Comparative Analysis of Machine Learning Models for Fraud Detection

Leslye Patricia NKWA TSAMO, Karel SODJINOUTI
National School of Statistics and Economic Analysis (ENSAE) Pierre Ndiaye

Introduction

With the increase of digital payments, businesses are facing an increase in online fraudd. These frauds, which are often difficult to detect are challenging the detection systems. In this context, the electronic business platform Xente has made available an anonymeous dataset for the first GalsenAI Hackaton Troughout his project we will evaluate multiple supervised algorithms to determine their effectiveness in detecting fraudulent transactions, using this dataset.

Dataset Overview

Feature Type	Feature Name	Description
Object	TransactionId	Transaction Identifier
Object	BatchId	Identifier attribute to transactions which have in common some specifications
Object	AccountId	It identifies the person who receive the payment
Object	SubscriptionId	Subscription or contract identifier
Object	CustomerId	Identify the person making the purchase
Object	CurrencyCode	it representes the transaction currency
Int	CountryCode	representing the country associated with the transaction.
Object	ProviderId	Identifier for the provider of the payment method
Object	ProductId	The specific identifier of the product or service being purchased.
Object	ProductCategory	The general category of the product, a higher level than ProductId.
Object	ChannelId	The channel through which the transaction was made.
Float	Amount	The nominal amount of the transaction, in the original currency
Int	Value	ambiguous variable
Object	TransactionStartTime	The exact timestamp (date and time) when the transaction began.
Int	PricingStrategy	The pricing plan associated with the transaction.
Int	FraudResult	Target Variable describing wether the transaction was fraudulent or not

Table 1. Features Overview with Data Types

The dataset contains numerical and categorical features related to online transactions, with a binary target variable, `FraudResult`. Figure 1 shows the distribution of numerical features, where long tails and extreme values suggest potential fraud patterns.

Preprocessing

The dataset exhibited a significant class imbalance, with fraud cases constituting only **0.2018%** of all transactions. This long-tail distribution challenges model training, leading to biased predictions towards the majority class (non-fraud).

Techniques Used:

- **Outlier Removal:** Although there is extreme outliers, we maintain them because it could be fraud cases.
- **Feature Engineering :** We use ***TransactionStartTime*** to create 7 others features: *DayOfWeek, WeekOfYear, IsWeekend, Month, Day, Hour, Crénaux-horaire.* We removed **CurrencyCode** and *CountryCode* because they have only one modality. TargetEncoder was apply on categorical features and RobustScaler on numerical one.
- **Train-Test Strategy:** Since we are using grids, we performed cross-validation. We used 85% of the data for the trainset.

Models and Techniques

We tested the following models:

1. Random Forest
2. XGBoost (Gradient Boosting variant)
3. Support Vector Machine (SVM)
4. K-Nearest Neighbors (KNN)
5. AdaBoost

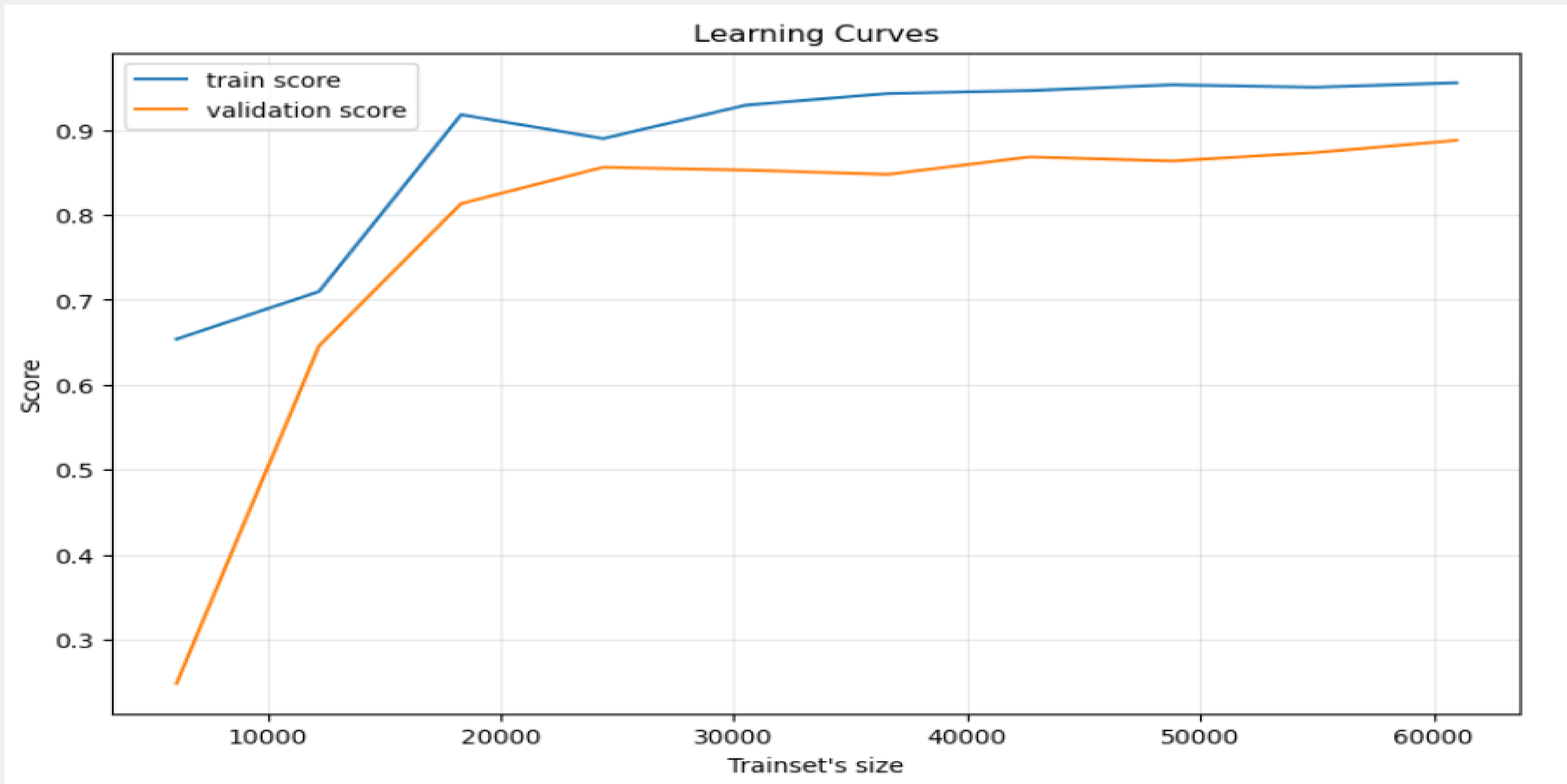
Hyperparameter Tuning: RandomizedSearchCV with 5-fold cross-validation

Performance Metrics

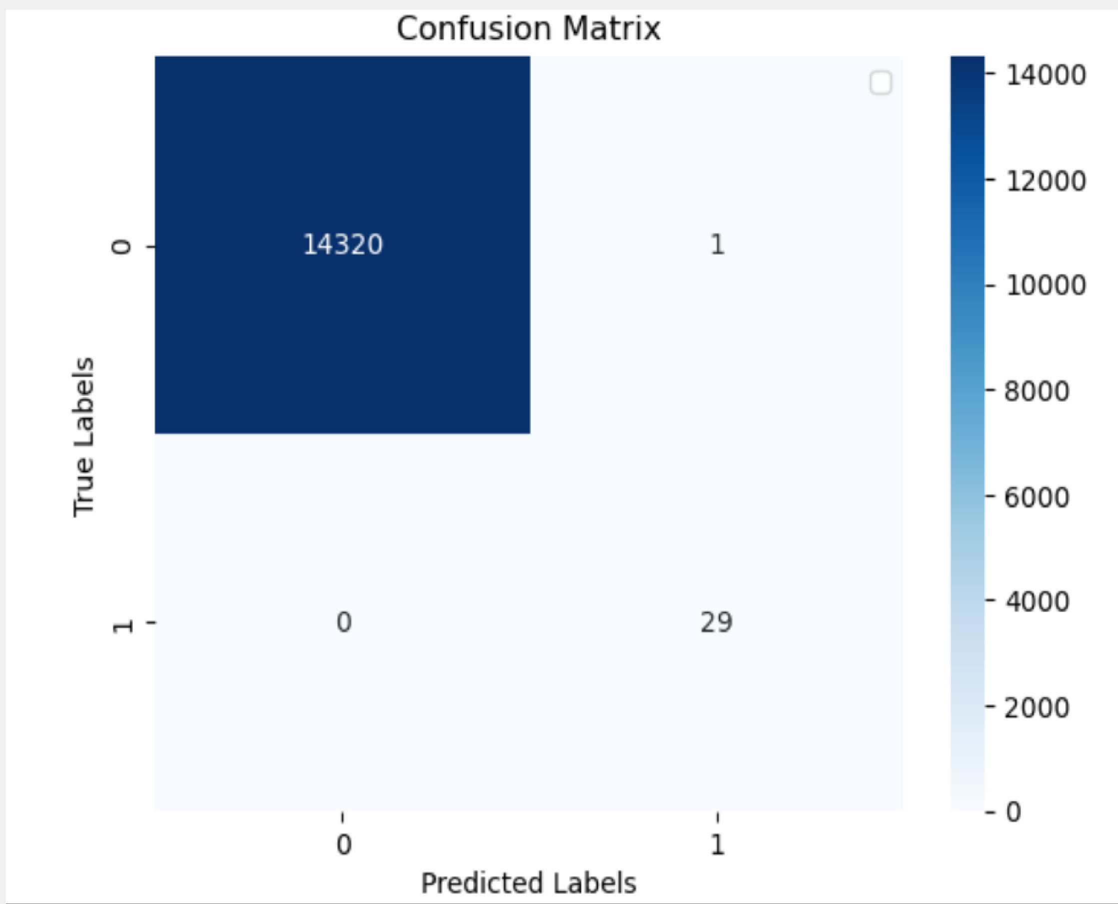
Model	F1 Score	Recall	Precision
Random Forest	**0.97**	**1.00**	0.98
XGBoost	0.72	0.79	0.66
Support Vector Machine	0.88	0.90	0.87
K-Nearest Neighbors	0.71	0.55	**1.00**
AdaBoost	**0.97**	0.97	0.97

Table 2. Performance Comparison of Models (After Tuning)

This table compares key performance metrics of all tuned models. The Random Forest Classifier and AdaBoost achieved the highest F1 score; Random Forest had the best recall, while K-Nearest Neighbors had the best Precision.



The training and validation curves illustrate the classification performance of the best model : Random Forest. It shows that the model generalize well on the validation set hence the recall of 1.



The confusion matrix shows that the Random Forest Classifier detects all fraud cases with only one false negative, demonstrating its effectiveness in identifying rare fraud events.

Conclusion and Future Work

Random Forest leads with the highest F1 Score (0.97) and perfect Recall (1.00), though with slightly lower Precision (0.98) AdaBoost shows excellent balance with 0.97 F1, 0.97 Recall, and 0.92 Precision.

Random Forest and AdaBoost emerged as top performers. While AdaBoost offered better balance in F1 and precision, the Random Forest excelled at identifying rare fraud cases.

Future enhancements may include:

- Testing CircularEncoder for sustainable prediction
- Applying cost-sensitive learning to reduce false positives
- Configure XGBoost differently, as it strangely does not perform very well
- Use deep learning models as Neural Networks

References

[1] J. "Online Payment Fraud Detection," from GalsenAI first Competitionhttps://drive.google.com/file/d/1k5uDrKS_KPASqB-xvNbPOAjhepiyODyJ/view?usp=drive_link.

ContactS: Leslye NKWA, SODJINOUTI Karel