Exploratory data analysis visuals

As observed, asking questions was the most common type of message being sent out by interns, whilst the least common message being sent out was of the communication type. Communication could be the least common type of message sent out because of the very general definition it holds, meaning that not many messages may be classified into this category. For example, messages that do not fulfil the appropriate criteria to be classified as any other type of message may fall into this category. Asking questions may be the most frequent type of message being sent out by students because it is implied that the only means of acquiring assistance from mentors would be to ask for advice through the messaging application. The next most frequent type of message being sent out was of the type, m\_making\_design\_choices which may imply that interns were most frequently using the messaging platform to express their decisions, as opposed to perhaps using consultants’ advice on their projects because not many messages of the customer consultation type were sent. In the next plot, 4 is the most common score, thus suggesting that it may have been the average score achieved by interns. The number of students achieving a score above average is slightly lesser than the number of students achieving below average scores indicating only a slight skew in the data to the left which does not significantly impact results in a statistical sense.

Analysing at a further depth, we can observe that the distribution of messages in terms of frequency of type of messages sent remains relatively consistent whilst the scores of interns vary independently from this distribution, indicating yet another potential sign of a lack of correlation between these variables. A more reliable measure can be used, which is the correlation matrix. Affirming our previous observation, the correlation matrix does indeed show relatively low coefficients for all variables when compared to outcome score. Furthermore, the moderately low correlation scores between most of variables avoids the issue of multicollinearity which means that should a variable appear to affect outcome score, its individual effect is easily distinguishable. Much more concrete results can be found using the feature importance functions of the corresponding appropriate models.

Modelling

Decision tree important features: j\_performance\_parameter\_requirements, m\_ask, j\_comm, m\_making\_design, m\_exp, extremely high gini scores, accuracy is low, sensitive to overfitting

Rf classifier: importance of features found but accuracy score is low. Effect of important features on score