**HW2**

Top of Form

**Q1**

100 Points

Overview and Assignment Goals:The objectives of this assignment are the following: -- Experiment with various classification models for a realistic, complex problem. -- Think about dealing with data with different attribute types: categorical and numerical (ratio). -- Think about dealing with imbalanced data, with the goal of optimizing the F1 metric.Detailed Description:A practical application that we have talked about is determining whether or not to make someone a loan. In this assignment, the goal is to develop predictive models that can determine someone’s credit risk, where 0 is high risk, and 1 is low risk. Then loans could potentially be made to just the low risk people. Since the dataset is imbalanced the scoring function will be the F1-score instead of simple accuracy. Remember that the F1 score is the harmonic mean of precision and recall.

The setup of the assignment is roughly the same as HW1. You are provided with labels for the training set, and only examples with no labels for the test set (the files are available on the Miner2 platform as usual). You will provide predictions on Miner2, which will be scored using the F1 measure this time. The training set has 24720 examples of Class 0 and 7841 examples of Class 1 (32561 total). We are providing no information about the distribution of labels in the test set, which contains 13305 examples. The training file has 13 columns, with the first 12 being the features and the 13th the label. The testing file has 12 columns corresponding to the features.

The features are, in order: id - unique identifier **F1 - Continuous value** describing number of years since last degree was completed **F2 - Continuous value** indicating hours worked per week 3 - Categorical Value **F4 - Categorical Value** indicating type of occupation **F5 - continuous value** denoting gains **F6 - continuous value** denoting loss **F7 - Categorical value** denoting marital status **F8 - Categorical value** denoting type of employment (e.g., Self) **F9 Categorical Value** denoting education type **F10 - Categorical Value** denoting different race **F11 - Categorical** - Female/Male

Rules:

-- This is an individual assignment. Discussion of broad level strategies are allowed but any copying of prediction files and source codes will be treated as an honor code violation

-- Feel free to use the programming language of your choice for this assignment.

-- You can use any libraries and templates you wish to for dealing with this problem. However, you should be able to explain the methods you use and why you chose them in detail. You may want to think about feature selection, engineering, and which classification techniques are likely to be useful on this dataset.

-- You are allowed 10 submissions in a 24 hour cycle.

Deliverables: Valid submissions to the miner2.vsnet.gmu.edu website Gradescope submission of report and source code:

Create a folder called HW2\_LastName, and put all the source code there. You should include a readme file in your code folder. The readme file should include information on how to run your code.

Submit (on Gradescope) a 3-page, single-spaced report in PDF format describing details regarding the steps you followed for developing the classifier for predicting the product review sentiments. Be sure to include the following in the report:

1. Please state your miner2 user name, mason userid (eg: gmason76), and your best public score for your submission (at the time of writing the report) at the beginning of your report.
2. A detailed, but readable description of your approach, and how and why you chose the parameters, features, etc. Specifically, you should include information on how you performed model selection and which classifier(s) stood out, along with your best understanding of why that might have been the case. You should also discuss whether you excluded any specific features and why, and also whether you tried any methods of feature selection or engineering. Finally, you should discuss your approach to dealing with the class imbalance problem (including anything you tried but that didn't work well).
3. Your methodology for choosing the final approach and associated parameters. Any graphs or tables illustrating key experiments you did in the process of choosing your final model.
4. Also submit an archive of your HW1\_LastName folder (.zip or .tar.gz) via Gradescope.

Make sure you answer all the questions. It is always a good idea to organize your report. Feel free to use subtitles in each section such as Introduction, Approaches, Experimental Results and Conclusion. You can also summarize results in a table and use bold font to highlight key results. If you are using something outside of the class notes, that's fine. Explain briefly to show your understanding to avoid losing points.

Grading: Grading for the Assignment will be based on your implementation (20%), report (50%) and ranking results (30%).

Bottom of Form