Overview

Senior Data Scientist - Technical Assessment

Welcome

Welcome to the **Senior Data Scientist Technical Assessment** at **SunCulture.** This assessment is designed to evaluate your technical expertise, strategic thinking, and ability to translate data into meaningful business outcomes.

We're not only interested in what you can build, but also how you approach complex business problems, structure your solutions, and think about deploying them in real-world environments.

This is your opportunity to demonstrate senior-level thinking across data science, machine learning, and AI strategy.

Assessment Overview

The assessment is structured into three parts:

- **Part 1** — Exploratory & Feature Engineering (Optional/Pre-Task) (if used). Lays the foundation for understanding the dataset and creating derived features for downstream tasks.
- **Part 2** — Predictive Modelling & Business Application. Tests your ability to apply machine learning to segmentation, marketing strategy, risk mitigation, and model deployment.
- **Part 3** — AI & Self-Service Data Platforms. Explores your strategic thinking, curiosity, and vision for applying AI to democratize data access within a business.

Data Source

You have been provided with a **multi-sheet Excel file**: `Senior Data_Scientist_Assessment_Data.xlsx`
This file contains 8 interconnected tables with information on:

- Customers
- Accounts
- Products
- Leads
- Users
- Departments

Tech Stack Expectations

We know many talented folks use different tools — including R, SPSS, Stata or other statistical languages — for analysis. But for this assessment, we're going all in on **Python**.
Most of our data stack, pipelines, and APIs are powered by Python, and we'd like to stay aligned with that.
So, forgive us for our *strict Python requirement* — it's not personal, it's just our stack 😄.
✅ Python only | ❌ R not supported for this assessment.

Submission Guidelines

- **Timeline**: Complete the assessment within 24 hours of receiving it.
- **Deliverables**:
    - A structured Python-based solution (**.ipynb** and/or **.py**)
    - Any lightweight deployment artifacts (e.g., **Dockerfile**, **docker-compose.yml**) where relevant
    - Clear documentation and justifications for all decisions.
- **Evaluation Criteria:**
    - Clarity and depth of reasoning
    - Technical accuracy and rigor
    - Business alignment and strategic thinking
    - Innovation and practicality in deployment/AI integration

Part 1

SunCulture

Part 1: Exploratory Analysis & Executive Insight Generation

This section tests your ability to proactively generate business value by transforming raw data into a clear, actionable story for executive stakeholders.

Task 1.1: Exploratory Data Analysis & Storytelling

Context: You're preparing for a quarterly business review with country managers from Kenya, Uganda, and Côte d'Ivoire. The leadership team wants to understand underlying patterns and opportunities beyond what's visible in our standard PowerBI dashboards.

Your Task: Using the provided relational dataset, conduct an exploratory analysis in a Jupyter notebook that uncovers **4-5 compelling business insights or potential risk patterns**. Focus on telling a data-driven story that could inform strategic decisions.

Requirements:

- **Code Quality:** Your notebook should be well-structured, readable, and demonstrate professional coding practices (not just one-liners).
- **Visual Storytelling:** Choose visualization types that effectively communicate your insights.
- **Business Context:** Frame your findings in terms of business impact (revenue, risk, operations, growth).
- **Actionable Insights:** Each finding should suggest potential actions or further investigation.

Deliverables:

1. **Jupyter Notebook (.ipynb):** A well-structured notebook that documents your analysis journey, including data preparation steps, the analysis, and the final insights.
2. **Visualization:** Include custom visualizations chosen specifically to support your narrative, demonstrating effective data storytelling.

Task 1.2: Executive Summary Presentation

Deliverable: Create a **5-slide PowerPoint presentation** summarizing your key findings for the executive meeting, including speaker notes for each slide.

Part II

Part 2. Predictive Modelling

This section evaluates your ability to apply machine learning to solve real-world marketing and credit strategy challenges, transforming data into direct business actions.

Task 2.1: Customer Segmentation & Profiling

Objective:
Use the available features (including any relevant features created through Feature Engineering) to conduct a comprehensive **customer segmentation analysis**, profiling distinct customer groups in the dataset.

Your Tasks:

1. **Data Preparation & Feature Engineering**: Create any new features you believe are relevant for segmentation. Justify your choices.
2. **Model Selection & Implementation**:
    a. Select and implement an appropriate unsupervised learning algorithm for segmentation.
    b. Justify your choice of algorithm
    c. Detail how you determined the optimal number of clusters.
    d. Discuss the **limitations of the dataset** for this task
3. **Segment Profiling**: Profile the resulting segments in **business terms**. For each segment, describe the typical customer using the available features

Task 2.2: Data-Driven Marketing Strategy

Objective:
Leverage the segmentation model to design a targeted, cost-effective marketing campaign.

Scenario:
The marketing team wants to launch a new premium loan product in Kenya. This product is best suited for stable, mid-to-high-income individuals. They have a limited budget and need to focus their efforts.

Your Tasks

1. From the segments identified in Task 2.1, select the two most promising segments for this premium loan product in Kenya.
2. For each selected segment, propose:

   a. A marketing message tailored to the segment's profile.
   b. A channel strategy for effective engagement.
3. Justify your recommendations by linking the segment's characteristics (from 2.1) to the product's requirements.

Task 2.3: Model Deployment & Lifecycle Management

Objective:
Assess the practical experience in deploying, maintaining, and continuously improving machine learning models in a production environment, including collaboration with DevOps or engineering teams.

Your Task:
In a one-pager, outline your approach to deploying a machine learning model to production, including the following points:

- **Model Packaging & Deployment Choices**
  - Describe how you would package a trained model
  - Highlight your preferred deployment strategy and why.
- **Collaboration with DevOps / Engineering**
  - Explain how you would work with a DevOps or platform team to deploy the model, monitor it, and manage infrastructure.
  - Share, if possible, a sample Dockerfile and docker-compose.yml illustrating your deployment setup for a simple Python-based model.
- **Model Maintenance & Iteration**
  - Describe how you monitor model performance
  - When and how you decide to refactor, retrain, or re-deploy a model.
  - How you would integrate CI/CD practices for model updates.
- **Handling Business Conflicts & Overrides**
  - Explain how you would handle scenarios where human credit assessments contradict model predictions.
  - How would you incorporate such feedback loops into the model improvement process?

Task 2.4: Credit Risk Mitigation Analysis

Objective: Provide data-driven, actionable recommendations to support the Credit Collection Team in mitigating default risk.

Your Tasks: Using only the provided dataset:

- Generate **at least three** actionable insights or recommendations.
- Ensure your recommendations are specific, measurable, and practical to implement.
- Focus on how these actions can improve collection rates and reduce default risk.

Task 2.5: Strategic Proposal for Risk Modeling

Objective: Design a forward-looking strategy to improve Credit Collections Rate by leveraging predictive modeling to identify and proactively manage high-risk accounts.

Your Task:
Draft a **one-to-two page strategic proposal** for a project to build a **Post-Sale Credit Risk Scoring Model** focused on early detection of potential loan repayment defaults.

Your proposal should clearly and concisely address the following sections:

1. **Business Problem**: Briefly articulate the issue of customer default, its operational and financial impact, and why improving collections is critical to business performance.
2. **Proposed Solution**: Describe how a predictive Credit Risk Scoring Model can help identify high-risk customers early, inform tailored collections strategies, and reduce write-offs.
3. **Data & Features**: Identify 3–5 new or enriched data sources and explain why each is critical for improving model accuracy and predictive power.
4. **Methodology**: At a high level, outline the modeling approach and describe how the resulting risk score would be integrated into business workflows
5. **Success Metrics**: Define how the project's business impact would be measured — e.g., improvement in on-time repayment rate, reduction in default rate, lift in collections efficiency, or cost savings.
6. **Next Steps**: Outline immediate POC steps, including data exploration, feature engineering, initial model development, validation, and stakeholder engagement.

Bonus Points (Optional)

- Provide a minimal working example (e.g., GitHub repo or snippet) showing how you would deploy a simple model using Docker and docker-compose.
- Outline how you would integrate model monitoring and alerts.
- Mention specific tools or platforms you've used

Part IIII

Part 3: AI & Self-Service Data Platforms

Objective:

To assess your ability to think strategically about the future of data access and AI integration in SunCulture — particularly how AI can democratize data insights and empower non-technical users through self-service analytics.

Task 4.1: Vision for an AI-Powered Self-Service Data Platform

Scenario:

Our organization envisions a future where business teams — from sales to operations — can access insights, run analyses, and visualize metrics through natural-language interaction, without depending on data teams.

As a senior data scientist, you are expected to contribute thought leadership and help shape this future.

Your Task:

In **1–2 pages**, describe your perspective and potential contribution to such an initiative. Your response should cover:

1. Conceptual Understanding
    a. What does an AI-powered self-service data platform mean to you?
    b. What business problems does it solve, and what challenges might it introduce?
2. Experience & Examples
    a. Have you been involved in any related projects (e.g., AI chat for analytics, natural-language query interfaces, data catalog automation)?
    b. If yes, briefly describe your role and share a link (e.g., GitHub, publication, demo, or architecture diagram).
    c. Proof-of-Concept (POC) Proposal
3. If you were to build a small POC using the dataset from **Task 2**, how would you approach it?
    a. Outline key components
    b. What open-source or cloud tools would you leverage
4. Business Readiness & Adoption
    a. What would it take for a business to successfully adopt such a platform?
    b. What cultural, data-governance, and infrastructure changes are required?
    c. How should success be measured — technically and organizationally?
5. Future Enhancements:

a. How would you evolve the platform over time to improve usability, accuracy, and trust?
b. Any practical thoughts on integrating feedback loops and human-in-the-loop validation?

Bonus Points (Optional)

- Provide a lightweight demo, notebook, or architecture diagram showing how a user could query business data through natural language.
- Discuss ethical considerations (e.g., data privacy, hallucination control, and explainability).