Jerimey Simons

CISA 4313 Fall 2023

Project Progress Report #3

Report: Importance of Creating a Supervised Machine Learning Model for Credit Card Fraud Detection.

Credit card fraud is a pressing issue in today's digital age, with a significant impact on both consumers and businesses. The rise of credit card usage, coupled with the worldwide connection of internet, has created an environment where fraudsters can exploit vulnerabilities in payment systems and compromise sensitive financial information. As outlined by a report from NBCnews.com (Jay, 2023), U.S. consumers have accumulated a staggering $43 billion in additional credit card debt during the second quarter of this year. This surge in credit card reliance, accompanied by the increasing connectivity of the digital world, has amplified the opportunities for potential victims.

According to Security.org (Security.org, 2023), a staggering 65 percent of credit and credit card holders have fallen victim to Fraud at some point in their lives, representing approximately 151 million Americans. Alarmingly, those who opt for automated payments and store their credit card information in web browsers are at a higher risk of falling prey to Fraud.

The impact of credit card fraud is not limited to individuals; it has broader economic consequences. As evidenced by a large-scale operation in New Jersey, where 18 individuals were charged and convicted in a $200 million international credit card fraud scheme, the cost of such fraudulent activities is borne by every American consumer (Justice.gov, 2013).

Furthermore, credit card fraud is not confined to a specific age group. Deloitte's survey reported by Vox.com (Ohlheise, 2023) suggests that Gen Z Americans are three times more likely to experience Fraud than boomers, illustrating the need for proactive measures to protect consumers across all age groups.

In a list of Money Lost, Credit Card Fraud ranked 10th on Social Catfish's report (Socialcatfish.com, 2023), with online scams raking in an alarming $264,148,905 from 2020-2022 due to online Fraud. The financial losses incurred by victims of credit card fraud are a stark reminder of the urgent need to address this issue comprehensively.

As the world is buzzing over the development of Artificial Intelligence, machine learning models are potential keys to the threat and offer a powerful tool to combat credit card fraud efficiently. Supervised machine learning algorithms can be trained on historical data to recognize patterns, anomalies, and potential fraud indicators.

My motivation in solving the problem of Credit Card fraud stems from the fact that I have also been a victim of Fraud. Unbeknownst to me, I checked my credit card statement as I do every morning, and something just did not add up. Fortunately, my bank was very accommodating, and I was able to have my account adjusted. I was also lucky enough that my credit report was not affected. However, not every victim of credit card fraud has a Rosey story where one call resolves their problem, which is why more fraud-detecting machine learning applications are needed, as bad actors are organized and are making considerable profits.

Description of Data Set:

For the project I will be using the Credit Card Fraud Detiction Dataset 2023 provide by Kaggle.com

Link: https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023

Size of the Data Set: 550,000

Description: "This dataset contains credit card transactions made by European cardholders in the year 2023. It comprises over 550,000 records, and the data has been anonymized to protect the cardholders' identities. The primary objective of this dataset is to facilitate the development of fraud detection algorithms and models to identify potentially fraudulent transactions."

Variables/Features:

id: Unique identifier for each transaction

V1-V28: Anonymized features representing various transaction attributes (e.g., time, location, etc.)

Amount: The transaction amount

Class: Binary label indicating whether the transaction is fraudulent (1) or not (0)

Related Work:

User Danang Hapis Fadillah on Kaggle site report 99. Accuracy with the following model by using key steps in a machine learning pipeline. Initially, it conducts data resampling by randomly selecting 5% of the observations from the original dataset. The subsequent phase involves dimensionality reduction and visualization using Uniform Manifold Approximation and Projection (UMAP) after scaling the features with StandardScaler. Following this, the dataset is split into training and testing sets, with 70% for training and 30% for testing, while ensuring class distribution balance through stratified sampling.

A preprocessing pipeline is then established, incorporating imputation through SimpleImputer and feature scaling using StandardScaler, particularly designed for numeric data. This pipeline is applied to the numeric columns in the training set via a ColumnTransformer.

The machine learning model pipeline is constructed, consisting of the preprocessing steps, Principal Component Analysis (PCA) with a specified number of components, and an XGBoost classifier as the algorithm. The model is subsequently trained on the training set.

The code proceeds to evaluate the model's performance on the test set, reporting the accuracy score and generating a confusion matrix and classification report. Additionally, a cross-validation assessment is conducted using the F1 macro score as the evaluation metric, with the results visualized through a plot.

In essence, the code encompasses data preprocessing, dimensionality reduction, model training, and evaluation, employing techniques like UMAP, PCA, XGBoost, and cross-validation to construct and assess a predictive model.

Reference Link: https://www.kaggle.com/code/dananghapisfadillah/umap-pca-xgboost-99-accuracy/comments

Machine Learning Algorithms Used:

1. **Decision Tree Classifier:**
   - Training Accuracy: 100%
   - Testing Accuracy:
     - Precision, Recall, and F1-score for both classes are 1.00.

- Overall accuracy: 100%

2. **Support Vector Machine (LinearSVC) Classifier:**
   - Training Accuracy: 96%
   - Testing Accuracy:
     - Precision, Recall, and F1-score for both classes are around 0.96.
     - Overall accuracy: 96%
3. **Random Forest Classifier:**
   - Training Accuracy: 100%
   - Testing Accuracy:
     - Precision, Recall, and F1-score for both classes are 1.00.
     - Overall accuracy: 100%

Evaluation Metric Preference:

The evaluation metric used in this case is the classification report, which includes precision, recall, and F1-score for each class, as well as overall accuracy.

Model Performance Evaluation:

Decision Tree Classifier:

- Training Accuracy: 100%
- Testing Accuracy: 100%
- The model seems to perform exceptionally well on both the training and testing datasets. There is a possibility of overfitting, as the accuracy is very high.

Support Vector Machine (LinearSVC) Classifier:

- Training Accuracy: 96%

- Testing Accuracy: 96%
- The model performs well on both training and testing datasets, and the accuracy is consistent. There is a slight difference between training and testing accuracy, suggesting a good fit.

Random Forest Classifier:

- Training Accuracy: 100%
- Testing Accuracy: 100%
- Similar to the Decision Tree model, the Random Forest model performs very well on both training and testing datasets, indicating a potential risk of overfitting.

Diagnosis of the Models:

Decision Tree Classifier:

- Potential overfitting due to perfect accuracy on the training set.

Support Vector Machine (LinearSVC) Classifier:

- The model appears to generalize well, as there is a small difference between training and testing accuracy.

Random Forest Classifier:

- Similar to the Decision Tree, there is a risk of overfitting, as the model achieves perfect accuracy on the training set.

Results:

- The models seem to perform well, with high accuracy on both training and testing datasets.

- Overfitting is a concern for the Decision Tree and Random Forest models due to their perfect accuracy on the training set.
- The Support Vector Machine model appears to be a good fit with balanced performance on training and testing datasets.

Overview and Evaluation of Neural Network Model

Dataset Overview

Dataset Size:

- The dataset contains 568,630 entries with 31 columns.
- Missing Values:
- There are no missing values in the dataset.

Data Types:

- The dataset consists mostly of float64 features, two int64 columns (id and Class), and no categorical variables.

Target Distribution:

- The target variable 'Class' is binary (0 and 1) and seems balanced, with 284,315 entries for each class.

Statistical Summary:

- The statistical summary provides details on the mean, standard deviation, minimum, maximum, and quartiles for each numerical feature.

Duplicates:

- No duplicates were found in the dataset.

Data Preprocessing

Scaling:

- The 'Amount' feature has been scaled using StandardScaler.

Train-Test Split:

- The dataset has been split into training and testing sets with a ratio of 70:30.

Model Overview

Neural Network Architecture:

- The neural network consists of one hidden layer with 10 units, ReLU activation, dropout regularization (dropout rate of 0.5), and an output layer with 1 unit and sigmoid activation for binary classification.

Model Compilation:

- Binary cross entropy is used as the loss function, and the Adam optimizer is employed.

Model Training:

- The model is trained for 50 epochs with a batch size of 10.

Model Evaluation

Training and Testing Accuracy:

- The model achieved a high accuracy of approximately 98.19% on the training set and 98.24% on the testing set.

Precision, Recall, and F1-score:

- Precision: 98.38%

- Recall: 98.11%
- F1-score: 98.25%
- Loss Visualization:
- Loss and accuracy are visualized over the epochs, showing a good convergence.

Conclusion:

Compared to other models (Decision Tree, LinearSVC, Random Forest) provided, the neural network demonstrates competitive performance. While the Decision Tree and Random Forest models achieved 100% accuracy on both training and testing datasets, the neural network's slightly lower accuracy suggests a more balanced performance without potential overfitting. Fine-tuning hyperparameters, exploring different architectures, and increasing the dataset size could be considered for further improving the neural network's performance.

In summary, the neural network model performs well in fraud detection, providing a good balance between training and testing accuracy. It is a competitive alternative to traditional machine learning models.

References

Security.org Team. (2023, January 31). *2023 Credit Card Fraud Report*. Retrieved from https://www.security.org/digital-safety/credit-card-fraud-report/

U.S. Department of Justice. (2013, February 5). *Eighteen People Charged In International, $200 Million Credit Card Fraud Scam.* Retrieved from https://www.justice.gov/usao-nj/pr/eighteen-people-charged-international-200-million-credit-card-fraud-scam

Jay, M. (2023, September 12). *Inflation is driving up consumer credit card debt by billions of dollars. NBC News.* Retrieved from https://www.nbcnews.com/business/consumer/credit-card-debt-rising-as-student-loan-payments-restart-rcna104442

Ohlheiser, A. W. (2023, September 21). *Gen Z falls for online scams more than their boomer grandparents do.* Vox. Retrieved from https://www.vox.com/technology/23882304/gen-z-vs-boomers-scams-hacks

Obi, O. (2023, September 27). *State of Internet Scams 2023*. Social Catfish. Retrieved from https://socialcatfish.com/scamfish/state-of-internet-scams-2023/