**Data Cleaning**

1. Run 1st Cell for Importing all the libraries

2. Run 2nd Cell for Importing Dataset and checking the shape of data

3. Run 3rd cell for Checking the first 5 entries of dataset

4. Run 4th and 5th cell for checking shape of the column and column names

5. Run 6th cell for checking all the null values in our dataset

6. our dataset is huge that's why we run 7th Cell is for showing all the maximum values of dataset

7. Run 8th Cell for all the column names in given dataset

8. now again checking shape of dataset in 9th cell

9. again checking all the null values but in this time, we want to check our output on its maximum values so we using the shape of dataset & there are 275 columns for output run the 10th cell

10. now all of the null values shown, so remove those columns that have maximum null values because if we put this null column in our model, we get the wrong data. for removing all the null columns run the 11th cell

11. this time we again check the shape of dataset

12. now there are only 40 columns remaining so we again checking all the null values of that columns for this run the 13th cell

13. we need to create one column called skills using branch and other branch column so we merge these 2 columns and replacing the 'other' to blank for this run the 14th cell.

14. we need to create another column called current company for that we merge the company name, company name.1 and college /company name columns but these columns have some null values so we fill this na values in selected columns because we didn't keep all the null values, we need to assign any value to those null cells so we fill na with blanks and merge all these columns and create the column for this run this 15th cell.

15. now we drop all of those columns that we use for merging and keep the new columns for this run 16th cell.

16. for cell 17,18,19 we apply the date and time format on created column and separate the full data and store in Date column and we seperate year from date and store in new column called year for all this step run these 3 cells

17. now we fill all the na values in our dataframe so run 20th cell

18. we rename the current academic year to academic year for this run 21th cell

20. for cell 22,23,24,25 we checking all the unique values of academic yr column and assign those unique values to some values and now we print this new column for all of this run these all 4 cells

21. for finding the year of graduation we do these calculations on 27,28,29,30,31,32,33,34,36,37,38,39 cells for this run all of this cell

22. if any duplicates on email column we run 40th cell.

23. now we export that clean data in csv for this run the last cell.

**Predictive Model**

1. Run the 1st cell for installing tpot in our environment

2. Run 2st Cell for Importing all the libraries

3. Run 3rd Cell for Importing Dataset and checking the shape of data

4. Run 4th cell for Checking the first 5 entries of dataset

5. our dataset is huge that's why we run 5th Cell is for showing all the maximum values of dataset

6. Run 6th cell for checking all the null values in our dataset

7. fill all the null values for this run 7th cell

8. Run the 8th cell for changing the int type for order data column

9. we need to create one column called full name using first name and last name column so we merge these 2 columns for this run 9th cell

10. searching unique value counts of full name column and store it into Names column for this run 10th cell

11. similarly we done for college for this run 11th cell

12. again we create one column called know about event using how did you come to know about this event and specify in others for this run 12th cell

13. remove the unwanted columns for this run 13th cell

14. checking the null values in 14th cell

15. checking the data frame in 15th cell

16. in 16th cell converting the categorical column into numerical

17. we convert the ticket type column into numerical using one hot encoding similarly we done for college name, designation and full name for this run 17th to 24th cell.

18. now we merge all the of this ticket type, college name, designation and full name data frame into one and created one master data frame for this run the 25th cell.

19. we checking the shape of df and master df in 26th and 27th cell.

20. now we split the data into x and y in 28th cell.

21. now we split our data into train and test and checking the shape of train and test for this run the 29th cell

22. for tpot all of the instruction given in model so run accordingly

23. now the best pipeline is provided by tpot and we run this pipeline and predict our column and print the prediction in 32 to 40th cell

24. now we export the confusion matrix on 41,42th cell

25. now we compare the score actual vs predicted and also checking the accuracy of model in 43rd and 44th cell

26. we export the feature importance and print some useful feature importance in 45th and 46th cell.