**ISYE 7406 Description of the Fraud Prediction Project**

**The Project Description is long. I appreciate your patience in reading the entire Description carefully and thoroughly.**

As a part of the assigned work for this course, you are required to participate in a team and conduct a course project. The purpose of this project is to encourage you to explore an avenue related to, but not limited to, the material in the class. To perform well in the course project, you may need to read additional literature and broaden your knowledge base. The project will enable you to gain hands-on experience about developing data science-enabled default prediction.

Introduction

In this course project, you are given a simulated credit card transaction dataset containing both legitimate and fraudulent transactions over a two-year period, from January 1, 2019, to December 31, 2020. The dataset includes transactions made by 1,000 customers at a pool of 800 merchants. Your task is to develop a fraud detection model capable of identifying fraudulent transactions while minimizing false positives, which can lead to unnecessary transaction declines and inconvenience for legitimate customers.

The dataset was generated using the Sparkov Data Generation tool, an open-source simulator developed by Brandon Harris. This simulation was conducted for the entire two-year duration, with the resulting transaction records combined and formatted into a standard structure. The complete dataset is provided at the start of the semester, including labeled transactions for both fraud and non-fraud cases.

Fraudulent transactions in real-world financial systems are rare but costly, making their detection a critical challenge. In this dataset, fraud cases account for a small proportion of overall transactions, mimicking the real-world class imbalance observed in financial fraud detection. Your model will be evaluated based on its ability to accurately detect fraudulent transactions while minimizing false positives and false negatives. The evaluation will be guided by a utility function incorporating the trade-offs between correctly identifying fraud, missing fraud cases, and incorrectly flagging legitimate transactions.

A financial institution typically weighs detection accuracy against operational costs. The following utility function will be used to assess model performance:

Total Utility = ∑#TP (S - L) - ∑#FN C - ∑#FP P

where:

* S is the savings from preventing fraudulent transactions.
* L is the operational cost associated with blocking or investigating a flagged transaction.
* C is the cost incurred due to undetected fraud.
* P is the penalty for incorrectly flagging a legitimate transaction.
* #TP, #FN, #FP refer to the number of true positives (correct fraud detections), false negatives (missed fraud cases), and false positives (false alarms), respectively.

The balance between fraud prevention and customer experience is a key consideration in financial fraud detection. A method that flags too many transactions as fraudulent may inconvenience legitimate customers, whereas a model with too few detections may allow significant fraud losses. Your goal in this project is to develop a fraud detection system that optimizes this trade-off and maximizes the total utility. For the purpose of this project, consider (S-L)=$50, C = $100, P = $5. The total utility should be evaluated using the test dataset. There is no constraints on the number of models you build for this project, however, you should provide good reasoning behind the choice of such set of models.

Datasets and variable descriptions

There are two data files to be uploaded for this project. They are:

1. **FraudTrain.csv**: This is the transaction data containing approximately 1.3 million records collected from January 1, 2019, to December 31, 2020. It includes both fraudulent and legitimate transactions and is used for model training and validation.
2. **FraudTest.csv**: This is the transaction data containing 555,718 test points. This dataset must **not** be used for model training or validation. It is strictly reserved for final model evaluation to assess test error.

Each transaction record in the dataset contains the following features:

* **trans\_date\_trans\_time**: Timestamp of the transaction.
* **cc\_num**: Credit card number associated with the transaction.
* **merchant**: Name of the merchant where the transaction occurred.
* **category**: Category of the merchant (e.g., travel, entertainment, home).
* **amt**: Transaction amount in USD.
* **first** / **last**: First and last name of the cardholder.
* **gender**: Gender of the cardholder.
* **street, city, state, zip**: Cardholder's address details.
* **lat, long**: Geographical coordinates of the cardholder’s location.
* **city\_pop**: Population of the cardholder’s city.
* **job**: Cardholder’s reported job title.
* **dob**: Cardholder’s date of birth.
* **trans\_num**: Unique identifier for each transaction.
* **unix\_time**: Transaction timestamp in Unix time format.
* **merch\_lat, merch\_long**: Geographical coordinates of the merchant’s location.
* **is\_fraud**: The fraud label (1 for fraudulent transactions, 0 for legitimate transactions).

**Important Notes:**

* The **is\_fraud** column is the target variable.
* The dataset simulates real-world fraud detection scenarios where fraudulent transactions are rare.
* Your fraud detection models need to be evaluated by its total utility.

Project logistics and evaluation

##### Teaming: A team should include 2-5 students, i.e., no fewer than 2 students and no more than 5 students.

**Report and Presentation:** Each team should submit one written report and make a set of Power Point slides for project presentation.

**Format of your report**: You need to start with a single-page executive summary (no more than one page) to state what you have done and the insights you gained from doing the project, that is, *anything* that you feel you have a better understanding because of doing this project. The main text of the report should clearly present your approaches, justification, results, and conclusion. You can briefly recap the problem but there is no need for any lengthy repetition of the problem described in this project description, unless you opted out of the provided projects.

There is no page limit for the final report. Nevertheless, keep in mind that the grade is given upon the quality instead of the length of your report. Also, please **do NOT include** any code or pseudo-code in your report.

When you have formatting and writing style questions, you can consult the publications in *IISE Transactions* for guidance. This is not to ask you to write a paper to be submitted to *IISE Transactions*. Instead, it is to give you an idea how professionals in our field write a formal technical report (a published paper is a technical report by itself), i.e., how they lay out their problem, argue their cases, and reach their conclusions.  Papers in the *Transactions* also show you how to present your figures, how to present your tables, and how to cite a source and list the references.

*Please make sure to write your report in a single column format.*

**Format of your Power Point slides:** Your set of slides will start with a title slide, which includes the title of your project and the team members. The total number of slides, including the title slide, MUST be equal to, or fewer than, **ten** (**10**). Inclusion of each EXTRA slide will receive one point reduction in the project score, until the project score reaches zero. Suppose that you have 30 slides. It gives you – 20 points. Even if your team has done everything else perfectly, gaining +25 points, your project score will still be five (5).

**Timeline:**

**Formation of team**: The list of students in your team is due on **Feb 28, 2024, by 5PM**. Please email your team formation (names and GTID of your team members) to my email address, make sure each team only send out one email.

**Project presentation:** April 15, 17, and 22.

**Report/Slides submission**:

* The written report and the slides of your presentation are due by **5PM on April 23**. One submission per team, including a written report and a file of slides.
* **Electronic submission is required**, so the electronic time stamp will be used to determine whether the report is submitted on time. The report and slide files should be submitted in GT Canvas.
* **You must submit both report and presentation in PDF files**. Canvas will not take other file formats.
* A late submission will be penalized 0.5 points for the first hour it is late and an additional 0.5 points per hour it is late. The time is counted starting from 5 PM on the project due date. This means that if your submission happens after 5PM and before 6PM, your project score will be deducted 0.5 points, and if it is after 6PM and before 7PM, your project score will be deducted 1 point, and so on. A complete submission is defined as submitting ALL required files in the required format. Anything other than a complete submission is considered a “late submission.”

**Grading:** A total of 25 points are allocated to this project, emphasis will be given to the sophistication of your methodology as demonstrated in your reasoning/arguments and results (15 points). Please note that *sophistication* of your methodology does not equate *complexity* of a method. If you believe a simple method will do the best for your problem, it is Okay but you still need to provide convincing arguments and good test results to support your claim. In addition to the sophistication consideration, other considerations for grading include the organization and clarity of your report and presentation (5 points for the report and 5 points for presentation slides).

Mandatory point deductions:

-1 per extra slide beyond 10;

-1 executive summary longer than one page;

-1 report not in a single column format;

-0.5 per hour the report/presentation/outcome is late.