Instructions

- This homework assignment is worth 63 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: January 27, 2023 by 11:59 pm.**

For this homework assignment and for future one, we will work on the challenge presented in the `data mining cup 2019`. Please read the task and get familiar with the data. For this week homework assignment, answer the following:

Exercise 1

(5 points) Create a s3 bucket to store the data files (`train.csv` and `test.csv`). Using pandas, read the both data files and called them `train` and `test`.

Exercise 2

(3 points) Report the number of observations in the `train` and `test` data-frames. Also, create a frequency table of the target variable (`fraud`).

Exercise 3

(20 points) Create at least two visualizations that may show interesting relationships between the input variables and the target variable. Make sure to describe the visualizations.

Exercise 4

(35 points) In this part, you will build and compare the performance of two different models (from the ones that were covered in DATA-445 or other models that you may be familiar). If you decide to build a model that has not been discussed in DATA-445 or any other analytics course at Grand View, you will need to provide documentation. Do the following:

(a) Using the `train` data-frame, define at least five variables as the input variables (that you may consider important to predict `fraud`). Split the data into two data-frames (taking into account the proportion of 0s and 1s in `fraud`): `training` (80%) and `testing` (20%).

(b) Using the `training` data-frame, build your first model. Then, predict the likelihood of `fraud` on the `testing` data-frame. Estimate the cutoff value that makes the model the closest to the perfect model based on the precision-recall curve. Use the `precision_recall_curve` function to compute precision-recall curve. Using the optimal cutoff value, create the classification report.

(c) Using the `training` data-frame, build your second model. Then, predict the likelihood of `fraud` on the `testing` data-frame. Estimate the cutoff value that makes the model the closest to the perfect model based on the precision-recall curve. Use the `precision_recall_curve` function to compute precision-recall curve. Using the optimal cutoff value, create the classification report.

(d) Using the results from part (b) and (c), what model would use to predict customer `fraud`? Be specific.