CoGrammar

Welcome to this session:

Skills Bootcamp - Tutorial

The session will start shortly...

Questions? Drop them in the chat.
We'll have dedicated moderators answering questions.

Skills Bootcamp Data Science Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. We will be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

CoGrammar

Skills Bootcamp Data Science Housekeeping

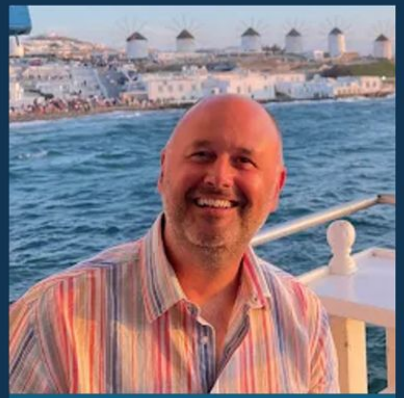
- For all **non-academic questions**, please submit a query: ***www.hyperiondev.com/support***
- **Report a safeguarding incident: *www.hyperiondev.com/safeguardreporting***
- We would love your feedback on lectures: ***Feedback on Lectures.***
- Find all the lecture **content** in your **Lecture Backpack** on GitHub.
- If you are hearing impaired, kindly use your computer's function through Google chrome to enable captions.

CoGrammar

Safeguarding & Welfare

We are committed to all our students and staff feeling safe and happy; we want to make sure there is always someone you can turn to if you are worried about anything.

If you are feeling upset or unsafe, are worried about a friend, student or family member, or you feel like something isn't right, speak to our safeguarding team:



Ian Wyles
Designated Safeguarding Lead

Simone Botes

Nurhaan Snyman

Rafiq Manan

Ronald Munodawafa

Tevin Pitts

Scan to report a safeguarding concern

or email the Designated Safeguarding Lead:
Ian Wyles
safeguarding@hyperiondev.com

CoGrammar HyperionDev

Skills Bootcamp Progression Overview

✅ Criterion 1 - Initial Requirements

Specific achievements within the first two weeks of the program.

To meet this criterion, students need to, by no later than **01 December 2024 (C11)** or **22 December 2024 (C12):**

- **Guided Learning Hours** (GLH)**:** Attend a minimum of 7-8 GLH per week (lectures, workshops, or mentor calls) for a total minimum of **15 GLH**.
- **Task Completion:** Successfully complete the **first 4 of the assigned tasks**.

✅ Criterion 2 - Mid-Course Progress

Progress through the successful completion of tasks within the first half of the program.

To meet this criterion, students should, by no later than **12 January 2025 (C11)** or **02 February 2025 (C12):**

- **Guided Learning Hours** (GLH)**:** Complete at least **60 GLH**.
- **Task Completion :** Successfully complete the **first 13 of the assigned tasks**.

CoGrammar

Skills Bootcamp Progression Overview

✅ Criterion 3 – End-Course Progress

Showcasing students' progress nearing the completion of the course.

To meet this criterion, students should:

- **Guided Learning Hours** (GLH)**:** Complete the **total minimum required GLH,** by the **support end date.**
- **Task Completion : Complete all mandatory tasks**, including any necessary resubmissions, by the end of the bootcamp, **09 March 2025 (C11)** or **30 March 2025 (C12).**

✅ Criterion 4 - Employability

Demonstrating progress to find employment.

To meet this criterion, students should:

- **Record an Interview Invite:** Students are required to record proof of invitation to an interview by **30 March 2025 (C11)** or **04 May 2025 (C12)**.
  - **South Holland Students** are required to proof and interview by **17 March 2025.**
- **Record a Final Job Outcome :** Within 12 weeks post-graduation, students are required to record a job outcome.

CoGrammar

Learning Outcomes

- Engineer relevant features from procurement data, including temporal variables (Year & Month)
- Apply PCA for dimensionality reduction, improving model performance and efficiency
- Implement One-Class SVM for anomaly detection to flag suspicious procurement patterns
- Handle imbalanced data using SMOTE to improve model predictions
- Use GridSearchCV to optimize hyperparameters for better model accuracy
- Evaluate model performance using a Confusion Matrix, understanding false positives/negatives

Lecture Overview

- Data Science for Social Good
- Build project

CoGrammar

What is the primary goal of using data science in government procurement analysis?

A. To increase government spending
B. To detect patterns and anomalies that may indicate corruption
C. To reduce the number of government contracts
D. To make procurement processes more complex

CoGrammar

What is the primary goal of using data science in government procurement analysis?

A. To increase government spending
B. **To detect patterns and anomalies that may indicate corruption**
C. To reduce the number of government contracts
D. To make procurement processes more complex

CoGrammar

Which of the following is a common indicator of corruption in procurement?

A. Many suppliers competing for contracts
B. Repeated awards to the same supplier under unclear conditions
C. Short contract durations and low prices
D. Public availability of contract details

CoGrammar

Which of the following is a common indicator of corruption in procurement?

A. Many suppliers competing for contracts
B. **Repeated awards to the same supplier under unclear conditions**
C. Short contract durations and low prices
D. Public availability of contract details

CoGrammar

What is an anomaly detection method commonly used in fraud detection?

A. Linear Regression
B. Isolation Forest
C. K-Means Clustering
D. Decision Trees

CoGrammar

What is an anomaly detection method commonly used in fraud detection?

A. Linear Regression
B. **Isolation Forest**
C. K-Means Clustering
D. Decision Trees

CoGrammar

Project

- Detecting Corruption in Government Procurement Using Data Science

CoGrammar

Problem Statement

- Why is this important?
  - Government procurement is prone to corruption due to large contract values and lack of transparency.
  - Irregularities such as overpricing, bid rigging, and favoritism are common.
  - Detecting corruption is challenging due to hidden patterns in the data.

Data Pipeline & Dataset

- Data Generation: Since real data is difficult to obtain, we created a synthetic dataset simulating procurement transactions.
- Key Data Features:
  - Contract value, number of bidders, contract duration
  - Country, supplier, risk score
  - Award date (used to derive trends)

CoGrammar

Data Science Approach

- How do we detect corruption risks?
  - Anomaly Detection:
    - Isolation Forest & One-Class SVM flag suspicious transactions.
  - Risk Classification:
    - A Random Forest model predicts corruption risk levels (Low, Medium, High).
  - Feature Importance Analysis:
    - Identifies key risk indicators in procurement.

Model Development & Evaluation

- Preprocessing: Encoding, scaling, and PCA for dimensionality reduction.
- Handling Class Imbalance: SMOTE to balance high-risk vs. low-risk cases.
- Hyperparameter Tuning: GridSearchCV for optimizing model performance.
- Evaluation Metrics:
  - Accuracy, classification report, confusion matrix visualization.

CoGrammar

Let’s Code

CoGrammar

Let’s Breathe!

Let’s take a small break before moving on to the next topic.

CoGrammar

Why is SMOTE used in data preprocessing for corruption detection?

A. To remove duplicate records in the dataset
B. To balance the dataset by oversampling minority classes
C. To improve dimensionality reduction with PCA
D. To eliminate anomalies before training the model

CoGrammar

Why is SMOTE used in data preprocessing for corruption detection?

A. To remove duplicate records in the dataset
B. To balance the dataset by oversampling minority classes
C. To improve dimensionality reduction with PCA
D. To eliminate anomalies before training the model

CoGrammar

Which machine learning model is commonly used for classification in corruption risk prediction?

A. K-Means Clustering
B. Random Forest
C. Principal Component Analysis (PCA)
D. Linear Regression

CoGrammar

Which machine learning model is commonly used for classification in corruption risk prediction?

A. K-Means Clustering
B. Random Forest
C. Principal Component Analysis (PCA)
D. Linear Regression

CoGrammar

In feature importance analysis, which feature might strongly indicate procurement fraud?

A. The number of bidders
B. The length of the contract description
C. The supplier's logo color
D. The time of day when the contract was signed

CoGrammar

In feature importance analysis, which feature might strongly indicate procurement fraud?

A. **The number of bidders**
B. The length of the contract description
C. The supplier's logo color
D. The time of day when the contract was signed

CoGrammar

How does Isolation Forest detect anomalies in procurement data?

A. It clusters data points and assigns labels to normal and fraudulent transactions
B. It isolates anomalies by randomly partitioning data until outliers stand alone
C. It uses deep learning to classify fraud cases
D. It predicts future contract values based on past data

CoGrammar

How does Isolation Forest detect anomalies in procurement data?

A. It clusters data points and assigns labels to normal and fraudulent transactions
B. **It isolates anomalies by randomly partitioning data until outliers stand alone**
C. It uses deep learning to classify fraud cases
D. It predicts future contract values based on past data

CoGrammar

What is a key limitation when using synthetic data for corruption detection?

A. It is too large to analyze effectively
B. It may not fully capture real-world corruption patterns
C. It always leads to biased results
D. It requires government approval before use

CoGrammar

What is a key limitation when using synthetic data for corruption detection?

A. It is too large to analyze effectively
B. **It may not fully capture real-world corruption patterns**
C. It always leads to biased results
D. It requires government approval before use

CoGrammar

Summary

- ★ Machine learning can identify irregular spending patterns in procurement.
- ★ Feature engineering enhances model performance.
- ★ PCA & anomaly detection help reduce noise and highlight corruption risks.
- ★ Ethical AI is crucial for fair, transparent data usage.

CoGrammar

Q & A SECTION

Please use this time to ask any questions relating to the topic, should you have any.

Thank you for attending

CoGrammar

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education