**Group Project: MIS-64037- Advanced Data Mining and Analytics**

**Part A: Description**

Machine learning has had fruitful applications in finance well before the advent of mobile banking apps, proficient chat bots, or search engines. Given high volume, accurate historical records, and quantitative nature of the finance world, few industries are better suited for artificial intelligence. There are more uses cases of machine learning in finance than ever before, a trend perpetuated by more accessible computing power and more accessible machine learning tools. Today, machine learning has come to play an integral role in many phases of the financial ecosystem, from approving loans, to managing assets, to assessing risks.

**Underwriting**:

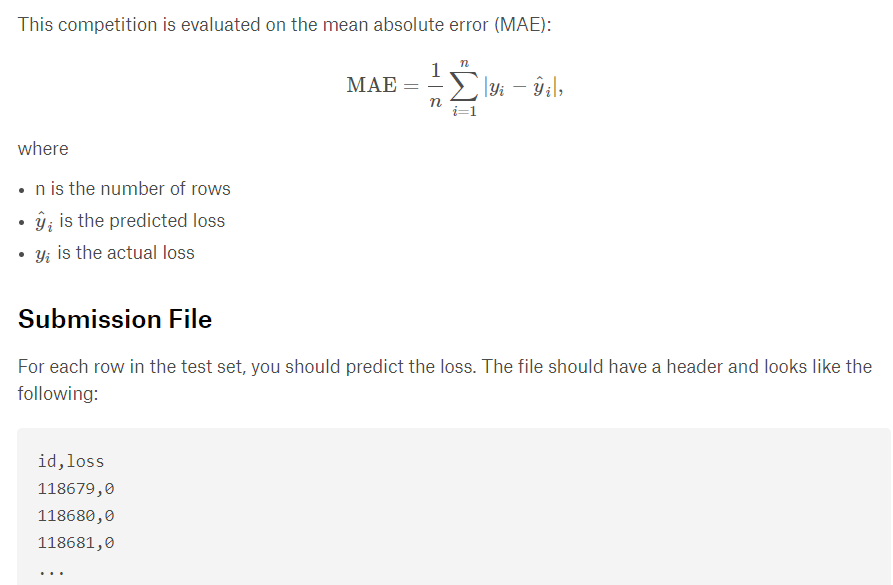
Underwriting could be described as a perfect job for machine learning in finance, and indeed there is a great deal of worry in the industry that machines will replace a large swath of the underwriting positions that exist today. Especially at large companies (big banks and publicly traded insurance firms), machine learning algorithms can be trained on millions of examples of consumer data (age, job, marital status, etc…) and financial lending or insurance results. The underlying trends that can be assessed with algorithms, and continuously analyzed to detect trends that might influence lending and insuring into the future.

**Part B: Assignment**

**Objective:**

This project asks you to determine whether a loan will default, as well as the loss incurred if it does default. Unlike traditional finance-based approaches to this problem, where one distinguishes between good or bad counterparties in a binary way, we seek to anticipate and incorporate both the default and the severity of the losses that result. In doing so, we are building a bridge between traditional banking, where we are looking at reducing the consumption of economic capital, to an asset-management perspective, where we optimize on the risk to the financial investor.

**Evaluation:**



**Project Deliverables:**

The following items needs to be delivered

1. **Project report (25 marks):**

This is your end of project delivery document. It is a document that summarizes different aspects of the project work. It includes the following sections:

1. Project Goal

2. Overview of data, including data exploration analysis

3. Details of your modeling strategy (i.e. what technique and why)

4. Estimation of model’s performance

5. Insights and conclusions

You can include snapshots of your R code and the outputs in the report (recommended). You have to submit a single document per group in PDF format. The first page of the document should include a table with a list of the names of the group participants and a very brief summary of contribution of each team member.

**2. R codes and script (25 marks):**

Submit the commented R script file that contains your work.

**3. Prediction’s File (25 marks):**

The reason to build any model is to be able to use it! In your project, once you have constructed your model, you need to use your model to predict the loss for a list of customers that I have provided to you in the test dataset. You will be submitting a file containing your prediction results in terms of the loss for each of those customers. I will use this to judge the accuracy of your model.

**4. Presentation (25 marks):**

Finally, you need to prepare a short presentation (10 minutes) sharing your insights from your project. This can be submitted either as a recorded presentation or you can present it in our synchronous class (I will send a separate announcement with regards to the date).

**Data**

The training and test datasets are available to download from the course portal under Project folder ‘**train\_v3.zip’** and‘**test\_\_no\_lossv3.zip’**. The target variable is called ‘loss’ which is the last column in the training dataset.