Homework 9 - Decision Trees (DTs)

Lending Club Data (100 points)

Due Date: Monday April 5 at 11:59 pm

Instruction:

- This is a group-work assignment!
- 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.

Question 1 EDA (20 points)

In this exercise I want you to do a quick EDA on the Lending club data set. The data is available on the Github repository for the course. All the variable names are self explanatory except the dti which stands for disposable debt to income ratio. Also note that loan_status=0 means default and 1 means good condition (no default).

Show me what you have learned from the previous EDAs you did in HW2, HW3 and, HW8. Try to come up with an interesting story (hypothesis) using this data set. Treat this exercise as a real world project. Many times the managers have no idea what they want from the data!! your job is to be as creative as possible and come up with informative charts and tables.

Import the lendingclub.cvs as df.

Question 2 Decision Trees Classification (80 points)

For this exercise, I want you to use loan_status as the target variable and answer the following questions.

1. What are the proportions of Good condition loans vs defaulted ones in the data set? Is the target variable (relatively) balanced or (relatively) imbalanced? **(5 points)**
2. Along with the target variable, define your feature space (X) and split the data into test (20%) and train set (80%) **(5 points)**
3. From sklearn.svm import the relevant function for DTs classification. Do the followings: **(20 points)**
    1. Train the DT classification model using its default inputs. (5 points)
    2. Make classifications on the test set and save them as y_hat (5 points)
    3. Use the built-in classification report function from sklearn. Report the Accuracy, precision, recall and f1 score along with the confusion matrix. Interpret all of these statistics. Do you trust the accuracy of the model? why? (10 points)
4. **Pruning the tree**: plot the accuracy_CV vs alphas from the cost complexity pruning path. Report the optimal value for alpha. **(10 points)**
5. Re-estimate (Re-fit) the DT classification model with the optimal parameters from the gridsearch method. Save the predictions as y_hat_optimized **(5 points)**
6. Report the optimized classification metrics and compare them with the outputs from part 3.3 in Question2. Do you notice anything strange? what is going on here? **(5 points)**
7. Estimate the optimized accuracy_test using 5 fold cross validation. **(5 points)**

8. Visualize a classification tree with the following hyper parameters: max_depth=4, min_samples_leaf=50. **(20 points)**
    1. In your decision tree, what is the best feature to start with and where does the algorithm put the cut off point? (5 points)
    2. Interpret what you see! take one path for example and go down the tree. If income < ... and fico < .... and .... then default or good condition (5 points)
    3. How many terminal nodes do you see? why the number of terminal nodes is less than 16 in this example? (5 points)
    4. Are you satisfied with the gini numbers at the terminal nodes in general? explain why? (5 points)

Good luck and enjoy machine learning!