**Job-athon Document (Approach)**

* **Checking the data: -**

1. Checking the data for duplicates in ID for both train and test set
2. Checking the data for null values (Credict\_Product feature had null values only)

* **Feature Engineering: - (**did not have time for all of the below step)

1. Check for feature correlation (PCA if high correlation between features exists or using L1/L2 to deal with it.
2. Deal with outliers and null values
3. High outliers lie in the Avg\_Account\_Balance feature.
4. The Null values in Credit Product feature can be taken care with another model where we predict if the customer has a null value or Yes/No value after removing Is\_Lead as a feature. Can be removed by doing EDA as well by taking into account Age, Gender or Vintage and understand which is the dominant value for different quartiles of other features.
5. We will check for feature importance, the feature which are least important can be removed by checking the feature importance method of any algo (I have used Logistic Regression to get the feature importance)
6. Feature normalization / standardization on Age, Vintage and Salary

* **Choosing Algorithms: -**
  1. Tried two algorithms Cat boost and Random Forest, Cat boost performed better marginally and used that in the test set.