**Project pipeline:**

The following is the recommended project structure:

1. Problem Definition
   1. Define the problem you’re trying to solve.

In the world of Instagram, Instagram`s influencers, and advertising companies there is a constant dilemma which content will go viral and why. For example, is it better to advertise food for humans with animals or on the beach?

The worth of this industry is about 15X109 $ worldwide. Therefore, it is crucial to know in advance what will go viral and get the best influence. In order to provide this knowledge to advertising agencies, we will build a site that will receive a photograph as input and the output will be the engagement rate.

* 1. Define the scoring to be used and a reasonable target.

We will use Mean Absolute Error as the scoring method, the objective MAE is 0.02 % the accuracy of the model should be above 95% so the clients will be calm while relying on us.

* 1. Define the data you’d like to have. Compare it to the data you actually have. Both in quantity and features.

We will gather and scape posts and photos of some Instagram`s influencers with the number of likes, comments and followers. We don`t need any more information as for today.

* 1. Define the conditions to use the algorithm.

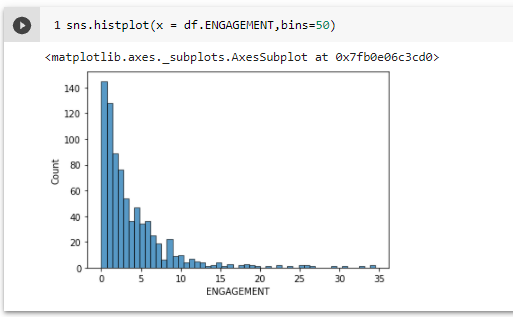
The algorithm will be the engine in web or app. It should be able to handle 1000 entries/hour.

1. Data receiving:

Receiving data by either getting it by scarping it from Instagram or receiving the data in excel.

1. Data Exploration:
   1. Using OpenCV to transform images to strings that describes it.
2. Present data. Explain why it was presented in the manner chosen.  
   For example, Use Histograms / Violin Plots to present data which is statistical and infer its Dynamic Range, Spread, Histogram of Distances, etc.

*The data is very imbalance:*

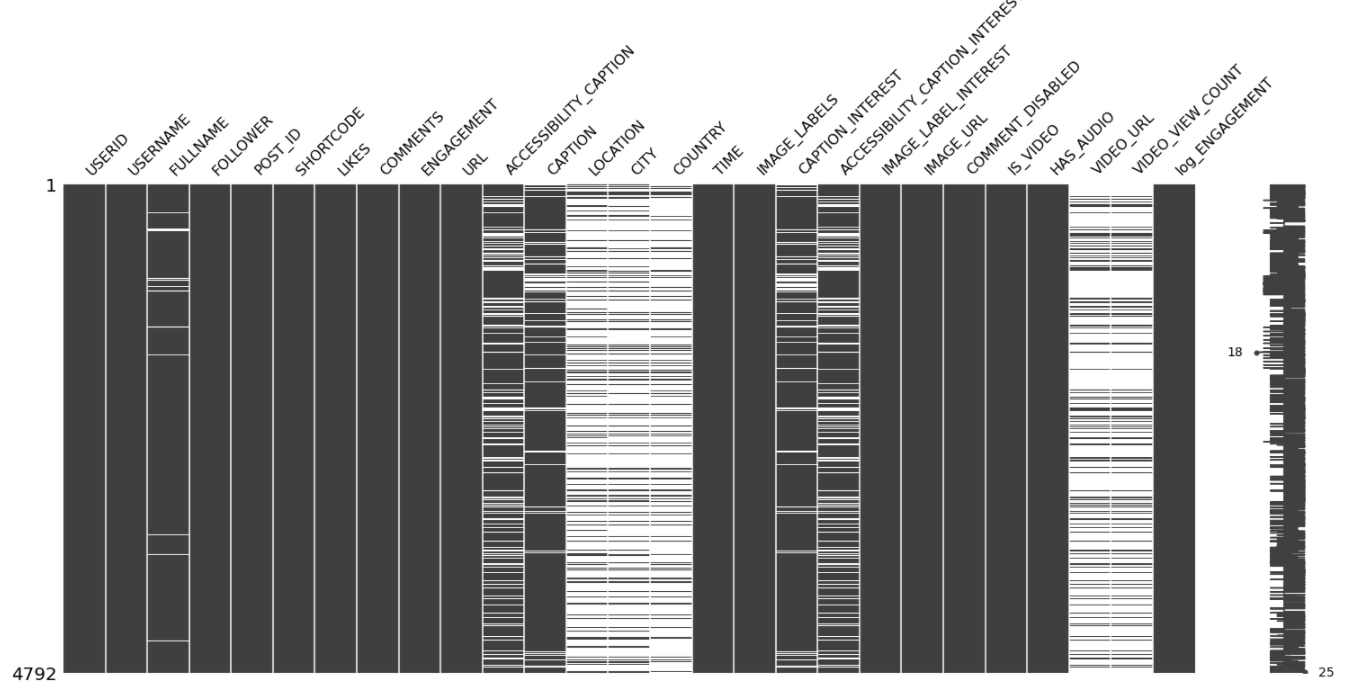


1. Make some conclusion on the data based on the visual information.  
   For example, look at the corr() plot and say something about the linear relationship between different features.
2. Try estimating the difficulty of achieving the target score.
3. Data Pre-Processing
   1. Handle missing data. Explain the method used (Removal, replacement, etc…).

*Currently we replaced all null with NULL. We could not remove the lines with missing data because 1. We didn’t have enough data 2. The columns are important.*

*Missing data on columns:*

'ACCESSIBILITY\_CAPTION','CAPTION','LOCATION', 'CITY', 'COUNTRY','CAPTION\_INTEREST','ACCESSIBILITY\_CAPTION\_INTEREST

**

* 1. Remove outliers. Explain the model used to infer an outlier.

*There are little post with extreme Engagement and Video counts (Sigma), what should we do with it?*

* 1. Handling the tuples hashtags in different columns.

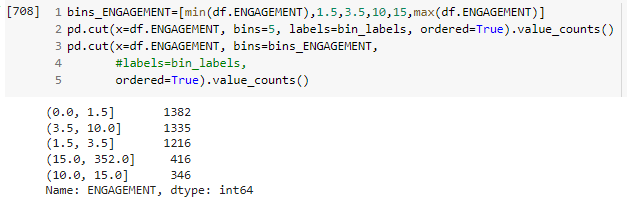
1. Feature Engineering
   1. If data requires, build new features from existing features. Analyze their contribution.
   2. Do estimation about feature importance. Choose the features relevant to the task.

*The most important features are:*

* + 1. *Followers*
    2. *Video count – should we remove it (as I cause data leak to y)*
    3. *User name*
    4. *ACCESSIBILITY\_CAPTION*
    5. *CAPTION\_OBJECT*
    6. *IMAGE\_LABELS*
    7. *CAPTION\_INTEREST*
    8. *Time -> hour minutes*
    9. *LOCATION*

1. Regression / Classification
   1. Choose the methods which are applicable to your problem. Explain your choices.

*We create 5 buckets:*



* 1. Build the testing environment. Explain the choices made (About the size of the test set, which type of cross validation, Grid search, etc…).

*Currently we have over fitting issue*

*We will handle it with more data and improving the algorithm (grid source)*

* 1. Apply the chosen method. Compare the scoring. Try different scoring and explain results. Use cross validation to search for optimal hyper parameters.
  2. Try a method you excluded on (a). Compare it to the results. Was it a good decision to exclude it?
  3. Go back to step (2) and use the insight gained.

1. Summary
   1. Present summary of results. Declare whether you failed or succeeded to reach the objective.
   2. Make some remarks about what it will take to move the project into production. Address things like the scalability of the model, how would you handle new data, etc...
   3. Provide some recommendations for future work (Extensions).
2. Presentation  
   It is recommended to add a presentation to the project. It is **mandatory for the final project**. The presentation should be short and focused.
   1. Slide 01 - Project name & authors.
   2. Slide 02 - Present the problem and objective with target score.
   3. Slide 03 - Present data (With a link to source).
   4. Slide 04 - Present the Pre-Processing process.
   5. Slide 05 - Present the feature engineering process.
   6. Slide 06 - Present the data for the predictor on its final state.
   7. Slide 07 - Present methods used as predictors.
   8. Slide 08 - Present main results and comparison of methods.
   9. Slide 09 - Present important “Trick” or contribution to results.
   10. Slide 10 - Present the summary and if the objective was achieved.
3. Presentation to Class  
   Make sure to rehearsal the presentation. Aim for 15-17 minutes to leave some time for questions.

Remarks

* Do the project in Jupyter Notebook (Google Colab).
* You may summarize it into a Google Slides presentation (Mandatory for final project).
* You may do it as individuals or teams. I suggest that the team will be no more than 2 members.
* The structure above is the suggested one. You may do a subset of it according to the time budget. Yet make sure you follow the structure as a guideline. Make sure you at least get to the analysis section with your time budget.
* A good project measured by the quality of the work and not necessarily if you achieved the scoring goal.
* The final project is expected to go deeper in each section.
* I recommend choosing a single data set to be used along all projects.

Elad and Raanan requests:

1. Image connected to post
2. Fill missing data on columns:

'ACCESSIBILITY\_CAPTION','CAPTION','LOCATION', 'CITY', 'COUNTRY',

'CAPTION\_INTEREST','ACCESSIBILITY\_CAPTION\_INTEREST