**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:**

In this project, you will be running the underwriting department of a bank and will decide who would be approved and who would be rejected. 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. You will submit deliverables for three different scenarios.

**Scenario 1**

In this scenario, we assume that your bank has a total capital of $1.4B for giving out loans. Loans are all fixed term (5-years), and the annual interest rate is 4.32% for all approved customers. To simplify the problem, we assume that the interest rate is not calculated as a compound rate. That is to say, for example if Mrs. White is taking a loan of $20,000. She will return $20,000 (the capital) plus 5\*4.32%\*20,000=4320 after five years if she does not default. If she default at 80%, it means that she would pay back only 20% of the capital 20,000\*20%=$4,000 and zero interest (i.e. the loss is $16,000 for your bank).

You are given the training dataset which contains a list of variables and the target variable that is “loss”. “loss” defines the percentage of the loan at which the customer was defaulted. If “loss” is zero you can imagine that the customer has fully paid back the capital and interest. If the “loss” is greater than zero, it means that the customer has defaulted. “loss” is expressed in percentage so if loss is 10, then it means that the customer has paid pack 90% of the capital but zero interests.

Based on this data, you will need to train model(s) to decide which customer listed in the “test\_scenario1\_2.csv” file (a total of 25471 customers) you would approve and which one you would reject. Obviously, your goal is to maximize the profit for your bank. Based on your decisions, I will calculated the total return after five years (which consists of profits from customers who you approved and paid back the capital and interest and losses from those who you have approved and have defaulted).

For this part, you will submit a csv file with 25471 rows each for one customer and a single column containing 1 and 0. 1 means approved, 0 means rejected.

Note that there is a column “requested\_loan” which shows the requested loan amount in US dollar in . This column does not exist in the train dataset. I leave it up to you how to use this column.

**Scenario 2**

Exactly similar to scenario 1 but in this case your bank budget to give loans is $450M . Again you need to submit a csv file with 25471 rows each for one customer and a single column containing 1 and 0. 1 means approved, 0 means rejected.

**Scenario 3**

In this case, you can see each customer is proposing an interest rate (column “Proposed\_Intrest\_Rate”). So the interest rate varies for different customers. For this scenario, we assume your bank has $1.4B available to give loans. The requested loan amounts and proposed interest rates are included in the file “test\_scenario3.csv”

For this part, you will submit a csv file with 25471 rows each for one customer and a single column containing 1 and 0. 1 means approved, 0 means rejected.

**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 3 files with the following names

GX\_S1.csv: binary approve/disapprove for each applicant for scenario 1   
GX\_S2.csv: binary approve/disapprove for each applicant for scenario 2   
GX\_S3.csv: binary approve/disapprove for each applicant for scenario 3 ,

Where X is you group number. So for example for group 3 we have G7\_S1.csv, G7\_S2.csv, and G7\_S3.csv.

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.