COMP 4983: Project Description (Competition)

[20% Project]
[10% Peer Evaluation]

Overview:

In this project, you will work in groups of 4 students, to address one of the key challenges in the automobile insurance industry - determining the risks associated with an insured customer and charging the appropriate insurance premium.

The goal of the project is to predict the insurance claim amount (outcome) of an insured customer given a series of unlabeled features which includes characteristics of the insured customer and insured vehicle.

You will implement the entire machine learning workflow in Python, which after data ingestion, includes multiple iterations of data preprocessing, model training, model testing/tuning and model deployment. There are no restrictions on the use of library functions and you are strongly encouraged to explore beyond the machine learning concepts and models covered in class.

Groups:

In this part of the lab, you will meet with your assigned groupmates for this course. To start, you should as a group, establish a group ground rules contract (which you do not need to submit) and begin work on the project.

Datasets:

For purposes of this project, there are a total of three (3) datasets: training set, test set and competition set.

You will use the training set to iteratively train, test and tune your model. The test set will be used at project checkpoints to assess the performance of your trained model. The competition set, which you do not have access to, is reserved for assessing the performance of your final deployed model and ranking among teams. The team with the best performing final deployed model on the competition set will be crowned the winner of this project.

You may download the following three (3) files from BCIT Learning Hub (Content | Project Material | Datasets):

- *trainingset.csv*
  This is the training set. Each row in the dataset represents the characteristics of the insured customer and insured vehicle. The first column contains the row index and the claim amount is provided in the last column.

- *testset.csv*
  This is the test set. Each row in the dataset represents the characteristics of the insured customer and insured vehicle. Note, however, that the claim amount is not provided in the test set. You will use the features in the test set to predict the claim amount and submit the predicted claim amounts for performance evaluation.

- *submission.csv*
  This is the sample file format for submission of the predicted claim amounts (using the test set) for performance evaluation.

Getting Started:

To get you started, your first task should be to perform data exploration and preprocessing on the training set. Data visualization and input transformations should be considered. The goal is to gauge the complexity of the dataset as well as search for dependencies between feature values and the claim amount, prior to determining the machine learning model to use.

Upcoming Due Dates:

All work submitted is subject to the standards of conduct as specified in BCIT Policy 5104. Late submissions will not be accepted.

October 25, 2022*: Data Preprocessing and Exploration (training set)
* recommended first iteration due date (group check-in; no deliverable)
[morning session: lecture; afternoon session: lab & project check-in]

November 1, 2022:	Trained Model Assessment Checkpoint #1 (test set)
November 8, 2022:	Trained Model Assessment Checkpoint #2 (test set)
November 15, 2022:	Trained Model Assessment Checkpoint #3 (test set)
November 18, 2022:	Trained Model Assessment Checkpoint #4 (test set)
November 22, 2022:	Final Deployed Model Assessment (competition set), Competition Ranking (competition set)
November 29, 2022:	Project Presentation
November 29, 2022:	Peer Evaluation