Table of Content

Background of Study

This study aims to enhance student outcomes for “The Study Group”, a key educational provider that partners with universities to support international students. Despite the comprehensive services being offered, there is no system to identify students at risk of failing, withdrawing, or not progressing. The Study Group seeks to device an early warning system through data analysis to identify at-risk students, enabling timely interventions. This analysis explores patterns and markers for the following business questions.

Is it possible to identify at-risk students before they fail or withdraw? What variables and thresholds serve as markers for identifying at-risk students? Are there stages or milestones of the learning journey where there’s an increase in the number of learners battling?

Project Development Process

Data Collection: I initially worked with the Module data and Course data provided by The Study Group. In subsequent meetings, Language Proficiency data was also made available.

Data Cleaning: Using the three datasets mentioned, I employed Python to clean the data by removing records for students without a defined progression status, as their status could not be determined from the available data. I also corrected the university names that were not properly written e.g “University of Strathclyde” was written as “Strathclyde University”. I opted for Python because it offers a flexible, and scalable environment for data analysis and machine learning. Unlike Excel, which is great for smaller datasets and basic analysis. Compared to R, python offers more extensive ecosystem for visualisation with libraries like Matplotlib and Seaborn.

Additionally, I used Microsoft Excel to clean the Language Proficiency data and convert the various exam types to the CEFR standard because the conversion scale for the various exams was gotten from over 6 different websites, so Microsoft Excel appeared to be most effective for the cleaning.

Data Wrangling: I used the IF function in python to group the numerous degrees chosen by students into broader categories, such as degrees that have key words like Architecture and Engineering were grouped into "Architecture and Engineering," those with Art and Humanities to "Humanities" , those with Management, Social, Finance, Ecomomics to "Management and Social Sciences," those with Mathematics, physics, chemistry, Biology, computer, technology to "Science and Technology," and the rest to "Others" (mostly Legal studies) because there were over 2617 Unique degrees. Universities were also categorised based on their global rankings from QS University Global Ranking. Universities that ranked between 1 to 200 were grouped as High ranked, 200 to 500 as Mid ranked and 500 to 1000 as low ranked, those not on the list, as not ranked and Not partnered.The Study Group centres were also grouped based on their location, South England, North England, Ireland, Scotland, Midlands and others(online, Holland).

Exploratory Data Analysis (EDA): I conducted both univariate and multivariate analyses on the datasets to identify trends, relationships and extract insights. I used Matplotlib to explore the Categorical variables, and Seaborn to explore the Numeric Variables and Stacked Barplot to explore the relationship between each variable and

student's outcome. I chose Matplotlib because It integrates seamlessly with other libraries like Seaborn, allowing me to create complex plots with ease.

Data Visualizations:

- **Bar Chart:** A label Bar Chart was used to visualize percentage distribution of all the categorical variables in both Datasets because bar charts provide easy comparison between variables. For example, the distribution of students across different The Study Group centres, Academic years, Nationality, Continents, Progression University, Progression Degree, Gender and Outcome. Each unique variable was ascribed a unique colour for easy visualisation at glance.
- **Histogram Boxplot:** A histogram boxplot was used to visualize the distribution of the numerical variables in the datasets. Histogram boxplot makes it easy to visualize the mean and mode. For example, the Average and mode across students for PresentCount, AbsentCounts, Authorized Absent, CompletedModules, CreditWeigtedAverage, NumberOfModules was easily visualized.
- **StackedBarplot:** This was used to establish the relationship between each variable and student's outcome because it provides the numbers and percentage at quick glance into the dataset. For example, student's outcome based on Gender, Nationality, Centres, degree category, course completion, Present count, credeitweightedaverage.

Machine Learning: I developed three machine learning models; Decision Tree Classifier, Random Forest Classifier, and AdaBoost Classifier. Among these, the Random Forest Classifier demonstrated the best performance.

Technical Overview of the code

Rationale Behind Choice of Tools

- **Programming Language and Environment:** I opted for Python because it offers a powerful, flexible, and scalable environment for data analysis and machine learning. Unlike Excel, which is great for smaller datasets and basic analysis, Python handled the 3 datasets efficiently and allowed for complex workflows. Compared to R, Python has a more extensive ecosystem for machine learning, with libraries like scikit-learn. Python also provides

seamless integration with various data manipulation libraries like pandas, visualization tools like Matplotlib and Seaborn. This made Python ideal for this project, from data cleaning to model building.

- **Libraries used:** The following are the libraries used in the analysis.
    - **Pandas** for data loading and manipulation due to its powerful data structures and ability to handle large datasets efficiently, this library enabled me to upload the course dataset and Language Proficiency dataset to python and merge them together for further analysis.
    - **Matplotlib:** Matplotlib enabled me to build a helper function called `labelled_barplot`, which allowed for the easy visualization of well-labeled percentage distributions of categorical variables on bar charts. Its flexibility also made it simple to integrate with Seaborn, which I used to plot both a boxplot and histogram on the same scale. This combination provided clear, informative visualizations that supported my analysis effectively.
    - **Seaborn:** I used Seaborn to create boxplots and histograms that were not only pleasing to the eyes but also provided deeper insights through features like KDE (Kernel Density Estimation) and the automatic handling of statistical summaries. Its integration with Matplotlib allowed me to maintain consistency across different plots making it a powerful tool for my analysis.
    - **Scikit-learn:** I used Scikit-learn because it provided a comprehensive suite of tools for machine learning, including classifiers like Decision Tree, Random Forest, and AdaBoost, which were essential for building and evaluating my models. Additionally, Scikit-learn offered a variety of metrics such as the confusion matrix and accuracy score, enabling me to easily compare the performance of different models.
- **Microsoft Excel:** This was adopted for preprocessing Language Proficiency data, leveraging its user-friendly interface for manual data transformations.

Data Wrangling Technique:

University Grouping: In order to streamline the universities chosen by the students and make it fit for EDA and machine learning they were grouped into High Ranked, Mid Ranked, Low Ranked, Not Ranked and Not Partnered based on **QS word**

University Ranking. University ranked among 1 - 200 were grouped as high ranked, 200 to 500 were group as Mid - Ranked, Over 500 were grouped as low ranked. This grouping assisted a lot during the EDA as it enabled me to understand how students chose their universities.

The Study Group Centres Grouping: The Centres were grouped into North England, South England, Midlands, Scotland, Ireland and Others. This helped to easily visualize the performance of students based on centres.

Exploratory Data Analysis:

Barplot: Label Barplot was used to determine the percentage distribution of all the categorical data in both module and course dataset. The categorical data include Centres, Course level, Season, Academic year, Module Outcome, Gender, Continent, status, University Ranking, Degree category.

Boxplot: I used histogram boxplot to explore the distribution of numerical variable mostly the attendance data (present count, absent count, etc) and score data (CreditweightedAverage,TotalModules, etc) in order to determine the average and modal value of each variable.

Stack Barplot: I used the stacked barplot to establish relationship between each of the variable and the Risk Status. Like Gender Vs Risk Status, Centre Vs Risk Status, Continent Vs Risk Status, Degree vs Risk Status, English Proficiency Vs Risk Status.

Pattern, Trends and Insights:

The analysis of student data revealed significant patterns in enrollment, performance, and risk factors that can inform strategies to identify and support at-risk students. Centers like Sheffield, Durham, Sussex, and Leeds have the highest enrollments, with the majority of students in foundation courses. Despite high pass rates, nearly 20% of students needed to retake exams. The most at-risk groups include males, students in International Year 1, those with beginner English proficiency, and students from Asia and Europe. Particularly concerning are the Midlands and South England centers, where over 40% of students are at risk. Low attendance and credit averages also correlate strongly with increased risk markers.

Key markers for identifying at-risk students include gender, course level, region, and English proficiency. For instance, 65% of students who withdrew and 85% of those terminated were male. Additionally, Asian students accounted for 90% of withdrawals and 97% of terminations. The data analysis also highlighted specific risk periods, such as the summer where students appear to have failed more. This recurred in 2020, 2021 and 2022. Identifying students who are likely to pass but not progress to their desired degree may involve focusing on those with marginal scores, low attendance, and specific course levels.

Interventions should target high-risk centers like Teesside, Huddersfield, Kingston, and London. Special support is needed for students with low attendance and poor English proficiency, especially those from Asia. Enhanced monitoring of students in Humanities and Legal Studies, as well as those in foundation courses, is essential. The analysis suggests that early, targeted interventions, such as academic support, attendance monitoring, and English language classes, could significantly reduce withdrawal and termination rates, thereby improving overall student success.