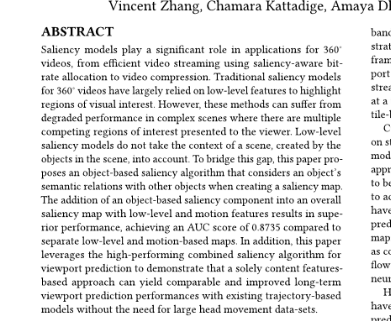
**Thesis Summary**

**Mapping Learning Journeys for K12 Programmers**

With the explosion in educational data, educators are looking towards educational data mining (EDM) techniques to uncover insights about students and how they learn. However, few EDM approaches provide a comprehensive view of individual student learning journeys.

To bridge this gap, this seminar proposes a decision-tree based approach to identify student learning journeys on the online coding platform, Grok Learning. The approach creates interpretable decision-trees that show pathways to success/failure, and identifies important slides that students should focus on to perform well. This seminar also discusses the impact of this approach on EDM, and proposes recommendations for further research.

Do some path analysis on decision trees -

****

**Contribution 1: Exploratory data analysis (Sem 1) – can probably omit – write this part last (tonight)**

* **Produce the following**
  + Score distributions (histograms) showing scores are quite skewed – why we need to discretise based on the population distribution for scores.
  + Distributions showing the number of problems completed
  + Outcome distributions
    - Most commonly occurring sequences (may have overlap with below)

**Contribution 2: Different types of students revealed through clustering (Week 4) – exploratory data analysis**

* **Density plot (i.e. saliency map) showing** 
  + Final scores vs number of slides completed

**Results**

* + There are categories of students
    - Students who perform well but do not complete any slides
    - Students who perform well and complete slides (linear relationship)
    - Students who do not perform well and complete slides
    - Students who do not perform well and do not complete any slides
* **To do**
  + Compute the density plot **(tonight)**
  + Evaluate the density plot using some measures
  + Explain the final score
  + Try clustering the students based on these two features into different clusters.
  + Evaluate using different methods and compare with Sophia’s work.

**Contribution 3: Course dropout prediction (Week 5)**

* Defined three types of dropout behaviours: early dropout, late dropout, no dropout
* Dropout distributions, as measured by weeks and types are relatively spread out.
* Course dropout prediction was framed as a classification task
  + **Classifier:** Multinomial logistic regression
  + **Validation:** 10-fold cross-validation with stratification
  + **Features:** Only used features from the first 1, 2 and 3 weeks of the course.
    - Aggregate
      * Number of problems completed in first 1,2 and 3 weeks.
      * Number of slides completed in first 1,2 and 3 weeks.
      * Number of slides attempted but not completed in first 1, 2 and 3 weeks
      * Number of problems attempted but not completed in first 1, 2 and 3 weeks
  + **Conclusions**
    - Student’s interactions for the first week can predict course dropout relatively well, with accuracy that is comparable to if we had taken a longer subset (e.g. up to three weeks) or even the entire interaction sequence.
    - **What this really means:** Student course dropout behaviour is highly determined by how they interact with slides in the first few weeks of the challenge.
    - **What does this mean for educators:** For educators, this emphasises the importance of intervening early to ensure that students can progress throughout the course.
  + **Work that needs to be done** 
    - Correlative analysis between the different features esp. between slides and problems.
    - **Use different classifiers (tonight)**
      * Multinomial logistic regression
      * Decision trees
      * Naive Bayes

**Contribution 4: Module outcome prediction (tonight)**

Predict the outcome of the final problem in the module, using all previous slides (and then previous problems too i.e. include Ns, Fs, Ps).

**Use different classifiers (tonight)**

* Multinomial logistic regression – average for 2018 challenge is roughly 80-85%
* Decision trees
* Naive Bayes
* **Conclusion:** Logistic regression performs better than DTs and NBs.

Also need to work out

* More detailed version of slide interactions – extract attempts at slides too, problem run, run in-line

**Contribution 5: Slide value importance**

Quantified the value of each slide in determining whether a student would successfully pass the last problem of each module, and whether they would dropout (no submission for the last problem). The educational value of this is that it helps educators and course content creators understand how their content affects performance for that module. This section also compares and validates these results with Grok designers.

**To do (tomorrow)**

* Compare with different feature importance methods, and choose the most appropriate method – also use CFS (correlative feature selection).
* Rank the different features using the different methods, and try to reconcile among different methods.
* Information gain
* **Evaluate the slide importance with Grok designers**
* **Compare this with what students actually complete**

**Contribution 6: Evaluate across all challenges and analyse differences/similarities**

* Detail what was required to make it extensible to all challenges.

**Contribution 7: Conference Paper**