Analysis on Open University learning data

Utilize or leverage learning dataset from Open University to improve outcomes

- student data
- engagement with VLEs
- assessment data
- course information

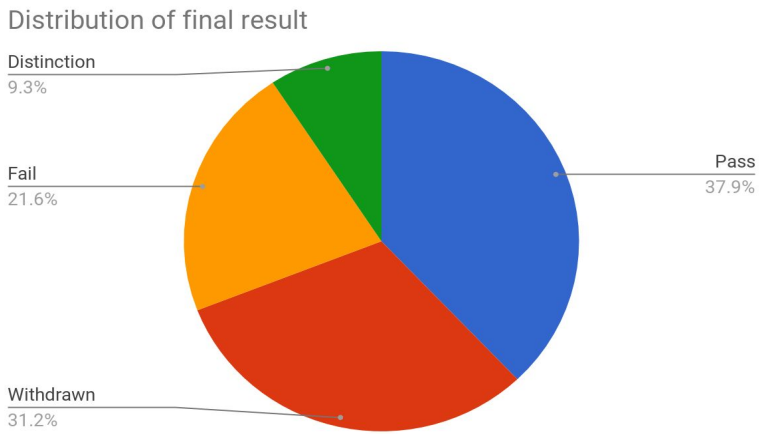
- student scores, more pass grades
- track completion (reduce withdrawals)
- better course content
- student diversity

Areas explored by data analysis on the given data sets :

- Biggest Problem Space in Virtual learning at Open University
- Dimensions that have the highest failure rate or withdrawal rate
- Solutions to minimize failure

Data Summary

- **Problem**: 31.1% of students withdraw the course, 21.6% of students fail. Thus <50% students pass
- **Very high correlation between number of clicks per day and the final result of students**
- 7 modules are being offered in 22 presentations (3.14 presentations per module)
- 32593 registrations from 28785 students (1.13 registrations per student)
- Students are distributed over the geographical regions and not concentrated over a few areas
- 83.46 % students have A level education or lower (secondary school leaving qualification)
- 99.3 % students are aged 0-35. Hardly any >55
- 30% of students submit assessment late

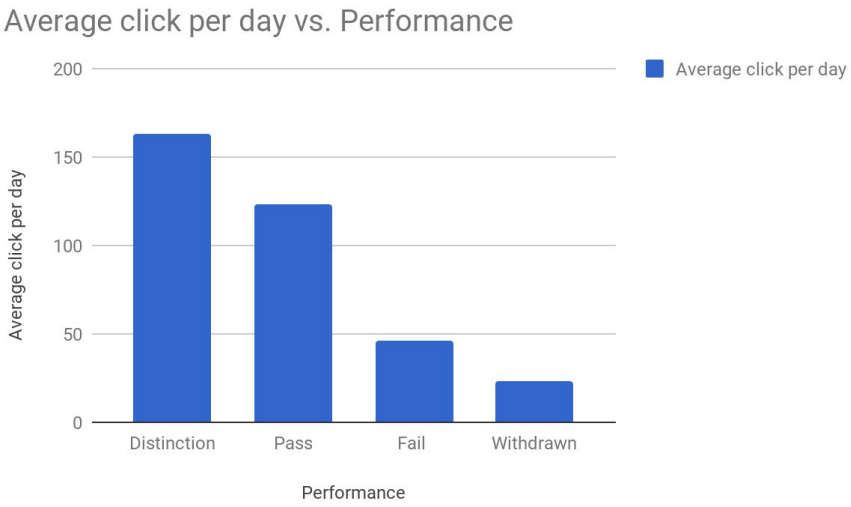


Methodology

1. Exploratory data analysis
    a. Pre-processing, data cleaning (one hot encoding, categorical and ordinal labelling, parsing, fill missing values)
    b. Data visualization
    c. Deriving insights: late submissions, correlations between variables
2. Model training and prediction
    a. Merging datasets, grouping fields, split train-test data
    b. Classifiers used
        i. DecisionTreeClassifier(max_depth=5),
        ii. RandomForestClassifier(max_depth=5, n_estimators=10, max_features=2)
    c. Test accuracy and validation
3. Conclusions and recommendations
    a. Identifying key issues
    b. Propose solutions

Insights

- Interactions with VLEs seems to be the most important factor in determining student performance
  - **High performing students are way more likely to click VLEs**
- Using gender, region, age in the decision tree did not significantly improve results, possibility overfitting



Recommendations

Increasing engagement with VLEs should lead to improvement in results

- Planning VLEs releases
  - Comparing VLEs released on weekdays v/s weekends
  - Choosing time of day to release VLEs when there is higher chances of getting clicked, thereby increasing student engagement
- During course
  - Asking multiple choice questions in between the course interaction will help students to engage more with the presentation
  - More visual content to drive engagement
- After class
  - Sending email to students when course to drive re-engagement
  - A/B testing with VLE content, placement and schedule

Future Work

- More fine-grained VLE representation for the model. Since VLE interaction is the single most representative factor in our model so far, next logical step would be the following.
  - Leveraging type of VLE information. ie. treating resource v/s oucontent v/s dataplus as different VLEs for our model instead of sum of clicks across all VLE types.
  - Intuition to be validated: Visual content is more likely to get engagement compared to text
  - Leveraging time of publishing of VLEs and time of VLE interaction to understand patterns (weekday v/s weekend or time of the day) for different types of VLEs.
- More accurate scoring
  - Using the actual score instead of just pass/fail information
  - Understanding performance of the student over the time of the module presentation, instead of just using the grade at the end for a better performance representation.
- Running analysis on cloud instance
  - We trained a simple classification tree, including the above richer information would help in training a regression tree. Such analysis would have to run on a cloud instance