Credit Card Fraud Detection for imbalanced dataset

**Literature Review:**

Over the past years, predictive analytics approaches are gaining more and more popularities within fraud detection researches due to their powerful capabilities of extracting hidden information from large loads of data with decent accuracy[1].However, in credit card fraud situations, the number of positive (fraudulent) cases is much smaller than the number of negative cases which creates a problem of imbalanced classification, where one class is very much smaller than the other class. To tackle this situation a lot of research has been and is still being carried out where researchers are suggesting ways to improve fraud classification results by reducing the class imbalance in the training data.

One of the approaches implemented is to increase the number/proportion of positive(fraudulent) cases (oversampling) and the other is to reduce the number/proportion of negative cases (undersampling) [2,3,4]. Although undersampling can achieve certain results, it will result in some loss of hidden information from the missing negative cases, which would impact the classifiers performance. For these reasons, oversampling has become the current research focus for overcoming the problem of imbalance.

Al Majzoub *et al.* [5] proposed a method of synthesizing minority oversampling (SMOTE) for imbalanced classes. Recently , Yee *et al*. [6] used GAN to oversample credit card fraud and showed that, overall, it is better than the traditional SMOTE method. In 2018, Mohammed *et al*. studied the scalability of several Machine Learning algorithms when dealing with problems involving imbalanced datasets [7]. A hybrid balancing mechanism combining random under-sampling with borderline-2 synthetic minority over-sampling technique (SMOTE) worked satisfyingly well in the experiments described in the paper. In 2020 H.Tingfei *et al* implemented VAE (Variational Autoencoder) to perform oversampling which gave diverse samples of the minority class, yielding overall strong results[8]. In this research I will be comparing few of the different techniques and approaches to see which one suits the best for credit card fraud dataset.

Reference:

1] Kamusweke, Knox, Mayumbo Nyirenda, and Monde Kabemba. Data Mining for Fraud Detection in Large Scale Financial Transactions. No. 1729. EasyChair, 2019.

[2] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and N. Seliya, ‘‘Examining characteristics of predictive models with imbalanced big data,’’ J. Big Data, vol. 6, no. 1, p. 69, Dec. 2019.

[3] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane, and N. Japkowicz, ‘‘Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance,’’ in Proc. IEEE Int. Conf. Data Mining (ICDM), Singapore, Nov. 2018, pp. 447–456.

[4] S. Tyagi and S. Mittal, ‘‘Sampling approaches for imbalanced data classification problem in machine learning,’’ in Proc. ICRIC, Cham, Switzerland, 2020, pp. 209–221.

[5] H. Al Majzoub, I. Elgedawy, O. Akaydin, and M. K. Ulukök, ‘‘HCABSMOTE: A hybrid clustered affinitive borderline SMOTE approach for imbalanced data binary classification,’’ Arabian J. Sci. Eng., vol. 45, pp. 3205–3222, Jan. 2020.

[6]O. S. Yee, S. Sagadevan, and N. H. A. H. Malim, ‘‘Credit card fraud detection using machine learning as data mining technique,’’ J. Telecommun., Electron. Comput. Eng., vol. 10, nos. 1–4, pp. 23–27, 2018.

[7] Mohammed, Rafiq Ahmed, et al. "Scalable machine learning techniques for highly imbalanced credit card fraud detection: a comparative study." Pacific Rim International Conference on Artificial Intelligence. Springer, Cham, 2018.

[8] H.Tingfei et al. “Using Variational Auto Encoding in Credit Card Fraud Detection”.IEEE Open Access 2020.