Data Scientist Case Studies

Task 1: Credit Risk Data Scientist Case Study

You are provided with a credit risk dataset containing 15,000 records. Your objective is to analyze the data, develop predictive models, and deliver actionable insights to enhance risk management and inform business strategy.

Dataset file: credit_risk_case.xlsx

Dataset Description

The dataset contains the following columns:

Column	Description
id	Unique identifier
member_id	Member ID
loan_amnt	Loan amount
funded_amnt	Funded amount
funded_amnt_inv	Amount funded by investors
term	Loan term (months)
installment	Monthly installment amount
emp_title	Job title
emp_length	Employment length
annual_inc	Annual income
delinq_2yrs	Number of delinquencies in the last 2 years
open_acc	Number of open accounts
total_acc	Total number of accounts
total_pymnt	Total payment
total_pymnt_inv	Total payment to investors
last_pymnt_amnt	Last payment amount
tot_coll_amt	Total collection amount
tot_cur_bal	Total current balance
total_rev_hi_lim	Total revolving credit limit
default	Target variable (0: non-default, 1: default)

Tasks

1. Analyze the overall statistics and default rate of the dataset.
2. Check for missing or inconsistent data and apply appropriate data cleaning steps.
3. Perform feature engineering to create new variables and enhance model performance.
4. Build a basic machine learning models to predict default (at least 2 model).
5. Identify the most influential factors affecting default (feature importance).
6. Evaluate your model using cross-validation and summarize the results.
7. Visualize your model results and key findings.

Task 2: RAG Case Study

You are provided with a customer feedback dataset from our product's support system, containing around 50,000 entries. Your task is to build an efficient sentiment analysis assistant capable of running on limited hardware resources, enhance both RAG-based and standard (non-RAG) pipelines, improve their performance, and derive actionable insights from the data.

Dataset file: musteriyorumlari.xlsx

Dataset Description

The dataset contains the following columns:

Column	Description
ID	Unique feedback identifier
Score	User rating (1-5)
Title	Feedback title
Feedback	Full feedback text
Timestamp	Submission time

Tasks

1. Topic-Based Retrieval & Sentiment (RAG Baseline)
    - Given a specific input (such as a product feature or keyword), build a basic retrieval-augmented pipeline to find all comments in the Feedback column related to that input. The script should be able to run whenever the input changes according to the user's request.
    - Perform sentiment analysis on the retrieved comments. Generate both categorical and numericalscores for this analysis.

2. Explore and compare RAG-style and non-RAG style pipelines in terms of performance andaccuracy.
3. Optimize your solution for low-resource hardware and large-scale data.
4. Extract meaningful insights and visualize results for business stakeholders.

Additional task — Submission requirements

1. High-Level Architecture & Project Structure

- **Propose a final system design** that supports both retrieval approaches, respects hardware constraints, and integrates your optimizations.
- **Provide a diagram** of your repository tree/structure showing how code, data, models, tests and documentation are organized.

2. GitHub Repository

- Instead of delivering Jupyter notebooks, **share a public or private GitHub repository** containing:
  - Code for topic-based retrieval, both pipelines, optimizations and benchmark scripts.
  - Modularized scripts (notebooks may be included as .ipynb, but all runnable scripts and entry points must be present).
  - README.md with clear run instructions, environment / dependency file (e.g., requirements.txt or environment.yml), and example commands to reproduce experiments.
  - Link to any large artifacts (models, embeddings, datasets) or instructions to regenerate them.

3. Presentation

- Problem statement, approaches, key findings, data insights, recommended architecture.

Good luck!