Topic - Credit Card Fraud Detection

Aim - The primary aim of this assignment is to develop a robust machine learning model capable of effectively identifying potential instances of credit card fraud. In the modern landscape of financial transactions, the detection of fraudulent activities holds paramount importance in safeguarding both consumers and financial institutions against illicit practices.

Objective -To achieve this aim, the primary objective is to gather a comprehensive dataset containing various features pertinent to credit card fraud detection. This dataset will serve as the foundational basis for the development and training of the machine learning model. Additionally, it is imperative to undertake rigorous data cleaning and preprocessing procedures to ensure the quality and reliability of the data. This involves tasks such as handling missing values, addressing outliers, and encoding categorical variables, among others.

Subsequently, the focus shifts towards the construction of the machine learning model using state-of-the-art techniques and algorithms. Through a meticulous process of model selection, training, and evaluation, the goal is to create a predictive model capable of accurately distinguishing between genuine and fraudulent transactions. The model will undergo thorough testing and validation procedures to assess its performance metrics, including accuracy, precision, recall, and F1-score.

The ultimate objective is to deploy a robust and reliable credit card fraud detection system that can be seamlessly integrated into existing financial infrastructure. By leveraging the power of machine learning, we aim to enhance fraud detection capabilities, mitigate financial losses, and bolster trust and confidence in the credit card
ecosystem. This endeavour aligns with broader efforts to harness technology for the betterment of society and the advancement of financial security and integrity.

Literature Review -

The references mentioned encompass a broad spectrum of topics within the realm of machine learning and fraud detection, offering valuable insights and practical guidance for researchers and practitioners alike.

1. **"A Hands-On Machine Learning with Scikit-Learn, TensorFlow, and Keras"** by Aurelien Geron serves as a comprehensive guide to machine learning, providing a blend of theoretical concepts and practical examples. With a focus on key algorithms implemented in scikit-learn, TensorFlow, and Keras, the book caters to both beginners and seasoned professionals in the field.

2. "Seth Weidman's **'Deep Learning from Scratch'"** delves deep into the fundamental concepts and frameworks of deep learning, particularly emphasising neural networks. In addition to neural networks, the book elucidates essential preprocessing techniques and visualisation concepts using seaborn and matplotlib, offering a holistic understanding of the subject matter.

3. **"Credit Card Fraud Detection using Machine Learning Algorithms"** by Vaishnavi Nath Dornadula sheds light on the challenges faced by engineers and researchers in building effective fraud detection models. By providing insights into common issues and offering strategies to tackle them, the publication serves as a valuable resource for those working in the field of fraud detection and financial security.

4. Emmanuel Ileberi's publication on **"A Machine Learning Based Credit Card Fraud Detection Using the GA Algorithm for Feature Selection"** focuses on the critical aspect of data cleaning processes essential for fraud detection datasets. Addressing the challenges posed by unbalanced datasets, the paper highlights the significance of resampling techniques and offers guidance on effective feature selection and data preprocessing strategies.

Together, these references contribute to a deeper understanding of machine learning techniques, fraud detection methodologies, and the challenges associated with building effective models in this domain. By leveraging the insights provided in these publications, researchers and practitioners can enhance their knowledge and skills in the field of fraud detection and machine learning.

Methodology:

Collecting and loading the data:

The first steps in this type of problem is to gather a reliable source for the dataset and the dataset was collected from kaggle website where this dataset was also used for various research publications also.
Secondly to clean the dataset ‘creditcard.csv’. This dataset was a csv file so the first step is to load the dataset in pandas to see what the datasets have to offer about its various features. In this step after the dataset was loaded we checked various columns, features and target variables.
Due to privacy concerns, the original features, i.e from V1 to V28, have been transformed using dimensionality reduction techniques like Principal Component Analysis (PCA) to maintain anonymity while still capturing relevant information about the transactions. The other features Time and amount were original from the time of collection. The target variables were 1 and 0 where 1 is where the fraud is done and 0 for being no fraud.

Pre-Processing and Visualization:

The next steps for these are to check if there are any missing values or outliers on any relation between the features. These steps are very much important for training machine learning models as we can get insight from visualising and gathering relation between features. We first checked for Nan values and found that there were some missing values and then we replaced the missing values with the mean of that row as it is the most appropriate approach. It is done using the pandas library. We used the data.describe() method in pandas to see various relations among the features.

We used the matplotlib and seaborn library to visualise the amount in histogram and used seaborn's boxplot to see the potential outliers. Although there are one or two outliers, they are negligible.

We also checked the relations among the features using seaborn's pairplot to plot the relation among the features in a graph.

The next step in the preprocessing comes the resampling part. The dataset was very much imbalanced and it may lead to underfitting and negligence of the other class. Hence we used a resampling technique using SMOTE (Synthetic Minority Over-sampling Technique) function. The SMOTE is a resampling technique where it finds the class with minority instances and increases the number of instances in the minority class until it matches the number of instances in the majority class.SMOTE ensures that the synthetic samples created maintain the underlying distribution of the minority class, preventing overfitting and improving the generalisation capability of the model.

After resampling we have equal instances for both the classes and therefore we can train the model using simple algorithms as it will be able to detect the patterns.

Training-Testing The Model:

Next step is to build the machine learning model. To build the model we have use the Logistic regression as it it the most appropriate for the classification after the resampling been done .The logistic regression is a non linear model where after being going through the linear algorithm y = w* X + b goes through a nonlinear function that is the sigmoid function to transform the output to a non linear output within a range 1 and 0 . And we want to predict in 1 and 0 hence this is a great algorithm for training the model.
We first split the data into training and testing samples and then trained the model using sklearn's LogisticRegression() function. After that we tested the model using the sklearn's predict function.
 THe prediction was calculated and used accuracy score and F1 score to test the model and it turned out to be a great result after we got a **0.97** on the accuracy score.

Conclusion:

In conclusion, the machine learning model developed for credit card fraud detection has demonstrated outstanding performance, as evidenced by its high accuracy and F1 score of 0.97. The utilisation of advanced evaluation techniques such as the confusion matrix and ROC curve further corroborated the model's efficacy, with the ROC curve exhibiting an impressive accuracy of 0.99. These results indicate that the model was adept at recognizing patterns within the dataset and accurately classifying transactions as either fraudulent or legitimate.

The robustness and reliability of the model are underscored by its ability to effectively handle the inherent challenges posed by imbalanced datasets, a common issue in fraud detection tasks. By leveraging techniques like SMOTE for oversampling and careful evaluation metrics selection, we ensured that the model's performance was not compromised despite the class imbalance.

In summary, there are several avenues for improving our credit card fraud detection model.
Explore advanced feature engineering techniques to capture subtle fraud patterns. Experiment with ensemble learning methods like Random Forests and Gradient Boosting for enhanced predictive power. Incorporate anomaly detection algorithms alongside traditional classifiers to detect unseen fraud patterns.We can continuously monitor and update the model using real-time data streams to adapt to evolving fraud schemes. Conduct thorough model interpretability analyses to understand the decision-making process and identify influential features.

By implementing these recommendations, we can enhance the effectiveness of our fraud detection model and contribute to bolstering financial security for both consumers and businesses.

Reference:

1. Nath Dornadula, V., Gudikandula, P., & Sampath, R. (2016). Credit Card Fraud Detection Using Machine Learning Algorithms. In 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES) (pp. 1-5). IEEE.

2. Bhattacharyya, P., & Kalita, J. K. (2018). A review of machine learning approaches to credit card fraud detection. Journal of King Saud University-Computer and Information Sciences, 30(4), 412-428.

3. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: a realistic modelling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784-3797.

4. Bhattacharyya, P., & Kalita, J. K. (2018). Credit card fraud detection using machine learning: A review. Proceedings of the 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT) (pp. 148-152). IEEE.
5. .Ngo, D. H., Tran, T. D., Nguyen, H. T. N., & Vu, T. (2020). Credit card fraud detection using machine learning: A comparative study. In 2020 10th International Conference on Information Communication and Management (ICICM) (pp. 131-135). IEEE.