Midterm Project Grading Rubric

Item	Points
Data wrangling - Clear description of how the data was merged, quality assessed, any cleaning performed, how were missing values handled, what preprocessing was done, was any data omitted and why etc. Describe in detail everything you did to get the data to the desired format?	10
Exploratory Data Analysis - A thorough analysis with a minimum requirement of at least three types of relevant plots, and summary statistics. Clear description of findings and how the EDA informed your modelling choices. The points you get will depend on the sophistication and rigor of the EDA you performed.	10
Feature engineering – Clear description of the feature engineering you performed and your logic behind the features you engineered. The points you get will depend on the sophistication and degree of feature engineering you performed.	10
Modelling - Description of choice of models, Minimum requirement is implementation of at least four different models (at least one ensemble). Address other modelling issues such as overfitting, data leakage etc. Address difficulties encountered (actual code/modelling issues) and how corrected etc. The points you get will depend upon the sophistication/rigor of the modelling.	30
Presentation – A professional presentation that highlights key aspects of all the topics mentioned above. The presentations should be max 10 to 12 minutes. No need to present code (you can pull up the code if you want to highlight something specific that others could learn from). Spend some time addressing difficulties encountered and how rectified, so others can learn from how you solved the issues you faced. The presentation has to be complete in that it addresses everything mentioned above. The points you get will depend upon the completeness and quality of the presentation.	7
Report - A professional and succinct report of the project that should contain the Python code and output along with written narrative. It should NOT just be code and output without any written commentary. A person with a basic level of data science knowledge should be able to read your report and understand what the objective of the project was, what you did, why you did certain things and the results and implications of your analysis. Think about it as a research paper. Please suppress warnings and avoid printing out large amounts of data in the report (e.g., printing the entire data frame etc.) to keep the report clean and short. The points you get will depend upon the quality of the report.	8
Active participation in the Kaggle competition – You are expected to actively participate in the competition, by actively trying to win it with multiple (at least 4) improving submissions.	10
Peer Evaluation – Each person in the group will anonymously grade the others participation and contribution out of 15 points, by filling out the form at https://forms.gle/eGJNTEbNeohJTHBT8 . Your points will be the average of your peer assigned points.	15

Deliverable for project

Please turn in the presentation and the report on Canvas.

Steps for doing the Dropout Prediction Challenge

1. Join data by student ID.
2. Clean data and perform feature engineering.
3. Separate the data into training and test based on test IDs I have provided in testids.csv file. We will call the test dataset Kaggletest just to differentiate it from other test datasets we will create later in step 5. Set the kaggletest dataset until step 6.
4. Add the Dropout labels column to the train dataset by joining to the data in droppoutTrainlabels.csv file.
5. Fit various models to the training dataset by going through the usual process of further splitting this dataset into train and test and measuring performance.
6. Predict using the Kaggletest dataset. This will result in a column with 1s and 0s in the prediction object you create.
7. Combine this prediction column with the student ID column in the kaggletest dataframe and create a new data frame that has two columns Student ID and Dropout
8. Write the dataframe created above to a csv file and submit to Kaggle. Note: the submission file should exactly look like the sample submission file I have provided. Once you submit to Kaggle you will get the accuracy (F1 measure) that determines your position on the leaderboard.