STINT Spring Hackathon: Rationale

Introduction

The general idea at the basis of the algorithm we developed is to improve the efficiency of the process which currently matches STINTs to businesses, rather than limiting ourselves at mimicking it.

Because of this, we decided that training our project employing ML or similar methods would not produce the results we are aiming for, since these methods are based on understanding the criteria and parameters that humans have used up to now to make the matchings. Rather, we will create our own parameters, giving them the weights and importance that was statistically determined to be the most suitable.

Another advantage of this method is that it allows the code to be extremely more flexible and easily adaptable to new implemented variables. This means that if, in the future, new factors will become relevant to the match-making process, they will be very easily implemented, simply by rearranging the various weights given to the other parameters, and creating new functions to calculate the new ones.

The philosophy behind the algorithm is to result in the best possible experience for both the business and the student. This will, of course, be achieved by matching the most suitable student to a specific business. There are multiple variables at play which might influence the algorithmic decision-making process, the key ones being the ones specified in the following pages. All of these will be weighted and added together to produce a final parameter of desirability (Ξ) — more on this below.

Before delving into the variables that were identified to be the most crucial ones, it is necessary to describe the process of variable normalisation which was employed throughout the project. Normalisation is essential to produce a sensible and meaningful result, and each of the constant parameters of the normalising functions has been chosen by considering the way it modifies the shape of the normalised curve, and, most importantly, by analysing significant data samples on MATLAB and calculating the standard deviation, mean, and overall distribution of the resulting normalised parameters.

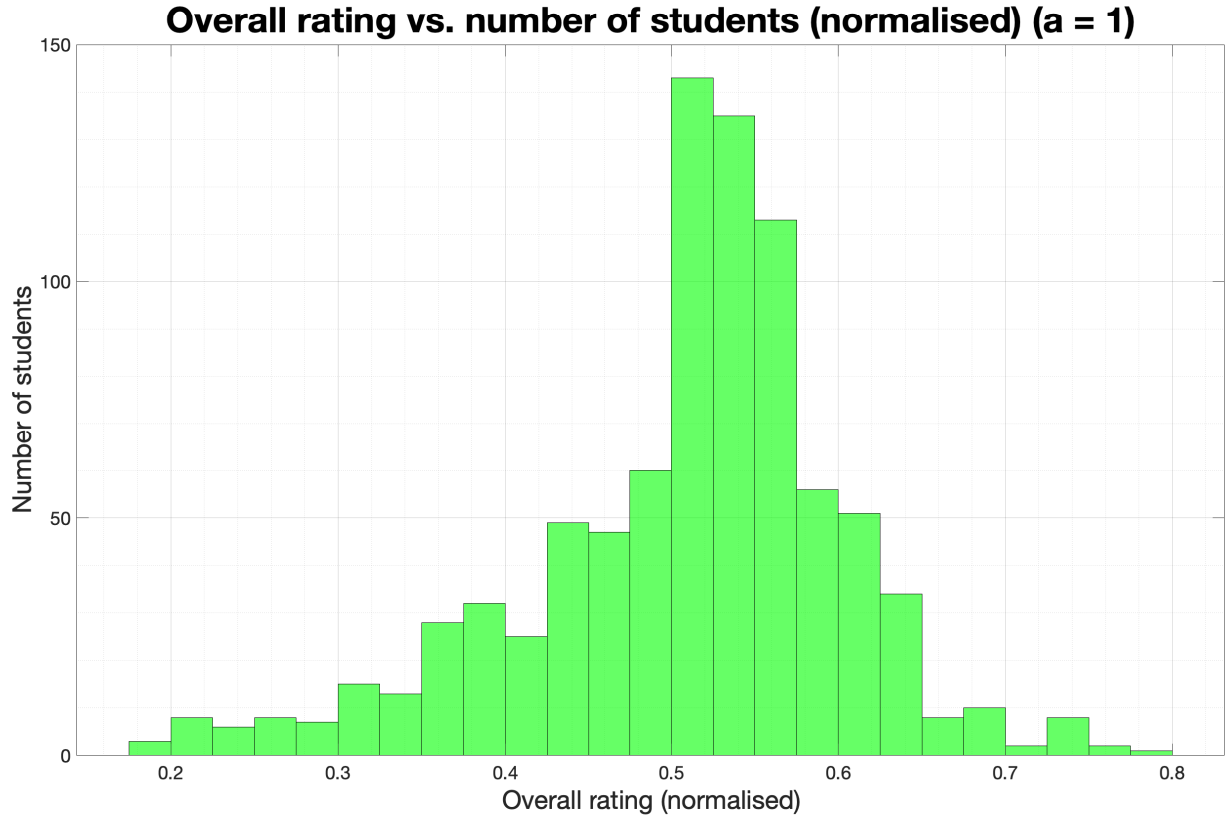
Overall rating

The first variable taken into consideration is the student's overall rating, $\mathbf{R}$, calculated as:

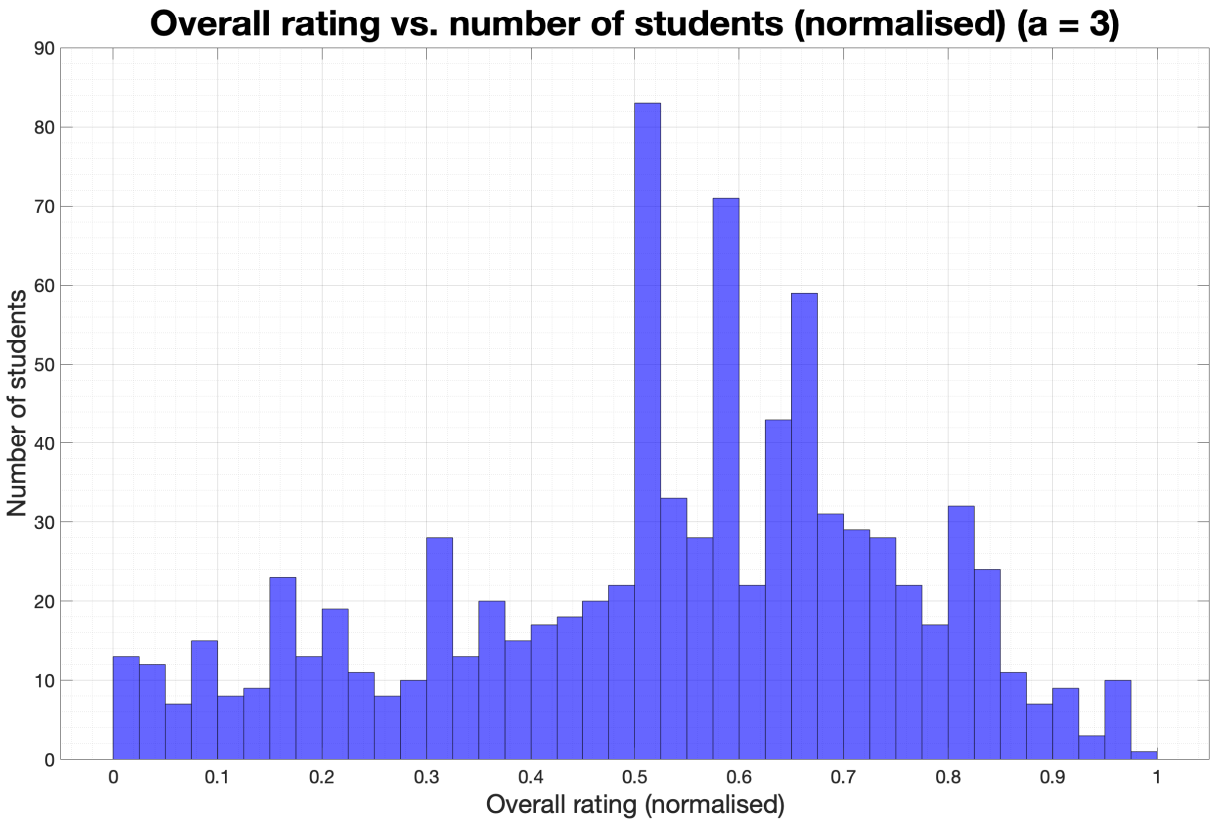
$$\mathbf{R} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \frac{\left(\mathbf{R_i} - \overline{\mathbf{R}}\right)}{\mathbf{N}},$$

where $\mathbf{N}$ is the total number of STINTs completed by the student in any job category and $\mathbf{R_i} - \overline{\mathbf{R}}$ is the difference between the single rating received by the student from the employer for a single job and the average rating the employer gives to employees. This value will vary between ± 4, and will therefore be normalised to vary between 0 and 1. The parameter $\mathbf{R}$ must be implemented in order to increase the statistical significance of business reviews, since the ratings that businesses give on average may drastically change from one to the other. The overall rating parameter gets around this issue by taking into consideration whether the individual rating is below or above average, rather than its absolute value, which would be, on its own, statistically insignificant.

Lastly, in order to normalise $\mathbf{R}$, we let $\mathbf{R}$(Normalised) $= \dfrac{\mathbf{e^{3R}}}{\mathbf{e^{3R}+1}}$. This is because, by definition, the average value of $\mathbf{R}$ will be around 0, and such a normalisation makes it so that the biggest changes in $\mathbf{R}$(Normalised) will occur for values of $\mathbf{R}$ close to 0, the average, to better differentiate between students above and below it.



The value of the coefficient (a) which multiplies $\mathbf{R}$ in the formula above is chosen in order to have the normalised curve acquire an appropriate slope in the area around the mean, so as to set the rate at which $\mathbf{R}$(Normalised) varies with $\mathbf{R}$.



Distance and Duration

Other two crucial variables that allow the algorithm to make an informed and appropriate decision are the distance between the student and the business, $\mathbf{D}$, and the duration of the STINT itself ($\mathbf{T}$). These two variables have to be considered together in a cohesive way, since, clearly, travelling, for example, 5 km for a one-hour-long STINT is not as desirable as it would be to travel 5 km for a four-hour-long STINT. The aim of this particular implementation was to increase efficiency, and make it easier for people closer to the job location to be matched with the employer.

A formula was implemented to allow for the creation of a single variable, $\mathbf{D_T}$, which relates these two quantities in the most appropriate way, and normalises them at the same time:

$$\mathbf{D_T} = \frac{\mathbf{e^{-(D-1.9T)}}}{\mathbf{e^{-(D-1.9T)} + 1}},$$

where $\mathbf{D}$ is the distance between the business and the student (in km), and $\mathbf{T}$ is the time duration of the STINT (in hours).

The initial idea was to set a number of conditions to filter out students that lived too far away relative to the duration of the STINT itself. However, by doing this, the criteria would be applied too harshly, meaning that, for instance, a student living 7.1 km away would not be considered, while someone 6.9 km away would. To solve this issue, the formula above was implemented, which allows for an extremely smooth and continuous transition between the values of distance and duration that make the student suitable for the job, and those that make him/her unsuitable.

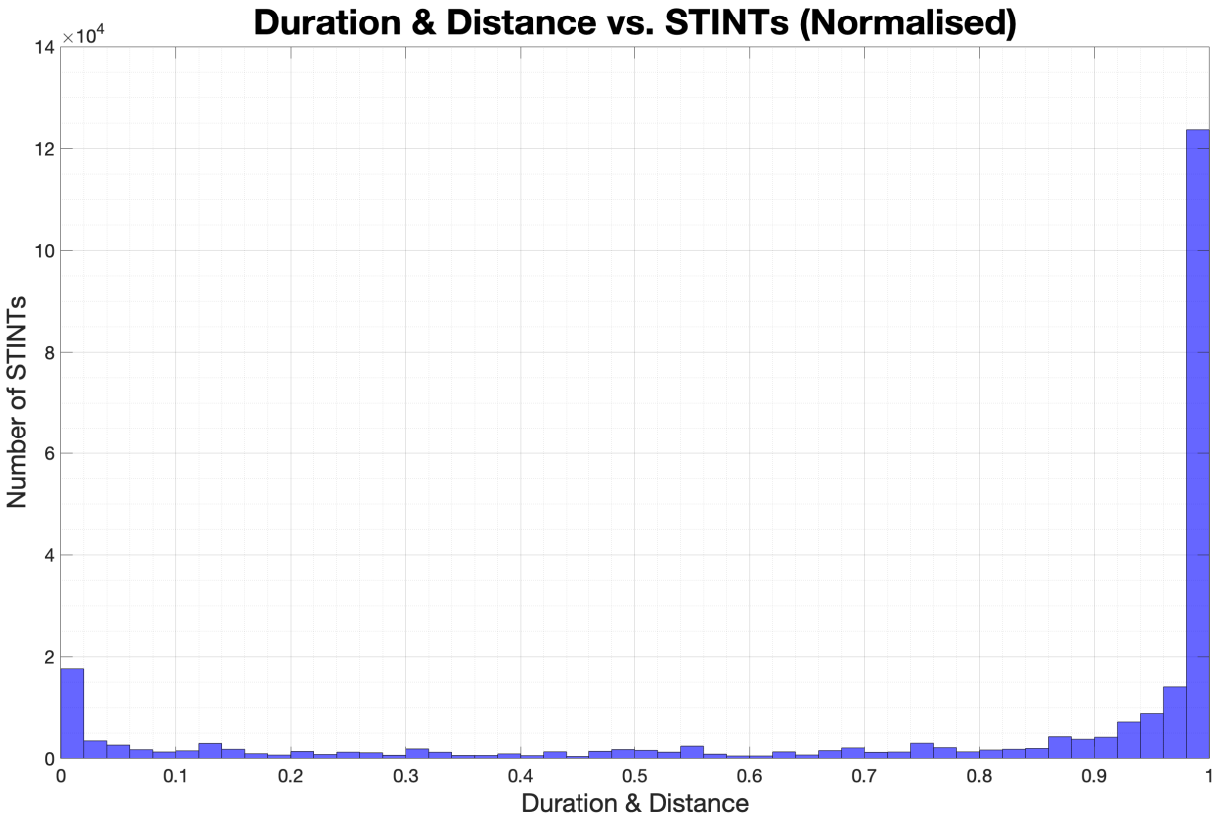
Now, to give an idea of how this formula works, different cases are analysed, showing distance, duration and the overall suitability for the STINT based solely on the two factors.

	Duration	Distance	Suitability (D_T)
Case 1	2	2	0.8581
Case 2	2	6	0.0998
Case 3	4	2	0.9963
Case 4	4	6	0.8320
Case 5	4	9	0.1978
Case 6	6	6	0.9955
Case 7	6	11	0.5987

	Duration	Distance	Suitability (D_T)
Case 8	6	14	0.0691
Case 9	8	14	0.7685

As it can be easily seen, when a student is too far away from the business' location, even if the duration of the stint is high enough, the suitability will be very low.

This was implemented because of the idea that, over a long distance, it is risky for the business to accept a student, since there are many factors that could cause his arrival to be delayed. This effectively means that, over short distances, this formula enhances the experience of the student (as he/she will be assigned to a STINT close to his/her location), whereas, over longer distances, the formula enhances the business' experience, in that the students who are too far away (and are therefore at a higher risk of delays etc.) will not be assigned to it.



As can be easily deduced from the graph, this geofencing feature does not exclude too many students, as most of them have high values of this parameter.

STINTs completed by the student in a specific role

This parameter — referred to henceforth as $\mathbf{N_S}$ — takes into account the number of STINTs completed by the student in a specific job category (i.e. bartending or waiting on tables). This variable was normalised between 0 and 1 via the following equation:

$$\mathbf{N_S}(\text{Normalised}) = \frac{\mathbf{e^{0.8(N_S-4)}}}{\mathbf{e^{0.8(N_S-4)}+1}}.$$

This makes it so that the biggest change in $\mathbf{N_S}$ occurs around the value of 4. The 0.8 factor which multiplies the exponents reduces the rate of change of this parameter, so that there is a smaller difference between students who completed just one or two STINTs more or less.

Rating for a specific job role

The student's overall rating for a specific job role, $\mathbf{R_S}$, calculated as:

$$\mathbf{R_S} = \sum_{\mathbf{i=1}}^{\mathbf{N_S}} \frac{\left(\mathbf{R_i} - \overline{\mathbf{R}}\right)}{\mathbf{N_S}},$$

where $\mathbf{N_S}$ is the total number of STINTs completed by the student in the specific job category and $\mathbf{R_i} - \overline{\mathbf{R}}$ is the difference between the single rating received by the student from the employer for a single specific job and the average rating the employer gives to employees for that specific job role (which might be different from the parameter used above to calculate the overall rating, $\mathbf{R}$, in the case that the employer offers more than one kind of job type). $\mathbf{R_S}$ was normalised in the same way in which $\mathbf{R}$ was, i.e.:

$$\mathbf{R_S}(\text{Normalised}) = \frac{\mathbf{e^{3R_S}}}{\mathbf{e^{3R_S}+1}}.$$

STINTs completed by the student

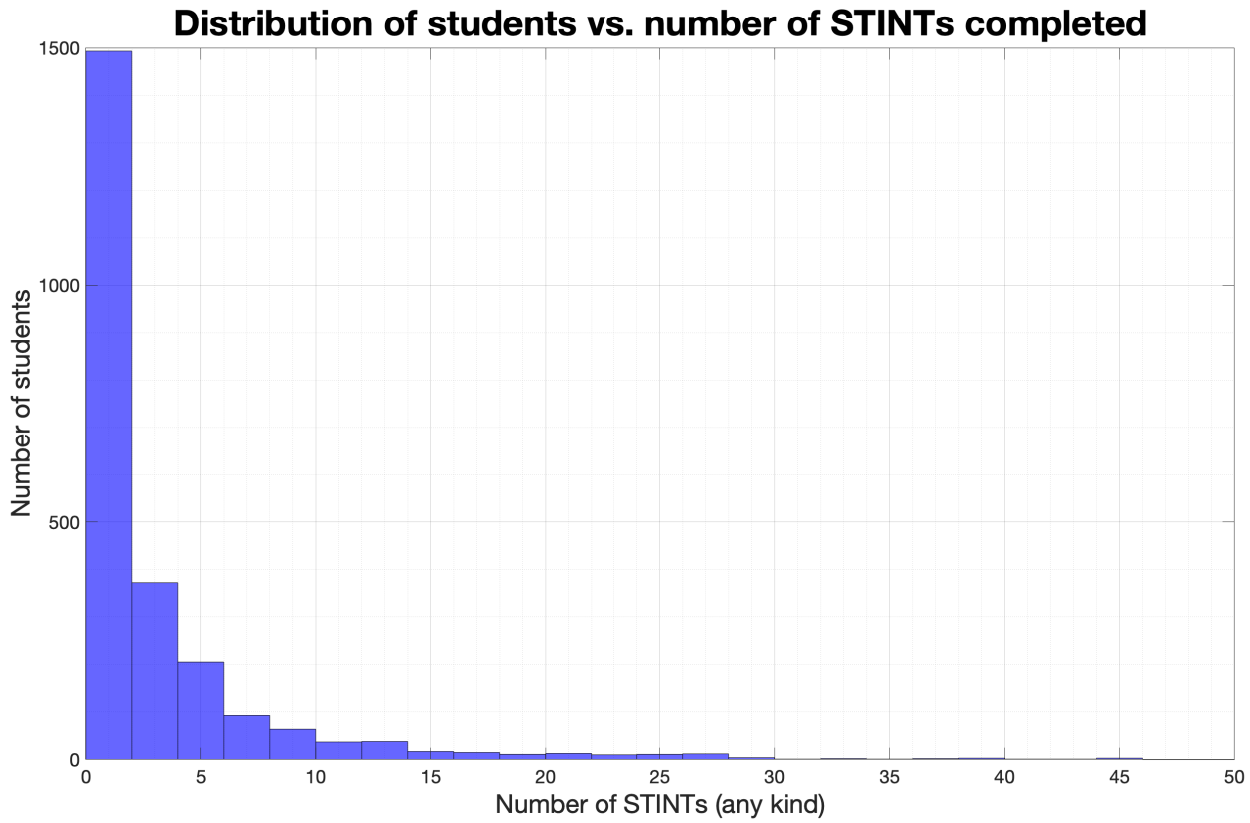
This parameter — referred to henceforth as $\mathbf{N}$ — takes into account the number of all STINTs completed by the student (regardless of job category).

While previously, when calculating $\mathbf{N_S}$, only the student's experience regarding a specific job area was considered, the parameter $\mathbf{N}$ can be used to analysed and model the overall experience of the student. This is crucial, in that it allows the student to be considered for new jobs also based on his loyalty to STINT. This feature, albeit undoubtedly necessary, must, at the same time, be considered less than $\mathbf{N_S}$ in making a matching choice.

This variable was normalised between 0 and 1 via the following equation:

$$\mathbf{N}(\text{Normalised}) = \frac{\mathbf{e}^{\mathbf{2.8(N-1.8)}}}{\mathbf{e}^{\mathbf{2.8(N-1.8)}} + \mathbf{1}}.$$

This makes it so that the biggest change in $\mathbf{N}$ occurs around the value of 1.8. The 2.8 factor which multiplies the exponents increases the rate of change of this parameter, so that there is a bigger difference between students who completed just one or two STINTs more or less. The normalisation for $\mathbf{N}$ is slightly different than that of $\mathbf{N_S}$. This is because most students will have a higher $\mathbf{N}$. Below the distribution of $\mathbf{N}$ is shown.



Desirability

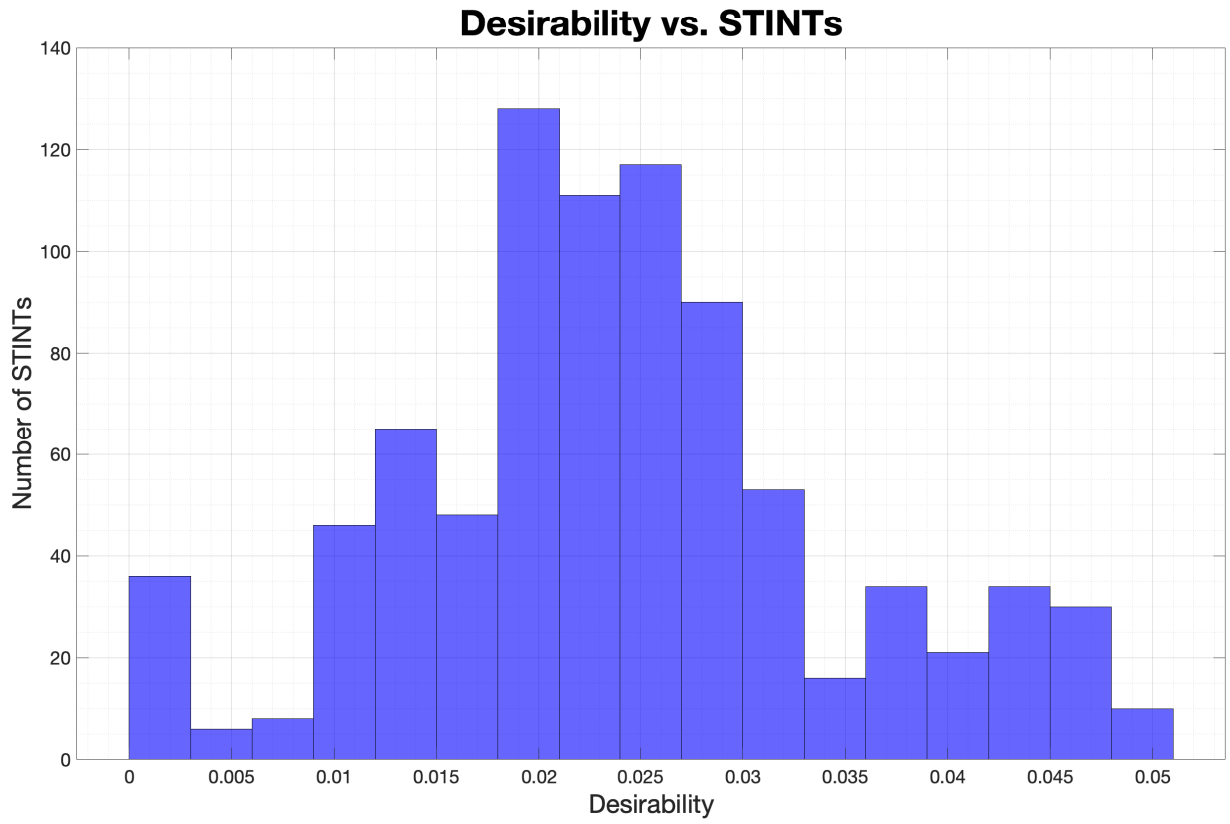
Desirability is the final parameter employed to rank students according to their qualities compared to a specific business' needs. This is defined as:

$$\Xi = \mathbf{N}\frac{\left(\alpha \cdot \mathbf{R_S} + \beta \cdot \mathbf{N_S} + \gamma \cdot \mathbf{R} + \delta \cdot \mathbf{D_T}\right)}{\alpha + \beta + \gamma + \delta},$$

where $\mathbf{R_S}$ is the rating achieved by the student in the specific job type being taken into consideration, $\mathbf{N_S}$ is the number of STINTs completed by the student in the specific job type being taken into consideration, $\mathbf{R}$ is the overall rating achieved by the student (see above), $\mathbf{D_T}$ is the "distance and duration" variable described in the corresponding paragraph above, and, lastly, $\mathbf{N}$ is the total number of STINTs completed by the student in any job category. Having established the relative relationships between the greek-letter coefficients as follows:

$$\alpha = \beta > \gamma > \delta,$$

the following weights were chosen: $\alpha = 3$, $\beta = 3$, $\gamma = 2$, $\delta = 1$. Clearly, a higher desirability parameter will result in a higher likelihood that the student will be matched with the corresponding business. The distribution of desirability across a very wide range of students is graphed below:



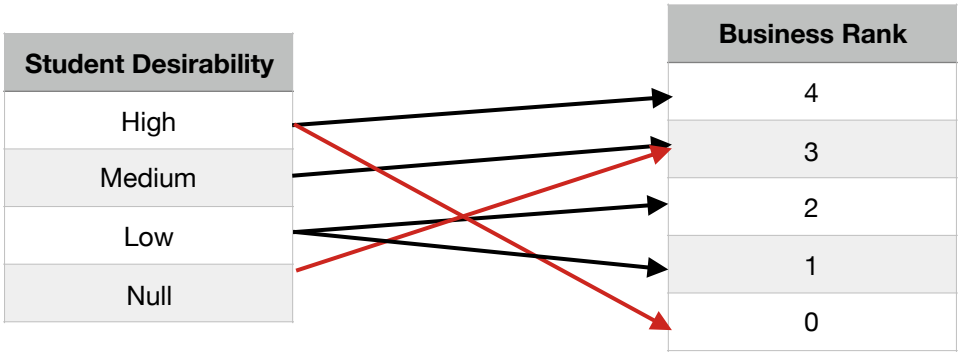
This desirability parameter, will not, however, apply to all student-business matchings. This is done in order not to penalise students who have yet to complete a reasonable number of STINTs: since their $\mathbf{R}$, $\mathbf{R_S}$, $\mathbf{N_S}$ and $\mathbf{N}$ parameters will be fairly low, their desirability Ξ would be, too.

In order to address this issue, the following steps are taken, which apply to students who have only completed between 0 and 2 STINTs, and businesses who have only hosted between 0 and 4:

- New students are assigned to average-ranking businesses (level 3) (cf. Business ratings), so that they'll have enjoyable and productive first experiences, making them more willing to work for STINT again in the future. The reason why they are assigned to average businesses is that, this way, they will still like working at that business (as opposed to working at a low-ranking one), whilst not risking STINT's reputation with high-ranking businesses, which are statistically more demanding. Such a resolution will reduce the liabilities linked with new students of unknown skills, and will be implemented by having **one of the five students assigned to average businesses be a newbie**.

- New businesses, whose business rank will be null, will be assigned **only very good students** for the first 4 STINTs they post, so that they will have a good first impression of the students, resulting in a higher likelihood of their continuing to use the app.

A rough summary of the matching between students and businesses is shown below. The features highlighted above are shown as red arrows:



Fast-tracking

Apart from desirability (with the due exceptions mentioned above), another parameter was introduced in order to best match businesses and students. In fact, at the very beginning of the algorithm, before ordering workers based on their respective desirability, a student will be "fast-tracked" to a certain business (i.e. he will be matched with a business regardless of other parameters) if both the following conditions are met:

- The student has already worked before with the specific business.
- The average of the ratings the student has received from that specific business is higher than the average rating that business gives to all its STINTs.

This "fast-tracking" feature is implemented in order to match businesses with students that already worked with it, and whose performance the businesses were satisfied with. This way, the creation of a desirability variable is no longer needed, since it is known that that specific business was already satisfied with that student in the past, meaning it would likely be happy with that happening again. Furthermore, this enables the student to work multiple times in the same workplace, contributing to improving their job skills, allowing for even higher business satisfaction. Lastly, in case the conditions for "fast-tracking" apply, this feature makes the match-making process much faster and more reliable.

The distribution of business ranks is shown below. This was used in order to divide businesses in even categories (High, medium, low and null), which were used for fast-tracking and desirability.