THE RISE OF INTERNET INFLUENCERS

Use-Case: Influencer Sponsorship

CAN YOU PREDICT WHO GETS THE DEAL?

- You've just been hired as a Data Scientist at DurhamPulse, a fast-growing platform that helps global brands partner with digital influencers.
- Every day, thousands of creators apply for sponsorship deals — but only a few are chosen.
- The Marketing Strategy Team wants to streamline this decision process using data science … they are saying that .. "We can't manually review every creator profile. Can your team help us **predict which influencers are most likely to receive sponsorship deals** — using their profile data, engagement metrics, and content quality?"
- Your insights will:
    - Help the brand partnerships team prioritize high-potential creators
    - Improve sponsorship ROI by targeting data-backed choices
    - Speed up the campaign onboarding pipeline

DATA INFLUENCERS

You've been given a dataset of 5,000 influencers.
Each record includes information such as:

- Number of Followers
- Engagement Rate
- Niche (Fashion, Gaming, Beauty, etc.)
- Growth Rate (recent follower growth)
- Video Quality (SD, HD, or 4K)
- Average Views per Post
- Whether they received a sponsorship deal or not

Deliverables:

Jupyter notebook
+support files (if needed)

Using this dataset, your role is to:

- Understand the data (EDA)
- Train a baseline machine learning model
- Build a deep learning model using MLP
- Apply regularization to improve generalization
- Reflect on how a CNN could help if you had influencer image data

Column Name	Data Type	Example Value	Description
Influencer_ID	Integer	1023	Unique identifier for each influencer
Followers	Integer	145000	Total number of followers on their main platform
Engagement_Rate	Float	0.056	Average engagement (likes/comments per follower) on recent posts
Growth_Rate	Float	0.021	Follower growth rate over the past month (as a percentage)
Niche	Categorical	Fashion, Tech,...	The content category the influencer is most known for
Video_Quality	Categorical	SD, HD, 4K	Typical video resolution used in content creation
Avg_Views_per_Post	Integer	30000	Average number of views per post/video
Sponsorship_Status	Binary	0 (No), 1 (Yes)	Target variable: whether the influencer received a brand sponsorship deal

PART A – Exploratory Data Analysis

A1. Load & Inspect Dataset (0.5 pt)
- Load the CSV
- Show shape, column names, data types, and head()
- Identify and explain any null values or irregularities

A2. Visual & Statistical Insights (0.5 pt)
- Plot the distribution of the target variable:
- Identify one potentially correlated feature with the target and explain

PART B – ML Baseline Model

B1. Data Preparation (1 pt)
- Encode categorical feature(s) and Normalize numerical features
- Split dataset into X_train, X_test, y_train, y_test (80/20)

B2. Build ML Model (1 pt)
- Train 3 different ML classifiers
- Print accuracy, precision, recall, and F1-score
- Plot chart comparing the models, select the one (your baseline) and use in C3

B3. Interpretation (1 pts)
- Plot the confusion matrix
- Explain the matrix in 2–3 lines

PART C – MLP Deep Learning Model

C1. Build MLP with Keras (1 pt)
- Build an MLP with:
  - Input layer
  - Dense(64, relu)
  - Dropout(0.3)
  - Dense(32, relu)
  - Output layer (sigmoid activation for binary classification)
- Use binary_crossentropy as the loss function and accuracy as a metric

C2. Train & Visualize (0.5 pt)
- Train model with validation split (e.g., 20%) for at least 20 epochs
- Plot accuracy and loss curves (train vs validation)

C3. Compare Results (0.5 pt)
- Compare final accuracy/F1 of the MLP model vs the baseline model (Part B)
- In 2–3 lines, explain: "Did deep learning improve results? Why or why not?"

PART D – Regularization & CNN Thought

D1. Regularization in MLP (2 pts)
- Add both:
  - Dropout(0.3)
  - L2 regularization to one or more layers
- Retrain model and plot new curves.

D2. Regularization Effect (0.5 pt)
- Compare the new vs old performance
- In 2–3 lines, explain: "Did regularization reduce overfitting? How do you know?"

D3. CNN Thought Experiment

D 3.1. Imagine you also had image data for your case: (0.5 pt)
- *In 1–2 sentences, describe what kind of image you would use and why.*

D3.2. Propose a simple CNN architecture (code only, no execution required). (0.5 pt)

D 3.3. What Would CNN Detect? (0.5 pt)
- What kind of visual patterns would the Conv2D filters learn?

PART E – Final Report to Marketing Team

Created a decision-support table (e.g., who to sponsor vs. not).

Influencer ID	Predicted Sponsorship	Follower Count	Engagement Rate	Sponsorship Recommendation
INF-045	0.91	85,000	6.8%	Yes
INF-201	0.88	40,000	8.5%	Yes
INF-117	0.63	15,000	3.1%	Maybe (manual review)
INF-309	0.41	30,000	1.8%	No
INF-410	0.25	10,000	2.2%	No

Threshold >=60% < 80% (To review)
>=80% (Yes)