a) What is your improvement strategy? What techniques would you apply to improve the quality of the output?

Some of the following steps were not taken due to lack of information/ time constraints.

1. Many columns had a lot of 0s in them. For example, there are records with n\_unique\_tokens=0 and num\_imgs= 0 and num\_videos =0, which is not sounding okay. This means there is neither text, nor image nor videos. This is suspicious as there can’t be a page with just hyperlinks or something else which is not captured properly. These records can be deleted.
2. 16737 records have ‘kw\_min\_min’ = -1. This is suspicious. This could be an error while capturing the content using NLP. Minimum share of the worst keyword can be 0 and not -1. But I’m not sure what is the cause of -1 and hence no action was taken. With more knowledge, I would be able to replace the -1s with 0s or median value of the column.
3. Though there is very low correlation between features and the target, some of the features are highly correlated. For example: LDA\_02 & data\_chanel\_is\_world, self\_reference\_max\_shares & self\_reference\_avg\_shares, kw\_max\_min & kw\_avg\_min. Model performance could improve by keeping only one of the features.
4. Dimensionality reduction was not performed. It could improve model performance.
5. An ensemble of ensemble models can be tried so that the scores by each model will considered and an average will be given.
6. Wanted to use GridSearchCV but due to performance issues, it could not be done.
7. Hyperparameter tuning is leading to long runs (performance issues) and hence it was not done but it should be done before selecting the final model.
8. Other feature selection techniques can be tried like: RFE, Forward stepwise selection.

b) What is the differentiating factor of your solution, how would it rank over solutions from other data scientists?

1. Before starting with data modelling, extensive EDA was done to understand the data but due to lack of clarity and lack of time, many methods could not be implemented.
2. Skewed data was identified, and steps were taken to normalize them.
3. Simplicity of code and usage of simple models.