DATA-448 IN-CLASS ASSIGNMENT FEBRUARY 15, 2023

Consider the `turnover.csv` data file (posted under the In-Class 11 assignment link). This file contains basic employment information of employees from some company. The goal is to build a binary classification to predict employee turnover. **In Python**, answer the following:

1. (3 points) Using the pandas library, read the csv data file and create a data-frame called `turnover`.
2. (6 points) Change `sales`, and `salary` from labels to dummy variables.
3. (6 points) Engineer the interactions/features in-class 9 assignment (the ones from the decision tree).
4. (5 points) Using `satisfaction_level`, `last_evaluation`, `number_project`, `average_montly_hours`, `time_spend_company`, `Work_accident`, `promotion_last_5years`, `sales` (dummy variables), and `salary` (dummy variables) and interactions/features (from part 3) as the input variables and `left` as the target variable, split the data into two data-frames (taking into account the proportion of 0s and 1s) `train` (80%) and `test` (20%).
5. (10 points) Using the `train` data-frame, do the following:

   (i) Split the data into two data-frames (taking into account the proportion of 0s and 1s) `training` (80%) and `testing` (20%).

   (ii) Build a random forest model with 500 trees and `max_depth = 3` on the `train` data-frame. Extract and store the importance of all the predictor variables.

   Repeat (i)-(ii) 10 times and compute the average importance of each of the features.
6. (8 points) Using `train` data-frame (with the top 5 features based on their importances) build a random forest model (with 500 trees and the maximum depth of each tree equal to 3). Then, use this model to make predictions on the `test` data-frame. Use the provided `precision_recall_cutoff.py` (posted under the In-Class 10 Assignment link) file to estimate the optimal cutoff value. Compute the classification report of this model.
7. (8 points) Using `train` data-frame (with the top 6 features based on their importances) build a random forest model (with 500 trees and the maximum depth of each tree equal to 3). Then, use this model to make predictions on the `test` data-frame. Use the provided `precision_recall_cutoff.py` (posted under the In-Class 10 Assignment link) file to estimate the optimal cutoff value. Compute the classification report of this model.
7. (3 points) Using the results from part 6 and 7, what model would use to predict `left`? Be specific.