BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

# Group Coursework Submission Form

## Specialist Masters Programme

| | |
|---|---|
| **Please list all names of group members:**<br>(Surname, first name)<br>1. Badhe, Tanaya<br>2. Kapoor, Karishma<br>3. Ramzin, Ivan | 4. Joshi, Rutva<br>5. Mankame, Piyusha<br>**GROUP NUMBER:** 5 |
| **MSc in: Business Analytics** | |
| **Module Code: SMM768** | |
| **Module Title: Applied Deep Learning** | |
| **Lecturer: Philippe Blaettchen** | **Submission Date: 08/03/2023** |

**Declaration:**

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

**Marker's Comments (if not being marked on-line):**

**Deduction for Late Submission:** □

**Final Mark:** %

# Question 1

$$hit\ rate(conversion\ rate) = \frac{fradulent\ claims\ correctly\ identified}{the\ total\ number\ of\ fradulent\ cases}$$

Total number of fraudulent cases consists of fraudulent claims, that are correctly identified, and those, that system identifies as non-fraudulent.

Hit rate indicates the model's ability to catch fraud. The higher, the better.

$$detection\ rate = \frac{cases, marked\ as\ fraud\ by\ the\ system}{all\ cases}$$

Detection rate indicates how comprehensive the model is in detecting fraud.

**Trade-off between them:** a model with high hit rate may also have a high detection rate, indicating that it correctly identifies most fraud cases while minimizing false positives. However, a model that is optimized for high hit rate may sacrifice detection rate, leading to more false positives, which can be costly for insurers and damage customer trust.

**Evaluation of model in monetary terms:** we can estimate the potential losses due to fraud that would be prevented by the model and compare that to the cost of implementing the model and any associated false positives (FP) or missed fraud cases. For example, if the model is estimated to prevent $1 million in fraudulent claims each year but has a cost of $500,000 and generates $100,000 in false positives, then the net benefit would be $400,000.

197/ 300 words

## Question 3

The logic underlying this model is that fraudulent transactions tend to occur soon after an account is created. Hence, transactions that occur too long after the account creation date are less likely to be fraudulent. By varying the value of t, the model can balance the *tradeoff* between detecting more fraudulent transactions and reducing the number of false positives.

The output shows the hit rate and detection rate values for each 't' value. We can see that as 't' increases, both hit rate and detection rate decrease. This is because as we increase the threshold value for flagging transactions as suspicious, we are reducing the number of transactions flagged as such, and as a result, we are also missing some fraudulent transactions (reducing hit rate). However, at the same time, the proportion of flagged transactions that are actually fraudulent (detection rate) also decreases.

By analyzing the hit rate and detection rate values for different 't' values, we can choose a sensible value for 't' based on our priorities. For example, if our priority is to minimize the number of missed fraudulent transactions (higher hit rate), we might choose a smaller value for 't' (such as 1 or 2). However, this might also result in a higher number of false positives (lower detection rate). Conversely, if our priority is to minimize false positives (higher detection rate), we might choose a larger value for 't' (such as 50 or 100). However, this might also result in a higher number of missed fraudulent transactions (lower hit rate).

254 words

# Question 7

**Decision Tree**: A popular machine learning model that can be used for classification tasks like fraud detection. They are easy to interpret. However, they can suffer from overfitting and may not generalize well to new data.

**TensorFlow**: An open-source framework for building and training machine learning models. It can handle large datasets and complex models with ease.

**Autoencoder**: Used for anomaly detection tasks like fraud detection. Effective at detecting anomalies and complex patterns, but challenging to train and computationally intensive.

**Transparency**

**Decision trees** are highly transparent that provide visual insight into how models make decisions by displaying variable importance in data splitting.

**TensorFlow** can be less transparent as they involve many layers of hidden nodes that are not directly interpretable by humans. Model's prediction process can be hard to comprehend.

**Autoencoders** can be less transparent because they are trained to compress data into a lower-dimensional space, which makes it harder for data processing.

**Importance of Explainability in Fraud Detection- Risks of Unexplained Flags:**

- **Lack of transparency:** Makes it difficult for users to understand why a certain incident was flagged as fraudulent.
- **Bias:** It can be difficult to identify and address any biases that may exist in decision-making process.
- **Legal and ethical issues:** If a tool flags an incident as fraudulent and there is no clear explanation as to why, it can be difficult to defend the decision in a legal or ethical context.
- **Inefficiency:** When there is no clear explanation for why a tool flagged an incident as fraudulent, it can be difficult to determine the best course of action. This can lead to inefficiencies in the fraud detection process, resulting in increased costs and potential loss of revenue.

284/300 words

# Question 8

Shift can use below techniques to improve the performance of fraud detection systems. **Ensemble methods** such as bagging, boosting, and stacking can combine multiple models to achieve better results. **Cost-sensitive learning** assigns different costs to different types of errors, depending on the severity of the consequences, and can help to prioritize accurate detection. **Data augmentation** generates new samples by applying transformations to existing data, which can increase the size of the minority class and improve model performance. **Anomaly detection models** normal behavior and flags deviations from that behavior as potentially fraudulent, making it useful for detecting novel or unseen types of fraud. **Active learning** iteratively trains the model on a small subset of the data and selects additional examples to label based on their potential to improve performance, reducing the amount of labeled data required to achieve desired results.

139/150 words

** Please refer to the Jupyter Notebook for Question 2, Question 3, Question 4, Question 5 and 6.