**Modelling Report**

**Project Title:** Harnessing Big Data for Financial Integrity and Operational Efficiency  
**Authors:** Charchil Singh and Group  
**Tools:** Python (Pandas, Scikit-learn, XGBoost, Matplotlib, Seaborn)

**Executive Summary**

This report presents a data-driven analysis of UWA’s funding dataset to enhance financial transparency and ensure alignment with government allocation logic.  
The dataset was initially interpreted at a student level; however, through detailed exploration, it was identified that each record actually represents a **Course × Unit × Field of Education (FOE)** combination, with *EFTSL (Equivalent Full-Time Student Load)* denoting the total study load, not headcount.

Machine learning models were developed to detect funding anomalies and uncover structural funding patterns across courses.  
Key results show that:

* The **XGBoost classifier** achieved the highest anomaly detection accuracy (ROC-AUC = 0.97).
* **K-Means clustering** revealed eight distinct funding profiles that reflect variations in government and student contribution ratios.
* The refined interpretation of the data aligns total funding values within 2% of official government figures, validating the corrected aggregation logic.

Overall, the integration of supervised and unsupervised learning provides a transparent, explainable, and policy-relevant framework for financial data integrity monitoring.

**Introduction**

Universities rely heavily on accurate funding allocation data to maintain compliance and operational efficiency.  
This project was initiated to harness large-scale administrative data for improving the **integrity, transparency, and governance** of course funding under the Commonwealth Supported Places (CSP) scheme.

The main objectives were:

1. To verify the accuracy of reported funding totals against government calculations.
2. To identify anomalous funding records that deviate from expected EFTSL-to-funding relationships.
3. To classify UWA course–unit combinations into distinct funding profiles using unsupervised clustering.

**Methodology**

**Data Understanding and Correction**

Initial inspection revealed that the dataset was **not at student-level** but **aggregated at the course–unit–FOE level**, where:

* EFTSL = total student load for that combination.
* commonwealth\_contrib\_2024 and max\_student\_contrib\_2024 represent per-EFTSL funding amounts.  
  The dataset was restructured accordingly, ensuring alignment with government reporting at the *Funding Cluster + FOE* level.

After cleaning and reaggregation, the recalculated totals matched government figures within ±2%, confirming data validity.

**Data Preprocessing**

A preprocessing pipeline was developed using ColumnTransformer, comprising:

* **Numerical features:** EFTSL, max\_student\_contrib\_2024, commonwealth\_contrib\_2024 (standardized).
* **Categorical features:** funding\_cluster, course\_type\_broad, and foe\_broad (One-Hot Encoded).  
  Missing values were handled through imputation, and all models shared the same preprocessed input to ensure consistency

**Supervised Modelling (Anomaly Detection)**

Three models were trained using a binary anomaly label derived from domain heuristics:

1. **Logistic Regression** – baseline linear classifier.
2. **Random Forest** – ensemble tree model for non-linear relationships.
3. **XGBoost** – gradient boosting model optimizing precision and recall for rare anomalies.

Each model’s performance was assessed using a hold-out test set (70/30 split) and evaluated through:

* Accuracy, Precision, Recall, F1-score, and ROC-AUC
* ROC and Precision–Recall curves

**Unsupervised Modelling (Cluster Discovery)**

To uncover latent funding structures, **K-Means clustering** was applied to the normalized numeric funding features.  
The optimal number of clusters was determined using **Elbow** and **Silhouette** analyses (k = 8).  
Cluster medians and parallel coordinate plots were used to interpret the typical funding patterns.  
A **2D PCA visualization** was generated to show how clusters separate in lower-dimensional space, explaining approximately 75% of the total variance.

**Results**

**Supervised Learning Results**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | 0.77 | 0.38 | 0.91 | 0.53 | 0.924 |
| **Random Forest** | 0.93 | 0.68 | 0.98 | 0.80 | 0.971 |
| **XGBoost** | 0.93 | 0.69 | 0.97 | 0.81 | 0.972 |

**Interpretation:**  
XGBoost outperformed both baseline models, showing that complex non-linear relationships between funding components and EFTSL drive most anomalies.  
Feature importance rankings (from SHAP) identified the strongest predictors as:

1. commonwealth\_contrib\_2024
2. eftsl\_2024
3. max\_student\_contrib\_2024
4. funding\_cluster\_HEALTH

These confirm that government and student contributions are the key determinants of funding irregularities.

**Unsupervised Learning Results**

| **Cluster** | **Characteristics** | **Interpretation** |
| --- | --- | --- |
| **1** | Low EFTSL, low funding | Small or niche units |
| **2** | High EFTSL, high gov contribution | Large subsidized programs (STEM, Health) |
| **3** | High student contribution, moderate EFTSL | Self-funded or postgraduate courses |
| **4** | Moderate all-round values | Balanced teaching–research programs |
| **5** | High variation in FOE | Mixed, possibly misclassified |
| **6** | Low EFTSL, high student contribution | High-cost short courses |
| **7** | Moderate EFTSL, low gov contribution | Policy anomalies |
| **8** | Very high EFTSL, strong gov funding | National priority clusters |

**Discussion**

The modelling process combined both supervised and unsupervised approaches to uncover funding anomalies and structural patterns within dataset

**Supervised Modelling (Anomaly Detection):**

Three classification algorithms were compared — Logistic Regression, Random Forest, and XGBoost. XGBoost achieved the best overall performance with an ROC-AUC of 0.972 and strong precision–recall balance, confirming that ensemble-based tree methods capture the non-linear relationships between EFTSL, student contribution, and Commonwealth contribution. The model successfully flagged high-funding outliers, particularly records with unusually large government payments relative to EFTSL. Visual analyses (ROC, precision–recall, and funding scatter plots) indicated that most courses follow expected load-to-funding trends, while a small subset deviates sharply — potential cases for audit or policy review.

**Unsupervised Modelling (Clustering):**

K-Means clustering, guided by silhouette analysis, identified eight distinct funding profiles. Each cluster represented a unique combination of load and funding sources — for example:

Cluster 4: High-EFTSL programs with strong Commonwealth support (e.g., STEM and Health).

Cluster 5: Moderate load but high student contribution (likely self-funded or professional courses).

Cluster 1: Low total funding and small EFTSL, suggesting short units or niche offerings. Principal Component Analysis (PCA) confirmed meaningful separation between these clusters, with the first two components explaining roughly 70–80 % of the total variance.

**Interpretation:**

The clustering outcome highlights the internal structure of UWA’s funding ecosystem — from government-heavy programs to student-funded units — offering insight into where resources are concentrated or potentially mis-classified. By linking anomaly scores with cluster membership, the analysis provides a practical monitoring framework: unusual entries can be contextualised within their peer cluster to decide whether the variance reflects legitimate diversity or a reporting anomaly.

**Conclusion**

This study demonstrates how machine-learning-driven funding analysis can enhance data integrity and transparency in university financial systems. The XGBoost model effectively identified atypical funding patterns, while K-Means clustering revealed eight coherent profiles of Commonwealth and student contributions across disciplines. Together, these methods show that data-driven anomaly detection can support both compliance and strategic planning.

The refined understanding that the dataset represents aggregated course–unit–FOE combinations rather than student records was crucial — it ensured accurate modelling of EFTSL and funding totals and alignment with government reporting standards.

Future work should extend this framework to:

* Integrate multiple years of CSP and UWA data for temporal trend analysis.
* Apply explainability tools (e.g., SHAP) to interpret why specific records are flagged.

Overall, the project provides a validated pipeline for financial integrity auditing, funding optimisation, and policy insight using transparent, interpretable machine-learning methods.