CS 583 — Deep Learning December 11, 2024

CS583 Final Project Report

Lecturer: Zhaozhuo Xu *By: Varad Hattekar, Shivanshu Vinay Singh*

1 What problem do we want to solve?

Fraudulent credit card transactions pose significant financial risks to consumers and financial institutions worldwide. Detecting fraudulent transactions in real time is crucial to minimize financial losses, enhance security, and maintain consumer trust. However, this task is challenging due to the highly imbalanced nature of the data, where fraudulent transactions are rare compared to legitimate ones. This report outlines the development of a credit card fraud detection model using deep learning techniques, addressing the challenges of data imbalance and real-time application.

Relevance of the problem :

Credit card fraud is a widespread issue that affects millions of transactions around the world, leading to significant financial and reputational damage to financial institutions and consumers alike. As digital transactions continue to increase, the urgency for effective and rapid fraud detection solutions becomes increasingly critical.

1.1 Problem Definition

- **Immediate Detection:** The project focuses on detecting fraudulent transactions as soon as they occur, enabling real-time response mechanisms to halt transactions before any financial damage can be done. This capability is crucial for reducing losses and protecting consumers.
- **Handling Imbalanced Data:** Fraudulent transactions represent a very small fraction of all transactions, making it difficult for traditional detection systems to identify them accurately. This imbalance leads to a higher likelihood of false positives. The model aims to address this challenge using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to better capture the rare fraudulent cases.
- **Operational Efficiency:** The detection system must integrate seamlessly into existing transaction processing workflows without causing significant delays or disruptions. The model should operate efficiently within the constraints of real-time transaction environments, ensuring a smooth user experience for consumers and businesses alike.
- **Consumer Trust:** By improving the accuracy and speed of fraud detection, this project also aims to enhance consumer trust in digital payment platforms. As e-commerce continues to grow, consumers' confidence in secure payment systems is critical for the continued adoption and success of online transactions.

2 What datasets did you use?

2.1 Data Source

The dataset used in this project was sourced from Kaggle. It has the following key characteristics:

- Contains anonymized credit card transactions from European cardholders.
- Includes 31 features:
  - 28 anonymized features (V1 to V28).
  - 'Time' (transaction timestamp).
  - 'Amount' (transaction value).
  - 'Class' (target variable: 0 for non-fraudulent, 1 for fraudulent).
- Exhibits a highly imbalanced distribution, with only 0.17% of transactions classified as fraudulent.
- Requires special handling to address class imbalance for accurate modeling.

2.2 Data Processing

To prepare the data for modeling, the following preprocessing steps were performed:

- **Exploratory Data Analysis (EDA):**
  - Visualized class distribution using count plots to highlight the imbalance between fraudulent and non-fraudulent transactions.
  - Analyzed the distribution of 'Amount' and 'Time' using histograms.
  - Generated a correlation heatmap to identify relationships among features.
- **Feature Scaling:**
  - Normalized the 'Amount' and 'Time' features to bring them to a consistent scale.
- **Handling Class Imbalance:**
  - Applied SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the minority class, improving the model's ability to detect fraud.

3 What models have you tried?

3.1 Machine Learning Models Considered

Several machine learning models were considered for credit card fraud detection:

- **Random Forests:** A robust model for handling imbalanced datasets and capturing non-linear patterns.
- **XGBoost:** Known for its efficiency and high performance on tabular data.

- **Deep Learning (Neural Networks):** Suitable for learning complex patterns and non-linear relationships in high-dimensional data.

Based on the dataset's high dimensionality and the need to capture intricate relationships among features, a deep learning model was selected for this project.

3.2 Chosen Model Architecture

The selected model architecture is a Sequential neural network with the following configuration:

- **Input Layer:** Contains 64 neurons to represent the input features.
- **Hidden Layers:**
  - Two hidden layers, each with 32 neurons.
  - ReLU activation function to introduce non-linearity.
- **Dropout Layers:** Applied a 20% dropout rate to prevent overfitting.
- **Output Layer:** A single neuron with sigmoid activation to classify transactions as fraudulent or non-fraudulent.

3.3 Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the model's performance. The following aspects were explored:

- **Number of Layers and Neurons:** Experimented with different configurations to balance underfitting and overfitting.
- **Dropout Rate:** Adjusted the dropout rate to mitigate overfitting while preserving model performance.
- **Learning Rate:** Tuned the learning rate of the Adam optimizer for faster and stable convergence.

4 How to evaluate the performance of the model on your dataset?

4.1 Exploratory Data Analysis

The following observations were made during exploratory data analysis:

- **Class Distribution:** The dataset was highly imbalanced, with only 0.17% of transactions classified as fraudulent.
- **Histograms:** Analyzed the distribution of transaction 'Amount' and 'Time' to identify patterns in the temporal and monetary characteristics.
- **Correlation Heatmap:** Visualized correlations between features to identify relationships that might inform feature selection or model improvement.

4.2 Results of Modeling

The deep learning model was evaluated using various metrics to assess its performance in detecting fraudulent transactions.

4.2.1 Model Selection

- Several models were considered, including Random Forests and XGBoost, but the Sequential neural network demonstrated superior performance.
- The deep learning model effectively captured complex relationships in the dataset and handled the class imbalance better than simpler models.

4.2.2 Hyperparameter Tuning

- Optimized the number of layers and neurons to balance underfitting and overfitting.
- Adjusted the dropout rate to prevent overfitting.
- Tuned the learning rate for stable convergence with the Adam optimizer.

4.2.3 Area Under Curve (AUC)

- Achieved an AUPRC (Area Under the Precision-Recall Curve) of 0.83, indicating strong performance in distinguishing fraudulent transactions.
- The model maintained a good balance between precision and recall, critical for the minority class.

4.2.4 Confusion Matrix

- The confusion matrix showed a high number of True Positives (correctly identified frauds) and a relatively low number of False Positives.
- This indicates that the model minimizes unnecessary alarms while effectively capturing fraudulent cases.

4.2.5 Classification Report

- Generated metrics like precision, recall, and F1-score to evaluate the model's performance on the minority class.
- The F1-score reflected a balanced trade-off between precision and recall, making the model reliable for fraud detection tasks.

5 How does your model perform?

This project demonstrated the feasibility and effectiveness of using a deep learning-based model to detect credit card fraud. The key takeaways from the project are as follows:

- **Data Preprocessing:** Addressing the class imbalance with SMOTE and feature scaling significantly improved the model's ability to detect fraudulent transactions.

- **Model Performance:** The Sequential neural network model achieved an AUPRC of 0.83, demonstrating a strong ability to distinguish between fraudulent and non-fraudulent transactions while balancing precision and recall.
- **Insights from Exploratory Analysis:** Patterns in transaction amounts, times, and feature correlations informed the model selection and design.
- **Scalability:** The model's adaptability to handle high-dimensional data makes it suitable for real-world applications in fraud detection systems.

Future work could focus on:

- **Hyperparameter Optimization:** Further tuning of the model's architecture, such as the number of layers, neurons, and dropout rates, could yield better results.
- **Model Comparisons:** Exploring other advanced models like Random Forests, XGBoost, or ensemble approaches could provide additional insights into the effectiveness of different algorithms.
- **Real-Time Deployment:** Implementing the model in a real-time fraud detection system to flag suspicious transactions immediately and integrate with financial institutions' workflows.

Overall, this project highlights the importance of leveraging advanced machine learning techniques for fraud detection to enhance financial security and minimize losses.