'InvisiFraud': Fraud Intelligence System

Analytics Capstone Project by Sai Bhere

Introduction

With the advent of online transactions and widespread credit card use, fraud has become a serious and growing concern for both financial institutions and consumers. Traditional rule-based fraud detection systems are usually ineffective, leading to a large number of false positives, missed fraud cases, and huge financial losses. The opportunity is to employ modern machine learning and data analytics techniques to develop a more efficient, accurate, and scalable fraud detection system capable of identifying fraudulent transactions in real time, reducing false alarms and the risk of undetected fraud. The InvisiFraud system detects fraudulent transactions with high accuracy using CNN-based deep learning models and supervised machine learning.

Objective

InvisiFraud attempts to detect fraudulent credit card transactions with a detection accuracy of at least 90% using multiple machine learning methods. The research also uses a Convolutional Neural Network (CNN) constructed in TensorFlow to create a system that is both extremely accurate and flexible to new fraud trends. To solve the dataset's class imbalance, the system will experiment with under-sampling and over-sampling strategies.

Models

1. Logistic Regression: Logistic regression, a simple linear model, predicts the likelihood of fraud in transactions. It is a cost-effective, interpretable, and efficient option for high-dimensional data.

2. Decision Tree: Decision trees split data into features to create predictions, but if not calibrated properly, they risk overfitting. Their transparency helps with fraud detection and provides clear decision-making information.

3. Random Forest: Random Forest, an ensemble model, improves resilience and lowers overfitting by combining predictions from many decision trees. Its excellent accuracy and capacity to identify critical data elements make it an attractive option for fraud detection.

4. KNN (K-Nearest Neighbors): KNN, a distance-based classifier, assigns classes based on their proximity to neighboring data points. While obvious, it is best suited to smaller, evenly dispersed datasets due to its high processing cost with large datasets.

5. CNN (Convolutional Neural Networks): CNNs, which were originally designed for image recognition, have been repurposed for tabular transaction data to capture complicated temporal and spatial patterns. While computationally demanding and less interpretable, they excel at detecting complex fraud patterns that simpler models cannot.

Results

```
--------- classification report of logistic regression---------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     84984
           1       0.79      0.58      0.67       134

    accuracy                           1.00     85118
   macro avg       0.89      0.79      0.83     85118
weighted avg       1.00      1.00      1.00     85118

--------------- Accuracy of logistic regression ----------------
Accuracy:- 0.9990953734815198
Precision:- 0.7878787878787878
```

```
---------- classification report of decision tree---------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     84984
           1       0.79      0.58      0.67       134

    accuracy                           1.00     85118
   macro avg       0.89      0.79      0.83     85118
weighted avg       1.00      1.00      1.00     85118

-----------------Accuracy of decision tree----------------
Accuracy:- 0.999177612255927
Precision:- 0.7162162162162162
```

```
---------- classification report of random forest---------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     84984
           1       0.79      0.58      0.67       134

    accuracy                           1.00     85118
   macro avg       0.89      0.79      0.83     85118
weighted avg       1.00      1.00      1.00     85118

-----------------Accuracy of random forest----------------
Accuracy:- 0.9994948189572124
Precision:- 0.9504950495049505
```

```
---------- classification report of KNN---------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     84984
           1       0.79      0.58      0.67       134

    accuracy                           1.00     85118
   macro avg       0.89      0.79      0.83     85118
weighted avg       1.00      1.00      1.00     85118

-----------------Accuracy of KNN----------------
Accuracy:- 0.9984609660789492
Precision:- 0.8
```

```
---------- classification report of CNN---------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     84984
           1       0.79      0.58      0.67       134

    accuracy                           1.00     85118
   macro avg       0.89      0.79      0.83     85118
weighted avg       1.00      1.00      1.00     85118

-----------------Accuracy of CNN----------------
Accuracy:- 0.9994125801828051
```

Conclusion

The 'InvisiFraud' project successfully used machine learning and deep learning models to detect credit card fraud. Among the models examined, Random Forest had the best accuracy (0.9995). However, all models displayed high accuracy rates, highlighting their applicability for real-world applications.

Model Name	Accuracy
Logistic Regression	0.99909
Decision Tree	0.99917
Random Forest	0.99949
KNN	0.99846
CNN	0.99941

References

- Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications
- Kim, J., & Park, H. (2021). A comparison of machine learning algorithms for credit card fraud detection. Journal of Financial Engineering
- Zong, Y., & Lu, X. (2021). Detecting fraudulent transactions using machine learning models: A review. Journal of Financial Technology