BIG DATA IN FINANCE AND BANKING: EXAMINATION: DSA 8504

- **Lecturer: Dr. John Olukuru**
- **Assistant Instructor: Joan Ngugi**

Date: April 2025

Instructions

- There are 4 Questions in total adding up to 100 marks. Answer all the questions.
- Datasets for Question 1 and 4 are attached. Be careful not to mix them up.
- Instructions for submission is under each question highlighted in this **colour.**
- Please zip all your submission documents into one zip folder.
- Check the submission deadline date and time carefully on the e-learning submission deadline section of this exam.
- **All the best. Congratulations for making it this far!**

Question 1(10 MARKS) – PYSPARK IMPLEMENTATION

You have two CSV datasets with the following schemas:

transactions.csv – Transaction details:

- transaction_id (string)
- customer_id (string)
- transaction_date (string in format yyyy-MM-dd)
- amount (double)
- transaction_type (string: 'debit' or 'credit')

customers.csv – Customer information:

- customer_id (string)
- account_open_date (string in format yyyy-MM-dd)
- risk_segment (string: 'low', 'medium', or 'high')

TASK

Using PySpark DataFrame operations, perform the following steps:

- **Load Data**: Read the transactions.csv and customers.csv files into PySpark DataFrames. Convert the appropriate columns to proper data types (e.g., parse date strings to Spark DateType, cast amount to numeric type).
- **Join Datasets**: Join the two DataFrames on the customer_id column (inner join) to combine each transaction with the corresponding customer information.
- **Calculate Average Transaction**: For each customer, calculate the average transaction amount (include both debit and credit transactions in the calculation).
- **Classify Spend Category**: Based on the average transaction amount, label each customer as:

  low spender – if the average amount is less than 100,

  moderate spender – if the average amount is between 100 and 1000 (inclusive),

  high spender – if the average amount is greater than 1000.
- **Prepare Final DataFrame**: Create a DataFrame with the columns: customer_id, risk_segment, avg_transaction_amount, spend_category for each customer.
- **Save as Parquet**: Save the resulting DataFrame in Parquet format.

Submission: Submit Notebook and the dataframe parquet output

Question 2 (20 MARKS) – DESIGNING A REAL-WORLD CREDIT SCORECARD FOR RISK AUTOMATION

You have joined the **Portfolio Risk and Analytics Team** at **Jenga Microfinance as consultant**, a digital-first lender offering short-term loans. The business is scaling quickly, and the leadership team wants to move toward **automated credit decisioning** based on a risk scorecard.

They provide you with a historical dataset of **10,000 disbursed loans**. Each record contains demographic, behavioural, and financial features. Key variables include:

disbursement_date, income, loan_amount, loan_term, credit_history, phone_verified, email_verified, num_of_loans, dependents,employment_status, residence_type, marital_status,defaulted (target)

BUSINESS GOAL

Jenga Microfinance want to approve or reject loans instantly based on a **risk score**, with minimal manual overrides. The scorecard must be interpretable, align with regulations, and be deployable. Your role is to design the scorecard pipeline and recommend a strategy to **integrate it into the loan approval process**, with a proper training/testing framework, feature handling, and risk thresholds.

QUESTIONS

1. Strategic Data Splitting(3mks)

Jenga Microfinance current analysts split data randomly for training/testing. Explain why this is dangerous in a lending environment. What is the best **splitting strategy** based on **disbursement_date**, and explain how it helps improve model reliability.

2. Variable Treatment & Governance (8 marks)

Jenga Microfinance collects over 15 features, including sensitive fields like marital status and education.

a) Propose a logic for:

- Selecting variables using WOE/IV
- Handling categorical variables like **employment_status** or **residence_type**
- Binning continuous variables for modelling

b) List **two features** that you would **exclude** from the model and justify why — either from a **regulatory**, **fairness**, or **predictive** standpoint.

3. **Cutoff Strategy Design (5 marks)**

You've built a scorecard model using logistic regression. Now the business wants to deploy it to auto-approve loans.

a) You observe that: 70% of non-defaulters score above 650.80% of defaulters score below 580. Which **cutoff strategy** would you recommend for **Automatic approval**, **Manual review** and **Rejection of loan** applications based on the scores:

b) What **two types of validation** would you run to **monitor model performance over time** once deployed?

4. Bias, Drift & Ethical Traps (4 marks)

Jenga Microfinance is committed to **responsible AI**. You're asked to audit your scorecard.

a) Identify one potential **source of data leakage or bias** that could affect fairness.

b) Explain how **model drift** might appear in this lending context — and how to detect/respond to it.

Submission: Submit Answers for the above as pdf documents

QUESTION 3 : CREDIT SCORING LOGIC CASE – SCORE SCALING DECISIONS (15 MARKS)

Jenga Microfinance has just completed a logistic regression credit model that outputs **probabilities of default (p)**. Your team is responsible for converting those into a **standardized credit score**, which the bank will use to approve, review, or reject applicants in real time.

The engineering team proposes the following formula:

Score=Offset+Factor·ln(p1−p)

Where:

- p = predicted probability of default
- Odds=(1−p)/p
- **Factor** = controls how much score changes with odds
- **Offset** = aligns the score with business-defined baselines

- PDO = Points to Double the Odds
- Base Odds = odds at the base score (e.g., 20:1 → Score = 600)

1. Score Sensitivity Design (3 marks)

Jenga Microfinance leadership wants a scorecard where **small differences in customer risk are clearly reflected in score differences**.

- Should the team choose a **high PDO (e.g., 70)** or a **low PDO (e.g., 30)**?
- Explain how this choice affects **score sensitivity** and **interpretation**.

2. Offset Alignment (3 marks)

The Head of Risk insists that customers with odds of 20:1 (good:bad) should receive a score of exactly **600**.

- What is the purpose of **Offset** in this case?
- What would happen if Offset were incorrectly set too high or too low?

3. Score Meaning & Customer Profiles (4 marks)

Two customers apply for a loan:

- **Customer A:** $p = 0.4$
- **Customer B:** $p = 0.1$

a. Without computing, who will receive a higher score and why?
b. What does a higher score represent in terms of **default risk and odds**?

4. Cutoff & Policy Trade-offs (5 marks)

Jenga Microfinance is considering the following score thresholds:

- Score > 650 → Auto-approve
- Score 580–650 → Manual review
- Score < 580 → Reject

a. What business risk does **Jenga Microfinance** reduce by auto-approving only customers above 650?
b. What trade-offs might arise from putting **too many** customers into the “manual review” range?
c. How could the wrong choice of PDO or Offset misclassify customers across these thresholds?

Submission: Submit Answers for the above as pdf documents

Question 4 (Case Study- 55 Marks)

Background

You've been brought into the Fraud & AML Intelligence Unit at Upesi Digital Bank, a fast-growing fintech operating across East Africa. Upesi processes thousands of transactions daily. A recent risk audit revealed a number of concerning transaction patterns involving customers that don't appear risky based on traditional scoring methods.

The compliance team suspects the presence of:

- Sleeper accounts activated suddenly for fund movements
- Circular transactions (money sent in loops)
- Rings or clusters of collusion
- Accounts acting as invisible intermediaries in fraudulent flows

Your Mission

Your task is to use network-based thinking to detect unusual or suspicious customer behaviour, especially those acting as high-volume pass-troughs or holding unusual positions in the network.

You've been given:

- ➢ A list of customer profiles
- ➢ A list of transactions between them

Your job is to:

- ✓ Model these as a graph
- ✓ Identify anomalies using network logic and statistics
- ✓ Visualize and explain what you find

Provided Data

- customers.csv One row per customer: customer_id, name, risk_segment, nationality, is_business(1 – Business Account, 0-Individual Account)
- transactions.csv One row per transaction: sender_id, receiver_id, amount, timestamp

Strategic Guidelines (Read Carefully)

- You are expected to model customers as nodes and transactions as edges.
- Do not try to merge customer metadata directly into the transaction table.
- This can lead to duplication, loss of structure, and confusion when building the graph.

Instead:

- ✓ Use the customer file to enrich your nodes
- ✓ Use the transaction file to create your edges

This will help you preserve the relational structure of the network and ensure clean, scalable modeling.

What to Focus On

- Construct a directed graph from the transaction data
- Enrich the nodes (customers) with metadata like nationality and business flag
- Compute graph-based measures (e.g., degree, betweenness, closeness centrality)
- Use statistical measures (e.g., mean + 2×std) to flag anomalies
- Visualize subgraphs of the most suspicious clusters

Technical Stack

You may use:

- Any Graph and Visualization Tools of Your Choice

Expected Deliverables

- A clean, well-commented Jupyter Notebook (.ipynb)
- A DataFrame of flagged customer nodes with reasons
- At least one visual graph showing a suspicious cluster
- Which metrics you used and why
- How you chose your thresholds

Final Tip

Fraud doesn't scream — it whispers. You're looking for accounts that behave **unlike others** in the network. The strength of your project lies in your ability to **translate suspicion into logic**, and **logic into graph features**. You are encouraged to think creatively, challenge assumptions, and explore the **grey areas of suspicious behaviour**. Your responsibility is to flash out suspicious behaviour. Validation of actual Fraud is usually done with the Compliance investigate teams.

Submission: Submit Answers for the above as notebook. Where external visualization tools have been used, output the visuals as separate files or screenshots