imperial college union

Imperial College Union Communications Committee

24 April 2018

AGENDA ITEM NO.	9
TITLE	Data analysis project update
AUTHOR	Andrew Keenan Head of Student Voice & Communications
EXECUTIVE SUMMARY	Lack of staff attention and exposure to busy student volunteers has delayed this project. These are my proposals to bring this project back up to speed.
PURPOSE	To update the Committee on the data analysis project
DECISION/ACTION REQUIRED	Commentary and approval.

Data analysis update

1. Current situation

1.1. This project has not progressed at an acceptable rate and the opportunity for it to influence the Leadership Elections 2018 is lost. Unless we move fast, the window to engage students (as volunteers or staff) and to influence 2018/19 planning will also be lost.

1.2. The two roots of this delay are my own failure to bring this to the top of my own priorities, and an excessive reliance on student volunteers.

1.3. Below I address these roots; in brief, I believe some elements of the task can be begun immediately rather than it all waiting for student staff or volunteers; and my own method of maintaining progress on my objectives has changed.

2. Root cause: reliance on students

2.1. The president of the Data Science Society is interested in the proposal and has agreed to promote it to their membership, which is a positive step. However there is work we can do locally, or work we can engage student staff for short periods, to deliver in the meantime.

2.2. The brief written by Paul Beaumont is appended in full below. Applying Paul's expertise means we have begun to break down our non-technical desired outcome into specific steps, some of which may be able to be started soon.

2.3. The steps which I believe are possible to begin without input from a data science specialist are to:

2.3.1. Establish a small set of detailed hypotheses regarding the voting likelihood of an individual and their pattern of engagement with Imperial College Union

2.3.2. Establish a written, GDPR-compliant statement setting out our case for which data we consider appropriate to use for this purpose

2.3.3. 'Data wrangling' – working with Systems to establish exactly what data we have, its format and accessibility, how far back in time it reaches, and mapping what may need to be done to bring together disparate and differently-formatted datasets into one form

3. Root cause: my prioritisation of this project

3.1. The April/May period sees a lot of planning activity, centred round budgeting. Feedback from my team, trustees and officers has highlighted the importance of effective and detailed planning for 2018/19, moving beyond budgeting into communications, training, governance and democracy. This process has freshly highlighted the value that this analysis can bring.

3.2. The new tactics of having an ICU-wide Business Plan and encouraging SMG members to share their overarching objectives has raised the importance of this project, and helped me understand its importance to our goals of improving engagement with the communities within our membership. I now discuss these objectives weekly with the MD and at least monthly with my entire directorate, giving more exposure to this project.

4. Next steps

4.1. I have begun to move forward with those parts of the analysis task that are within our present capacity. Any suggestions from the Committee on speeding those up, or ensuring I am held to account for this objective, are welcomed.

Paper by Paul Beaumont:

A possible election turnout modelling approach

Aims

1. Imperial College Union wishes to use their existing data on members to ascertain if there is any predictive power between cohorts, students' measurable interactions with the Union, and voter turnout.
2. Through this process the Union will ideally identify traits about voter turnout amongst it's membership that run deeper than high level stratification by UG/PG, EU/non-EU.
3. Understanding this will allow the Union to focus communications on those students who have some mid-range chance of voting and hopefully this will lead to an increased voter turnout, through more targeted messaging.
4. This is a semi-technical document with a proposed approach, meant to guide a student or group into tackling this problem. It is written without access or consideration to the data itself.

Process

1. The main body of work will lie in setting out a data pipeline to process any input data.
2. After wrangling, feature creation and selection can be likely used as input to an array of models.
3. Depending on the exact desired outcome, a predictive model should be chosen and run (including testing on a control set to evaluate predictive power).

Data Pipeline

1. The data pipeline is the software that manages the use of data throughout the modelling process. Inevitably this will start with wrangling the data and passing it to the predictive model, and may also include using said model on this years' data and making a prediction about the probability of a member voting.
2. To ensure re-usability of this model in future years, the pipeline's initial input should be from some reproducible export of the main ICU database, such that minimal wrangling lives outside the pipeline.

Feature Engineering

1. Features should be engineered to try explain cohorts, students' measurable interactions with the union and the likelihood of voting.
2. Various hypotheses may be formed and features should then try and reflect as many of the hypotheses as possible, as well as sufficiently many "other" features that could support either counter factual results or unrealised relationships.
3. Hypothesis based features could include:

Hypothesis	**Potential Features to try**
Someone who is a member of more clubs is more likely to vote than someone who is not a member of any	- Is someone a club member - Is someone in a (non-free) club - Has someone been a club member in the last 2 years

	- How many clubs is someone in - How many clubs has someone been in in the last 2 years on average • It is worth noting that including too many correlated features may detract from the model
Postgraduates who attended IC as an undergraduate are more likely to vote than other Postgraduates	- Did someone registered on a PG course attend for their UG? (fill 0 to all UGs) - Their year of the PhD (i.e. to pick out if they're writing up/final year)
Someone who's (been at IC for more than one year) and voted in a previous election/engaged is more likely to vote in this election	- Voted previously - Submitted a SACA nomination previously
Mature students are less likely to engage with the Union	- Students' age
Students from certain countries don't engage fully with the Union	- Categorical feature for each country

4. Nulls in features should be dealt with via some standard imputation strategy, or by disregarding of the data.
5. Normalising the feature space will help with analysis of the model (see later).

Feature Selection, Training & Test Sets

1. Consideration should be given to the multi-collinearity of features, and suitable assessments should be made to remove highly correlated features (using Variable Inflation Factor assessment, or similar).
2. The pipeline will need to distinguish automatically between student data for training, testing (pre-"this year") and data for prediction ("this year").
    a. Both the training and test set should be formed from only pre-"this year" data.
    b. Pre-"this year", the data should be split into appropriate training/testing sets. This could be done by (say) taking all data pre 2017-18 as the training set, and 2017-18 as the testing set. 2017-18 could then be used for evaluating the predictive power of the model.
    c. 2018-19 data would then be for prediction.
3. Features should only be variables for which information is available in the 2018-19 data already.

Modelling

1. Whether a member votes in a particular Summer election cycle is a binary outcome: Yes, or No. This can be modelled by a Logistic Regression model.
2. The model should be trained on the training set only, with the test set left for assessing how good the predictive power is.
3. When validating the model an appropriate threshold (0.5) may be set to test the accuracy of the model. A Confusion Matrix may be used to identify how good the model is (on the test data).

4. When using the model for prediction, the output will be a probability. This can be used directly to rank students in the manner required by the Aims of this work.
5. A more complicated methodology would be to use cross-validation and then take an average of the models. This approach can probably be avoided for reasonably large training/test sets however.

Analysis

1. Using the model to make predictions for this year's data is sufficient to rank the membership into likelihood of voting.
2. The sign of the coefficient of the features in the model will identify the positively and negatively contributing traits to whether a member is likely to vote. If the features have been normalised, the size of the coefficient will give an indication as to the importance of the trait being described by the feature. These can then be used to inform targeted messaging in email communications.

Alternative/Extra Modelling Approach

1. The most important features could also be used as the basis for some clustering analysis that sought to group students by their likelihood to vote, and their shared characteristics (which may help with targeted marketing).