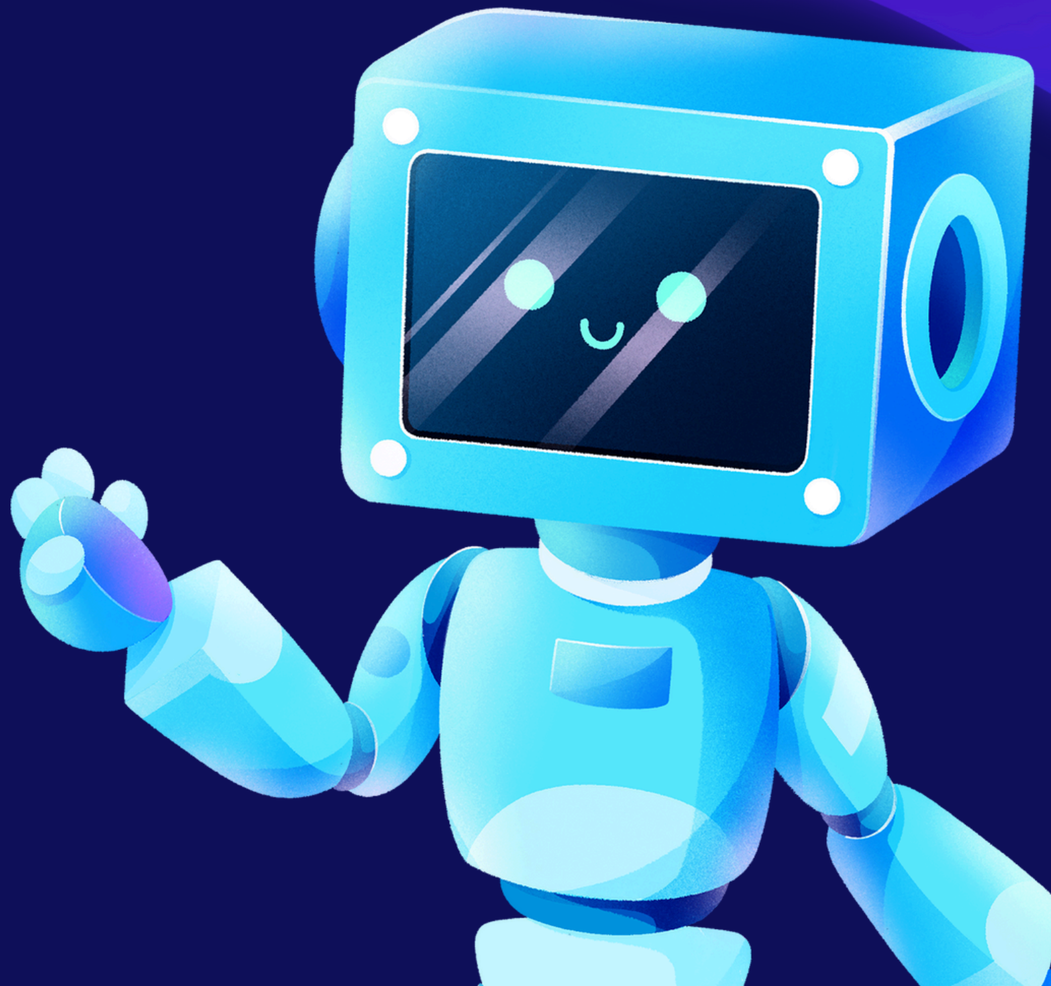
InterUni Datathon 2024

Case Brief



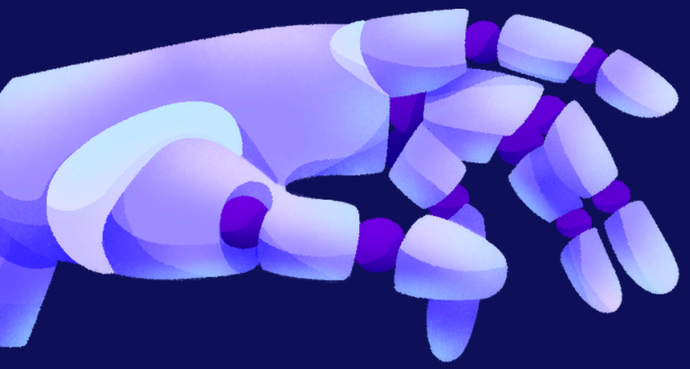
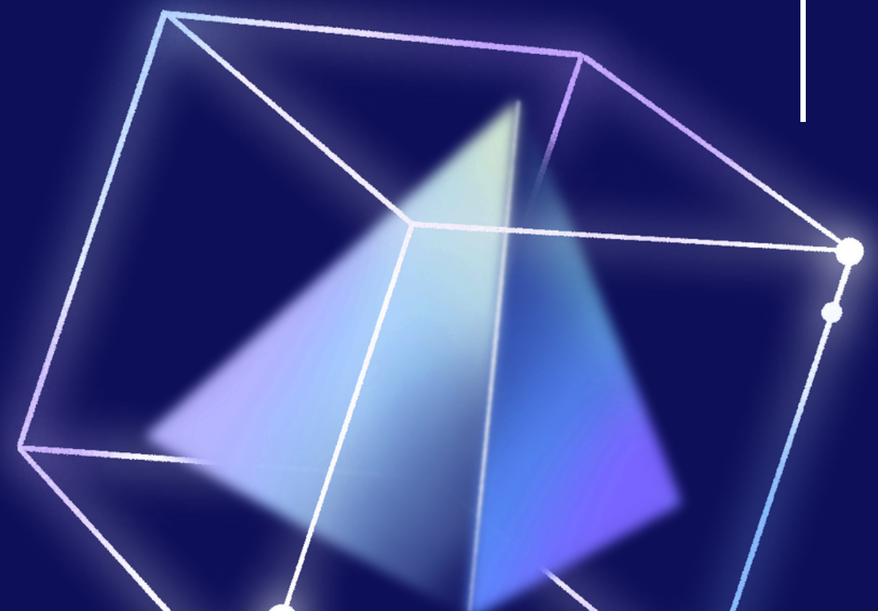


Table of Contents



Instructions

Rules

- 2 rounds of selections
  - 1st round: Ranked Kaggle Competition
  - Finals: 4 mins presentation, 2 mins Q&A, 4 winners

Submission

- Deliverables: submit a kaggle csv and a slide deck
- Submission deadline: **1pm, no submissions allowed after**
- **Submit both PRESENTATION AND KAGGLE**
- ***Submission link provided in case brief***

The Case

You are working as a Data Analyst at Skyline Financial Services (SFS), a leading digital bank that prides itself on using data-driven insights to drive business growth. Recently, the company has experienced a significant and **unexpected dip in monthly revenue**. An internal investigation has revealed a disturbing pattern of fraudulent activity slipping past existing detection systems.

Skyline Financial Services has tasked your team with investigating the root causes of this revenue loss. You have been provided with a dataset containing detailed user information, including demographics, spending behaviours, and transaction history. Your job is to **analyse this data** to not only **identify potential fraud** but also to **understand the underlying reasons** why certain users engage in fraudulent activities.

AIM

Develop a model that accurately identifies fraudulent users while also providing a clear explanation of the factors driving these fraudulent behaviors. Your findings will be crucial in developing new strategies to prevent future fraud and protect the company's revenue. The **targeted stakeholder** for your presentation will be a senior risk management executive who is familiar with data analytics but not necessarily a technical expert.

In your solution, you should consider:

- *How will your model detect fraudulent users in real-time?*
- *What are the key indicators or patterns that suggest fraudulent behavior?*
- *How can your model be used to not only detect but also prevent fraud in the future?*

You may use the following questions to guide you through your work:

- *Describe your model and explain why you chose this approach over other potential models.*
- *Discuss any alternative models you tried and why you believe your chosen model performs better.*
- *How did you handle missing or incomplete data in the dataset?*
- *How did you process and encode categorical data (e.g., 'Occupation' or 'EmailDomain')?*
- *How did you handle unbalanced data?*
- *How did you test and validate your model?*

Kaggle Competition

SUBMISSION ITEM 1

Melbourne Kaggle Competition: *This competition is to be used by only Melbourne Students.*

- https://www.kaggle.com/competitions/inter-uni-datathon-2024-vic

Sydney Kaggle Competition: *This competition is to be used by only Sydney Students.*

- https://www.kaggle.com/competitions/inter-uni-datathon-2024-nsw

Please submit your results on Kaggle before **1pm Sunday 15th September**. *This is a hard deadline so make sure to be on time!*

The datasets will be the exact same but we will take the top 3 from each state for the final judging.

Presentation Submission

SUBMISSION ITEM 2

Google Slides/Powerpoint to be submitted to the link before 1pm Sunday 15th September. *This is a hard deadline so make sure to be on time!*

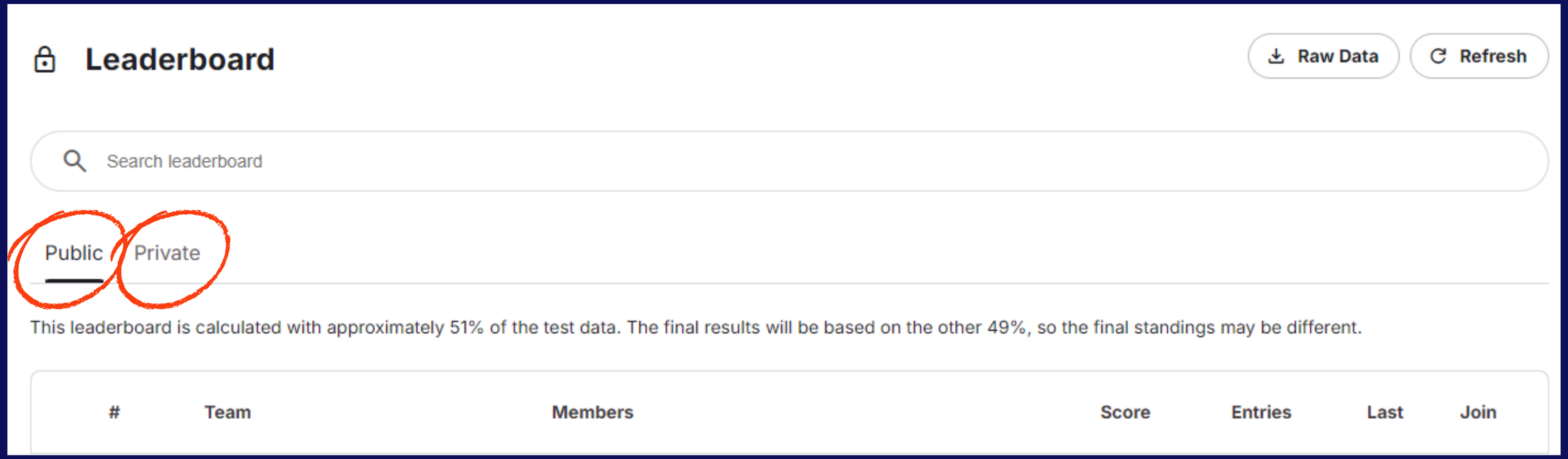
- https://forms.gle/6o3x598ENg5CSYkL8

Data Dictionary

Type	Name	Description
Input Variables	UserID	Unique identifier for each user who made the transaction
Input Variables	Age	Age of users in years
Input Variables	Terrorism	A binary flag with 1 indiciating potential links to terrorism-related activities. 0 indicates no links.
Input Variables	Income	Users income
Input Variables	Email	The user’s email address
Input Variables	Occupation	User’s occupation
Input Variables	EducationLevel	User’s highest level of education
Input Variables	MaritalStatus	Marital status of a user
Input Variables	NumDependents	Number of dependents the user has

Type	Name	Description
Input Variables	MerchantID	A unique identifier for the merchant or business
Input Variables	IsFraud	Target variable, 1 indicates fraudulent transaction, 0 if otherwise
Input Variables	GiftsTransaction	The user’s gift spending amount
Input Variables	TransactionNumber	Unique identifier for each transaction
Input Variables	TransactionType	The type of transaction
Input Variables	TransactionDate	Date when a transaction occurred
Input Variables	TransactionTime	The time when the transaction took place.
Input Variables	TransactionAmount	The monetary value of the transaction
Input Variables	TransactionLocation	Geographic location where transaction occurred

Type	Name	Description
Input Variables	UserTenure	Length of time user has held an account with the bank in months
Input Variables	Expenditure	The user’s spending or expenditure amount
Input Variables	DeviceType	The type of device used to complete the transaction
Input Variables	Latitude	The latitude coordinate of the user’s local address (Not TransactionLocation)



PUBLIC DASHBOARD

- DISPLAYS THE SCORE OF 51% OF THE TEST DATASET.
- VISIBLE TO EVERYONE IN THE COMPETITION.
- SHOWS RANKINGS BASED ON 51% PORTION OF THE TEST SET.
- WHEN USERS SUBMIT PREDICTIONS, THEY CAN ONLY SEE THEIR PUBLIC DASHBOARD SCORE.

PRIVATE DASHBOARD

- DISPLAYS THE SCORE OF THE REMAINING 49% OF THE TEST DATASET.
- HIDDEN DURING THE COMPETITION TO PREVENT OVERFITTING.
- FINAL RANKINGS ARE DETERMINED BY THE PRIVATE DASHBOARD SCORE AFTER THE COMPETITION ENDS.

Marking Rubric

Note

Top 6 teams (3 from Sydney and 3 from Melbourne) will be selected in the Initial Round Selection criteria.

The final winners are selected via our judging panel by using the Judging Criteria.

Initial Round Selection Criteria

Storytelling and Context

Creates a compelling and interesting story that flows with the analysis in the slide deck

Visualisations

Examines the effective use of visuals and supporting materials to enhance the presentation.

Results

The final mark will be based on the accuracy score calculated from the Kaggle evaluation.

Cleanliness of Code/Notebooks

We are looking for the following:

- *Concise Markdowns*
- *Clean Code*
- *Good Comments*

Judges Criteria

Presentation Quality

Presents concise visuals with clear delivery and messaging throughout the presentation, making it engaging and easy to follow.

Prediction

Uses insights derived from analysis to predict which of the ten customers from the test set will fraud based on their attributes, supporting the predictions with effective quantitative and/or qualitative reasoning.

Insights

Conclusions are clearly communicated and interpreted, effectively addressing the problem statement and answering the problem or question.

Analysis

Using clear quantitative analysis as evidence to support theories, along with effective visualizations to derive insights and efficient, relevant use of datasets to support answers, is essential.


Best of Luck!!!

DISCORD QR CODE