1. What steps did you take to prepare the data for the project? Was any cleaning necessary?
   * Performed the below activities
     + Tidy dataset concept : Combined investment gain and loss column
     + Lot of missing data with ? In columns
     + Null value present in one observation : Did data imputation
   * Feature Engineering : Added new columns using binning techinques
     + Age binning
     + Investment binning
     + Education Rank
     + Education type
     + working hours and edu rank (Edu rank multiplied by working hours)
     + country and continent
2. What algorithmic method did you apply? Why? What other methods did you consider?
   * Compared several ml models and identified XG boost as the best model
   * XG boost takes care of both bias and variance which makes it usually the best option in ML algorithms
   * Considered models
     + Logistic regression
     + Random Forest
     + KNN
     + Decision Tree
   * Tried a DNN keras model also which rendered similar results to XG. (Although this will perform better with more data)
3. Describe how the algorithmic method that you chose works?
   * Xgboost combines tree ensembles with bagging and boosting techniques to perform faster and accurate predictions
4. What features did you use? Why?
   * I used all the features, as I found none of the features having any detrimental effect to the model’s accuracy
5. How did you train your model? During training, what issues concerned you?
   * Number of missing data were a concern for me during the training
6. How did you assess the performance of your predictions? Why did you choose that method?
   * I used the below techniques to assess the performance
     + Confusion matrix : Highlevel Right/predictions
     + Classification report : Indepth results
     + ROC curve : TPR vs FPR comparison
7. Would you consider any alternative approaches for assessing performance besides the metric you selected?
   * I have been regularly using Classification report for many of my projects, I would continue to use the same
8. Which features had the greatest impact on income tagging? How did you identify these to be most significant?
   * Relationship, age and investment were the most significant features
   * I used RFE method to identify the top features
9. Which features had the least impact on income tagging? How did you identify these?
   * origin\_country was surprisingly one of the least effective
   * USA had lots of both above and below 50k observations
   * And most of the data had USA as country, so this feature couldn’t become a key differentiator