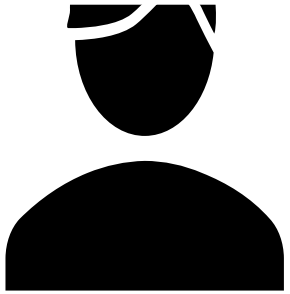
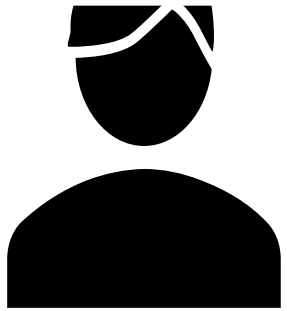
Leveraging Machine Learning For Credit Card Fraud Detection.

GROUP 11 MEMBERS



Collins Wanjao

Anita Bosibori



Bernard Kinyua

Janet Atunga

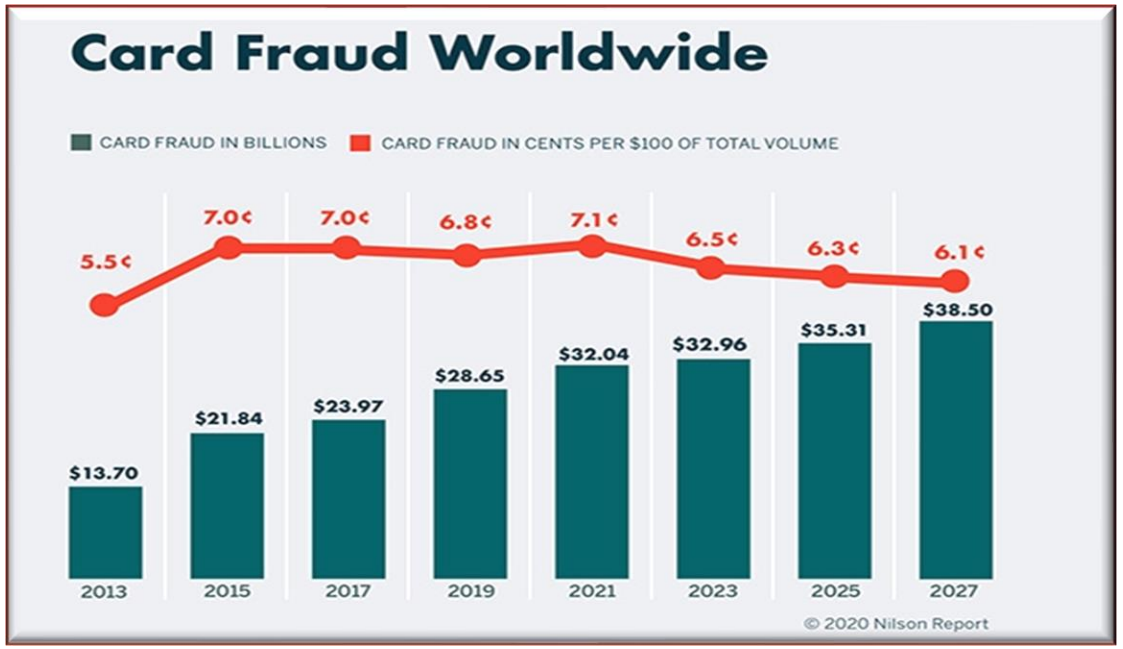
Joyleen Cherono

Ruth Kamau

Credit Card Fraud Definition

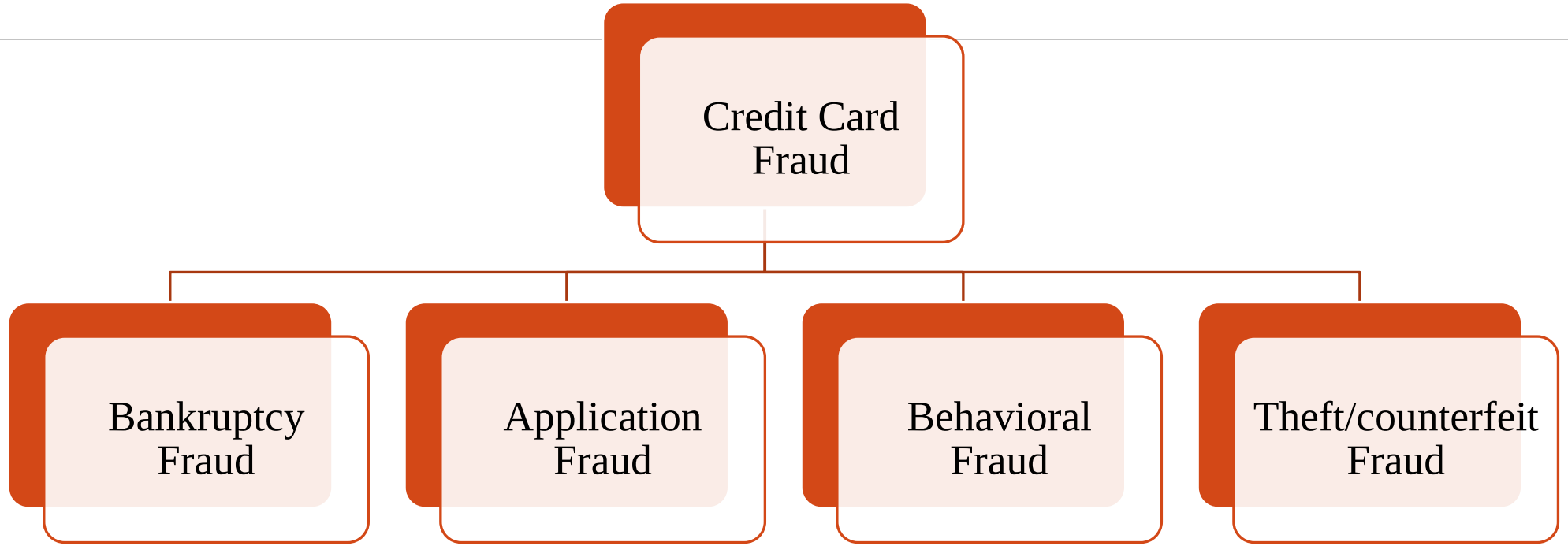
Credit card frauds are committed in the following ways:

- An act of criminal deception (mislead with intent) by use of unauthorized account and/or personal information
- Illegal or unauthorized use of account for personal gain
- Misrepresentation of account information to obtain goods and/or services.

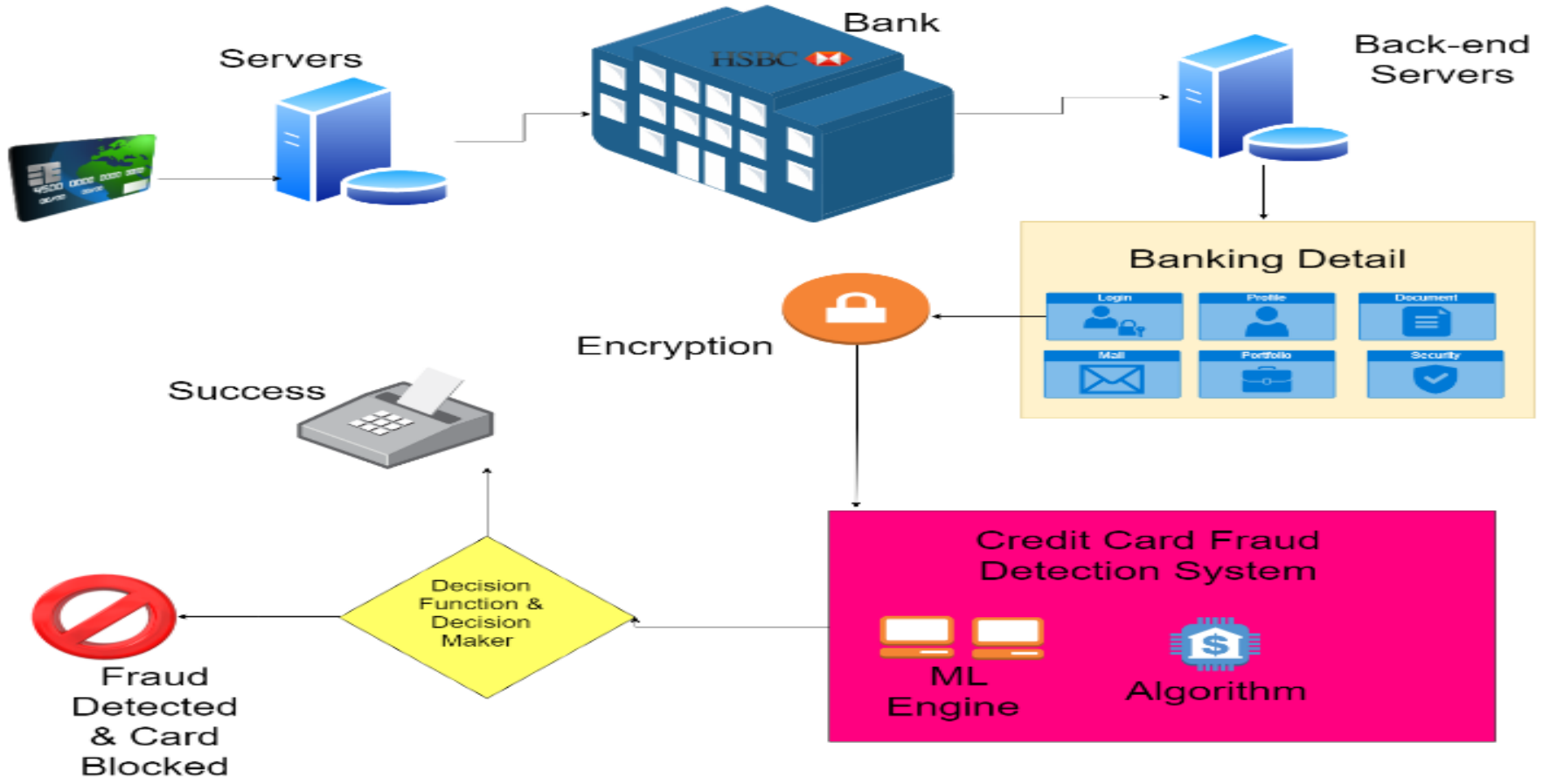


[1] Issue 1187, The Nilson Report, Dec. 2020. https://nilsonreport.com/publication_chart_of_the_month.php?%201=1&issue=1187

CREDIT CARD FRAUD TYPES



Credit Card Fraud Detection Scheme



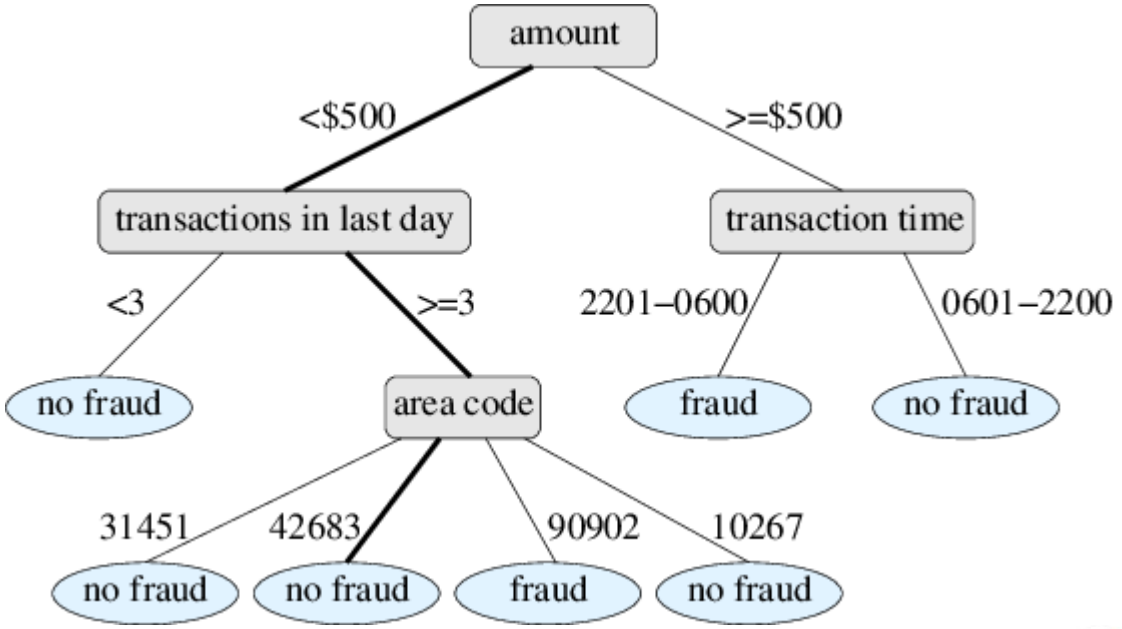
Machine Learning Methods

- The Decision tree (DT)
- K-nearest Neighbor(KNN)
- Logistic Regression
- Random Forest

The Decision Tree Classifier(DT)

- Classification rules, extracted from decision trees, IF-THEN expressions and all the tests have to be succeed if each rule is to be generated
- Separates the complex problem into many simple ones
- Resolves the Sub problem through repeatedly using

Example of DT implementation in fraud detection



Advantage

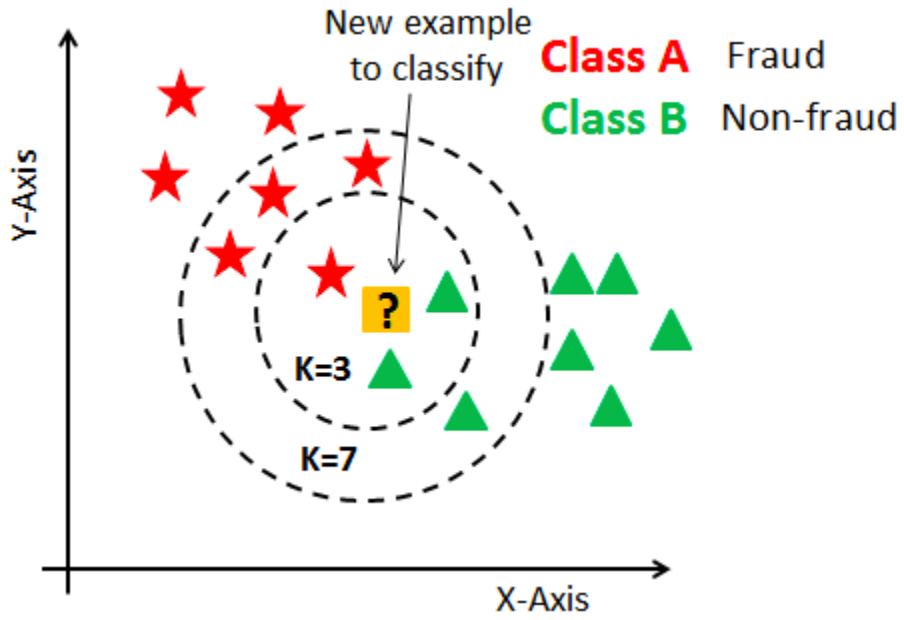
High flexibility/good haleness/ explainable/easy to implement/easy to display and to understand

Disadvantage

Requirements to check each condition one by one. In fraud detection condition is transaction.

K-nearest Neighbor(KNN)

- All the instances correspond to points in an n-dimensional feature space.
- Each instance is represented with a set of numerical attributes.
- Each of the training data consists of a set of vectors and a class label associated with each vector.
- Classification is done by comparing feature vectors of different K nearest points.
- Selection the K-nearest examples to sample in the training set.
- Assigning the sample to the most common class among its K-nearest neighbors.



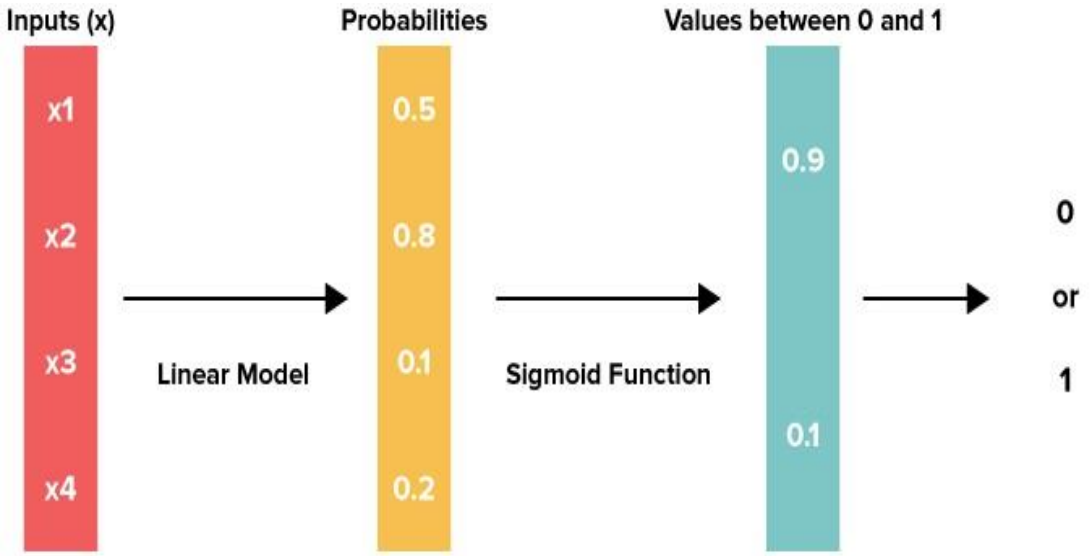
Advantage

There is no requirement of the predictive model before transaction classification in a dataset.

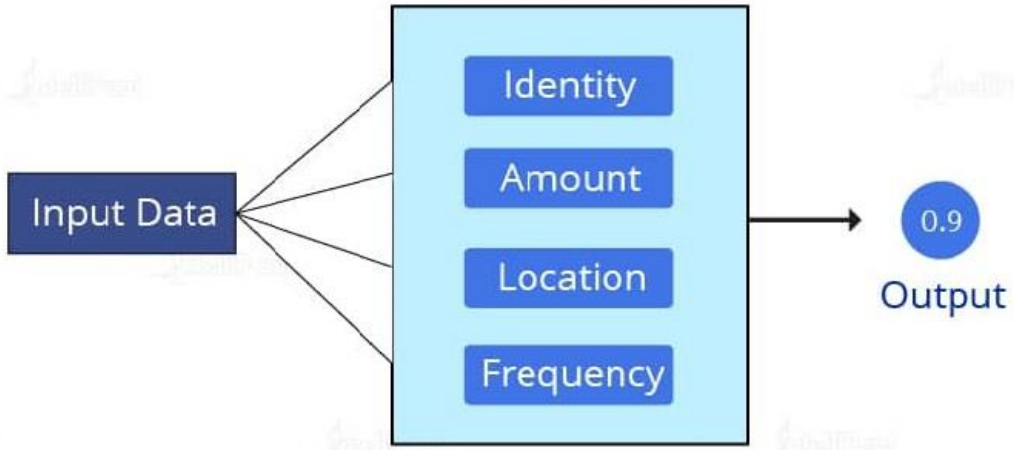
Disadvantage

The fraud detection accuracy depends on the measure of distance between two neighbors. This technique cannot detect the fraud at the time of transaction.

Logistic Regression(LR)



Simple example of LR work in Credit Card Fraud Detection Process



Advantage

It produces a simple probability formulae for classification. It works well with linear data for credit card fraud detection

Disadvantage

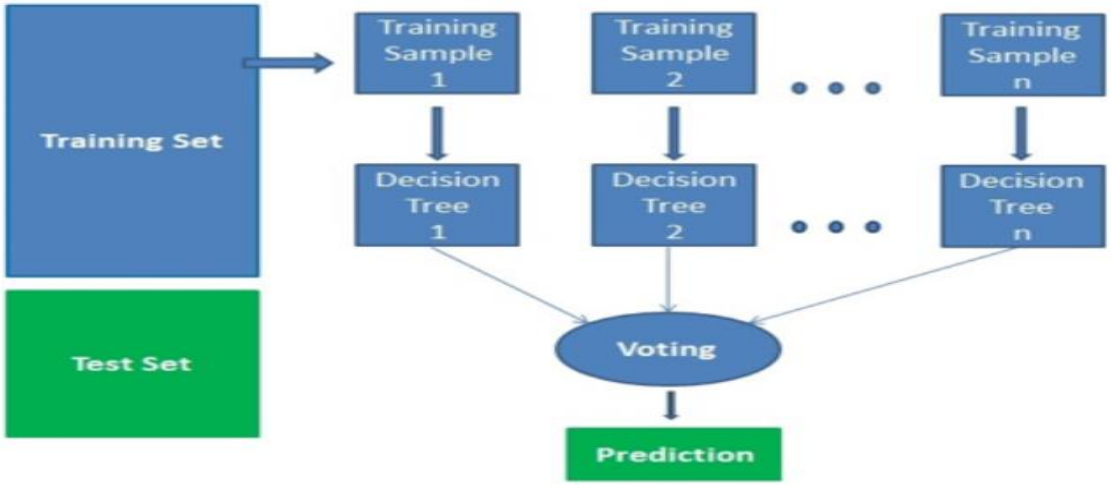
1.It can't be applied on non-linear data

2. It is not capable of handling fraud detection at the time of transaction

Random Forest Classifier(RF)

- All the instances correspond to points in an n-dimensional feature space.
- RF uses bootstrap samples from the training data to build multiple DTs.
- RF only uses a random subset of input features to split the data at each node in the tree.
- The predictions of each DT are combined and the most common output class (majority voting) is selected

Simple example of RF work in Credit Card Fraud Detection Process



Advantage

Its simplicity and quick implementation. It serves as a baseline for evaluating more advanced models, requiring minimal computational resources. While not effective for actual detection, it aids in cost-efficient initial assessment and guides further model development

Disadvantage

It lack of discriminative power, ineffectiveness in detecting fraud, and potential for generating false positives, risking customer inconvenience

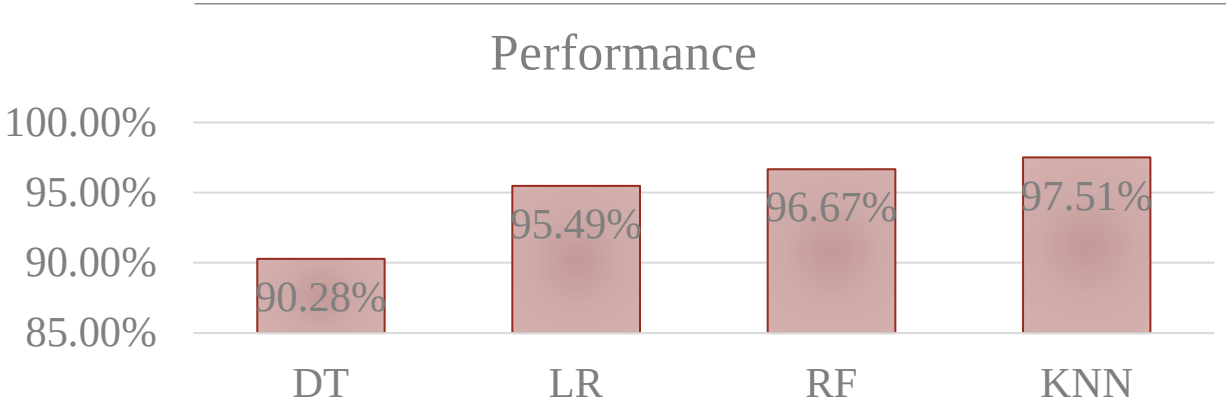
Evaluation Criteria

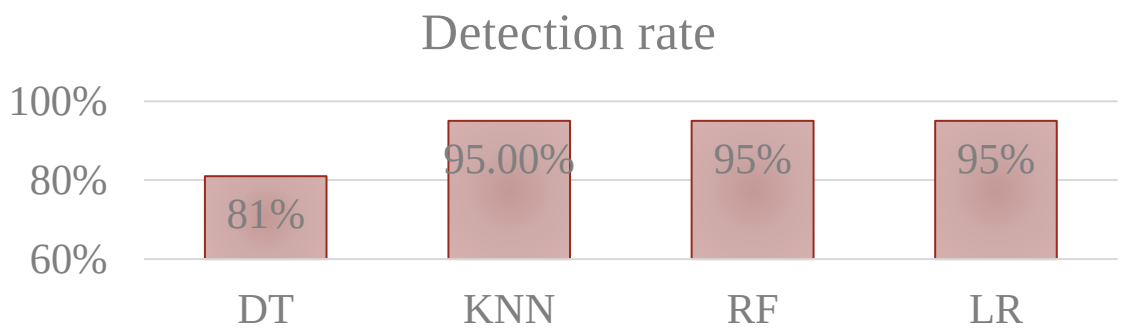
Measure	Description	Formula
Accuracy	The percentage of correctly predicated transactions	TN+TP/TP+TN+FP+FN
Detection Rate (Precision)	The number of classified fraud transactions that actually are fraud transactions, and gives the accuracy of the fraud transaction	TP/TP+FP
False Alarm Rate	The ratio of credit card fraud detected incorrectly	FP/FP+TN
True Positive Rate (Sensitivity or Recall)	The number of correctly classified fraud transaction and gives the accuracy of the fraud transaction.	TP/TP+FN
True Negative Rate (Specificity)	The amount of correctly classified non-fraud and gives the accuracy of the non-fraud transaction.	TN/TN+FP

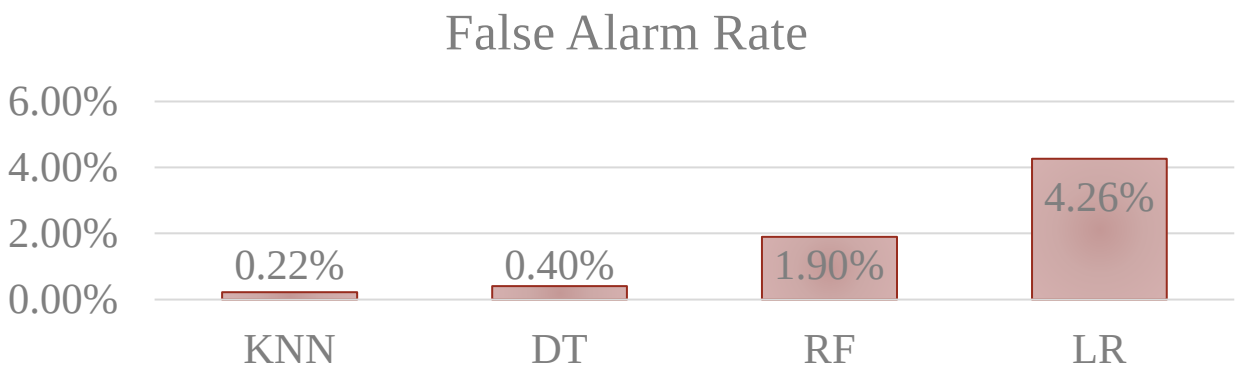
P = Positive N = Negative
FP = False Positive FN = False Negative
TP = True Positive TN = True Negative

[9] Singh A, Jain A. An Empirical Study of AML Approach for Credit Card Fraud Detection–Financial Transactions. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL. 2020 Feb 2;14(6):670-90.

Results


Performance
100.00%
95.00%
90.00%
85.00%
90.28%
95.49%
96.67%
97.51%
DT
LR
RF
KNN


Detection rate
100%
80%
60%
81%
95.00%
95%
95%
DT
KNN
RF
LR


False Alarm Rate
6.00%
4.00%
2.00%
0.00%
0.22%
0.40%
1.90%
4.26%
KNN
DT
RF
LR

Comparative analysis

Comparison of different techniques.

ML Algorithm	Cross Validation Score	ROC AUC Score	Recall Score (Fraud)	False Positives F	False Negatives
Logistic Regression (Vanilla)	99.09%	95.49%	95.00%	4.26%	0.01%
Logistic Regression (Tuned)	99.09%	95.49%	95.00%	4.26%	0.01%
Decision Tree Classifier (Vanilla)	99.81%	92.71%	86%	0.29%	0.02%
Decision Tree Classifier (Tuned)	99.87%	90.28%	81%	0.40%	0.03%
KNN Classifier (Vanilla)	99.94%	97.46%	95%	0.32%	0.01%
KNN Classifier (Tuned)	99.94%	97.51%	95%	0.22%	0.01%
Random Forest Classifier (Tuned)	99.46%	96.67%	95%	1.90%	0.01%

Challenges in Credit Card Fraud Detection

- Data Deficiency
- Highly Imbalanced Data
- Behavior Variations
- Cost-sensitive Problem

Current Solutions to the Challenges

Data Deficiency	Creation of Synthetic Data
Imbalanced Data	Oversampling and Under sampling methods SMOTE, etc
Behavior	Adaptive base learners and ensembles
Cost-sensitiveness	Implementing several algorithms and methods Feature engineering

Recommendations

- Creating a Hybrid Model of Various Techniques
- Improvement of Optimization Techniques of Existing CCFD Models
- Further exploration of hyperparameter tuning techniques, such as Bayesian optimization or grid search, may yield better-tuned models with
- Ensemble methods, such as stacking or boosting, could be explored to combine the strengths of multiple models and mitigate individual weaknesses.
- Continual monitoring and updating of models with new data and evolving fraud patterns are essential to maintain their effectiveness over time.

Any Questions?

Thank You For Attention