IML-2: Machine Learning Model Report
Credit Card Fraud Detection Using Logistic Regression

Introduction:

- Credit card fraud remains a significant concern in the digital age, posing financial risks to both consumers and financial institutions. In this study, we delve into the effectiveness of logistic regression as a modeling technique for identifying fraudulent transactions within a large-scale credit card transaction dataset.

Dataset Overview

- A publicly available csv dataset comprising 284,807 credit card transactions, each characterized by 31 features including time, transaction amount, and anonymized transaction features (V1 to V28) is used as a base for the purpose of this study.
- To address the data imbalance issue, we employed under-sampling to create a balanced dataset containing an equal number of normal and fraudulent transactions (492 each).
- Notably, the dataset exhibits a significant class imbalance, with fraudulent transactions representing a minority class. Addressing this class imbalance is crucial for developing a robust and unbiased fraud detection model.
- The dataset is partitioned into training and testing sets using an 80:20 split ratio.

Dataset Pre-processing

- The dataset comprises 284,807 transactions, encompassing 31 features, including time, transaction amount, and anonymized transaction features (V1 to V28).
- The dataset contains no missing values. However, it is essential to address the significant data imbalance, with only 492 (0.17%) instances of fraudulent transactions.
- The dataset undergoes extensive preprocessing, including exploratory data analysis, handling missing values, and scaling features to ensure uniformity, under-sampling is employed to address class imbalance, while feature engineering enhances the discriminatory power of the model.

Sampling and Balancing

- To mitigate the data imbalance issue, we employed under-sampling to create a balanced dataset containing an equal number of normal and fraudulent transactions (492 each). This balanced dataset ensures that the model receives adequate exposure to both classes, thereby enhancing its ability to discern patterns associated with fraudulent activities.

- Data balancing through under-sampling significantly contributes to the model's ability to generalize well to both normal and fraudulent transactions, thereby enhancing its efficacy in real-world scenarios
- 0 is a normal transaction and 1 is a fraudulent transaction.

Model

- Logistic regression is a simple yet effective algorithm for binary classification tasks like this
- Logistic regression models are trained on the preprocessed dataset, with careful consideration given to hyperparameter tuning and regularization techniques.

Model Training:

- All the independent variable columns are stored in the X variable and the single independent column [Class] is stored in the Y variable.
- Following data balancing, we partitioned the dataset into training and testing sets using an 80:20 split ratio.
- Testing data is used to determine the performance of the trained model, whereas training data is used to train the machine learning model.
- Subsequently, a logistic regression model was trained on the balanced training data.

Model Evaluation:

- The trained models are evaluated on both training and testing datasets to assess their ability to accurately classify fraudulent transactions.
- Accuracy Score
    - Measures the proportion of correctly classified examples in the training dataset, indicating the model's performance on seen data.
    - The logistic regression model attained an impressive accuracy score of 99.93% on the test data, indicating its proficiency in classifying transactions accurately.

- Confusion Matrix: The confusion matrix is computed to analyze the distribution of true positives, true negatives, false positives, and false negatives, aiding in diagnosing the model's performance across different classes.
    - True Positives (TP): 56,856 instances were correctly predicted as positive. These are the cases where the actual class was positive, and the model correctly predicted them as positive.

    - False Positives (FP): 8 instances were incorrectly predicted as positive. These are the cases where the actual class was negative, but the model incorrectly predicted them as positive.
    - False Negatives (FN): 32 instances were incorrectly predicted as negative. These are the cases where the actual class was positive, but the model incorrectly predicted them as negative.
    - True Negatives (TN): 66 instances were correctly predicted as negative. These are the cases where the actual class was negative, and the model correctly predicted them as negative.

- Precision(89.19%): It indicates that out of all the instances predicted as positive, approximately 89.19% were actually positive, while the remaining percentage might be false positives.

- Recall(67.35%): It indicates that approximately 67.35% of all actual positive instances were correctly identified by the model.

- F1-score(76.74%): This score indicates a good balance between precision and recall, suggesting that the model performs well in both identifying positive instances and avoiding false positives.

Visualization:

- Confusion Matrix Heatmap: A heatmap visualization of the confusion matrix provided an intuitive representation of the model's classification performance, facilitating insights into its strengths and limitations.
- Precision, Recall, and F1-score Histogram: A histogram depicting the precision, recall, and F1-score metrics offered additional clarity on the model's performance across different evaluation metrics.
- Precision, Recall, and F1-score Pie Chart: A pie chart visually summarized the distribution of precision, recall, and F1-score metrics, providing a concise overview of the model's overall performance.

Prediction:

- To validate the model's predictive capability, we conducted a sample transaction test. The model correctly identified a real transaction, reaffirming its reliability in discerning fraudulent activities.
- Sample Prediction Output: The transaction is Rea

Improving the model:

- Feature Engineering: Enhance model performance by creating new features or modifying existing ones to better capture the underlying patterns in the data.

- Algorithm Tuning: Optimize hyperparameters of the Logistic Regression model or experiment with different algorithms like Random Forest, XGBoost, or Neural Networks.
- Data Resampling: Address the class imbalance by using techniques like SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN to generate synthetic samples of the minority class.
- Ensemble Methods: Combine multiple models using techniques like Bagging or Boosting to improve the overall prediction accuracy and robustness of the model.

Discussion:

- The success of our approach highlights the importance of data balancing techniques in improving the performance of credit card fraud detection systems.
- By creating a balanced dataset, we ensure that the model receives sufficient exposure to both normal and fraudulent transactions, enabling it to learn meaningful patterns and anomalies associated with fraudulent activities.
- Logistic regression proves to be a robust and efficient algorithm for this task, offering high accuracy and interpretability.

Results:

- Our logistic regression model achieved impressive accuracy scores of 99.91% on the training set and 99.93% on the test set.
- Evaluation of the confusion matrix revealed the model's ability to correctly classify the majority of normal and fraudulent transactions.
- Precision, recall, and F1-score metrics further demonstrated the model's effectiveness in identifying fraudulent activities, with precision, recall, and F1-score values of 89.19%, 67.35%, and 76.74%, respectively.

Conclusion:

- In conclusion, logistic regression stands out as a dependable and interpretable tool for credit card fraud detection, offering a fine balance between performance and transparency.
- Through the strategic application of machine learning techniques, financial institutions can fortify their fraud detection capabilities and mitigate the risks associated with fraudulent transactions.
- Continuous research and development efforts are imperative to outpace evolving fraud tactics, safeguarding the interests of both financial institutions and consumers.