Title	**Machine Learning Course Work**
Module	**NIB7072 Machine Learning**
Module Lead	**Ms. Akshila Anurangi**
Start Date	**27th July 2024**
End date	**20th August 2024**

Question 1 (7.5 Marks) – [word count: 100 words]

Imagine you are a data scientist and a client reaches out to you to get your opinion on getting started with advanced analytics for their retail supermarket chain. Give them 5 example use cases they can implement and describe what data and variables they should start collecting to cater these use cases. (e.g. : if you are suggesting a demand forecasting use case, what data points they should have in hand to do this). You don't have to specify data sources for each use case. Think of the retail supermarket and give the data sets and its variables in general that could be used to get started with advanced analytics.

Question 2 (5 Marks) – [word count: 100 words]

The retail supermarket chain decided that they wanted to start their analytical journey with a customer churn prediction use case. Use your creativity to,

- define engagement objectives,
- and three key questions to address in the use case and deliverables we can provide for the management to retain / gain back the customers.

Question 3 (10 Marks) – [word count: 250 words]

Create a methodology to implement the churn prediction use case. You can discuss the following areas. Create 2- 3 Presentation slides to present this solution design to the retail supermarket management.

- Data sources to be used.
- How to define the target variable
- What features and synthetic features can be used for the model
- What are the suitable algorithms?
- How can hyperparameter tuning be done for the algorithms you mentioned?
- If the model has under-forecasting or over-forecasting issues, how would you mitigate those.
- How to evaluate the accuracy of the predictions
- How to identify the drivers of churn
- What actions can be taken to prevent customer churn?

Question 4 (7.5 Marks) – [word count: 150 words]

The supermarket chain is satisfied with the customer churn model you built and now wants to deploy it in the cloud. Discuss,

- key principles of MLOps they can adapt,
- Example technologies and toolkits they can use for this purpose.

Question 5 - Technical Report (50 marks) – [word count: 2000 words]

Sales Forecasting Problem: You are expected to forecast the item quantity sales at a daily level for the month February 2022 for the following departments at following stores.

Departments: Beverages, Grocery, Household

Outlets: XYZ, ABC

Come up with an approach to identify above criteria.

A. Conduct appropriate descriptive analysis on the dataset to validate/test hypotheses, used notebooks should be clean and should have proper descriptions under each analysis to summarize your findings.

B. Follow the stipulated folder structure in designing the model.

C. Build a pipeline by creating following under the src/ folder:
    a. Processing Data sources
    b. Primary keys (store | department | date)
    c. Target variable
    d. Sales related features
    e. Item related features
    f. Time related features
    g. Outlet related features

D. Create separate pipelines to generate the master table (containing all required synthetic features) and to fit the model.
E. Assess the MAPE at following granularity:
    a. Store | Department | Date
    b. Store| Date

F. Incorporate model management best practices (PEP-8 Standards, Modularized Code, Configurations, Functions/Classes)
G. Commit and deploy the model , you can deploy it as a web app (with a simple front end) or any other preferred way.
H. Create a technical report demonstrating your understanding in terms of the execution of steps 1,2,3,4,5,6,7

<u>Data Dictionary</u>

Outlet_info: Contains outlet related information

store: Unique identifier for an outlet
profile: Outlet definition based on the type of customers that shop
size: Outlet definition based on the layout of the outlet

transactions_info: Contains transactions related information

date_id - Date of transaction
item_dept – Item Department
item_qty - standard no of units of the transaction
net_sales - standard sales value of the transaction
store : Unique identifier for an outlet
item : Unique identifier for an item
invoice_num : Unique identifier for an invoice

Folder Structure

Level 1	Level 2	Level 3	Description
analysis/			contains scripts for analysis work
src/	utils*/	module specific utils	utils notebooks that are used across modules
	models*/	notebooks that train & use models	notebooks that train & use models
pipelines/	<module_name>		main scripts for each module
conf/			the main get_conf() notebook

Technical Report Structure

Following is a guideline for the technical document. Feel free to modify the structure while preserving the main sections.

1 Introduction
- 1.1 Statement of purpose
- 1.2 Document Structure

2 Methodology

3 Implementation
- 3.1 Data preparation
  - 3.1.1 Data sources
  - 3.1.2 Features Construction
  - 3.1.3 Target Construction
  - 3.1.4 Master Table Creation
- 3.2 Model Development
  - 3.2.1 Choosing the Algorithm
  - 3.2.2 Hyperparameter Optimization
  - 3.2.3 Feature Importance
  - 3.2.4 Model Accuracy Measures
- 3.3 Risks and Assumptions

3.4 Pipelines
3.5 Model Deployment

4 Findings and Conclusions

Question 6 - Explainable ML implementation with a summary Report (20 marks) [report word count: 1500 words].

you are required to explore the application of Explainable ML techniques to address a significant challenge in a chosen domain. Choose a problem domain that interests you (e.g., natural language processing, image processing, healthcare diagnostics) and find an appropriate, sufficiently complex dataset.

A. Define the problem you wish to address and the dataset you have chosen. Explain why this problem is significant and how model explainability is important for this. using a suitable approach to implement a model, use hyperparameter tuning to optimize the model (Other than manual, grid and random search. Recommendation is Bayesian hyperparameter tuning but you can use any option) keep a clean notebook / codebase for your implementation with appropriate comments / descriptions added into the steps you have followed.
B. Discuss and Implement appropriate explainability methods to interpret model predictions, provide actionable insights for the problem you are addressing.
C. Present your findings in a clear and concise report, demonstrating a comprehensive understanding of Explainable ML techniques, their implementation, and their implications for the chosen problem domain.
    a. Compare and contrast your chosen approach with Gen AI methods in terms of performance, interpretability, and applicability to your problem domain.
    b. Discuss any ethical considerations or limitations associated with your approach, including issues related to bias, fairness, and data privacy.

Marking Scheme

Question	Question Subpart	Allocated Marks	Description
Question 01 (7.5 marks)	**-**	**7.5**	Relevant use cases are given with a short description, Different Data Sources are given with variables that be that can be derived from each data source
Question 02 (05 marks)	**A**	**2.5**	Given engagement objectives are clear and relevant for the customer churn use case.
	B	**2.5**	Key Questions addresses the main pain points of the problem
Question 03 (10 marks)	**-**	**10**	The methodology should cover all important aspects of building the use case. The solution design should not be too technical, simple enough to be understood by a business person. Recommend preparing two or three presentation slides.
Question 04(7.5 marks)	**A**	**2.5**	Correct MLops principles are given
	B	**5**	Given technologies and toolkits are relevant and serve the purpose
Question 05 (50 marks)		**5**	Descriptive analysis is complete, all the necessary checks and imputations are performed on the dataset, clearly summarizing the findings.
		5	Given folder structure is followed when developing the code
		10	functions are defined correctly, doc strings are given, notebooks are chained properly, the code compiles without any errors, correct logics are used
		5	Main pipeline creates the master table and fits the model, all the notebooks are neat and clean, the used model is appropriate, if correct and novel models are used
		5	Appropriately calculates the MAPE at correct granularity
		5	Coding best practices are followed when developing the use case

		5	Using an appropriate cloud service the model is deployed
		10	Organization and clarity of the written report, Technical report is complete, a detailed description about the implementation and use case is given, and describes the application, purpose, creation or architecture of the use case implementation. Quality and clarity of visual aids, such as diagrams, charts, and examples of generated outputs or explanations. Adherence to academic writing standards and proper referencing.
Question 06 (20 marks)		**2.5**	Problem Definition and Dataset Selection - Clarity and significance of the chosen problem domain. Justification of why Explainable ML techniques are important.Quality and relevance of the selected dataset.
		5	Implementation of ML and Explainable ML Techniques , novelty of the used methods, Correct implementation of the chosen model, Quality and diversity of the generated outputs. Use of appropriate evaluation metrics for generated outputs, Critical analysis of the results and their implications.
		5	Documentation and Presentation : Organization and clarity of the written report. Quality and clarity of visual aids, such as diagrams, charts, and examples of generated outputs or explanations. Adherence to academic writing standards and proper referencing.
		2.5	Originality and creativity in the application of AI techniques to solve the chosen problem, A short literature review with referenced techniques, Innovation in addressing challenges or extending the scope of the project.
		2.5	Thorough evaluation of the performance, interpretability, and applicability of chosen techniques. Clear comparison with Gen AI approaches in terms of strengths and weaknesses. (only if applicable) Insightful discussion on the practical implications and potential industry applications.

		2.5	Identification and discussion of ethical considerations related to data privacy, bias, fairness, and transparency. Consideration of potential societal impacts of the techniques used.