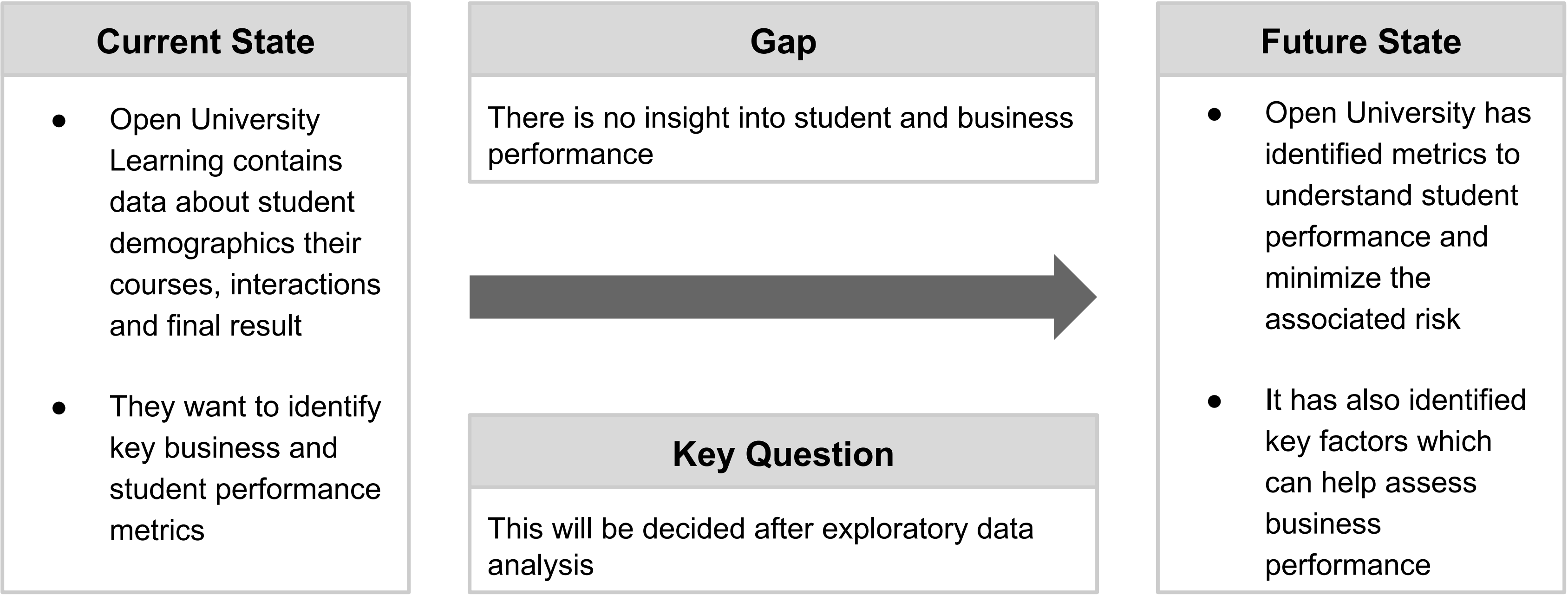
Open University Learning Analysis

Open University wants to understand student performance and module engagement to provide better learning experience



Current State

- Open University Learning contains data about student demographics their courses, interactions and final result
- They want to identify key business and student performance metrics

Gap

There is no insight into student and business performance

Key Question

This will be decided after exploratory data analysis

Future State

- Open University has identified metrics to understand student performance and minimize the associated risk
- It has also identified key factors which can help assess business performance

Student course data is used to perform a holistic analysis to understand business better

Problem Statement

Open University wants to analyze its data to understand metrics important to measure student and business performance

Data

Student demographic information, assessment details, student scores, engagement, course details about 30k students across 43 attributes

Analysis

- Univariate and bivariate analysis to understand relation between variables
- Feature selection and model building to predict student’s final result based on attributes obtained from data exploration

Findings

- Features like scores, sum of clicks, student demographics are important variables for student classification
- To understand module health, increasing withdrawal rate and falling scores and engagement is a good indicator
- Modules perform better during J session as compared to B session

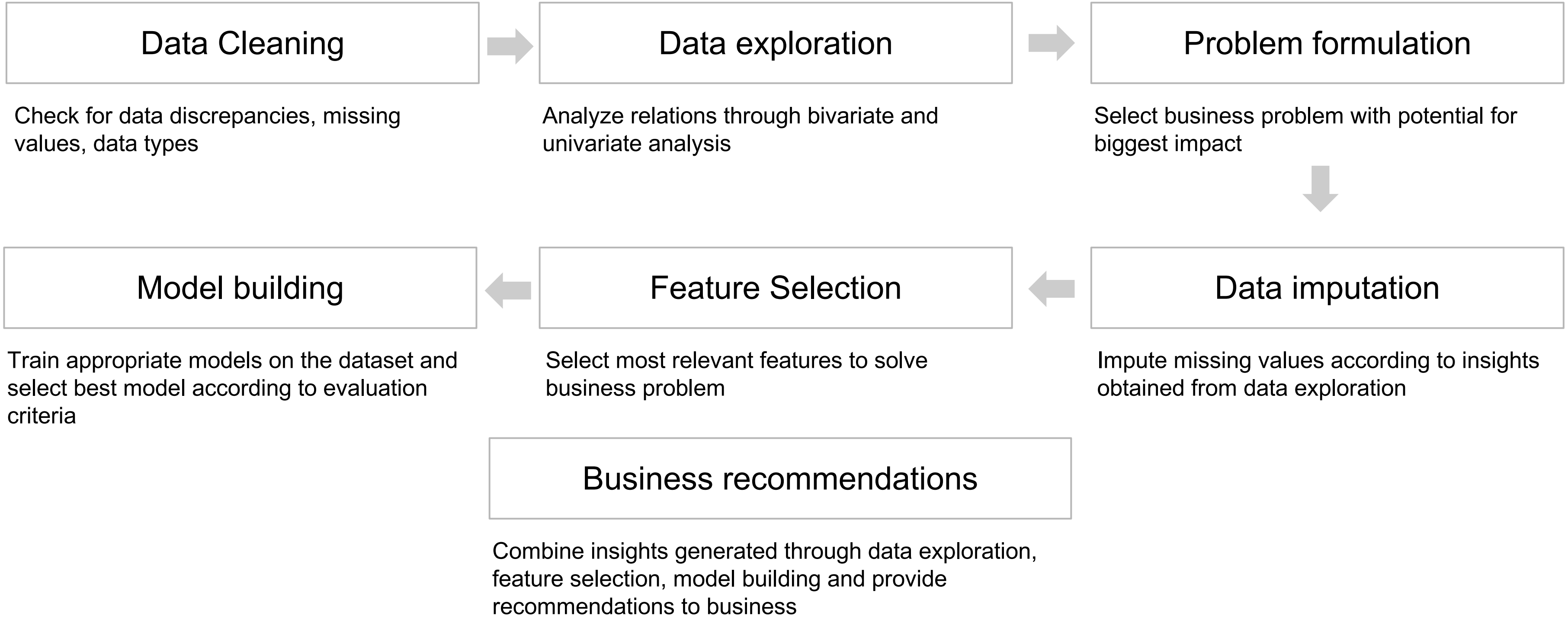
Recommendations

- Module BBB, GGG could be scraped due to low engagement and material of subpar quality
- Student who are being classified as failing or withdrawing the course should be paid extra attention from instructors

Improvements

- Classification models like AdaBoost, XGBoost, SVM can be employed to see if recall for Withdrawal, Failure further improves
- A more comprehensive analysis on student engagement on the basis of activity can be done to understand more clearly

A methodical data analysis can help glean insights to help Open University understand its business performance



Data Cleaning

Check for data discrepancies, missing values, data types

Data exploration

Analyze relations through bivariate and univariate analysis

Problem formulation

Select business problem with potential for biggest impact

Data imputation

Impute missing values according to insights obtained from data exploration

Feature Selection

Select most relevant features to solve business problem

Model building

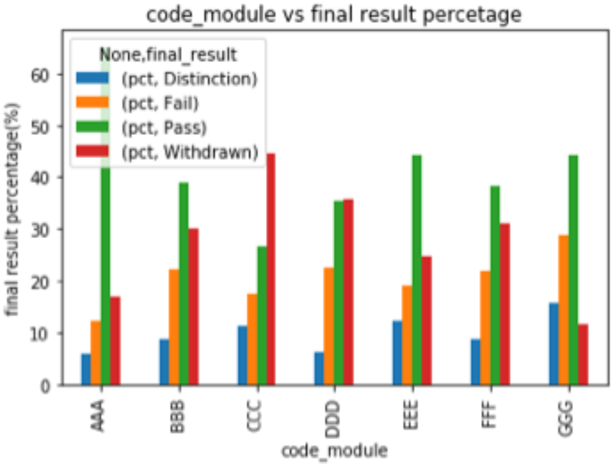
Train appropriate models on the dataset and select best model according to evaluation criteria

Business recommendations

Combine insights generated through data exploration, feature selection, model building and provide recommendations to business

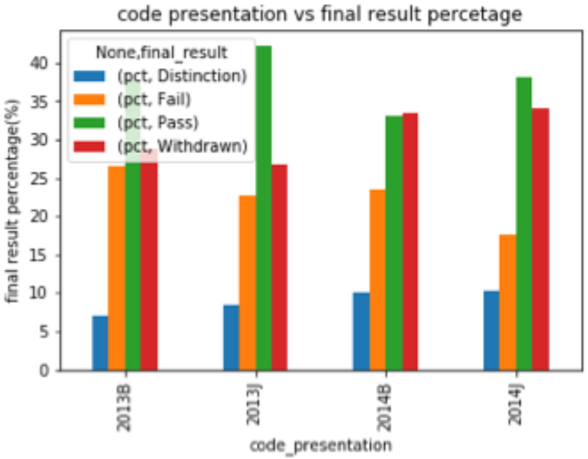
Student’s performance varies across courses and presentations due to module quality and student engagement

Student result percent for different course modules



- Module CCC -> high withdrawal rate of 45%; pass percentage being about 28%
- Module AAA -> high pass percentage of over 60%
- Failure percentage remains to be almost same across modules lying between 20-30%

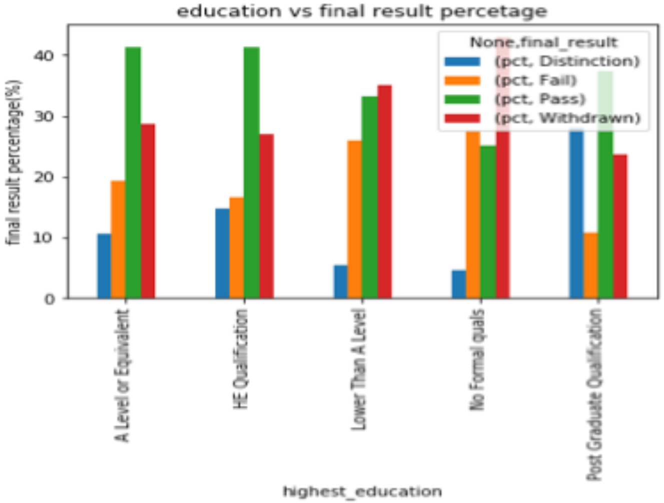
Student result percent for different course presentations



- Pass percentage shows fluctuating trend
- Maximum being 40% during session J
- Falling to 30-35% during session B
- Low quality of modules offered, inefficient instructors during session B

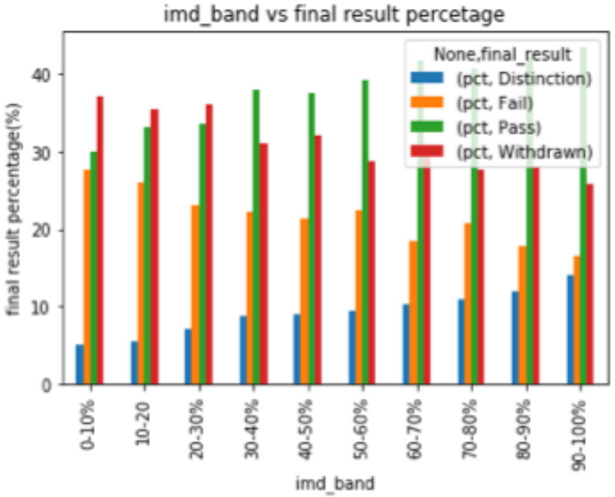
Student’s module and presentation performance is different for different demographic segments

Student result percent for different education background



- Student with no formal qualification or qualification lower than A level -> highest withdrawal rate of 45%
- Higher education level -> Higher pass percent
- Post graduate students lowest failure rates

Student result percent for different IMD bands

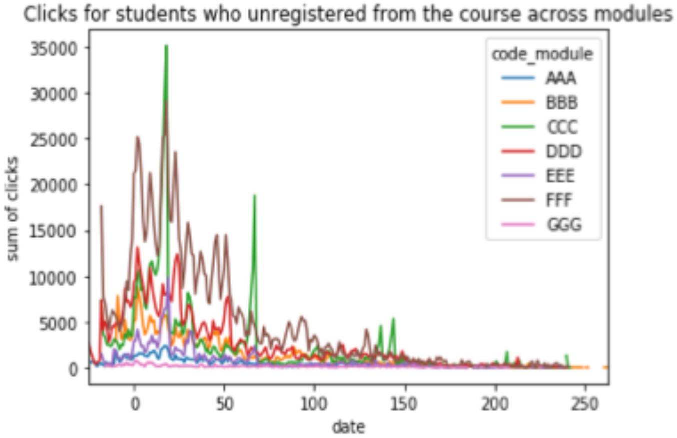


- Lower imd_band -> high withdrawal rate of 38%
- Higher imd_band -> withdrawal rate of 28%
- Pass percent increases gradually with increasing imd_band

** IMD: Index of Multiple Deprivation

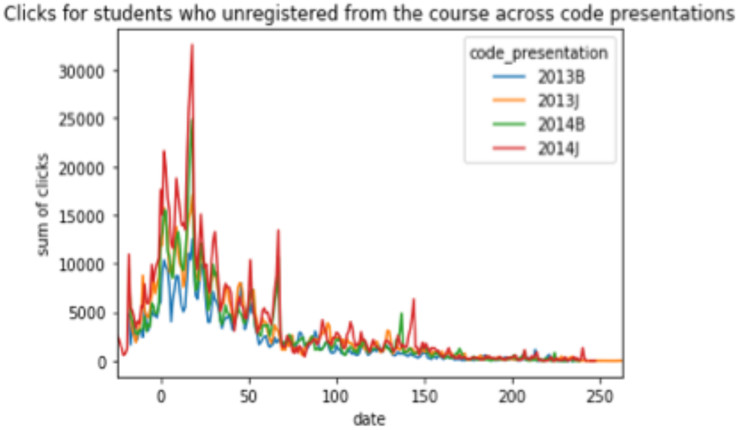
Student engagement varies over time for students who withdrew from course and students who completed the course across modules and presentations

Student engagement for students who withdrew from course across modules



- Modules CCC, EEE -> highest click activity initially but big dips afterwards
- CCC -> highest withdrawal rate
- Clicks fall after 50 days from start of module
- For students who completed -> no gradual dip, fluctuating trend, interaction till end of module

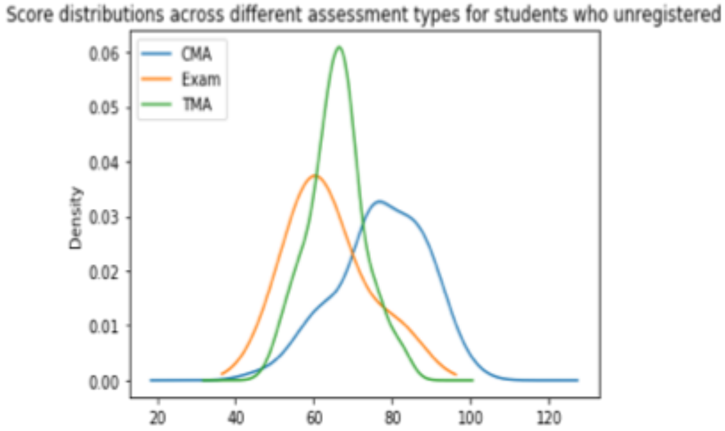
Student engagement for students who withdrew from course across presentations



- Presentation 2013J and 2014J -> maximum clicks
- Session J -> lower withdrawal
- Difference in module offerings, quality or demographics

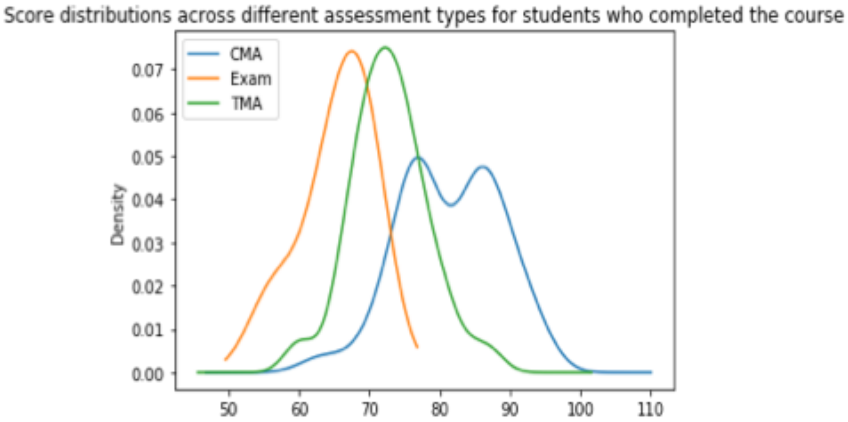
Score distributions across assessment types helps understand student performance for different customer segments

Score distribution of students who withdrew from the course



- Students scored maximum in CMA followed by TMA and Exams
- For CMA, distribution function was spread over a wide range of values from 40-100
- TMA most values were centered around 60

Score distribution of students who completed the course



- Marks centered around 65 and 70 for Exams and TMA respectively
- CMA has a widespread distribution with values from 60 to 100
- CMA has 2 distinct peaks for students who scored around 75 and students who scored 90

Student demographic information analysis, module engagement and scores can help assess student performance and has potential to create huge impact for business

Current State

- Open University Learning contains data about student demographics their courses, interactions and final result
- They want to identify key business and student performance metrics

Gap

There is no insight into student and business performance

Key Question

What are the key student demographics and performance that increases probability of student failure or withdrawal?
What are module related metrics that might affect student performance?

Future State

- Open University has identified metrics to understand student performance and minimize the associated risk
- It has also identified key factors which can help assess business performance

Data exploration insights can be used to impute variables and select relevant features for further processing

Data imputation

- Values for score imputed by mean score of each student
- Imdb_band missing values has higher pass percentage, imputed by average of higher imd_band, 60-70%
- Values like sum_clicks, date imputed by mode of column

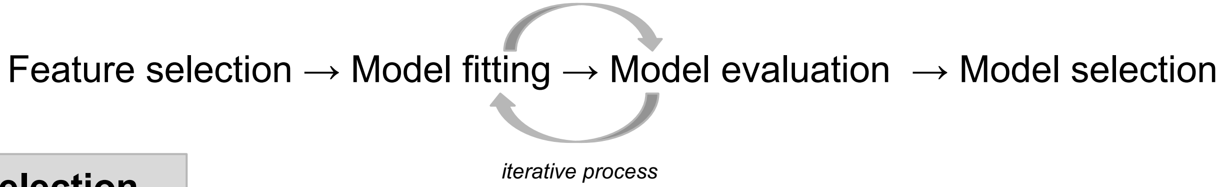
Data preparation

- StudentInfo, assessment, courses and click activity datasets combined to create master dataset
- Master dataset has final results according to scores of student for different demographics and assessment types
- Value counts for final result for students revealed class imbalance with 61% of dataset having students with pass result

Open University can incorporate variables learnt from data exploration to build solution to minimize student failure or withdrawal

Why classification?

- Classification model can help instructors understand variables which affect student performance and probability of student failing or withdrawing from the course
- This can lead to them paying extra attention to students at risk through more interaction, exercises



Feature selection → Model fitting → Model evaluation → Model selection

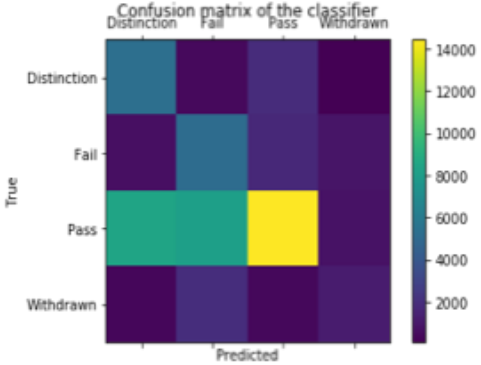
iterative process

Feature Selection

- Tree based methods employed for selecting relevant features
- Work well with class imbalance
- Combined with features obtained from EDA
- Selected features like score, sum of clicks, date submitted, module length and relevant demographics

Classification methods with high recall are good classifiers for predicting failure and withdrawal rate

Confusion matrix for Logistic Regression on downsampled data

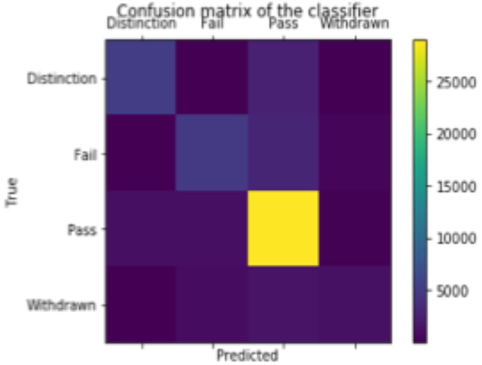


```
             precision    recall

Distinction       0.36      0.68
       Fail       0.33      0.61
       Pass       0.78      0.45
  Withdrawn       0.39      0.30
```

- Data downsampling was done due to huge class imbalance in dataset
- Logistic Regression was fit on training data size = 0.7
- Class Fail getting accurately recognized in most cases
- Class Withdrawn is getting predicted as withdrawn or fail in most cases

Logistic Regression on downsampled data is a better classifier

Confusion Matrix for Random Forest run on original data



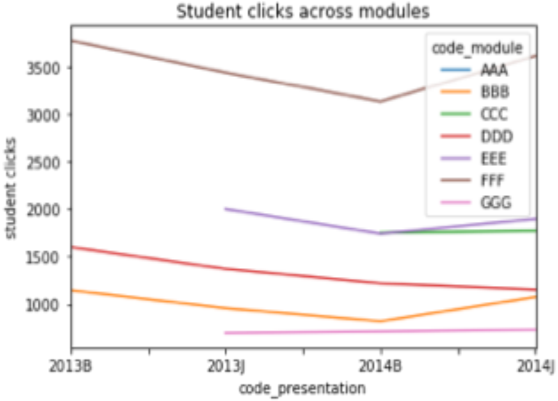
```
             precision    recall

Distinction       0.79      0.65
       Fail       0.68      0.58
       Pass       0.80      0.92
  Withdrawn       0.65      0.34
```

- Random Forest takes care of class imbalance
- Classifier was fit on original training data using gini as split criteria
- Class Fail getting accurately recognized in most cases but had a lower recall than Logistic Regression
- Class Withdrawn getting predicted as Pass in most cases which is a huge concern although recall is higher

Precision = tp/(tp+fp) **Recall = tp/(tp+fn)**

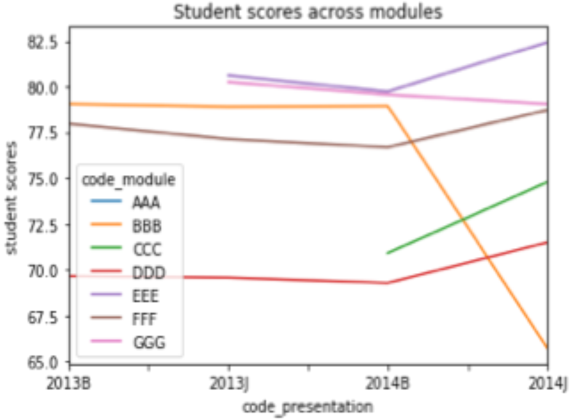
Student failures and withdrawals can also be a result of low quality modules

Student engagement over time across modules



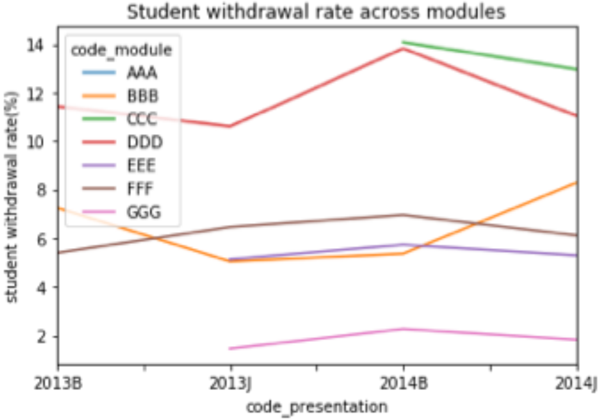
- Student engagement is higher during 'J' code presentation
- CCC module which was offered in 2014 has higher student engagement
- Modules like DDD have consistently falling student clicks

Student scores over time across modules



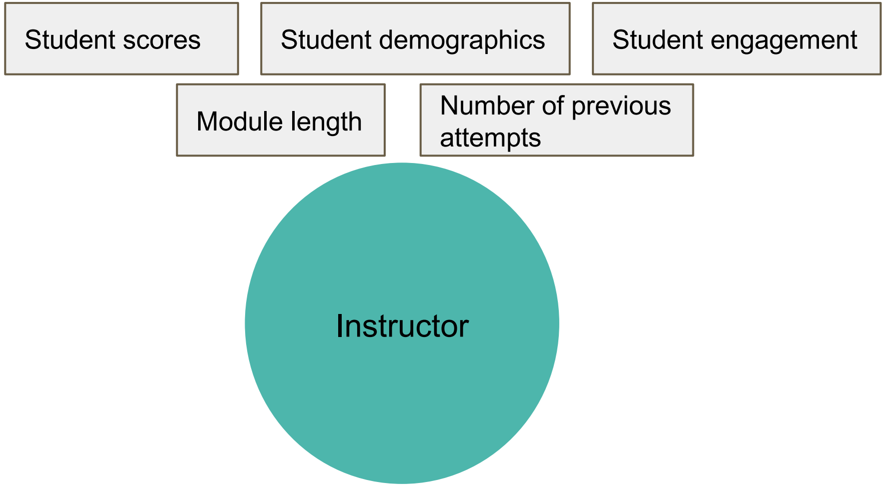
- Mean student score has taken a huge hit for BBB module in 2014J whereas scores for all other modules have increased.
- Although DDD has lower student engagement, mean scores are not affected.

Student withdrawal rate over time across modules



- As student scores have dipped significantly for module BBB, withdrawal rate has increased for the same
- DDD is exhibiting a fluctuating trend in terms of withdrawal rate

Student failures and withdrawals have a multifarious perspective and minimizing these is value adding for both instructors and stakeholders



Student scores

Student demographics

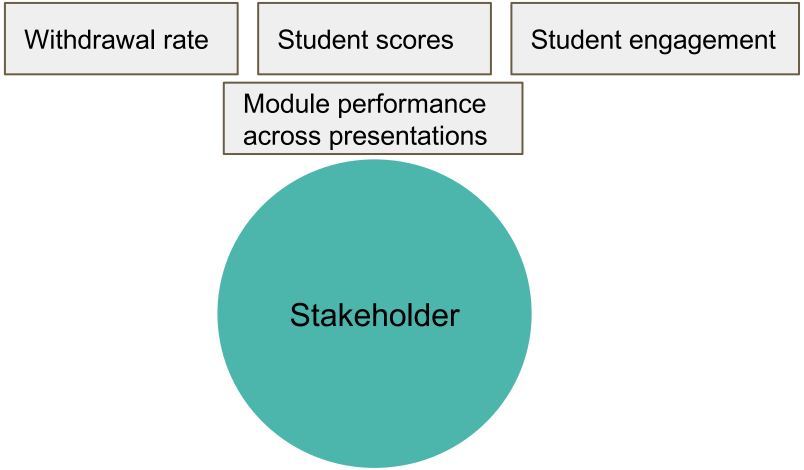
Student engagement

Module length

Number of previous attempts

Instructor

- Know probability of a student failing or withdrawing based on the demographics and scores in a particular assignment
- The classification model is proactive in nature and can help pull down the withdrawal and failure rate
- With the student identified, instructors can provide extra help



Withdrawal rate

Student scores

Student engagement

Module performance across presentations

Stakeholder

- Know module performance over time
- Can help eliminate low performing, low quality modules
- Can help redesign better modules and launch them during specific sessions when they have higher engagement

The impact of this analysis can be measured by implementing recommendations generated from the insights through appropriate testing

Classification model to measure student performance:

- Impact of employing classification model can be measured by falling failure and withdrawal rates
- A random subset of students can be classified on the basis of results obtained from classification model
- Their end of module performance when compared with students who were not classified can be compared to measure success of this analysis

Analysis to measure health of a module:

- Analyzing student scores, engagement and withdrawal rate for a module over time can help identify less-engaging/low quality modules
- The business impact of scraping certain courses could be measured by overall rising ROI (Return on Investment)
- Also, the overall student withdrawal and failure rates could be measured