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

- The dataset had around 19k records, multiple records for the same employee.
- I used different machine learning models for prediction. In the end, I defined a function that combines the predictions of all the different models and keeps the most occurred predicted value. Ex:- From 4 different models if I get the predictions 1, 0, 1, 1 then the final prediction will be 1 because it is the most occurred predicted value.

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

- There are only 2381 employees, but we have a total of 19104 records. That means we have multiple records for the same employee. we will keep the latest record only which is required to determine whether an employee is currently working or not.
- The target variable was hidden in the last working date column, so I had to extract the target variable from that column.
- Gender, education level, City were some categorical features, so I used dummy variables and replaced those categorical variables.
- I used standardscaler for Standardization of the data.
- Since we did not have any e different test sets so I did not divide the dataset into training and testing data because later on, I had to make predictions on the training data.

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

- I have used 7 different models like XGBoost, KNN, SVC, GaussianNB, RandomForestClassifier with GridSearchCV for training.
- In the end, I defined a function that combines the predictions of all the different models and keeps the most occurred predicted value. Ex:- From 4 different models if I get the predictions 1, 0, 1, 1 then the final prediction will be 1 because it is the most occurred predicted value.