CASS BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON
EST 1894

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Dubey, Naveen 2. Shetty, Rakshith 3. Semaleesan, Sneha	4. Saha, Soumyajit 5. Parashar, Tridev **GROUP NUMBER:** 9
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: %

Table of Contents:

Assignment Questions Answered

Notes: Questions 1, 3, 7 and 8 are answered in this PDF. Kindly refer to the code file for the answers corresponding to the code.

- 1. Detection Rates v/s Hit Rate 3
- 3. Value of “t” 5
- 7. Different approaches attempted 6
- 8. Potential approaches 7

Appendix 8

1. Detection Rate v/s Hit Rate

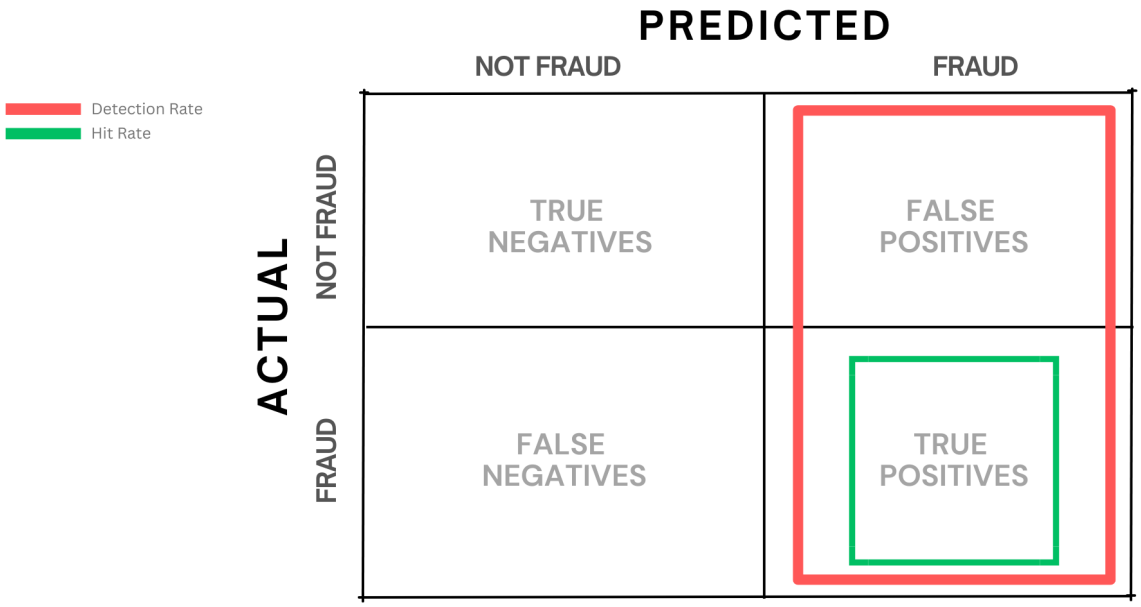
In this case study:

- **Detection Rate** is the proportion of frauds detected by the model to total claims.

$$Detection\ Rate\ = \frac{Number\ of\ True\ Positive\ and\ False\ Positive\ Cases}{Total\ number\ of\ claims}$$

- **Hit Rate** is the proportion of actual fraud cases out of all the cases that were detected as fraud by the model. Also called the positively predicted value (PPV), it is an indicator of the accuracy of the model.

$$Hit\ Rate\ = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positive\ and\ False\ Positive\ Cases}$$



Using the following formula, we can calculate the incremental savings from accurately determining frauds -

$$= Hit\ Rate\ (h)\ *\ Detection\ Rate\ (d)\ *\ (\Sigma\ (Claim\ Value\ -\ Investigation\ cost))$$

Where,

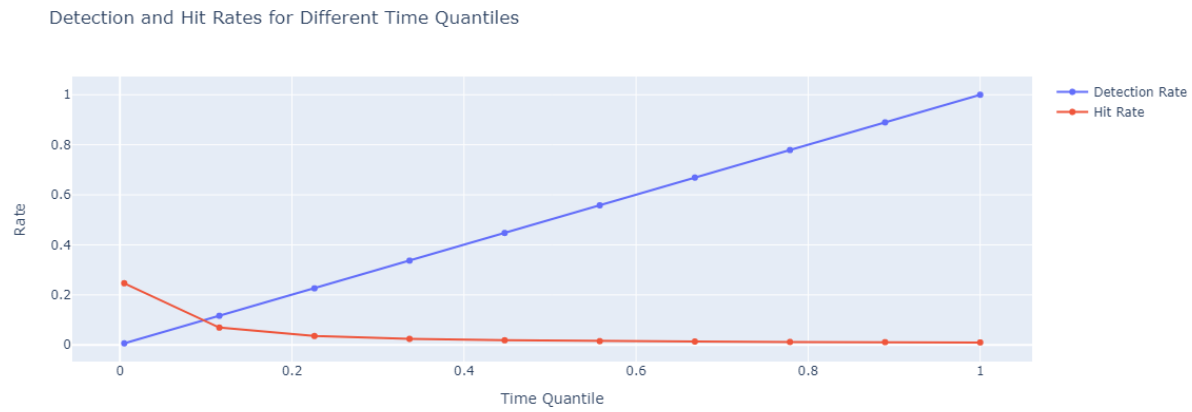
Claim Value = Individual claim value of each claimant

Investigation cost = Cost of resources spent to look into the corresponding individual claim.

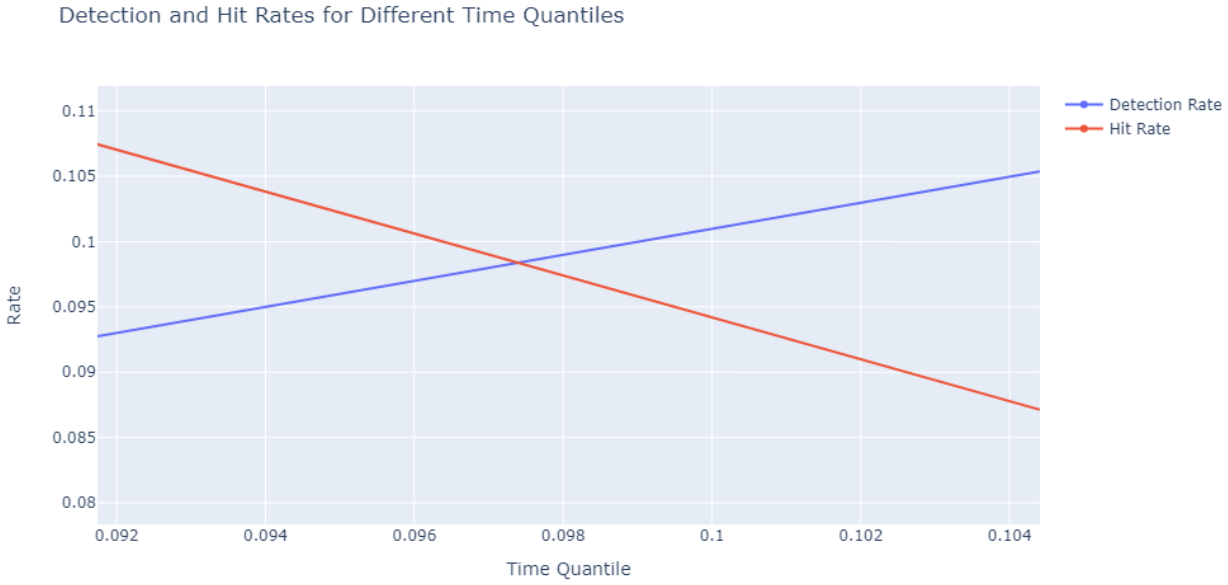
The detection rate is an important metric to identify potential frauds and save the company from bearing the loss of undetected frauds. However, a higher hit rate is important to avoid unnecessary costs of investigating the incorrectly predicted fraud cases(also known as false negatives). The trade off depends on the purpose of the business. If the insurers want to save

the cost of losing millions of dollars due to undetected frauds, then the focus should be on increasing the detection rate by decreasing the cases that were wrongly classified as not-a-fraud. If saving the implicit reputational loss i.e cost of the public discovering the investigation of falsely accused/ innocent claims is of importance, then the focus should be on increasing the hit rate. In addition, if the cost of loss from the claim is less than the cost of investigation, then also the hit rate should be focussed on rather than the detection rate.

3. Value of ‘t’:



The graph above shows the variation in hit rate and detection rate with respect to different time quantiles. The time quantile for a value “t” refers to the first “t” percentage of the observations once the data has been sorted in terms of the time between the date of insurance subscription and claim date. Here, we can see that the two curves intersect at an approximate value of 0.0975% (visible from the graph below, which is the zoomed version of the first curve) of the time quantile. Beyond this point, the hit rate drops and subsequently decreases which is not desirable in terms of cost. In continuation to our discussion from Question 1 (about the trade off between hit rate and detection rate), we choose the region before the intersection where the range of values for “t” is between 0 to 0.0975 (approximately). The final value of “t” should be chosen based on the desirable hit rate and detection rate set by the insurers.



7. Different approaches attempted

In the autoencoder model, we have considered three optimizers namely SGD with momentum, RMSprop and Adam. And in terms of dense layer activation functions, we have considered ReLU and LeakyReLU.

In terms of detection rate and hit rate, the KERAS tuner search for best hyperparameters gave us ADAM as the best optimizer. The values for detection rate and hit rate, in the leaky ReLU model are 42.11% and 9.84% respectively and in the ReLU model they are 54.39% and 8.29%. The threshold used for deriving these rates were observed from the MSE Plots (as shown in the Appendix) indicating the difference in MSE between non fraud and fraud cases wherein the MSE beyond which fraud cases are observed has been taken as the threshold.

In SGD, the gradient is updated with a momentum term and in RMSprop, the current gradient is scaled to converge faster. ADAM combines the approaches of the other two. In terms of the parameters, ADAM has the highest parameters and therefore least transparency in comparison to SGD and Adam. The change in activation function from ReLU to Leaky ReLU does not change the number of parameters and therefore there is no variation in transparency in this approach.

If there is no clear explanation regarding how a suspicious claim is flagged as a possible fraud, the following problems may arise -

- Reputational cost: Investigating a wrongfully suspected claim may lead to damage of public goodwill.
- Cost of investigation: More resources may be deployed to identify reasons for false flagging, leading to higher costs.
- Detection and Hit rate: Each increase in false positives increases the detection rate but reduces the hit rate. An unclear flagging has a higher chance of being a false positive than a true one and companies will always investigate fewer flagged events to save resources

8. Alternative Methods

In case of imbalanced datasets, due to presence of a significant difference in number of majority and minority classes (presence of the majority class is significantly greater than the minority class), ML algorithms may perform poorly in prediction of minority classes (in the case study, the minority class is a positive fraud claim). Two ways in which the model could be improved is by:

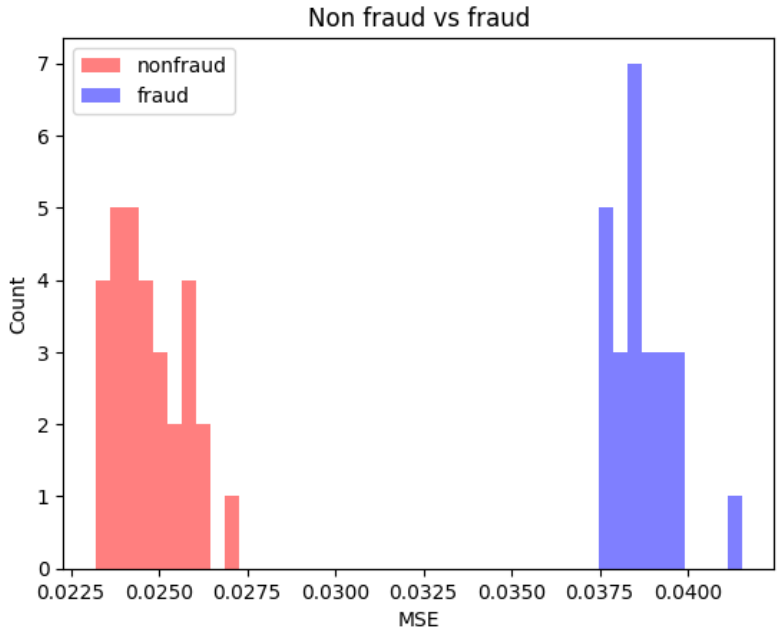
- Sampling:
  The dataset could either be:
  - Oversampling: By increasing the samples of the minority class through random duplication
  - Undersampling: By decreasing the samples of the majority class through random deletion

  Sampling may lead to issues of overfitting and a more advanced way would be using:
- Data Augmentation:
  By creating modified yet similar samples of the minority class using synthetic replication (e.g: using SMOTE)

Appendix

1. MSE Plot

a. For ReLU Activation Function



b. For Leaky ReLU Activation Function