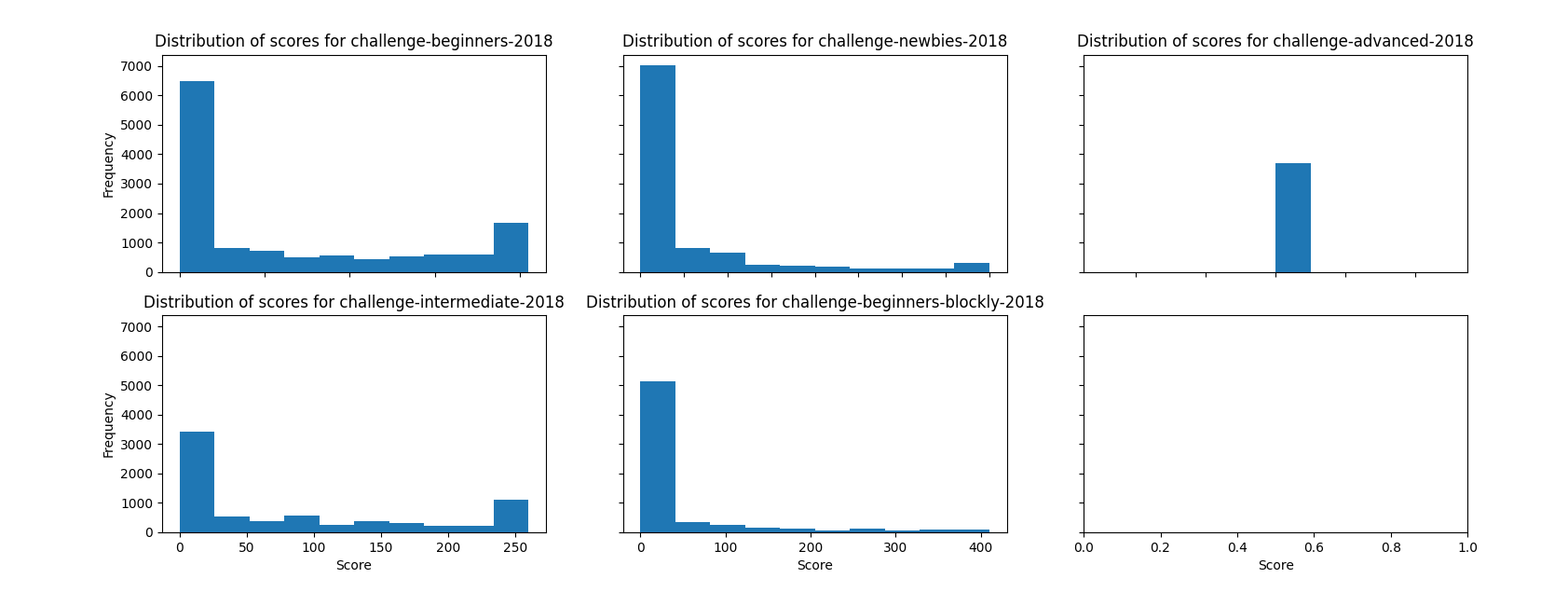
**Week 4 Update**

**Score distributions for all challenges**

Last week, we defined a student’s performance score in terms of the number of slides they successfully completed in a challenge. There is also data around a student’s performance in terms of marks scored. For instance, the challenge-newbies-2018 has a total score out of 400.

How the score is calculated: Students are given marks for each task. There is some penalty based on the number of attempts (we can ask Bryn and look at Sophia’s thesis) – overall performance measure and more granular. This score might be a bit skewed.

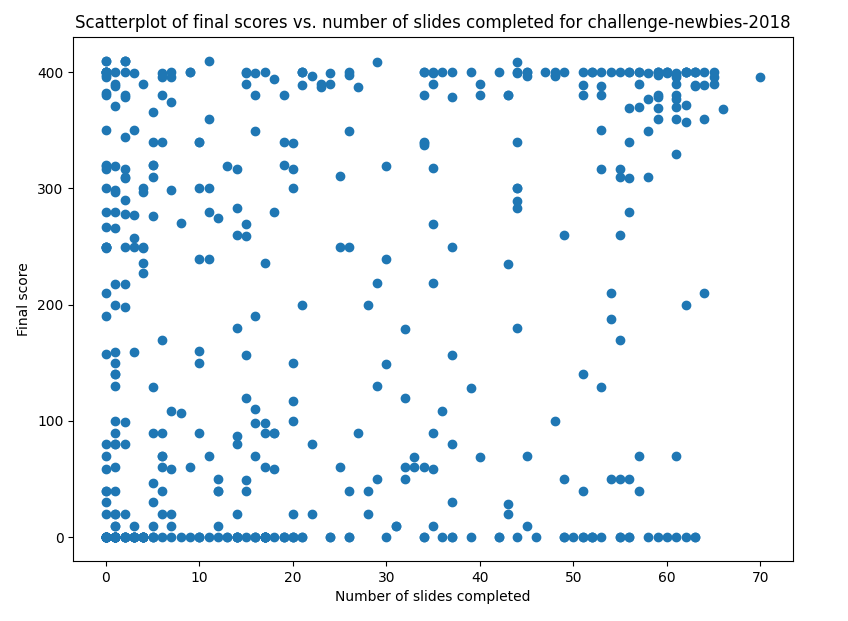
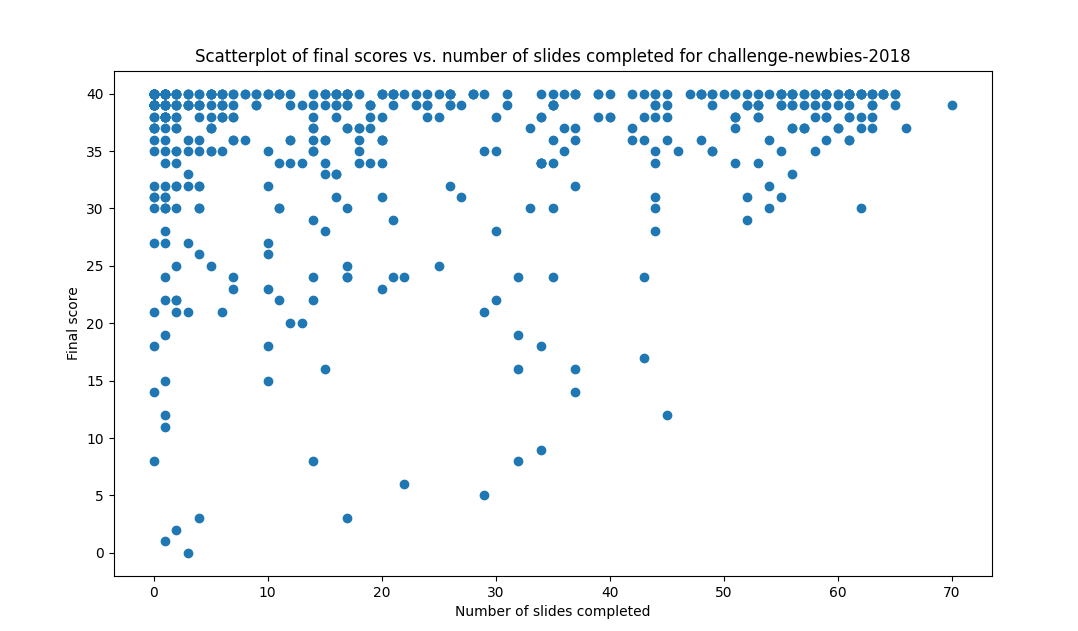
We can see that the distribution of scores for all challenges (besides challenge-advanced-2018) is right skewed, with a significant portion getting low scores.

****

**We may need to discretise the score into four ranges – 0-100, 100-200, 200-300, 300-400**

**Scatterplots**

The figure below shows that there is a concentration of students who have low number of slides completed and low final scores, and a concentration of students who have high number of slides completed and high scores. There is also a group of students who have low slides completed but high final score. Final scores in this context are defined as a scores extracted from the enrolments.csv file.

****

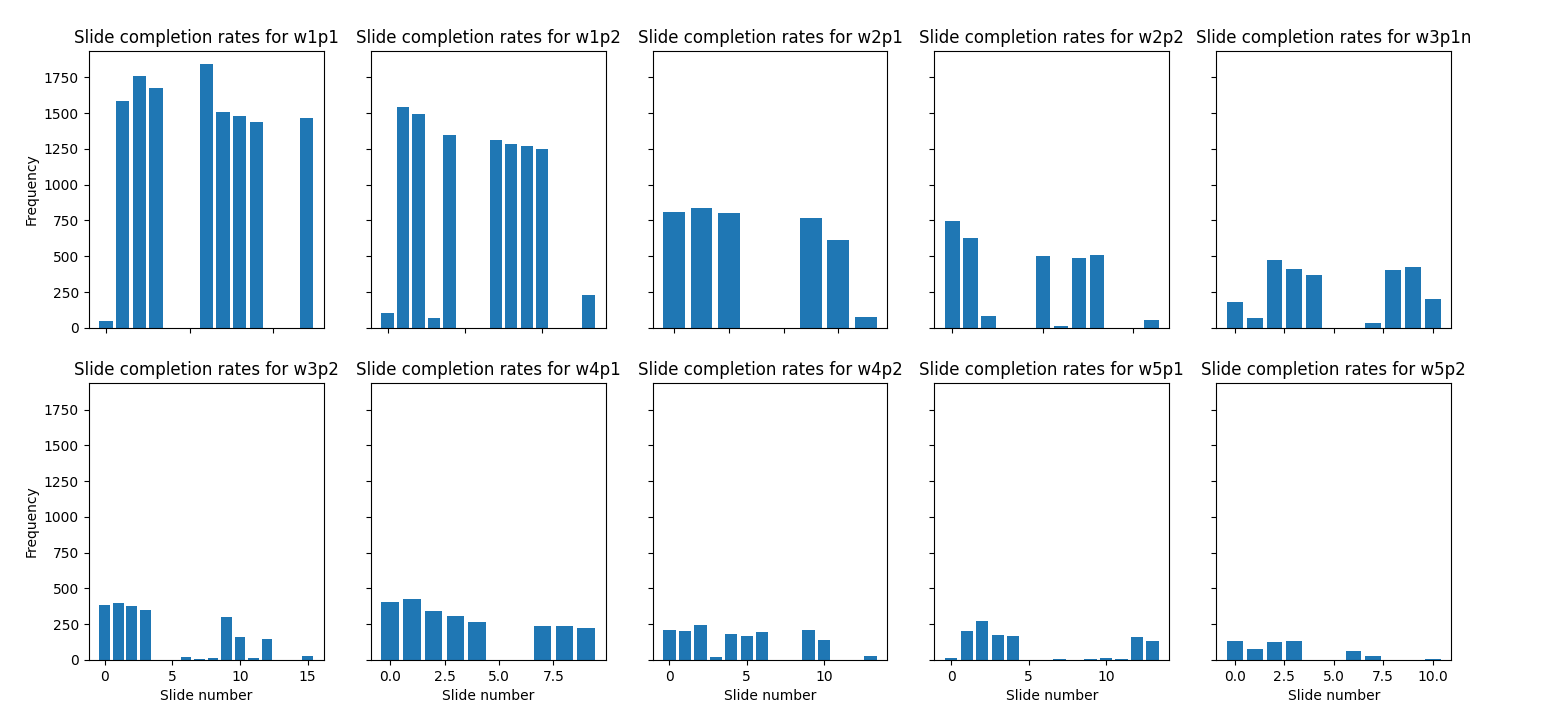
The figure above shows that there is a weak positive relationship between the number of slides successfully passed and the number of slides completed, but typically a higher proportion of students who successfully pass most or all slides. (NOTE: CAN COLOUR THESE DOTS)

**Slide completion rates for challenge-newbies-2018**

As we hypothesised, only some slides are attempted by most students. Slides that have low completion rates tend to be narration slides that require students to click on a narrate button, or when the current slide is too similar to the previous slide (have another in-depth look). Less-attempted slides were identified by first selecting an appropriate frequency threshold from the graphs below, and then excluding slides that had completion rates below that threshold.

For challenge-newbies-2018, the less-attempted slides are:

|  |  |
| --- | --- |
| **Module** | **Less attempted slides that can be excluded** |
| w1p1 | 0 |
| w1p2 | 0, 3, 13 |
| w2p1 | 7 |
| w2p2 | 2, 6, 11 |
| w3p1 | 0, 1, 7, 10 |
| w3p2 | 6, 7, 8, 11, 15 |
| w4p1 |  |
| w4p2 | 3, 13 |
| w5p1 | 0, 7, 9, 10, 11 |
| w5p2 | 7, 10 |



**Predictors**

**Tasks**

**Clean up code-base ASAP - Saturday**

|  |  |  |
| --- | --- | --- |
| **Task** | **Notes** | **Date** |
| Extract and plot dropout distributions | Note: Do across courses. |  |
| Work out a definition for dropouts based on the distribution | Note: Do across courses. |  |
| Create new interaction strings using only the important slides. | Note: Compare performance of new interaction strings with old, and visualise well.  Note: Do across courses. |  |
| Create a scatterplot of dropouts vs features (e.g. number of slides completed) – hopefully we get two distinct classes (Y/N). | Note: This might not be a scatterplot. Be precise with your language.  Note: Do across courses. |  |
| Create, run and evaluate simple classifiers for dropout | Note: Performance of classifiers might not be great if there are clusters. Explain why.  Note: Look at the data and then explain which model might be best – e.g. if sequence, then maybe LSTM.  Note: Do across courses. |  |
| Create, run and evaluate simple classifiers for performance | Note: Difference between performance measures (# of slides completed and final score considering submissions)  Note: Performance of classifiers might not be great if there are clusters. Explain why.  Note: Look at the data and then explain which model might be best – e.g. if sequence, then maybe LSTM.  Note: Do across courses. |  |
| **Create a classifier with number of submissions made and performance** | Note: Again, perform an initial clustering analysis on the data – are there different types of students e.g. students who make high number of submissions but don’t do well (i.e. spammers), students who make low numbers of submissions but do well, people in the middle etc. |  |
| **Create a classifier with number of submissions made and dropout** |  |  |

1. **Hypothesis: Is interactive slide behaviour a good predictor of the score? – prediction task, specifically final score (in terms of four ranges) and predict dropout (needs to be defined after we see the distribution).**
   1. **Try different methods**
   2. **Look at dropout distributions**

**AS SOON AS I HAVE RESULTS PUT IN DROPBOX – TRY SIMPLE CLASSIFIER – NB, DT – CLASSIFIER CHOICE IS NOT THAT IMPORTANT – VERY LIKELY WILL HAVE SIMIALR RESULTS – LOGISTIC REGRESSION**

**Depending on results, take a look at data. And then do a clustering task. The results are not high because there are these clusters of students.**

1. **Number of submissions that are made – low, medium, high – people who have low number of submissions are likely to either be really good or not engaged at all (how can we distinguish between these two groups), and then people who make high number of submissions, they tend to make minor changes, and then people in the middle group/medium are representative of the strength of the relationship.**
2. **Do this separately for courses. But don’t do for students who do several courses at the same time.**

**66% training and 33% for testing – make sure you don’t get this wrong**

**If you discretise the score, it will be a classification task not regression task. If you predict dropout it will be a classification task not a regression task.**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. **Time – people who make consecutive submissions in short periods of time. They don’t really modify the code and make superficial changes. Try to gain the change. We can measure code distance as an idea.**
2. **Across courses – this is for future work. Not easy to identify people who are enrolled in multiple courses.**

**Hypothesis: There are 3 groups of students**

* **Slides**
  + **Non-aggregate statistics (sequences)**
    - **Boolean binary interaction sequence, representing whether a student has completed certain slides, for modules and overall.** 
      * **Hypothesis:** Whether a student has completed certain slides can have an impact on their performance and propensity to dropout.
      * **Assumption:** If this was taken as the dataset, then this assumes that all slides are independent of each other.
    - **Boolean binary interaction sequence for key slides only, representing whether a student has completed certain key slides.**
  + **Aggregate statistics (summary)**
    - **Number of key slides (predetermined) that were completed in a module, across modules and overall**
      * **Hypothesis:** Certain slides are important and can lead to success in the course. Students who complete those slides tend to do better than students who don’t and are less likely to drop out.
      * **Can filter by module**
      * **Plot slide completion across the modules**
    - **Number of slides completed in a module, across modules, and overall**
      * **Hypothesis:** Students who complete more slides are likely to do better than students who complete fewer slides, and less likely to drop out.
      * **Can filter by module**
* **Problems**
  + **Number of overall submissions made (passed and failed submissions)** 
    - **Hypothesis:** Students who tend to make a high number of submissions are more likely to be spamming the system, are less likely to do well, and are more likely to drop out.
    - Plot this measure with final score (/400) – students who make too many submissions/superficial changes don’t really engage with the content. There are some hard-working diligent students who are just slow – hopefully we can see the difference in the two groups of people. Students who start too late – we can use the time as well to narrow/refine the groupings.
  + **Number of times a solution is submitted to a problem (passed and failed submissions counted) before completing a prerequisite slide.**
    - **Hypothesis:** Students who submit problems before completing a prerequisite slide are more likely to rush through the course, are less likely to do well, and are more likely to drop out.

**Things to predict**

* **Overall performance**
  + Final score for the course, expressed as a score out of ~400 – found in the enrolments.csv file.
  + Final score for the course, expressed as the total number of problems passed
* **Late-stage performance**
  + Score for the final module/s (or last attempted module), given that the students complete the preceding modules.
* **Dropout** 
  + Whether a student will continue with the course, given their history
  + P P P F F F P P N N N N N N N P P
  + How are we defining drop-out?
  + What does the data show? Plot the outcomes afterwards
    - Map out drop-out distributions
* **Type of learner** 
  + Based on clustering of students and their features into distinct learner types
  + This might be a bit stretched.