JAIPURIA INSTITUTE OF MANAGEMENT
LUCKNOW NOIDA JAIPUR INDORE

JAIPURIA INSTITUTE OF MANAGEMENT, NOIDA

PGDM / PGDM (M) / PGDM (SM)

FIFTH TRIMESTER (Batch 2024-26)

END TERM EXAMINATIONS, JANUARY 2026

SET-1

Course Name	Advanced Machine Learning	Course Code	**20841**
Max. Time	**2 hours**	Max. Marks	**40 MM**

INSTRUCTIONS:

a. All questions are compulsory to attempt

1. Read the following case and answer the questions given at the end:

RetailSphere's Data Dilemma: Turning Customer Data into Decisions

RetailSphere Pvt. Ltd. is a rapidly growing omnichannel retail company operating across India through physical stores, a mobile application, and an e-commerce platform. The company sells a wide range of products including groceries, consumer electronics, apparel, and home essentials. Over the past two years, RetailSphere has invested heavily in digital infrastructure and data collection, capturing transaction data, customer profiles, browsing behavior, and thousands of customers reviews every week.

Despite strong top-line growth, the senior management team has started noticing several worrying patterns.

First, the marketing team observes that a significant number of customers stop purchasing after a few months of initial engagement. These customers show early signs such as declining purchase frequency, increased returns, higher discount dependence, and negative service interactions. Management wants to identify such customers in advance so that retention offers and personalized interventions can be deployed proactively. However, historical data shows that customers who actually leave form only a small proportion of the total customer base, making predictions difficult.

At the same time, the merchandising team is struggling to understand the diversity of customer behavior. While some customers shop frequently across categories, others purchase occasionally but spend large amounts. A few customers are extremely price-sensitive, whereas others show strong brand loyalty. Treating all customers uniformly has resulted in inefficient campaigns and low response rates. The leadership believes that grouping customers based on behavioral similarity could improve targeting and strategic planning.

Meanwhile, store managers and category heads are under pressure to increase the average basket value. Analysts discover that customers often buy certain products together, but these patterns are not always obvious. For example, some low-margin items appear to drive the sale of high-margin products when purchased together. Management wants to uncover hidden product associations from transaction data to design better bundles, promotions, and layout strategies—both online and offline.

On the digital front, RetailSphere's app and website currently rely on generic "top-selling products" lists. Customers have complained that recommendations feel irrelevant and repetitive. Management wants a system that can suggest products based on similar customer preferences and historical interactions, even when explicit product ratings are limited. They are also concerned about scalability as the number of users and products grows rapidly.

Finally, customer experience teams monitor thousands of product reviews, feedback comments, and social media mentions daily. While dashboards show star ratings, they fail to capture the actual emotional tone of customer feedback. Management wants to automatically analyze textual reviews to distinguish satisfied customers from dissatisfied ones, identify problematic products or services, and connect these insights with customer behavior and retention patterns.

RetailSphere's analytics team has been asked to design robust, data-driven solutions to address these challenges using machine learning techniques already validated within the organization.

Case Questions:

Q1. RetailSphere's management believes that grouping customers based on similarities in purchasing behavior will improve marketing effectiveness. **(1+6+3 = 10 Marks)**

- Identify the type of machine learning problem involved in this scenario.
- Describe how the solution would be developed step-by-step, from data selection to interpretation of results.
- Explain how the output can be used to design differentiated business strategies.

Q2. RetailSphere wants to automatically analyze large volumes of textual customer feedback to understand customer satisfaction levels. **(1+6+5 = 12 Marks)**

- Identify the machine learning problem type involved in analyzing customer reviews.
- Describe the end-to-end process of building such a model, starting from raw text data.
- Discuss how insights from this analysis can be integrated with customer retention or segmentation strategies.

Q3. As part of its effort to improve cross-selling and bundle design, RetailSphere analyzes a sample of recent transaction data from its urban convenience-store format. Unlike traditional grocery staples, these stores focus on quick-consumption, lifestyle, and impulse-driven products, where co-purchase patterns are not always intuitive.

To achieve this, the analytics team extracts a sample of recent transaction data and performs market basket analysis using Python-based analytics tools. The objective is to identify:

- high-frequency products,
- strong product associations, and
- suitable product pairs for bundle offers.

The sample dataset and the Python output is given below:

Transaction Dataset (Sample Used for Analysis)

Transaction ID	Items Purchased	Transaction ID	Items Purchased
T1	Protein Bar, Sports Drink	T11	Ready-to-Eat Salad, Low-Fat Yogurt
T2	Ready-to-Eat Salad, Low-Fat Yogurt	T12	Protein Bar, Sports Drink
T3	Protein Bar, Peanut Butter	T13	Energy Gel, Sports Drink
T4	Sports Drink, Energy Gel	T14	Whole Wheat Bread, Peanut Butter
T5	Protein Bar, Sports Drink, Energy Gel	T15	Protein Bar, Sports Drink, Energy Gel
T6	Ready-to-Eat Salad, Cold-Pressed Juice	T16	Ready-to-Eat Salad, Cold-Pressed Juice
T7	Protein Bar, Sports Drink	T17	Protein Bar, Granola
T8	Peanut Butter, Whole Wheat Bread	T18	Sports Drink, Energy Gel
T9	Low-Fat Yogurt, Granola	T19	Protein Bar, Sports Drink
T10	Protein Bar, Cold-Pressed Juice	T20	Low-Fat Yogurt, Granola

(Total Transactions = 20)

Output (Derived after running Market Basket Analysis using Python)

Itemset	Support
{Protein Bar}	0.45
{Sports Drink}	0.40
{Energy Gel}	0.25
{Low-Fat Yogurt}	0.20
{Protein Bar, Sports Drink}	0.30
{Sports Drink, Energy Gel}	0.25
{Ready-to-Eat Salad, Low-Fat Yogurt}	0.15
{Whole Wheat Bread, Peanut Butter}	0.10

Association Rule Output

(Obtained using Python-based analysis after applying minimum support = 20%, minimum confidence = 60%, and lift ≥ 1)

Rule	Support	Confidence	Lift
{Energy Gel} → {Sports Drink}	0.25	1.00	2.5
{Protein Bar} → {Sports Drink}	0.30	0.67	1.68

Answer the following questions based on the above output: **(2+3+3 = 8 Marks)**

- Identify any two high-frequency products and justify your answer using support values.
- Interpret the association rule {Energy Gel} → {Sports Drink} using support and confidence in business terms.
- RetailSphere wants to create a bundle offer. Which of the following pairs would you recommend and why?
  (i) {Protein Bar, Sports Drink}
  (ii) {Whole Wheat Bread, Peanut Butter}
  Use support values to justify your answer.

Q4. As part of *RetailSphere's Data Dilemma: Turning Customer Data into Decisions*, the analytics team builds a customer churn prediction model to identify customers likely to stop purchasing in the near future.

The dataset consists of historical customer behavior variables such as purchase frequency, recency, average order value, return rate, discount usage, and complaint history. Since churned customers form a minority class, the team evaluates models using multiple performance metrics rather than accuracy alone. Five different classification algorithms are trained and validated using the same dataset and preprocessing steps:

- k-Nearest Neighbors
- Naïve Bayes
- Logistic Regression
- Decision Tree
- Random Forest

All models are evaluated on a held-out validation set.

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC
k-Nearest Neighbors	0.82	0.71	0.64	0.67	0.78
Naïve Bayes	0.79	0.66	0.69	0.67	0.81
Logistic Regression	0.85	0.74	0.72	0.73	0.86
Decision Tree	0.83	0.70	0.76	0.73	0.84

Random Forest	**0.88**	**0.79**	**0.81**	**0.80**	**0.91**

Model Performance Matrix
(Derived after running models in Python on the validation dataset)

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC
k-Nearest Neighbors	0.82	0.71	0.64	0.67	0.78
Naïve Bayes	0.79	0.66	0.69	0.67	0.81
Logistic Regression	0.85	0.74	0.72	0.73	0.86
Decision Tree	0.83	0.70	0.76	0.73	0.84
Random Forest	**0.88**	**0.79**	**0.81**	**0.80**	**0.91**

Hyperparameter Tuning Results for the Best Model
(Obtained using grid-based tuning in Python)

Hyperparameter	Values Tested	Selected Value
Number of Trees (n_estimators)	50, 100, 200	**200**
Maximum Depth (max_depth)	5, 10, None	**10**
Minimum Samples per Leaf	1, 5, 10	**5**
Criterion	Gini, Entropy	**Gini**

Post-tuning Performance (Random Forest):

- Accuracy: **0.89**
- Precision: **0.81**
- Recall: **0.83**
- F1-score: **0.82**
- ROC–AUC: **0.92**

Answer the following questions based on the above output: **(3+3+4 = 10 Marks)**

- Critically evaluate the performance matrix and recommend the most suitable model for deployment in RetailSphere's churn prediction system. Support your recommendation by comparing at least two evaluation metrics and explaining the business implications of those metrics.
- Explain why accuracy alone is not sufficient for evaluating churn prediction models in this case.
- Discuss the impact of hyperparameter tuning on the selected model's performance.