**Question 7**

Although the approaches we have attempted can be effective in preventing fraud, it can be problematic if no clear explanation for flagging a certain incident as possibly fraudulent is given.

First of all, the lack of clear explanation can lead to mistrust between the insurer and the insured. If a customer has a claim denied or is investigated for fraud without being given a clear reason, they may feel unfairly treated. This can damage the relationship between the insurer and the customer, which can lead to negative reviews online as well as a loss of customers (Pérez et al., 2005). False accusations can be damaging for both the customer and the insurer, leading to potential legal action and reputational damage (Akomea-Frimpong et al., 2016).

What is more, without a clear explanation, it can be challenging for customers to understand how the fraud detection system works (Wilson, 2009). The confusion and misunderstandings about how the system operates, and what constitutes as fraudulent activity are one of the major reasons of fraudulent activities. Additionally, given a fraudster with access to claim data and sufficient technological expertise, there is a possibility that they could replicate deep learning techniques to evade detection, or the model's accuracy may decrease due to changes in features over time, which makes it challenging for the model to learn effectively.

Besides, it can be hard for insurers with limited knowledge to identify why a certain incident was flagged and upgrade the system to make the necessary improvements as a fraud detection system can apply a various number of algorithms. Failure to improve the system will not only result in false accusation of fraud, but will also lead to potential fraud slipping through the cracks, which can be costly for the insurer and their customers (Benedek et al., 2022). False negatives caused by insufficiently developed model will result in many cases of fraud remain undetected, which may lead to features being less discernible and lower model accuracy.

Lastly, the model requires retraining when new datasets are added. With an increase in variables and datasets, the model will require significantly more computing power and time to train compared to other regression models.

**Question 8**

Imbalanced data is a common issue in insurance fraud detection which can negatively affect the performance of machine learning algorithms in detecting fraud (Abdallah et al., 2016). In order to improve Shift’s prediction of insurance fraud, several approaches can be used:

Resampling: This involves either over sampling the minority class or undercsampling the majority class. Under sampling approach can be applied to remove some data in the majority class and over sampling approach can be applied to replicate the data in the minority class (Chen, 2006).

Cost-sensitive Learning: Cost-sensitive learning involves assigning a cost value to misclassifications of various classes, based on a cost matrix that identifies the different types of errors. The cost matrix is designed to guide the model to minimize costs or maximize benefits (Sahin, 2013).

Anomaly Detection: Anomaly detection algorithms can be used to identify unusual patterns in data, which may indicate fraudulent activity, even if it is not well-represented in the training data (Hilal et al., 2022). By supplementing traditional machine learning approaches with anomaly detection techniques, it is possible to improve the accuracy of fraud detection models while minimizing false positives. Additionally, it can be used to identify previously unknown types of fraud, which can help to reduce the risk of fraudulent activity going undetected.

References

Abdallah, A., Maarof, M.A. and Zainal, A., 2016. Fraud detection system: A survey. *Journal of Network and Computer Applications*, *68*, pp.90-113.

Akomea-Frimpong, I., Andoh, C. and Ofosu-Hene, E.D., 2016. Causes, effects and deterrence of insurance fraud: evidence from Ghana. *Journal of Financial Crime*, *23*(4), pp.678-699.

Benedek, B., Ciumas, C. and Nagy, B.Z., 2022. Automobile insurance fraud detection in the age of big data–a systematic and comprehensive literature review. *Journal of Financial Regulation and Compliance*.

Hilal, W., Gadsden, S.A. and Yawney, J., 2022. Financial Fraud:: A Review of Anomaly Detection Techniques and Recent Advances.

Pérez, J.M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I. and Martín, J.I., 2005. Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance. In *Pattern Recognition and Data Mining: Third International Conference on Advances in Pattern Recognition,*

Sahin, Y., Bulkan, S. and Duman, E., 2013. A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), pp.5916-5923.

Wilson, J.H. (2009), “An analytical approach to detecting insurance fraud using logistic regression”, *Journal of Finance and Accountancy*, Vol. 85 No. 150, pp. 1-15.

Other findings about investigation cost:

Viaene et al. (2007) also focused on the costs of the investigation process rather than on minimizing the error rate (misclassification) and showed that the cost-sensitive fraud screening could be a profitable approach for property and casualty insurance companies. Bolancé et al. (2012) approached the problem as an operational risk and used value at risk as the risk measure, then carried out a nonparametric estimation of the loss risk. Similarly to Phua et al. (2004), Viaene et al. (2007) and Bolancé et al. (2012), Zelenkov (2019) proposed a cost-sensitive approach, but in this case the author proposed an example-dependent cost-sensitive AdaBoost algorithm which assumed different costs not just for the different misclassification errors (as in previous studies) but for each particular case as well.