Homework 6 - Logistic Regression

Credit Card data set (100 points)

Instruction:

- You are expected to submit the **.ipynb** file and the exported **.html**.
- Only one member in each group needs to submit the assignment. It will be automatically submitted for the rest of group members.
- This is a long assignment, start early!

Question 1 Logistic Regression (100 points)

In this exercise I want you to apply logistic regression model to the credit card data set which is available on the GitHub folder for HW6. This data set should be familiar as you have done an EDA on it in HW-3 EDA for Credit Card Default dataset. Import the *credit_card_clean.csv* as a data frame and call it df. I specifically want you to do the followings:

1. Change the type of the feature variables as you see fit! numerical variables vs categorical ones. You can use my answer key from HW3 as your reference.
   (5 points)
2. Define your target variable. What are the proportions of default vs non-default in your data set? Is the target variable balanced or relatively imbalanced?
   (5 points)
3. Use get_dummies(drop_first=True) function from pandas package to make the categorical variables into dummy variables. How many features you have now?
   (5 points)
4. Along with the target variable, define your feature space (X) and split the data into test (30%) and train set (70%)
   (5 points)
5. From sklearn.linear_model import the relevant functions for Logistic Regression. Do the followings: **(30 points)**
   1 Train the logistic regression model using its default parameters. (5 points)
   2 Generate the predicted probabilities and predicted classifications and save them as y_hat_probs, y_hat respectively. (5 points)
   3 Plot the histogram of y_hat_probs? Explain what you see? if you set threshold=0.80, what does the model always predict? What is the implication for recall? (5 points)
   4 Generate predicted classifications for two different thresholds (30% and 60% threshold). Save these new predictions as y_hat_30 and y_hat_60. Which threshold should you use if your goal is to avoid too many false negatives? Explain your answer. (10 points)
   5 Construct a data frame named df_predictions with 5 columns. y_test, and the four y_hats from previous part (5 points)
6. Borrow my_logistic_report() function from the python notebook of class 11. (**25 points**)
   1 Report the Accuracy, precision, recall and f1 score along with the confusion matrix for threshold =0.5. Interpret all these statistics. Do you trust the accuracy of the model? why? (15 points)

2 Now use threshold = 0.3 in the my_logistic_report() function. what happens to accuracy, precision, recall and f1 score? what happens to false negatives? is this consistent with you answer to question 5.4? (10 points)

7. Plot the ROC curve and report the AUC score. Is your model doing a better job than random prediction (no skill)? **(10 points)**
8. Estimate the accuracy_test using K-Fold Cross Validation technique (try K=5 and K=10) and name them as accuracy_CV5 and accuracy_CV10. Are these numbers close to accuracy score from part 6.1? report your numbers with 5 digits precision. Why do you think all these 3 numbers are very close to each other? **(15 points)**

Good luck and enjoy machine learning!