1. **A brief on the approach, which you have used to solve the problem.**

   I have approached this problem as a classification problem. We had to predict whether an employee will leave the organization in the next 2 quarters. In the dataset, there were multiple records for each employee for each reporting month. I first focused on creating unique records for each employee and calculating the target variable (attrition in 2 quarters) by using the joining date and last working day columns.
   Once the feature engineering part was done, I built a classification model to reach the final output.

2. **What data-preprocessing / feature engineering ideas really worked? How did you discover them?**

   As mentioned in answer 1, it was important to create a unique record for each employee. I took the latest record for each column for each employee except the total business value. The total business value was aggregated for each employee by calculating the mean business values for all months.
   The joining designation column was used in conjunction with the designation column. I calculated the difference between the 2 columns to show the number of promotions for that employee and used that column for building the model.
   For calculating the target variable, I took the difference between the last working date and joining date in days. For the employees who left the organization before 220 days(took a buffer of 40 days), were labelled as 1 and the rest were labelled 0.
   Finally encoding was done for categorical variables using the get_dummies function in pandas.

3. **What does your final model look like? How did you reach it?**

   I applied the Random Forest Classifier algorithm as Random Forest can be used to handle large datasets as well and provides good accuracy on cross validation.
   I divided the train dataset into train and test in 70:30 ratio and applied the model. The accuracy score and F1 score obtained were 78% and 73% respectively.
   To further enhance the model, I did Hyperparameter Tuning by using the Random Search CV provided by scikit learn and estimated optimal ranges of parameters for the Random Forest Model. After applying the model this time, the score and F1 score were improved to 80% and 76% respectively.