomes, leading to abstract and impractical suggestions.



Recommendations for Improvement:

- **Strengthen Justification and Discussion:** Clearly justify your choice of clustering algorithms and distance measures, and use internal validation methods with detailed interpretation to support your decisions.
- **Enhance Contextual Relevance:** Ensure your analysis is directly relevant to the problem context, and tailor your communication for non-technical stakeholders using clear, concise language and structured summaries.
- **Improve Evaluation and Validation Practices:** Perform thorough internal validation, generate and save manual validation outputs, and provide comprehensive interpretations to ensure a complete analysis.
- **Avoid Repetitiveness and Abstract Points:** Provide unique, practical, and actionable suggestions that are directly tied to the clustering outcomes, avoiding hypothetical or repetitive points.
- **Focus on Data Interpretation:** Go beyond technical descriptions by providing meaningful interpretations of clustering results, using relevant visual aids to support your findings and enhance understanding.

Question 3 – Classification model

In this question, students were asked to deal with class imbalance, construct a neural network to classify problematic quotes, and calculate relevant model success measures. Each part of this question was done adequately with average marks for each section close to a borderline pass.

What the Students Did Well:

- **Good Contextual Discussions:** Many students provided clear explanations and linked their discussions to the assignment context, particularly in defining methods and comparing models.
- **Metrics and Comparisons:** Students often identified appropriate measures and gave good comparisons with benchmark models. They also showed a strong understanding of why accuracy might not be suitable.
- **Iterative Improvements:** A number of students demonstrated iterative improvements, clear thought processes, and logical reasoning in their model development.
- **Clear Structure:** Submissions generally had a clear structure, making it easy to follow the steps taken and the rationale behind decisions.
- **Thorough Testing:** Some students thoroughly tested final model outcomes and demonstrated good logic between iterations, testing features and parameters.



What the Students Did Poorly:

- Lack of Specificity and Context: Many students provided generic solutions that lacked specificity or were not well-linked to the assignment context. SMOTE, for example, was often cited inappropriately.
- Inadequate Justifications: Justifications for chosen methods and measures were often lacking or insufficiently detailed.
- Misinterpretation of Metrics: There were frequent issues with misinterpreting metrics or using inappropriate metrics, such as relying on accuracy.
- Insufficient Iterations and Checks: Some submissions showed minimal iteration, random experimentation, or inadequate final model checks. Final model evaluations were sometimes based on training rather than test data.
- Poor Code Readability: A few students submitted code that was difficult to follow, lacked visualizations, or was not well-commented.

Recommendations for Improvement:

- Provide Specific and Contextual Methods: Ensure that methods are not only defined but also explicitly linked to the assignment context. Avoid generic strategies and provide detailed implementation plans.
- Thoroughly Justify Choices: Give clear and comprehensive justifications for the chosen methods and measures. Explain why certain metrics are selected and how they align with the objectives.
- Interpret Metrics Correctly: Focus on using appropriate metrics for the dataset and interpret them correctly. Avoid over-reliance on accuracy and understand the importance of metrics like precision, recall, and AUC.
- Increase Iterations and Checks: Conduct more directed iterations with clear logical steps. Perform thorough checks on the final model using test data, and ensure that different features and parameters are tested.
- Improve Code Clarity: Make code easy to read with proper sectioning, comments, and visualizations. Clearly document the thought process and ensure that the final outputs are clearly presented and explained.



Question 4 – Communication

In this question, students were asked to prepare an executive summary of your findings from the classification model, to be presented to executives of the movie production company. Both written and verbal responses were required. This question was poorly done, especially for 4(a) where most students provided an inadequate answer.

What the Students Did Well:

- **Clear Communication:** Many students demonstrated strong communication skills, providing clear and concise explanations.
- **Structure and Transitions:** Several students presented their information with a clear structure and logical transitions, making it easier to follow.
- **Use of Visual Aids:** Effective use of visual aids helped in explaining complex concepts.
- **Understanding of Technical Concepts:** A good number of students showed a strong understanding of technical concepts and were able to explain them clearly.
- **Linking to Objectives:** Some students successfully linked the model's success back to the objectives of the movie producers.
- **Discussion of Limitations:** Many students listed relevant limitations and proposed actions to address these issues.

What the Students Did Poorly:

- **Overly Technical Language:** A common issue was the use of technical jargon that was not suitable for the intended audience (movie producers).
- **Lack of Visual Aids:** Some presentations lacked visual aids, making it difficult to understand complex concepts.
- **Poor Audio Quality:** Issues such as loud background music and unclear audio were noted in some submissions.
- **Inadequate Explanation of Classifiers:** Several students did not adequately explain how the classifiers worked.
- **Missing Submissions:** There were instances of missing or late submissions.
- **Confusing Structure and Transitions:** Some presentations lacked clear transitions and structure, making them difficult to follow.
- **Excessive Focus on Technical Details:** Many students spent too much time on technical details that were not relevant to the audience, making the presentations less engaging.
- **Lack of Link to Objectives:** Some students did not effectively link their discussions back to the movie producers' objectives.



Recommendations for Improvement:

- **Simplify Language:** Use language that is appropriate for the audience, avoiding unnecessary technical jargon.
- **Enhance Visual Aids:** Incorporate more visual aids to help explain complex concepts in a simpler manner.
- **Improve Audio Quality:** Ensure that audio is clear and free of distracting background noise or music.
- **Explain Classifiers Clearly:** Provide a clear and simple explanation of how classifiers work, avoiding overly technical details.
- **Submit on Time:** Ensure that submissions are made on time and are complete.
- **Organize Content Logically:** Maintain a clear structure with logical transitions to help the audience follow along.
- **Focus on Audience Needs:** Tailor the presentation to the audience's needs, focusing on the most relevant information and linking back to their objectives.
- **Provide Context:** Always link back technical discussions to the overall business problem or objective to make it relevant for the audience.