Virtual Internships Data Project Report

By:

- Aden Siu
- Yohan Nanayakkara
- Shesh Murali
- Benjamin Hall
- Jaehong Kang

Table of Contents

Executive summary

This report investigates the relationship between communication patterns within student interns and mentor interactions and their performance in the virtual internship simulation. The project analyses the chat logs between interactions in the simulated biomedical engineering company, Nephrotext, to ultimately develop a predictive model for the purpose of predicting the final design report scores

Project Background

The concept of the virtual internships is that it offers web based simulations for these university students to experience real world professional scenarios as engineers and scientists. The context of the dataset is that interns worked within teams working within the fictional biomedical engineering firm to design a prototype device for patients suffering from kidney failure.The responsibilities of the interns and mentors encompassed the stages of the design process, for instance background information research understand stakeholder requirements, creating prototypes, conducting testing and assessments and rationalising their design choices. The teams operated on an online platform through an online chat tool, the communication between mentors and other students to facilitate inquiry and encore reflective dialogue. The communication between them was finally expressed in the final report or performance which was scored on a scale out of 8

Data Collection

The data consists of conversations of 75 groups where there are 5 separate groups, with 15 implementations of each group. The conversations have been annotated to indicate the presence or absence of essential concepts in the engineering discourse, particularly pertaining to design action and justifications. Moreover the dataset includes the outcome score, ranging from 0-8 for each student's final design evaluation completed during their internship. The objective of this project is to predict the performance on the report using the information about the topics discussed by students throughout the internship. These topics include experimental testing, design chives, asking questions, customer consultant requests, performance parameters requirements and justifying communication. All together there are 392 participants ; where mentors consist of 23 individuals and students accumulate up to 369 people.

Hypothesis

It is hypothesised that the more frequent chat activity leads to higher report scores. By discussing more topics often , showing effective communication, attaining more mentor interaction, and having everyone on the same page, (thus needing to ask less questions), will lead to more positive outcome scores

Objectives and Methods

The dataset was initially presented in an unfavourable way, limiting our understanding of the content. Therefore steps were taken including cleaning and processing the data, converting it to student and team level statistics and establishing multiple data frames isolating each component. All groups were working towards the same object, each group per implementation were extracted and contained the groups topics and mentor chat interactions to create a group level data frame.

Similarity, for student level we extracted all the data for each student, with the sum of their topic occurrences These two data frames were combined to make a mixed dataframe as well.

The preparation permitted us to conduct Exploratory data analysis (EDA) to help further our understanding of the impact of each component on final outcome scores. Methods such as correlation matrices, scripts for continuous variables, histograms and heatmaps were employed to gain understanding of exploratory variables. Additionally, we analysed chat utterances for each group to further evaluate the relationship between topic and outcomes scores. This was done by isolating non-stop words and performing frequent word analysis, named entity recognition (NER) analytics and keyword extraction

Furthermore, modelling techniques were applied to construct a model proficient in predicting final report marks based on the type of texts sent between interns and mentors, and the type of topic of the text. The explanatory variables became numerical, with the outcome score being ordinal (having a qualitative categorical property) which presented limitations in the modelling, regression modelling techniques did not produce appropriate accuracies and error scores when predicting outcome scores on the testing sets as regression does not deal well with categorical data. Therefore, decision trees and support vector machines were preferred for modelling

Analysis

The EDA involved creating visualisations such as heatmaps, barcharts and pair plots to examine the trends and relationships within the data. The pair plot showed no clear correlation between the number of messages and outcome scores, while bar charts highlighted an imbalance in the score distributions , with the majority scoring 4. The heat maps revealed significant topics impacting mean outcomes scores like experimental testing, customer consultant request, and performance parameters, requirements. For deeper insights Frequency word analysis, Named entity recognition analysis and keyword extraction on chat utterances, identifying the impact of specific vocabulary and communication patterns on outcome scores. Word clouds and frequency analysis illustrated the prevalence of different topics within the group discussions, like delving into the characteristics of the prototype to examine the relationship between the topics uttered and the outcome score of groups. Multivariate linear regression (MLR) identified significant features that impacted scores but had limitations in accuracy. Principal component analysis (PCA) helped reduce data dimensions revealing strong positive relationships between total messages,mentor interactions and outcomes scores. Stratified sampling and SMOTE were additionally used to balance the dataset and improve model performance. A total of 30 pipelines and 39 classification models were built using stratified sampling, SMOTE and principal component analysis to optimise accuracy and precision. The top performing models were identified by comparing confusion matrices and accuracy scored for each dataset. The top performing models heavily utilised sum of mentor interactions, sum of questions asked per team, and total word count.

Key findings

1. Multivariate linear regression: While regression techniques would have inevitably struggled with accuracy, our regression helped us identify the direction in which features might have correlated with outcome score, such examples were finding positive correlations between

groups discussing customer requests as well as discussion of performance requirements and outcome score and negative correlation with students asking questions.

2. Principle Component Analysis: This technique to reduce the dimensions within the data was used to help improve model accuracy and was able to be used to identify key and important features within the dataset, showing stronger relationships with mentor interactions, total messages sent and amount of questions asked.
3. Classification Modelling: Utilising an advanced array of pipelines, various classification models were created, unable to find stronger models than simply guessing 4 each time, less accurate but less skewed (and thus more valid) oversampled minority models showed a strong reliance on mentor interactivity and the amount of the questions asked

4. Regression machine learning models
   Various models, including multivariate linear regression, support vector regression and random forest regression were tested. Results indicated that simple regression models were insufficient while classification models provided more accurate but not as usefully applicable predictions

Conclusion

The research suggests that the content of communication is important for successful project outcomes. Our models demonstrate that effective teams seem to communicate more overall (text more frequently) and garner more mentor support, they needed to ask less questions as they likely knew more about what they had to do and worked well independently (in order for things to go smoothly in the chat, thus needing to ask less questions) and were able to spend more time considering what was needed from the customers and for performance parameter requirements. Finally, future models should explore the more nuanced relationships between mentor and student through topic modelling , sentiment analysis and potentially add additional variables to improve predictive accuracy as well as utilise more data balancing techniques such as undersampling and more intense oversampling.

Introduction

With the immense dataset of student performance and chat records at an educational internship simulation of a fictional biomedical engineering company, Nephrotex, our team aims to uncover correlations and severity of each recorded feature and how they relate to report marks of the student interns.

Virtual Internships

Figure 1 - Company Logo

Where the data comes from

The data comes from Virtual Internships, a company that conducts web-based simulations aimed towards university students, to allow them to replicate real world situations as scientists, engineers, artists, and workers. For this dataset, interns at a fictional biomedical engineering firm collaborate in teams to develop a prototype device aimed at aiding patients with kidney failure. Their responsibilities encompass various stages of the design process, such as researching background information, comprehending stakeholder requirements, crafting prototypes, conducting testing and assessments, and rationalising their design choices. Teams operate within an online platform, utilising a chat tool for communication. Additionally, each team is supported by a mentor assigned to facilitate inquiries and guide reflective dialogues.

How is was collected

This dataset contains chat records from 15 internships of Nephrotex, comprising 309 individuals. The conversations have been annotated to indicate the presence or absence of essential concepts in engineering discourse, particularly pertaining to design actions and justifications. Additionally, the dataset includes the Outcome score, ranging from 0-8, for each student's final design report completed during the internship. The objective of this project is to predict performance on the report using information about the topics discussed by students throughout the internship. These topics include:

- Experimental testing
- Design choices
- Asking questions
- Customer consultant requests
- Performance parameters requirements
- Justifying communication

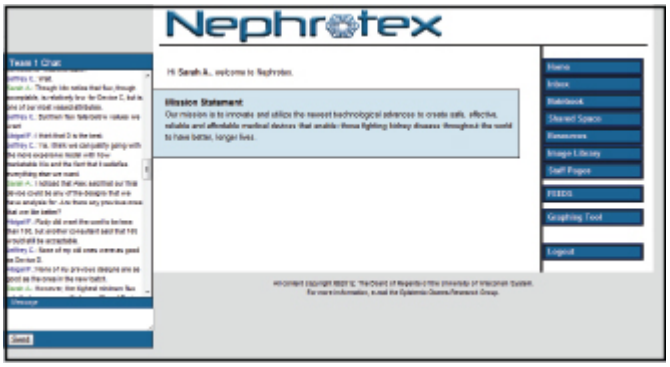

Figure 2 - Nephrotex online web communication

Objectives

The dataset is in an unfavourable format, limiting our understanding of the content. Steps that we have taken to address this is to clean and process the data, and convert the data to student and team level statistics and create multiple data frames isolating each component. Since all groups were working towards the same objective, we extracted each group per implementation and extracted the sum of the groups topic and mentor chat occurrences and created a group level dataframe. For student level, we extracted all the data for each student, with the sum of their topic and mentor chat occurrences.

This allows us to conduct exploratory data analysis to understand the severity of each component towards final outcome scores. Methods like heatmaps, correlation matrices , pair plots for continuous variables, histograms and other graphing techniques were conducted to give a greater understanding of the exploratory variables.

Furthermore, we analysed the chat utterances for each group, to further evaluate the relationship of topics with outcome score. By first isolating non-stop words, we performed Frequent Words Analysis, Named Entity Recognition (NER) Analysis and Keyword Extraction.

Upon this, we applied various modelling techniques to construct a model that is proficient in predicting final report marks based on chat topics and utterance, mentor impact and vocabulary use. The exploratory variables of the dataset are categorical, with the outcome score being ordinal, thus, we encountered several limitations with the modelling. The main limitation is that regression modelling techniques do not produce appropriate accuracies and error scores when predicting outcome scores onto the testing sets, which was subsequently expected. Therefore, modelling techniques in decision trees and support vector machines were preferred.

Hypothesis

Our main hypothesis is that more chat occurrences overall would lead to greater report marks. This entails that groups that talk about topics more, would address them better, subsequently portraying effective team communication. We also believe that more mentor interaction and questions would also correlate with final report marks. This is due to our belief that mentors would provide help when asked, thus, the more the teams ask for help, the more assistance they will receive towards building the prototype.

Data Quality/EDA

Data Quality:

To determine the quality of the data, the first step taken was to identify any missing or NaN values in the data. In looking into this, it was found that 3 participants in the virtual internships did not have an allocated role name. As the original dataset had almost 400 total participants, these 3 participants were somewhat insignificant, and were imputed by removing them from the dataset.

```
unique_id                              0
userIDs                                0
implementation                         0
Line_ID                                0
ChatGroup                              0
content                                0
group_id                               0
RoleName                               3
roomName                               0
m_experimental_testing                 0
m_making_design_choices                0
m_asking_questions                     0
j_customer_consultants_requests        0
j_performance_parameters_requirements  0
j_communication                        0
OutcomeScore                           0
wordCount                              0
dtype: int64
```

Figure 3 - Missing values within original dataset

The next quality issue checked for was for duplicate values. We found that in all 19180 rows of messages in the dataset, there were no duplicates found, and therefore no manipulation was required.

Data Preprocessing:

The goal of the project was to investigate how discourse features, and participant interactions in the chat group impacted the outcome score of students on an individual and group level. To achieve this, the dataset was split into a student level, and a group level data set. The student level data had all students, with their outcome scores, the total number of messages they sent relating to each topic, the total number of messages their group sent relating to each topic, the total number of words they sent, and the number of mentor interactions throughout the internship. This allowed for the investigation into the relationship between a student's individual score, the interactions of the individual and the interactions of the whole group. On the other hand, the group level data was created by taking the mean outcome score of each group, the total amount of messages they sent relating to each topic, and the total number of mentor interactions. This allowed analysis into how a group's average score is impacted by the group's activity in the chat group. These two data frames were what the group intended on completing exploratory data analysis on, as well as the modelling.

To conduct text analysis we looked at how we can simplify the text in the chat logs as there were numerous words with little importance as due to the text being chat logs consisting of conversations. Few tools we utilised to remove all the noise were implementing stop words, punctuation removal, lowercasing and text stemming. Stop words works by utilising the nlkt package which has a list of stopwords stored in it, tokenises each word which splits the text into individual words and then looping through the chat logs and removing each word that's classified as a stop word. This allowed us to create another column in our student dataset that included the amount of non-stop words in their texts.

Pipelines Used for Preprocessing

During our exploratory data analysis, we noticed that there were a lot more data points for outcome score 4 for group level, student level and our mixed group & student level data. This meant that our models would be skewed towards group 4. In order to improve the accuracy and validity of our models, before we modelled our data, we used the following processes. First of all, Our testing and training data was not split randomly, but utilising Stratify, we split the data according to its representation within the dataset. For example, in the student level data, there are 125 4's and 57 3's, representing 33% and 15% of the data respectively. So stratify keeps those ratios in its testing and training splits. In ours, we take 80% of each class to train and 20% of each class of outcome score to test it with. After splitting the testing and training sets, we did some further pre-processing. We applied a synthetic minority oversampling technique (SMOTE) to help balance our severely imbalanced dataset by creating new data points that have been duplicated from classes with fewer values. The only issue with this was that it meant datasets with too few data points would create errors with the pre-processing technique, so we eliminated any class with less than 7 data points. This cost us precious data but ensured the processing could run. We used this in favour of NearMiss (an undersampling technique) but decided against it to avoid too much information loss in our dataset given the tiny size of our minority classes. The key issue with doing this however, was that we were unsure if this was going to actually help improve the accuracy of our models, we'll discuss how we solved this issue in our modelling section.

Exploratory Data Analysis:
Many figures were produced as part of the exploratory data analysis, which allowed for careful examination of potential trends, relationships or issues within the data. To begin with, a pairplot was created to get a glimpse into possible trends between outcome score, the topics of conversation, and mentor interactions. It was clear from the pairplot that there was no clear correlation between outcome score and the number of messages students sent regarding each of the topics, as students who sent 0 messages for a topic, achieved outcome scores ranging from 0-8 for all topics. This meant that regardless of the number of messages sent regarding a topic, the outcome score was not impacted.

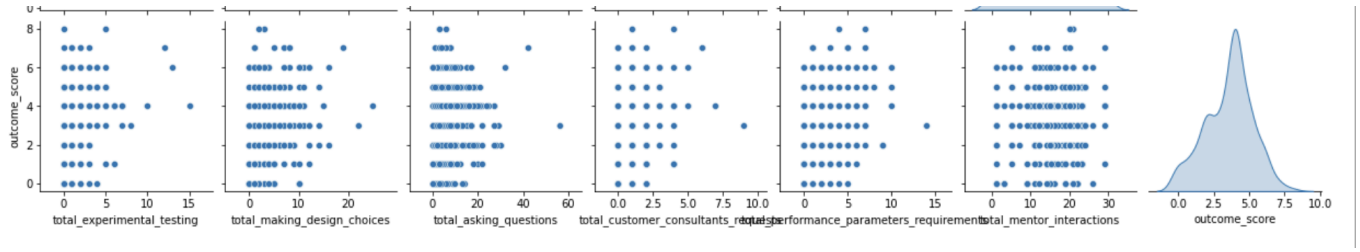

Figure 4 - Pairplots of variables and outcome score

```
both_data_sets
outcome_score
4.0    125
5.0     57
2.0     55
3.0     49
6.0     31
1.0     24
0.0     19
7.0      7
8.0      2
Name: count, dtype:
```

Another figure produced was a collection of bar charts each representing a different topic of conversation. The graphs compared how many messages were sent by

students who achieved scores ranging from 0-8 for a particular topic. From this figure, it was noticed that overall, there was no clear correlation between score and topic of conversation, as we would expect, further reinforcing our evidence from the previous figure. However, the figure helped identify an imbalance in the range of scores, which could pose a potential issue for later in the project. This meant that there were significantly less students who achieved a score of 8 or 0 than there were students who scored a 4. This pattern can be roughly observed by the increasing frequency of data points from outcome scores 0 to 4, but a decreasing frequency from scores 4 to 8. This would deprive us of an accurate model as the models would be trained on mostly students who mostly achieved an average score of 4, without sufficient data of students who achieved higher or lower scores of 0 and 8.

Figure 5 - Specific outcome score occurrences

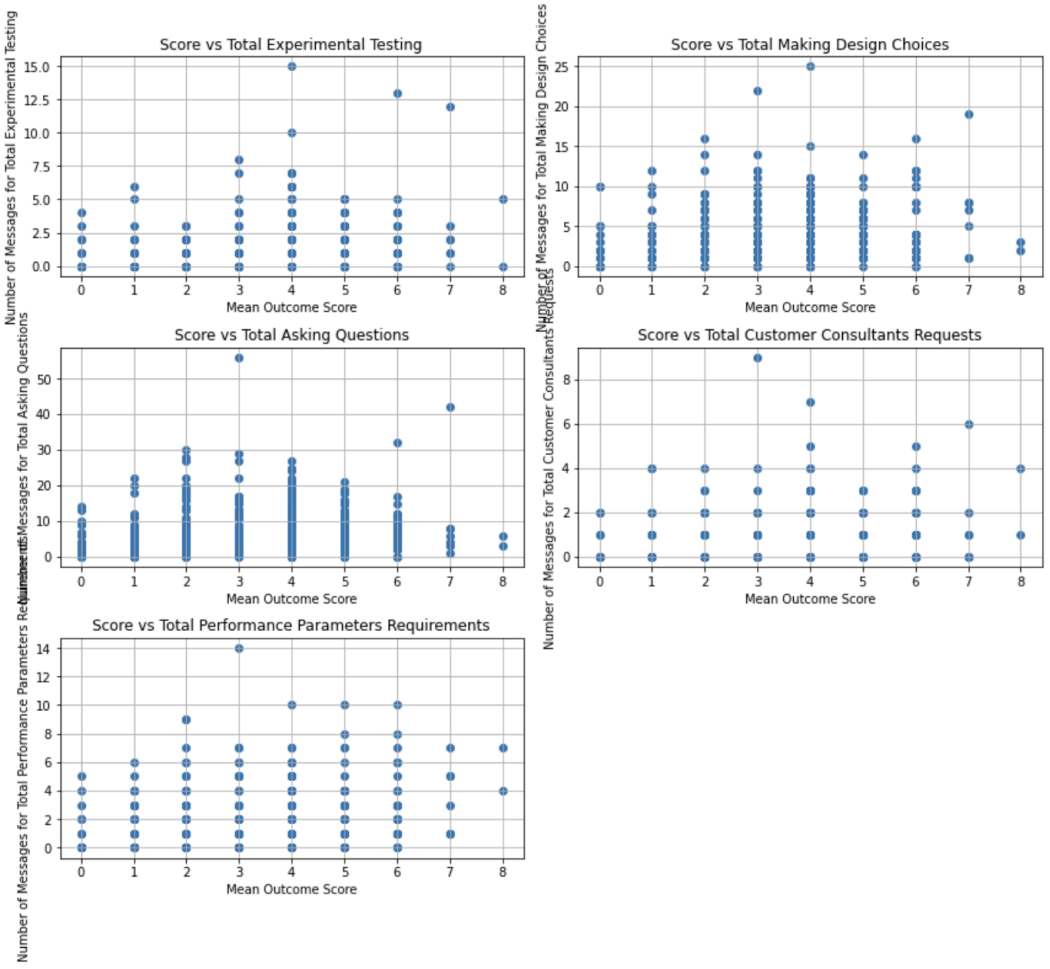

Figure 6 - Scatter plots of mean outcome score against variables

To analyse the correlations between a group's mean outcome score and the number of messages the group sent relating to each topic, a heatmap was created. This visualisation revealed to us the topics of conversation which had the most significant impact on mean outcome score for a group. These topics included 'Experimental Testing' (0.35), 'Customer Consultant Requests' (0.22) and 'Performance Parameters Requirements' (0.16). This suggests that talking about specific topics in a group setting led to better results in terms of scores. It may also mean that the number of overall

messages sent is not a key factor in calculating outcome score, but rather it is the contents of the messages which is important.

Furthermore, looking at the outcome score with regards to features we can see that a higher ratio of discussion towards customer consultancy requests and performance parameter requirements and a lower ratio of asking questions and texting about making design choices seemed to make up a significant portion of outcome score 8. Seeming to once again indicate, asking less questions correlates slightly with higher outcome scores.

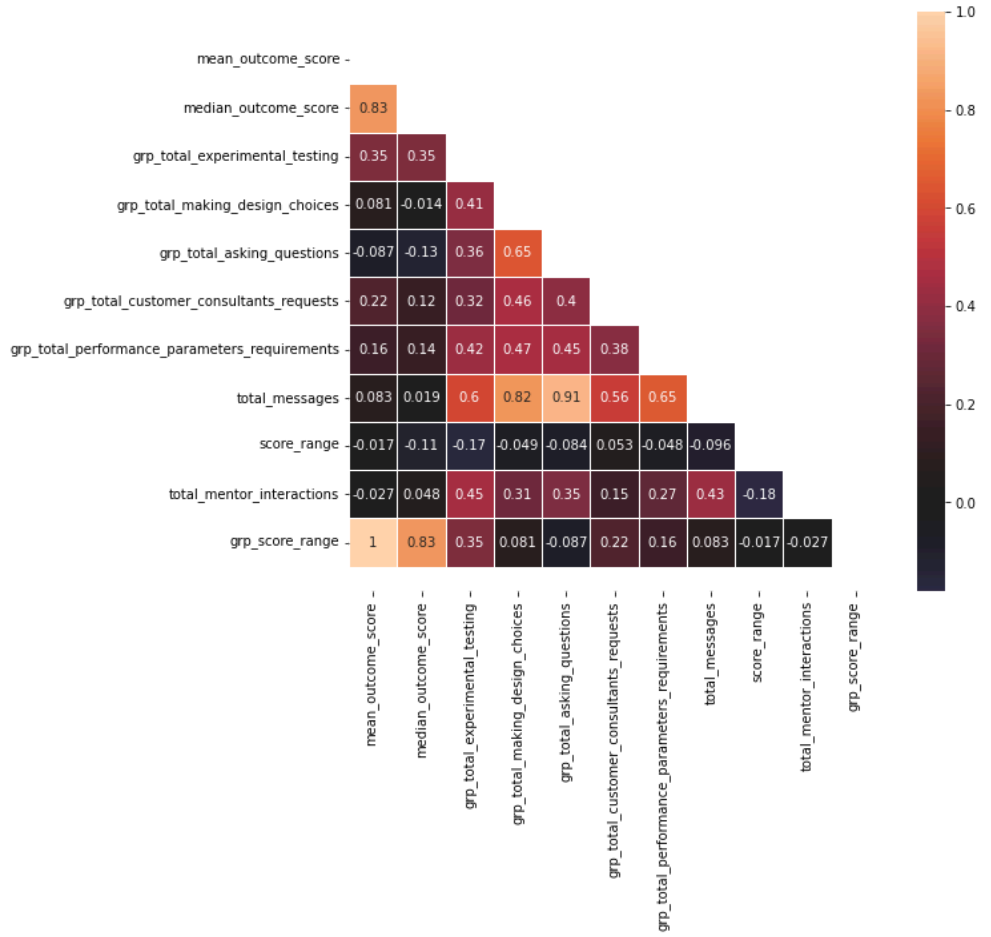

Figure 7 - Heatmap of group level dataset

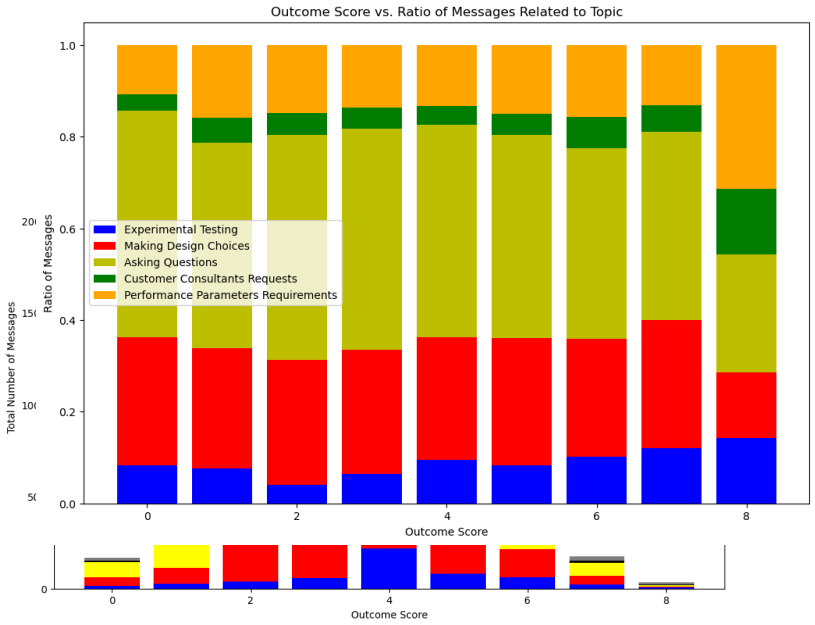

Figure 8 - Stack bar chart of outcome score and ratio of messages

Model Development

Multivariate Linear Regression (MLR)

This modelling technique was chosen due to its ability to utilise multiple variables and configure their relationships with the target variable. In this situation, our explanatory variables are grp_total_experimental_testing, grp_total_making_design_choices, grp_total_asking_questions, grp_total_customer_consultants_requests, grp_total_performance_parameters_requirements, total_messages and total_mentor_interactions, with response variable as mean_outcome_score.

Using multiple sklearn tools, was able to split the data into training and testing sets randomly(8:2) then instantiated and fitted the model onto the training set. Finally to evaluate the performance of the model, predict on the testing set and obtain R^2, RMSE and MAE scores.
As shown in figure 13, the scores are not ideal, hence, normalisation on explanatory variables was conducted in attempts to obtain a better score.
From figure 14, the scores obtained were not preferable and the prediction scores on the testing set worsened.
Although the models did not perform well, to observe the effects of the feature variables we conducted Analysis of Coefficients. As seen from figure 15 grp_total_performance_parameters_requirements has the highest magnitude of influence in predicting mean_outcome_score, followed by grp_total_asking_questions and grp_total_customer_consultants_requests as 2nd and 3rd. To further investigate the importance of each coefficients by taking into account their variability, we multiplied coefficient value by their standard deviation for each feature as shown in figure 16.

Overall, grp_total_performance_parameters_requirements, grp_total_asking_questions and grp_total_customer_consultants_requests variables have the highest effect in influencing scores. grp_total_performance_parameters_requirements and grp_total_asking_questions having a positive relationship. This indicates that groups that chat and aim to meet more customer consultant requests and performance parameter requirements obtain greater outcome scores. What went against our initial hypothesis is that grp_total_asking_questions has a negative relationship with mean_outcome_score. This indicates that more questions being asked results in worse outcome scores, elucidating the idea that teams that manage to work more towards a problem independently score better.

	R^2	RMSE	MAE
train	0.180963	0.678586	0.542419
test	0.067362	0.883967	0.699075

Figure 9 - MLR modelling on group level data

	R^2	RMSE	MAE
train	0.204194	0.671593	0.512927
test	0.032974	0.864949	0.683338

Figure 10 - MLR modelling on normalised group level data

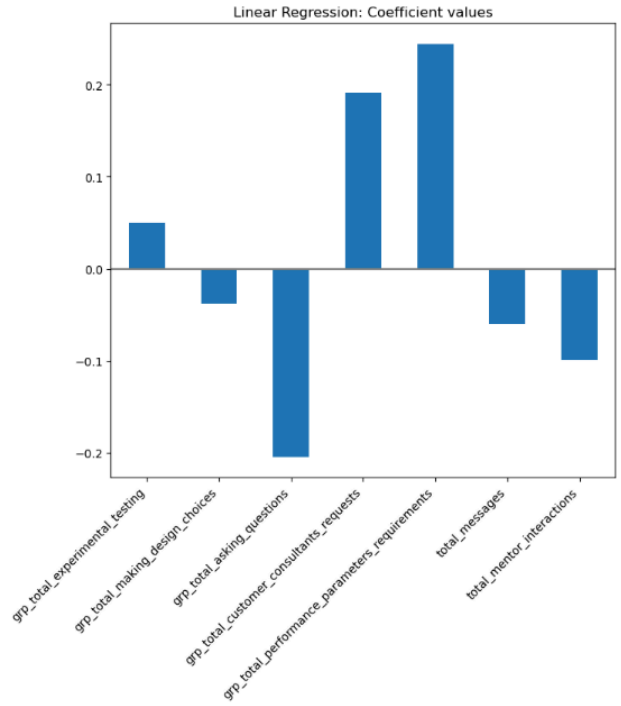

Figure 11 - Coefficient values on group level data

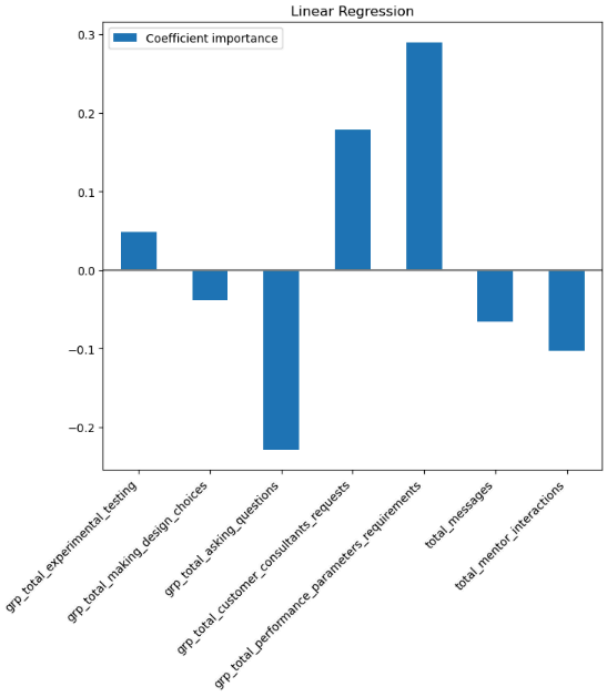

Figure 12 - Coefficient importance values on group level data

A further 15 Random forest regression and support vector machine models were created for each dataset but were deemed insufficiently accurate to be of any value or benefit to this report, especially as this is specifically a classification problem. Figure 17 displays one such model:

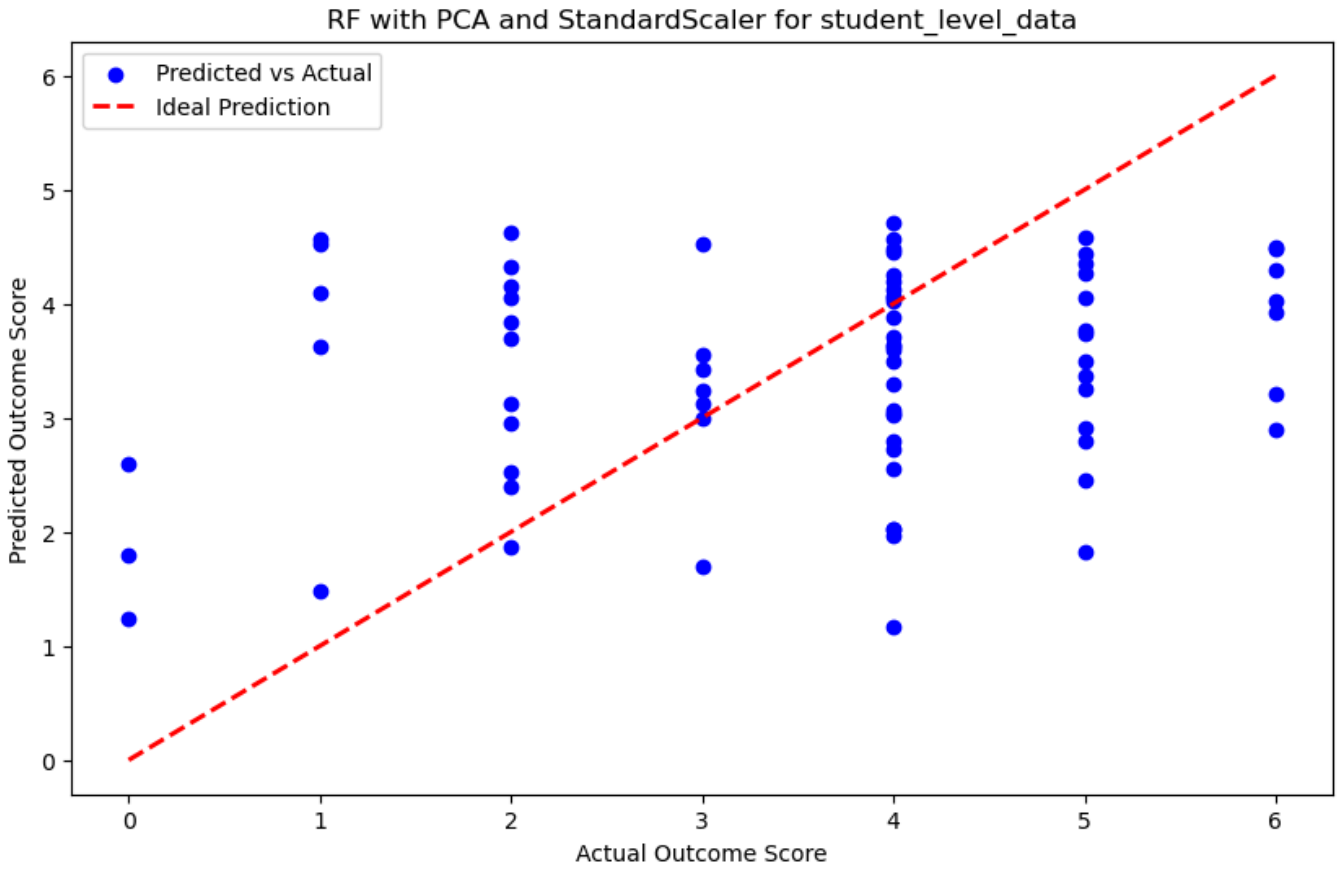

Figure 13 - Random forest with PCA and StandardScalar for student_level_data

Principal Component Analysis (PCA) with Logistic regression

PCA was conducted to reduce the dimensions of the data in hopes to obtain an accurate model. To set up this model, we first rounded all the mean values in the group level data to their whole number, allowing for computation of accuracy scores for later modelling evaluation.
To choose the number of components to use we first created a Cumulative Explained Variance Plot to assess how many components needed to exceed the 95% threshold. Figure 18, suggests that 3 is the optimal number of components.
To find similarities between groups in the group level data, figure 19 and 20 are scatter plots of the first 3 principal components. From figure 20, it shows that between each of the PC's, the patterns and structures are very weak, hinting that the variability captured by the principal components is not strongly associated with mean_outcome_score.

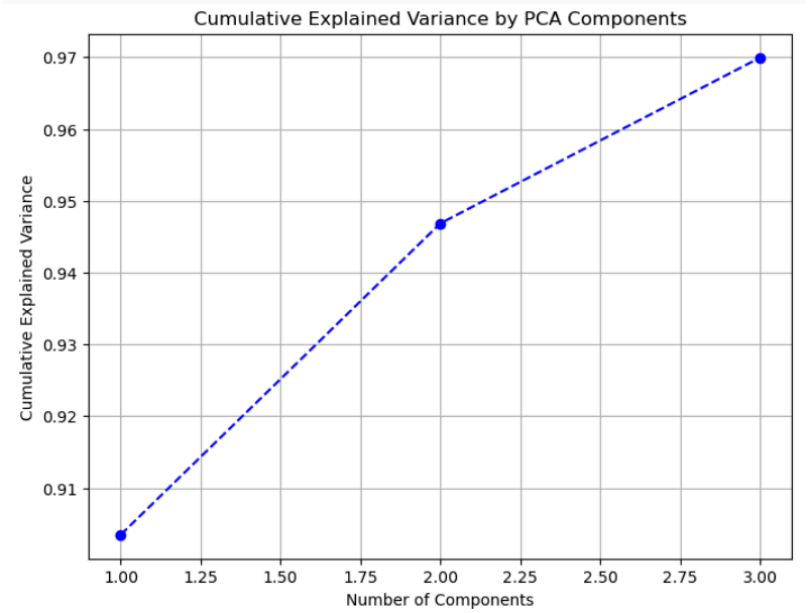

Figure 14 - Cumulative Explained Variance Plot on group level data

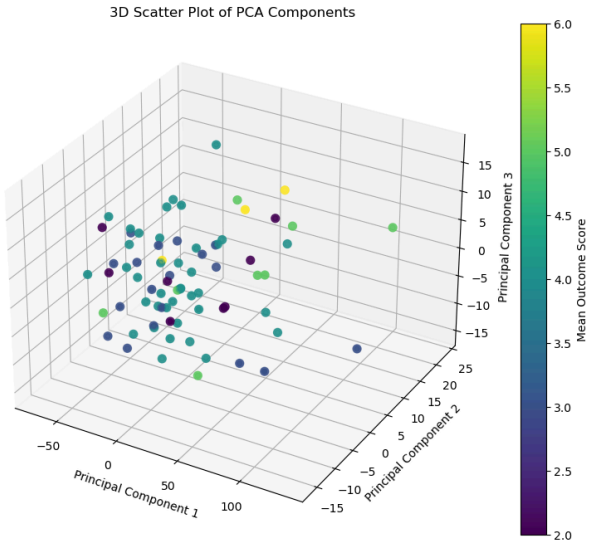

Figure 15 - 3D Scatter Plot of PCA components

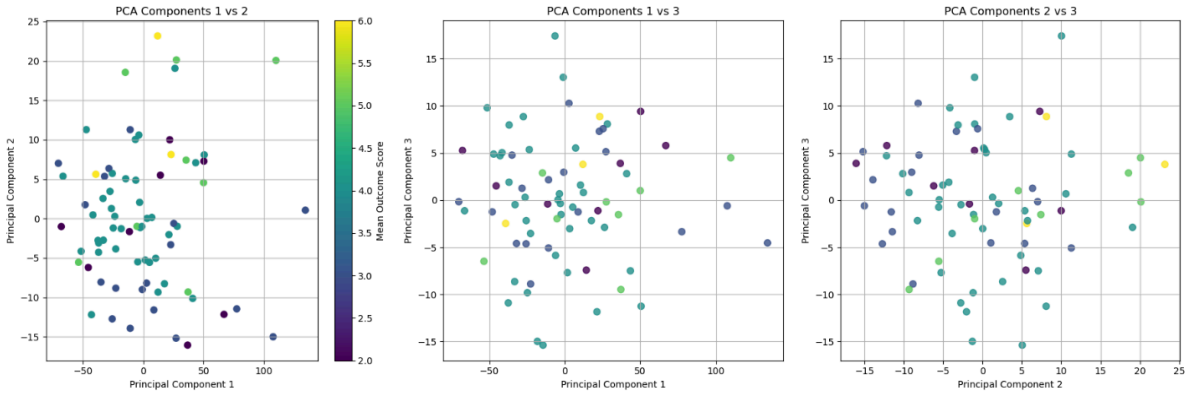

Figure 16 - 2D Scatter Plots of PCA components

Classification

Since our target variable was ordinal, we utilised various classification techniques to predict outcome score. We applied logistic regression, random forest, support vector machine classification and decision tree classification models to our 3 main datasets. Student, Group and Mixed. As we discussed in our pre-processing section, when building our models, we used a stratified 80-20 training and testing split, then we built pipelines to run our training data through to hopefully boost our accuracy. We built a total of 30 pipelines and built 39 classification models all together. The 30 pipelines consisted of variations of applying and not applying the model to our principal component analysis generated datasets with 2 or 3 components, using and not using SMOTE on our datasets, using each of our 4 classification models, with variations to SVM kernels, logistic regression max iterations, decision tree max depth and random forest n_estimators. This allowed us to quickly generate and compare various models, their confusion matrices, and their accuracies. We then created 3 new datasets containing the pipeline attributes, dataframe name, accuracy and precision. This allowed us to compare our models to see if we were gaining higher accuracy with our various modelling and preprocessing techniques.

Results:

Figure 21 presents the accuracy score and classification report of the PCA_3 Logistic regression model. 60% of the predictions made by the model were correct, performing below average. From the classification report, the model performs poorly for classes 2.0 and 3.0, while achieving relatively better performance for class 4.0.
From this model, to analyse the effect and importance each feature has on outcome score, we constructed a heatmap, as shown in figure 22. The strongest relationship results show that total_messages have a very strong positive relationship with PC1, total_mentor_interactions have a very strong positive relationship with PC3 and grp_total_asking_questions have a strong negative relationship with PC2.

```
Accuracy: 0.6
Classification Report:
              precision    recall  f1-score   support

         2.0       0.00      0.00      0.00         1
         3.0       0.00      0.00      0.00         5
         4.0       0.60      1.00      0.75         9

    accuracy                           0.60        15
   macro avg       0.20      0.33      0.25        15
weighted avg       0.36      0.60      0.45        15
```

Figure 17 - Accuracy and Classification Report of Logistic modelling with PCA

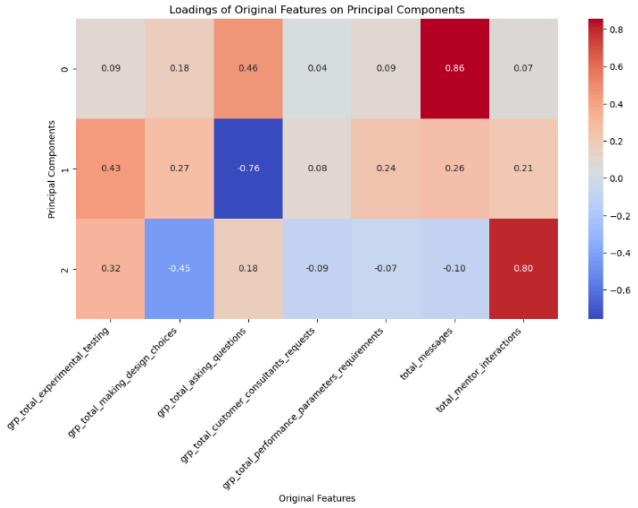

Figure 18 - Heatmap of feature influence on PC's

These results in tandem with the results from our multivariate linear regression earlier showing that rp_total_asking_questions has a negative relationship with mean_outcome_score and grp_total_performance_parameters_requirements, grp_total_asking_questions and grp_total_customer_consultants_requests variables have the highest effect in influencing scores. grp_total_performance_parameters_requirements and grp_total_asking_questions have a positive relationship with outcome score, are suggesting that groups that interact most with each other obtain greater outcome scores, leading us to believe that efficient teamwork leads to greater outcome scores and results. Furthermore, groups that receive more mentor interactions, have greater aid towards meeting their goals. This is parallel to our hypothesis that mentor aid is effective. Contrastingly, however, the more questions each teams have asked, lead to lower

outcome scores. This was highlighted during Multivariate Linear Regression, since mentor interactions seemed to have a positive influence on outcome scores, this leads us to believe that groups are asking more questions because they either aren't fully equipped with the knowledge to create a machine for kidney failure, or, the group is needing to clarify often with each other (asking more questions to each other) as they have a lack of cohesion and teamwork capabilities (lack of teamwork and cohesion intuitively would lead to lower outcome scores).

Having found the most accurate models for each data set, we were then able to investigate the confusion matrices of the top models for each.

Below are the top performing models for each dataset.

Dataset	Pipeline	Accuracy	Precision
both_data_sets	LR with PCA_3, 20000	0.35135135	0.16975194
both_data_sets	LR with PCA_3, 10000	0.35135135	0.16975194
both_data_sets	SVM with PCA_3	0.33783784	0.1141344
group_level_data	RF Only - Top 2 Featur	0.73333333	0.7037037
group_level_data	DT with PCA_3	0.73333333	0.68888889
group_level_data	RF with PCA_3	0.73333333	0.62666667
student_level_data	SVM with PCA_3	0.33783784	0.1141344
student_level_data	SVM with PCA_2	0.33783784	0.1141344
student_level_data	SVM Only	0.33783784	0.1141344
student_level_data	LR with SMOTE, 20000	0.33783784	0.1141344

Figure 19 - Tabe of models and their performance

Analysing their confusion matrices we can gather more insight into what they mean.

```
Dataframe: both_data_sets, Pipeline: LR with PCA_3, 20000, Accuracy: 0.3513
5135135135137, Precision: 0.16975194372454647
[[ 1  0  0  0  3  0  0  0]
 [ 0  0  0  0  5  0  0  0]
 [ 0  0  0  0 11  0  0  0]
 [ 0  0  0  0 10  0  0  0]
 [ 0  0  0  0 25  0  0  0]
 [ 0  0  0  0 12  0  0  0]
 [ 0  0  0  0  6  0  0  0]
 [ 0  0  0  0  1  0  0  0]]
Dataframe: student_level_data, Pipeline: SVM with PCA_3, Accuracy: 0.337837
83783783783, Precision: 0.1141344046749452
[[ 0  0  0  0  4  0  0  0]
 [ 0  0  0  0  5  0  0  0]
 [ 0  0  0  0 11  0  0  0]
 [ 0  0  0  0 10  0  0  0]
 [ 0  0  0  0 25  0  0  0]
 [ 0  0  0  0 12  0  0  0]
 [ 0  0  0  0  6  0  0  0]
 [ 0  0  0  0  1  0  0  0]]
Top Features for RF Only: grp_total_experimental_testing, grp_total_asking_
questions
Confusion Matrix (Top Features):
[[0 0 1 1]
 [0 3 1 0]
 [0 0 7 1]
 [0 0 0 1]]
```

Figure 20 - Confusion matrices for LR with PCA on both datasets, SVM with PCA on student level data and Top features for RF, respectively

For both student and mixed datasets, you can see that the models were prone to only suggesting that the answer should be 4. This is an example of why we used SMOTE, although it is interesting to note that because class 4 represented 33.8% of the data, that none of our models (including our ones where we oversampled minorities) could predict with any higher accuracy 33.8%. Our group level data however, was slightly more helpful, predicting with 73.3% accuracy predicting majority correctly for outcome scores 3, 4 and 5 based on total experimental testing and group total asking of questions. This highlights how dependent a student is on their group for getting good marks as the effort that a student puts in is relatively lost to chance but the effort that the group puts in is much more meaningful. In addition to that, it further promotes the idea that the fewer questions a group asks, perhaps the better the score, this would be supported by the fact that "group total asking questions" has quite a significant impact on predicting outcome score.

Below is the mixed dataset analysis with SMOTE:

```
Dataframe: both data sets, Pipeline: RF with SMOTE, Accuracy: 0.25675675675
675674, Precision: 0.25460687960687967
[[ 1  1  0  1  1  0  0  0]
 [ 0  0  1  0  2  2  0  0]
 [ 0  1  1  1  7  0  1  0]
 [ 1  2  1  2  4  0  0  0]
 [ 0  1  6  0 12  2  4  0]
 [ 1  1  2  1  4  3  0  0]
 [ 1  0  1  1  3  0  0  0]
 [ 0  0  0  0  0  1  0  0]]
```

Figure 21 - Confusion matrix for RF with SMOTE

What we can see from this is that applying SMOTE does mean the model is less skewed, less relying on change and hitting more of the diagonal making it a more valid model but also a significantly less accurate one. This possibly means that student level and mixed level data are much more difficult to predict outcome scores for as opposed to group level data. Further supporting the hypothesis that the effort one student puts in is not as important as the combined efforts of the group, we see that at a group level, it is much easier to predict outcome score than at an individual student level. The fact that the outcome score of 4 still dominates possibly means that there was need for even more balancing of the data. Perhaps NearMiss in combination with SMOTE would have led to higher outcomes.

The following confusion matrix shows the mixed group and student's data's highest SMOTE, outperforming all other Mixed data SMOTE models. It shows that it is made up of total mentor interactions and group total asking questions. Along with the results from the regression analysis, this once again suggests that total mentor interactions and total group questions asked have the biggest impact on a groups outcome score. With asking questions having a negative impact and mentor interactions having a positive impact, as seen in the regression analysis section.

```
Dataframe: Mixed_data, Top Features Pipeline: RF with SMOTE, Accuracy: 0.22972972972972974, Precision: 0.28236429707017946
Top Features for RF with SMOTE: total_mentor_interactions, grp_total_asking_questions
Confusion Matrix (Top Features):
[[0 0 2 0 2 0 0 0]
 [1 1 1 0 2 0 0 0]
 [1 0 2 0 6 2 0 0]
 [1 0 3 1 4 0 1 0]
 [2 0 3 2 8 8 2 0]
 [3 0 2 1 2 4 0 0]
 [0 0 1 0 1 3 1 0]
 [0 0 0 1 0 0 0 0]]
```

Figure 22 - mixed group and student's data's highest SMOTE

Conclusions

This research underscores the pivotal role of communication content in determining success of virtual internship projects. Findings suggest that the most effective teams engage in frequent communication and ask fewer questions. Suggesting a high level of independence is required to achieve a greater result. The exploratory data analysis conducted revealed that specific topics such as experimental testing, customer consultant requests and performance parameters had drastic effects on outcome score, Heatmaps showed positive correlation between these metrics and higher scores. Modelling highlighted the limitations of regression techniques due to the ordinal nature of the outcome variable. Principal component analysis (pca) helped identify key features such as total messages and mentor interactions as strong predictors of success. Thus another modelling technique such as logistic regression with pca achieved an accuracy of 60%. However, models based on group level data typically metrics on experimental testing and group questions provided better predictive capabilities and provided an accuracy of 73.3%. Overall, our results confirm that groups with higher mentor interactions and focused discussions on essential topics achieve better outcomes. Future research should explore advanced techniques like topic modelling and sentiment analysis to uncover deeper insights. Additionally, combining undersampling with oversampling could enhance model accuracy, and incorporating more variables might improve predictive performance. By incorporating these aspects, future models can offer more precise insights into the factors driving successful virtual internship outcome scores.