[2]: [Fraud Detection]

[Member 1 Ahmed Basha-20398547], [Member 2 Amany Azzam-20399133]
[Member 3 Areeg Mansour-20399122], [Member 4 Hanan Omara-20398559]

1 General Requirements

We want to make sure each of the team members takes care of one model or a pipeline. So that everyone can benefit from the task. You can treat a group-based project as a learning group given a specific task, but each of you still needs to implement your own DL pipeline.

The project proposal must clearly describe which member is taking care of what task and get it approved by TAs.

There are multiple goals for the group-based projects:

1) Get familiar with experiment design and dataset preparation for a specific deep-learning task.
2) Get familiar with SOTA models for the specific task.
3) Able to replicate SOTA models and apply them to new datasets.
4) Able to compare and conclude on the performance of SOTA models.
5) Collaborate with others to boost your knowledge on the specific task.
6) Presentation (we won't ask for a written report, we ask for a recorded presentation and a replicate package for all experiments, your code will tell the methodology, and your results will tell your findings).
7) For the replication package, we recommend once you have done all experiments, create a neat replication package in a notebook file (ipynb) and upload it into your google drive and onQ. Use sections and descriptions to make your replication package easy to follow. One notebook file would be fine for each team. The replication package is only required for the final evaluation (it is not required for the project proposal)

In your project proposal, you will provide four sections as described bellow.

2 Motivation and Problem Statement

The objective is to develop an accurate fraud detection system for a highly imbalanced labeled dataset, with 30 features. The system will analyze the dataset based on the paper ”Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation,” as a Baseline, identify patterns indicative of fraud, and flag potential fraud cases for further investigation. The project's success will be measured using accuracy, precision, recall, and F1-score metrics.

the pervasive issue of fraud in various industries, including finance, insurance, and e-commerce.Traditional methods like rule-based systems may not be effective in detecting new and evolving fraud patterns. Machine learning-based approaches are more flexible and robust to identify fraud in dynamic and imbalanced datasets. Detecting and preventing fraud is crucial for minimizing financial losses, maintaining customer trust, and protecting the reputation of businesses.

The main user groups that would benefit from the outcome of this proposed problem include:

1) Businesses (finance, insurance, e-commerce) to minimize financial losses, and maintain customer trust.
2) Government agencies to ensure compliance with regulations and prevent financial crimes.
3) Financial institutions (banks, credit card companies) to detect and prevent fraudulent transactions.

3 Related Work

“Credit Card Fraud Detection using Machine Learning and Data Science“ reference [4]. They used a dataset containing credit card transactions and evaluated the performance of various algorithms. The authors concluded that Random Forest outperformed other algorithms, achieving an accuracy of 99.6

“Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning” reference [3].introduced a Python library called imbalanced-learn that provides tools for handling imbalanced datasets. The library offers various resampling techniques and algorithms specifically designed for imbalanced data. This work is relevant to our project as it addresses the challenge of imbalanced datasets and provides practical tools that can be integrated into our system.

“Deep learning for anomaly detection: A survey” reference [2]. provide a comprehensive review of deep learning techniques for anomaly detection, including autoencoders, recurrent neural networks, and generative adversarial networks. This work is relevant to our project as it offers insights into the potential of deep learning methods for fraud detection and provides a valuable resource for understanding the state-of-the-art techniques in this domain.

“Feature engineering strategies for credit card fraud detection” reference [1].Focused on the importance of feature engineering in improving the performance of fraud detection systems. They proposed various feature engineering techniques, such as aggregation, encoding, and feature selection, and evaluated their impact on the performance of different classifiers. The authors concluded that feature engineering can significantly improve the performance of fraud detection systems.

4 Basic Data Exploration

The Bank Account Fraud (BAF) dataset is available on Kaggle and GitHub, provided by Feedzai for fraud detection and prevention. It contains 10,000 labeled instances of account opening applications with personal and financial data. The dataset is in CSV format with 30 features for each application. Additionally, there are six tabular datasets with one million synthetic instances generated using CTGAN, with five variants representing specific bias types and a "base" dataset. The datasets were sampled from a larger sample of 2.5 million instances, with different probability rates depending on the protected attribute group and label of the instance, to obtain desired group size and prevalence per month of the dataset, resulting in different types of bias. The generated data's representativeness was validated by comparing it to the original dataset distribution.

5 Experiment Design and Work Distribution

Each member of the team will implement his own methods on the pipeline and models that are already provided. We will divide the work among the four members as follows:

1. Hyperparameter Tuning:

Hanan, will be responsible for using the base dataset and performing hyperparameter tuning. This involves selecting the most appropriate hyperparameters to optimize its performance. and use techniques like Grid Search, Random Search, or Bayesian Optimization for hyperparameter tuning.

2. Preprocessing: Areeg, will be responsible for using the base dataset and applying preprocessing techniques. This involves cleaning the data, handling missing values, and transforming the data into a suitable format for machine learning algorithms. and will also explore feature scaling and normalization techniques to improve the performance of the model.

3. Handling Imbalance: Amany, will be responsible for using the base dataset and handling the imbalance in the data. This includes exploring techniques for handling imbalanced datasets, such as oversampling, undersampling, or using synthetic data generation methods like SMOTE. and implement the chosen technique and evaluate its impact on the base model's performance.

4. Model Evaluation and Best Approach: Ahmed, will be responsible for using the base dataset and its variants to evaluate the best approach among the three implemented by the other members. This involves comparing the performance of the model with hyperparameter tuning, preprocessing, and handling imbalance. and will use appropriate evaluation metrics, such as precision, recall, F1-score, or area under the ROC curve, to assess the performance of each approach. Based on the evaluation, I can identify the best approach for handling biased and imbalanced datasets.

References

[1] A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten. Feature engineering techniques for credit card fraud detection. *Expert Systems with Applications*, 2016.

[2] R. Chalapathy and S. Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[3] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 2017.

[4] P. Singh and D. Yadav. Credit card fraud detection using machine learning. *International Journal of Engineering Research and Technology*, 2019.