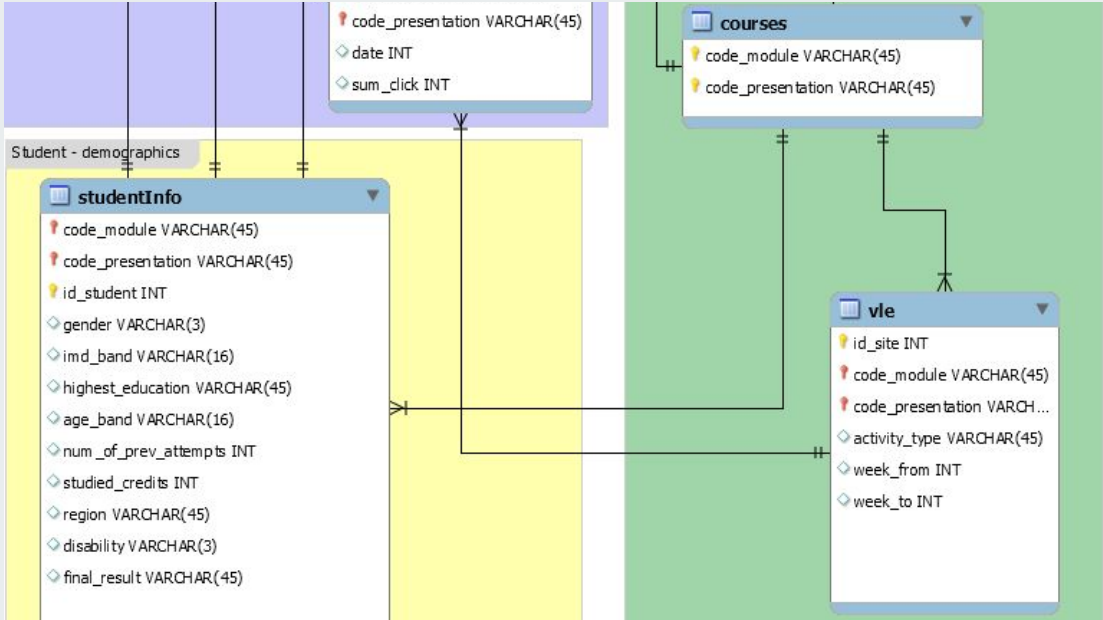
Project 4: Analytics of Open University

Group 6: Hail Nijo, Tyson Horeswell, Patric Beaven, Vijay Mani



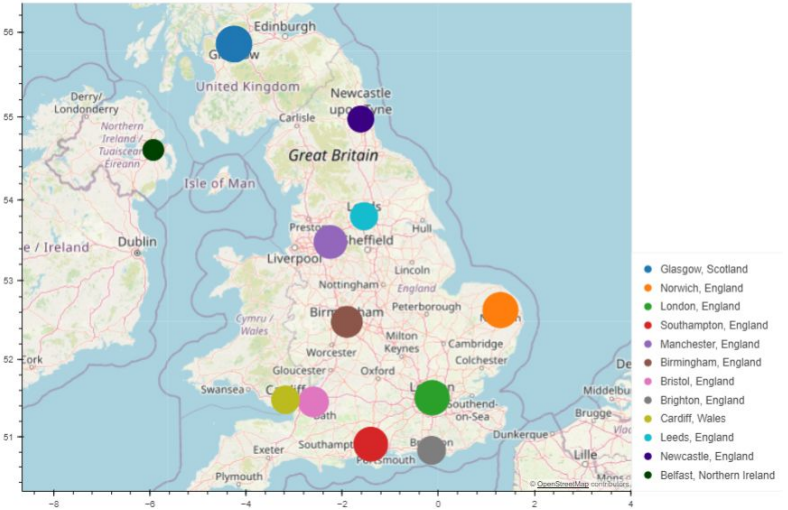
About Dataset

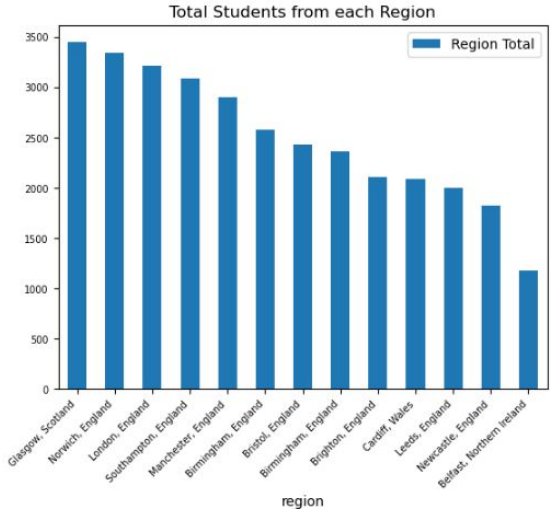
- Dataset contains 7 csv's which shows data about students, courses and their usage of the Virtual Learning Environment(VLE)
- Contains data from Student intake periods February and October in the years 2013 and 2014.
- Has data for 32,593 instances of students joining a course. 27,295 being completely unique students
- **Data Source:** Kaggle
  Title: Open University Learning Analytics
  URL: https://www.kaggle.com/datasets/mexwell/open-university-learning-analytics/data

Technologies Used in the Project

- SQLite - Portable database allowing everyone to use the same database small and fast
- Jupyter Notebooks - Easy to code, test and iterate
- SQLAlchemy - Connect to SQLite to create and query the database
- Pandas and Matplotlib - Analysis and Visualization
- SkLearn and Tensorflow - Machine learning
- Streamlit & Pandasai - App development

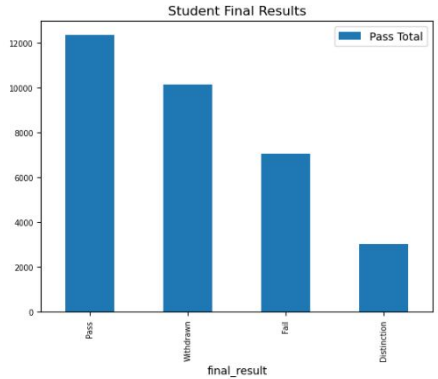
General Overview (Geographic)

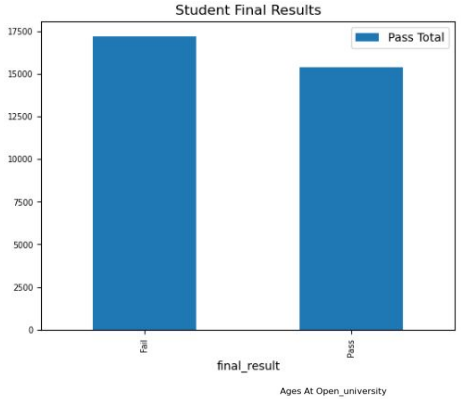




- Region with the most students was Scotland.
- Ireland had the least.

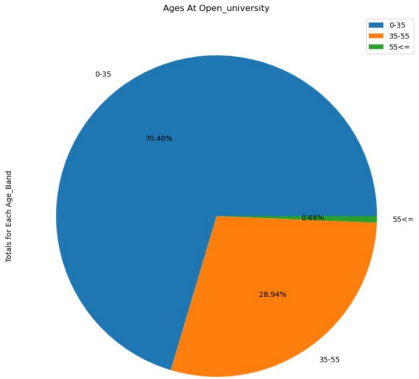
General Overview (Students)

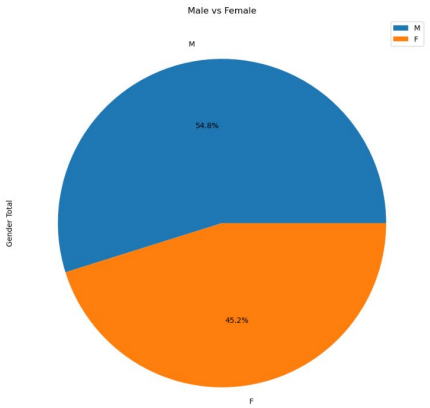




- When you count Withdrawn as a Fail and Distinction as a Pass, There is more fails than passes.
- Very High Withdrawal rate

- Slightly more Male (54.8%) than Female (45.2%) students
- 0-35 represent the overwhelming majority of students.





Questions Asked

- Which students are most likely to drop out of their course before it completes?
- Does IMD_BAND affect student success?
- What is the effect of Student Interaction vs their final score?

Which students are most likely to drop out of their course before it completes?

Jupyter notebook flow:

- SQLITE created from CSVs sourced from Kaggle.
- Explore, cleaning and merging datasets.
- Creating Bins.
- Create training and testing dataframes.
- Logistic regression method used for prediction

Which students are most likely to drop out of their course before it completes?

Disability has the highest positive correlation, that means that when the student has a disability, increases the probability to drop out from the course. Assessment score has the highest negative correlation, which means that when Assessment score increases then probability to drop out of the course decreases.

Accuracy score - 81.5

	Feature	Correlation
7	disability	1.069622
1	region	0.554217
8	final_result	0.220115
4	age_band	0.011615
2	highest_education	0.006146
5	num_of_prev_attempts	-0.055242
3	imd_band	-0.058045
6	studied_credits	-0.208220
0	gender	-0.283133
10	total_clicks	-0.733081
11	registration	-0.841382
9	assessment_score	-4.368085

Does IMD_BAND affect student success?

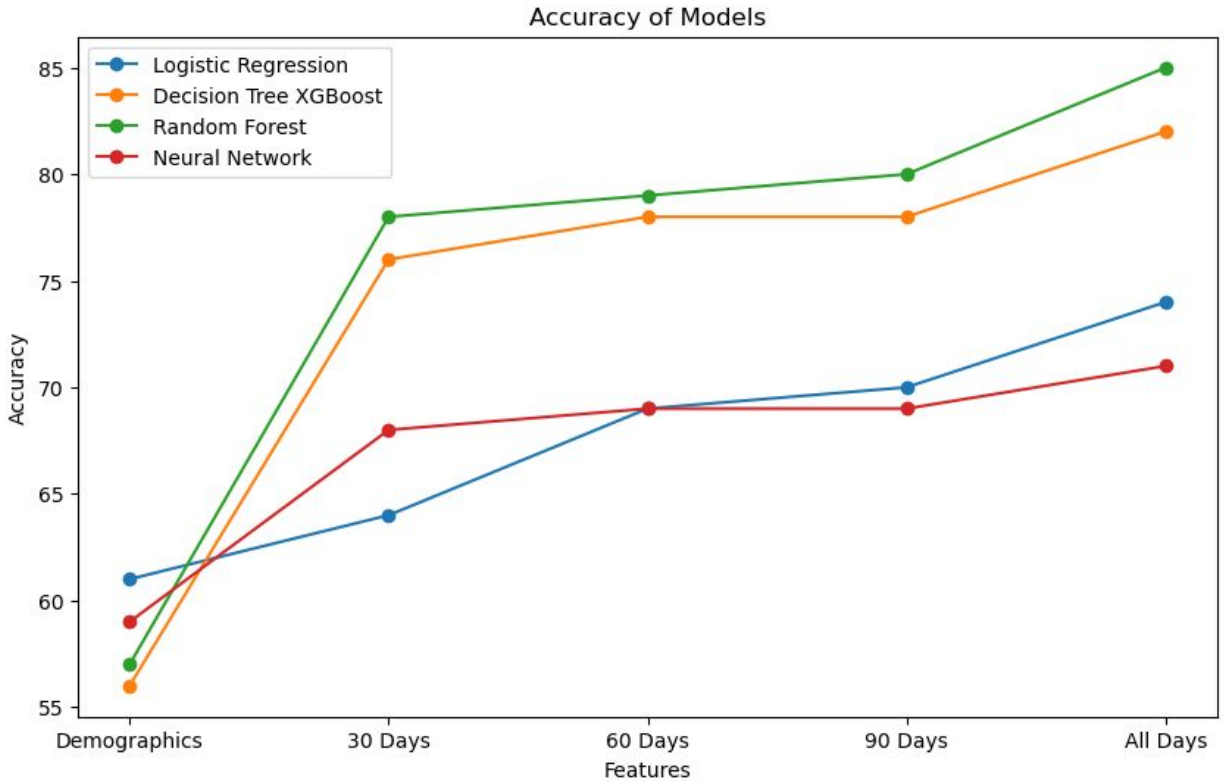
Target = Final Result (Pass or Fail)

Features in final models:

- IMD Band

- Course Module
- Region
- Highest Education
- Age
- Gender
- Previous attempts
- Study Credits
- Disability

- Registration Date
- Number of clicks
- Assessment scores



Does IMD_BAND affect student success?

- Multiple features impacted student success
- Only focusing on IMD Band to predict if a student would pass or fail a course resulted in low accuracy models
- Top three important features
  1. Number of clicks with class material
  2. Assessment scores
  3. Date of registration

Random Forest

Confusion Matrix

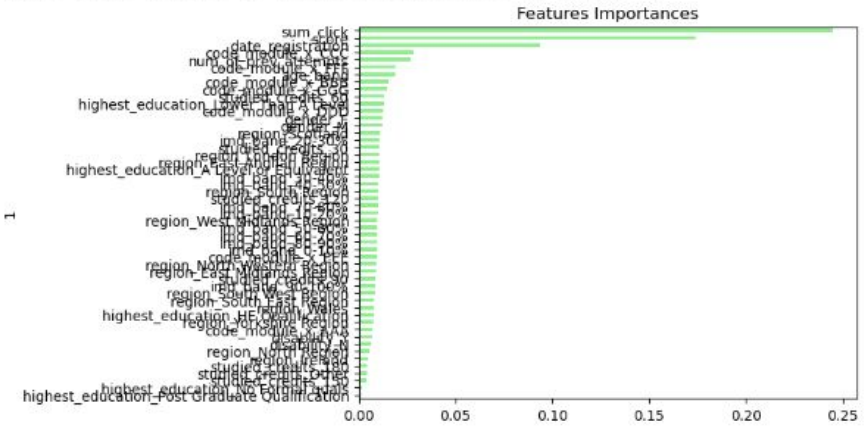
	Predicted 0	Predicted 1
Actual 0	3089	716
Actual 1	537	3805

```
Accuracy Score : 0.8462010556032895
Classification Report
              precision    recall  f1-score   support

           0       0.85      0.81      0.83      3805
           1       0.84      0.88      0.86      4342

    accuracy                           0.85      8147
   macro avg       0.85      0.84      0.85      8147
weighted avg       0.85      0.85      0.85      8147
```

```
Out[45]:
<Axes: title={'center': 'Features Importances'}, ylabel='1'>
```

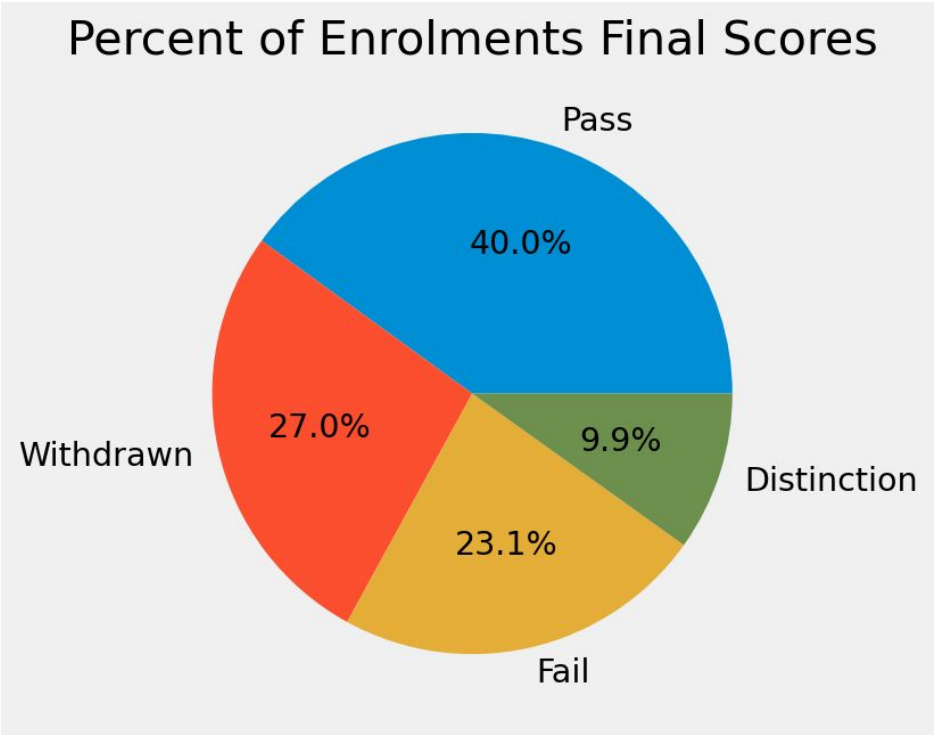


What is the effect of Student Interaction vs their final score?

Wanted to find out if the students' interaction, not their results, would predict their final score.

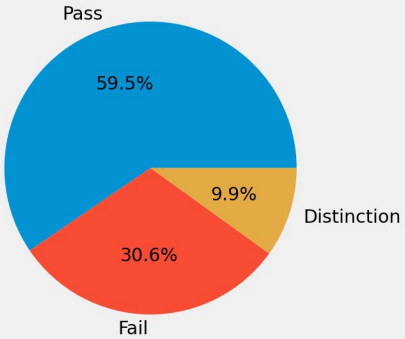
The following columns where used:

- Code_module (unique 7)
- Code_presentation (unique 4)
- Date_submitted (all values)
- **Date (all values)**
- ID_site (all values)
- Activity_type (unique 10)
- Registration (binned to 3)
- Total_clicks (binned to 6)
- Final _result (binned to 2)



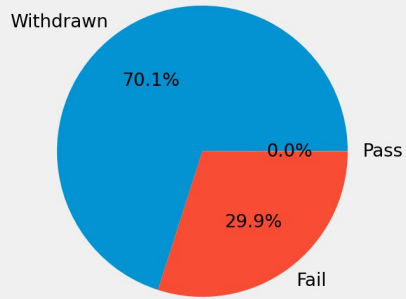
What is the effect of Student Interaction vs their final score?

Percent of scores of unregistered enrolments before the course started



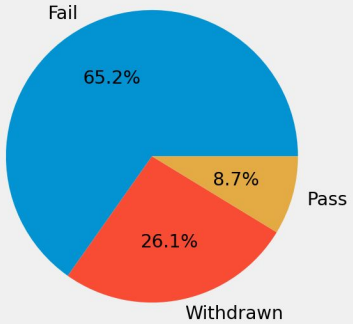
Pass	150
Fail	77
Distinction	25

Percent of Final Scores of enrolments that interacted with no assessments



Withdrawn	2337
Fail	998
Pass	1

Percent of Final Scores of enrolments that completed assessments but did not interact



Fail	15
Withdrawn	6
Pass	2

What is the effect of Student Interaction vs their final score?

I ran through a number of models to find the best model for the dataset the best 3 being:

- Decision Tree
- Random Forest
- KNN

However the following got over the threshold:

- Naive Bayes
- Logic Regression

	Model	Score
0	Decision Tree	93.10
1	Random Forest	93.09
2	KNN	87.03
3	Naive Bayes	76.47
4	Logistic Regression	75.06
5	Linear SVC	74.78
6	Support Vector Machines	73.64
7	Stochastic Gradient Decent	73.64

Limitations and Challenges of Dataset & Analysis

Data

- The age of the data is 10 years old 2013-2014 would be better to have more recent data
- The amount of data is only 7 courses over 4 presentations (sessions)
- There are only 27,295 students there would have been many more
- A limited amount of demographic information
- The size of the dataset was about 500 MB larger than github allows

Analysis

- We only had 2 weeks to analyse the data
- Some of the lexicon are harder to understand, or different from Australia, such as education levels

Web Deployment

- Pickle files of models and dataframes too large to store on github

References

Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).