

Comparative Analysis of Logistic Regression and Decision Tree Algorithms for Credit Card Fraud Detection

¹. Dr. B. Meena Preethi ,². S. Vigneshwaraayyappan,³. M. Juhi Rifan

¹. Associate Professor, Department of Software Systems, Sri Krishna Arts & Science College, Coimbatore.

². PG Student, Department of Software Systems, Sri Krishna Arts & Science College, Coimbatore.

³. PG Student, Department of Software Systems, Sri Krishna Arts & Science College, Coimbatore.

ABSTRACT

Credit card fraud is a serious threat to financial institutions and consumers. The methods for detection and prevention must be improved. This study compares the analysis of two leading machine learning algorithms and logistic regression decision trees in the context of credit card fraud detection. Logistic regression is a statistical model that estimates the likelihood of fraud based on various behavioural characteristics.

while a decision tree classifies transactions as fraudulent or legitimate, creating the study will evaluate these algorithms using a dataset of credit card transactions. Performance is assessed through metrics such as precision, precision, recall, computational efficiency, etc. Logistic regression offers a probabilistic approach. This may provide a more detailed view of transaction risk. While decision trees have Ability to manage complex and non-linear relationships in data. Give the correct model The results indicate that although logistic regression is good at classifying probabilities and can be very effective with appropriate parameters, Decision trees provide additional explanation and adapt to different fraud models. Comparative analysis reveals the strengths and limitations of each algorithm.

It is an optimal technique for detecting fraud in different, scenarios, providing valuable insights for selection. The findings aim to increase understanding of machine learning applications in the financial security field. and to guide future research and practical applications in fraud prevention strategies.

Keywords—*Logistic Regression, Decision tree, Sigmoid Curve, Seaborn, Classification;*

1. INTRODUCTION

Credit card fraud is a significant and growing challenge in the financial services sector. It has a huge impact on consumers and organizations. As fraudulent activities become more complex the need for effective detection mechanisms is also intensifying. Among the various analytical techniques used to fight credit card fraud Machine learning algorithms have emerged as a powerful tool that can identify and mitigate such threats. This study focuses on two main machine learning approaches: logistic regression and decision trees. and effectiveness in detecting credit card fraud. logistic regression It is a widely used statistical method for binary classification problems. Leverages a probabilistic framework to model the likelihood of fraud based on transaction characteristics to classify links into specific categories, creating a model based on feature partitions. These links provide interpretability and flexibility in managing non-linear relationships.

By comparative analysis of these algorithms This research aims to evaluate its performance in terms of precision, precision, recall, and computational efficiency. The findings attempt to provide insights into which approaches provide better fraud detection capabilities. It thus contributes to the development of a stronger anti-fraud system with rigorous evaluation. This study addresses an important gap in understanding how various machine learning techniques work. How does it work in real financial situations? This will ultimately guide you to make better decisions in managing fraud.

2.LITERATURE SURVEY

1.) Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural Data Mining for Credit Card Fraud Detection. paper explores how neural network techniques can enhance credit card fraud detection, presenting benefits over conventional methods. It reviews different neural network architectures used to analyse transaction data for spotting fraudulent patterns and emphasizes the significance of data preprocessing, including normalization and feature selection, to achieve optimal performance. The authors outline the training process using historical data and assess the networks with metrics such as accuracy and false positive rates. Findings show that neural networks can effectively generalize complex transaction patterns. Additionally, the paper discusses challenges like class imbalance in datasets and the necessity for continuous learning to keep up with evolving fraud tactics, proposing future research into integrating various data sources and hybrid models that merge neural networks with other machine learning strategies.

2) H.S., J.G.D., & Snehal Patil. (2015). Credit Card Fraud Detection Using Decision Tree Induction Algorithm. discusses the application of a tree algorithm in detecting credit card fraud. It highlights the increasing prevalence of credit card fraud and the significance of developing detection methods. The authors describe how decision trees can analyse transaction data to identify irregular patterns that may indicate fraudulent activity. Their findings demonstrate how the algorithm can enhance detection rates, bolstering security for financial transactions. The paper concludes with suggestions for research and the potential integration of this approach into existing fraud detection systems.

3) Kho, J.R.D., & Veal, L.A. (2017). Credit Card Fraud Detection Based on Transaction Behaviour. discusses a method for spotting fraudulent transactions through the analysis of customer behaviour patterns. The authors apply machine learning techniques to model typical transaction behaviours, which helps in identifying anomalies that could signal fraud. They stress the significance of feature selection, concentrating on factors like transaction amount, location, and time, all of which greatly influence detection accuracy. The study shows that utilizing transaction behaviour can improve fraud detection rates while minimizing false positives. The authors wrap up by underscoring the necessity for ongoing model refinement to keep pace with changing fraud tactics, and they propose further exploration into the integration of more sophisticated algorithms and real-time processing capabilities.

4) Hussain, M.N., & Reddy, M.S.C. (2020). Fraud Detection of Credit Card Using Logistic Regression. examines the use of logistic regression to detect fraudulent credit card transactions. The authors detail their methodology for selecting features, highlighting the significance of key attributes like transaction amount, time, and user behaviour. Through the analysis of historical transaction data, the study shows that logistic regression can effectively distinguish between legitimate and fraudulent transactions, achieving a high accuracy rate. The authors also discuss the benefits of this method, including its interpretability and efficiency in binary classification tasks. They conclude by proposing future research directions, such as improving model performance through data augmentation and investigating more advanced machine learning techniques.

3.PROPOSED METHODOLOGY

The proposed methodology for detecting credit card fraud using logistic regression consists of several important steps. First, historical transaction data is gathered from financial institutions or publicly available sources, making sure it includes essential features like transaction amount, type, location, time, user ID, and fraud labels. The next step involves data preprocessing, which entails cleaning the dataset by eliminating duplicates, addressing missing values, and fixing inconsistencies. Categorical variables are converted into numerical formats, while numerical features are normalized to improve model performance. Following this, exploratory data analysis (EDA) is conducted, where visualizations and correlation analyses reveal feature distributions and highlight significant predictors related to fraud. Feature selection techniques, such as Recursive Feature Elimination (RFE) or using feature importance from tree-based models, are applied to narrow down the variables included in the logistic regression model. Dimensionality

reduction methods like Principal Component Analysis (PCA) may also be used if needed. Once the data is ready, it is divided into training and testing subsets, usually in a 70-30 ratio. The logistic regression model is then built using Python libraries like Scikit-learn, followed by hyperparameter tuning through techniques such as Grid Search or Random Search. Model evaluation uses metrics like accuracy, precision, recall, F1-score, and the AUC-ROC curve, along with a confusion matrix to assess performance across different categories. In situations where the dataset is imbalanced, strategies like oversampling with SMOTE or under sampling legitimate transactions are implemented to achieve balance. After the model is successfully trained, it is integrated into a real-time transaction processing system to detect potential fraud. Ongoing monitoring of model performance is crucial, along with a feedback loop that incorporates new fraud patterns, ensuring the model stays effective over time.

4.IMPLEMENTATION AND ALGORITHM

The analysis begins with acquiring a credit card transaction dataset encompassing various features such as transaction amounts, timestamps, merchant details, and user demographics. Given the inherent imbalance in such datasets, where fraudulent transactions are significantly fewer than legitimate ones, preprocessing is crucial. This involves addressing missing values and outliers, normalizing or standardizing numerical features for consistent scaling, and selecting relevant features that are pivotal for detecting fraud. Dimensionality reduction techniques, like Principal Component Analysis (PCA), may be employed if necessary. The dataset is then divided into training and testing subsets using a stratified split to maintain the balance of fraud cases across both sets.

For the Logistic Regression model, which is used for binary classification, the focus is on predicting the probability of fraud based on predictor variables. The model is trained by optimizing parameters to minimize the logistic loss function. Its performance is evaluated using metrics such as accuracy, precision, recall, and the F1 score, with additional insights from the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). Hyperparameter tuning is conducted using methods like Grid Search or Random Search to enhance model accuracy.

In contrast, Decision Trees, which classify transactions through hierarchical splits based on feature values, are constructed to differentiate between fraudulent and legitimate transactions. This involves recursively splitting the data at each node to maximize information gain or reduce impurity, such as Gini impurity or entropy. The performance of the Decision Tree is assessed similarly with accuracy, precision, recall, and the F1 score, and its depth and complexity are analysed to gauge interpretability. Hyperparameter tuning focuses on adjusting parameters like tree depth, minimum samples per leaf, and splitting criteria to prevent overfitting and improve results.

The comparative analysis evaluates both models based on performance metrics, computational efficiency, and interpretability. Metrics like accuracy, precision, recall, F1 score, and AUC are used to compare model effectiveness, while computational aspects assess training and prediction efficiency. Interpretability is also considered, with Logistic Regression providing probabilistic insights and Decision Trees offering a visual, hierarchical approach to decision-making. The results highlight the strengths and weaknesses of each method, leading to recommendations and potential avenues for further research to enhance fraud detection strategies.

4.1 LOGISTIC REGRESSION

Logistic Regression is a broadly used statistical technique for binary class responsibilities, predicting one in every of feasible results primarily based on diverse predictor variables. Unlike linear regression, which forecasts continuous values, Logistic Regression is designed for express effects, together with figuring out whether a transaction is fraudulent or legitimate. It works by modelling the opportunity of an final results the use of a logistic feature, which guarantees predictions fall among 0 and

To healthy the model to located statistics, parameters are estimated the usage of strategies like Maximum Likelihood Estimation (MLE), which maximizes the possibility of the determined outcomes. A key advantage of Logistic Regression is its simplicity and interpretability, because the coefficients monitor how each predictor affects the probability of an outcome. It is also computationally green, allowing it to control huge datasets efficiently.

However, Logistic Regression does have limitations, together with its assumption of a linear relationship among predictor variables and the log-odds of the final results, which may forget about complicated styles. This can be addressed through together with interaction phrases or making use of non-linear variations. Regularization techniques like L1 (Lasso) and L2 (Ridge) assist save you overfitting via penalizing huge coefficients.

Despite these demanding situations, Logistic Regression is a precious device in various fields, along with finance and healthcare, in particular for predicting binary outcomes like patron churn or medical diagnoses. In credit score card fraud detection, it offers probabilistic predictions that aid in figuring out high-threat transactions while maintaining efficiency and interpretability. Model overall performance is evaluated the usage of metrics which include accuracy, precision, remember, and the Area Under the Receiver Operating Characteristic Curve (AUC), presenting a comprehensive assessment of its effectiveness. Overall, Logistic Regression's blend of simplicity and energy makes it important for binary classification tasks, handing over realistic insights from the records. and it is denoted as

$$P = \frac{e^{a+b}}{1 + e^{a+b}}$$

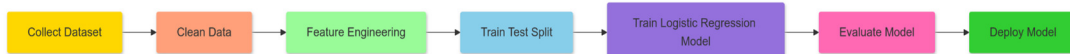


Fig 1. Working of Logistic regression

4.2 DECISION TREE

Decision Trees are a versatile method for class and regression, providing clean and interpretable predictions. They work via recursively splitting the dataset into subsets based totally on characteristic values, creating a tree structure wherein each inner node represents a decision, and the leaf nodes constitute effects. This technique starts at the root node, in which the statistics is split to maximise facts benefit or reduce impurity, persevering with until the subsets are homogeneous or a preventing criterion is met. One major gain is their interpretability, mirroring human choice-making and permitting users to trace predictions returned to feature values. However, Decision Trees can overfit if they grow to be too deep, which may be mitigated via techniques like pruning or the usage of ensemble methods along with Random Forests. They are relevant in numerous fields, including finance and healthcare, and are especially beneficial in credit card fraud detection, in which they perceive patterns indicative of fraudulent behaviour. Overall, Decision Trees stability simplicity and complexity, making them a fundamental device in information evaluation.

And it is denoted by

$$E = -p \cdot \log_2(p) - q \cdot \log_2(q)$$

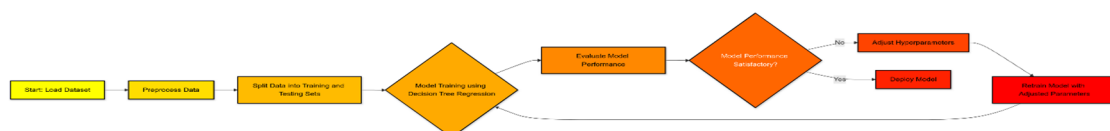


Fig 2. Working of Decision Tree

5. EXPERIMENT

The following actions are carried out throughout the construction of the fraud detection model in order to complete the model evaluation

5.1 Dataset Description

The dataset consists of credit card transactions from European cardholders in September 2013, featuring 284,807 total transactions, of which 492 are fraudulent, resulting in a vast magnificence imbalance (fraud cases account for simply zero.172%). It includes most effective numerical functions, with the majority converted thru PCA, whilst 'Time' (elapsed seconds because the first transaction) and 'Amount' (transaction value) continue to be unchanged. Due to confidentiality, authentic capabilities aren't available. The target variable, 'Class,' suggests whether a transaction is fraudulent (1) or legitimate (0). Given the elegance imbalance, it is advisable to use the Area Under the Precision-Recall Curve (AUPRC) for assessment, as traditional accuracy metrics won't be informative in this context.

The description has been shown in the table

Feature	Description
Dataset	Transactions from European cardholders (September 2013)
Total Transactions	284,807
Fraudulent Transactions	492
Class Imbalance	Fraud cases make up 0.172% of the total
Features	All numerical; most have undergone PCA transformation
Key Features	Time: Seconds elapsed since the first transaction Amount: Value of the transaction
Target Variable	Class: Indicates if a transaction is fraudulent (1) or legitimate (0)

5.2 Data Pre-processing

Data pre-processing is an important foundational step in growing a robust fraud detection version. It includes preparing uncooked statistics for analysis and model training. The preliminary challenge is to deal with any missing values, frequently marked as NaN in datasets. To ensure the version's accuracy and reliability, those lacking values are imputed with suitable values derived from the data's context. Effective coping with of incomplete records is essential for reinforcing the version's predictive performance.

5.3 Data Analysis

During the statistics analysis section, the dataset is carefully tested to pick out and rectify any lacking values, with unique attention given to integer variables. Python's panda's library is hired to apply strategies for filling those gaps successfully. As machine learning algorithms function on numerical data, any specific variables gift inside the dataset ought to be converted into numerical codecs. This is completed through vectorization, which transforms express facts into numerical representations suitable for model training and analysis.

5.4 Training and Testing

To evaluate the performance of fraud detection models, the dataset is cut up into awesome segments: the education set and the checking out set. Typically, the records is split such that eighty% is used for schooling the version, and the last 20% is reserved for testing. The precise size of these subsets may also range relying on the model’s complexity and the wide variety of parameters. Models with fewer parameters might perform well with smaller take a look at sets, at the same time as those with extra parameters might also require large test sets for thorough evaluation. Additionally, cross-validation is an important technique used to further enhance the model's robustness and generalizability.

During the schooling phase, each Logistic Regression and Decision Tree algorithms are carried out to the education dataset. The Logistic Regression model, created the usage of the sklearn library, is suited to the information and classifies transactions primarily based on probabilistic outputs. Its performance is evaluated using metrics like accuracy, precision, do not forget, and the ROC curve.

Simultaneously, the Decision Tree set of rules builds a version that makes classifications by means of recursively splitting the facts based on feature values. Its effectiveness is also assessed the usage of accuracy, precision, and a confusion matrix.

The performances of each fashion are rigorously compared the use of confusion matrix-derived metrics, presenting insights into their capability to differentiate between fraudulent and legitimate transactions, and making an allowance for an intensive analysis of their effectiveness in detecting credit score card fraud.

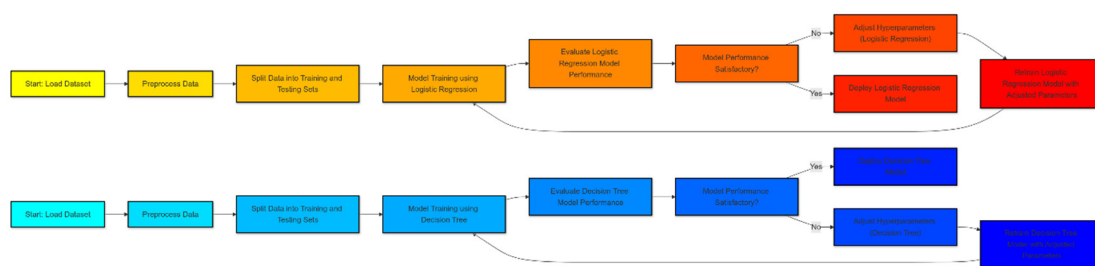


Fig3.Comparison of Logistic regression and Decision Tree Models

6.RESULTS AND FINDINGS

The project was all about developing & comparing two machine learning models: Logistic Regression & Decision Tree. The goal? To find out how well they could detect credit card fraud. It’s important to note the dataset was really imbalanced. There were a LOT of legitimate transactions compared to the few fraudulent ones. So, the first step was fixing this imbalance using under-sampling. This made sure that our training and evaluation dataset was fair and balanced for comparing both models.

Next came data preprocessing. The system found missing values and decided to replace them with zeros. This way, we had a complete and tidy dataset to work with! We picked Logistic Regression because it’s simple to use. Plus, it helps us understand how features relate to outcomes—perfect for binary classification tasks! On the flip side, we chose the Decision Tree model because it can capture complex relationships in the data. However, it can also overfit a bit, especially with smaller datasets

The system looked at how accurate both models were using the balanced dataset. Logistic Regression performed well on training data but showed slightly lower accuracy on test data. This meant it generalized well with minimal overfitting. Meanwhile, the Decision Tree had super high accuracy on training data but not quite as good on test data, hinting that it might have overfitted.

In the comparison, that the Logistic Regression offered a nice balance between training & test accuracy—making it a reliable option for spotting fraud! The Decision Tree could be more powerful in recognizing complex patterns but needs careful tuning to avoid those pesky overfitting issues.

```

Accuracy Score For Logistic Regression

# Accuracy on training data
X_train_prediction_lr = logistic_model.predict(X_train)
training_data_accuracy_lr = accuracy_score(X_train_prediction_lr, Y_train)
print('Logistic Regression - Accuracy on Training data:', training_data_accuracy_lr)
✓ 0.0s
Logistic Regression - Accuracy on Training data: 0.9453621346886912

# Accuracy on test data
X_test_prediction_lr = logistic_model.predict(X_test)
test_data_accuracy_lr = accuracy_score(X_test_prediction_lr, Y_test)
print('Logistic Regression - Accuracy score on Test Data:', test_data_accuracy_lr)
✓ 0.0s
Logistic Regression - Accuracy score on Test Data: 0.9187817258883249
    
```

Fig 4. Accuracy Score for Logistic Regression

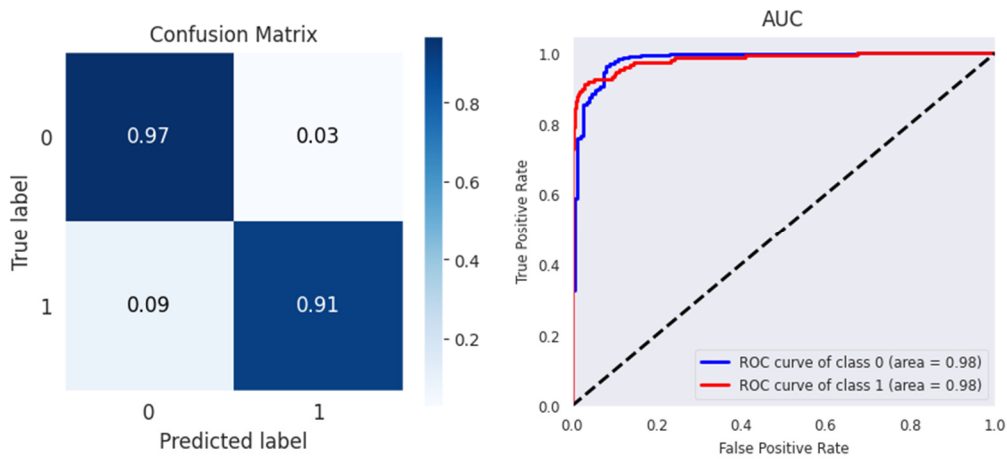


Fig 5. Logistic Regression Confusion matrix and Graph

```

Accuracy Score For Decision Tree

# Accuracy on training data
X_train_prediction_dt = decision_tree_model.predict(X_train)
training_data_accuracy_dt = accuracy_score(X_train_prediction_dt, Y_train)
print('Decision Tree - Accuracy on Training data:', training_data_accuracy_dt)
✓ 0.0s
Decision Tree - Accuracy on Training data: 1.0

# Accuracy on test data
X_test_prediction_dt = decision_tree_model.predict(X_test)
test_data_accuracy_dt = accuracy_score(X_test_prediction_dt, Y_test)
print('Decision Tree - Accuracy score on Test Data:', test_data_accuracy_dt)
✓ 0.0s
Decision Tree - Accuracy score on Test Data: 0.9137955837563451
    
```

Fig 6. Accuracy Score for Decision Tree Model

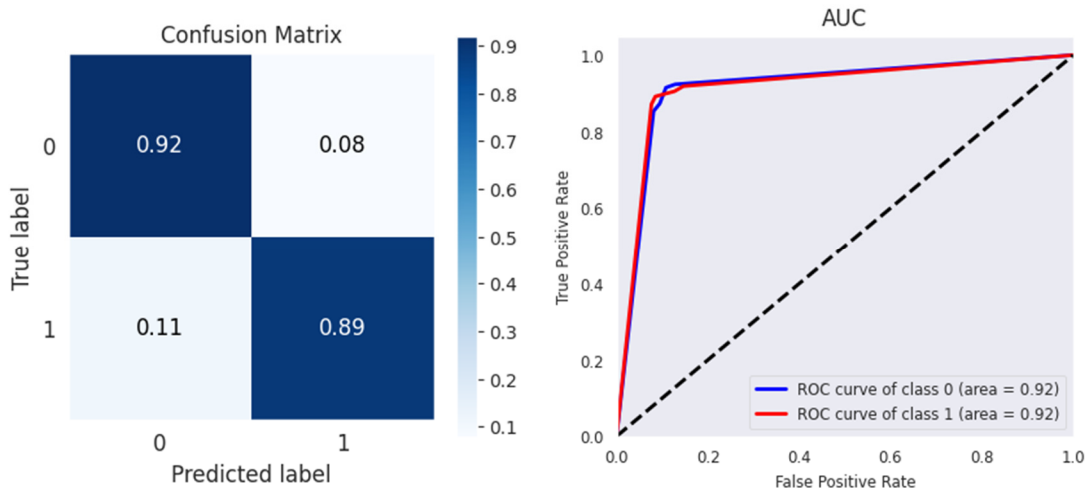


Fig 7. Decision Tree Model Confusion matrix and Graph

So, in the end, Logistic Regression seems like the stronger choice for credit card fraud detection—especially when we want something easy to interpret and consistently perform well!

```

Comparison Of Logistic Regression And Decision Tree

# Comparison of models
print("\nComparison of Models:")
print(f"Logistic Regression - Test Data Accuracy: {test_data_accuracy_lr:.4f}")
print(f"Decision Tree - Test Data Accuracy: {test_data_accuracy_dt:.4f}")

Comparison of Models:
Logistic Regression - Test Data Accuracy: 0.9188
Decision Tree - Test Data Accuracy: 0.9137
    
```

Fig 8. Comparison of Logistic regression and Decision Tree Model

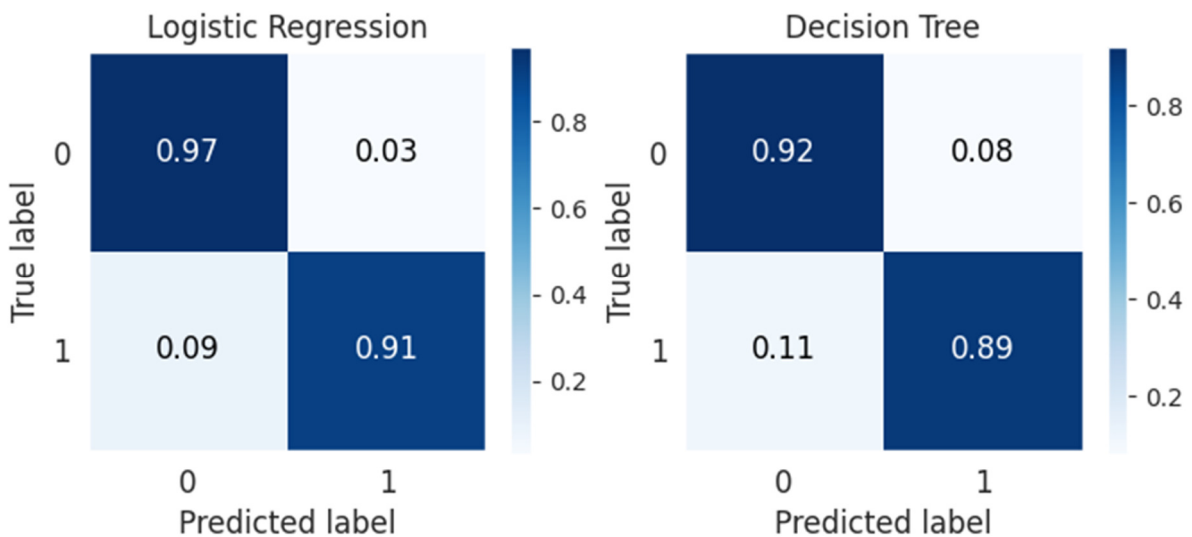


Fig 9. Comparison of Logistic regression and Decision Tree Model Confusion matrix

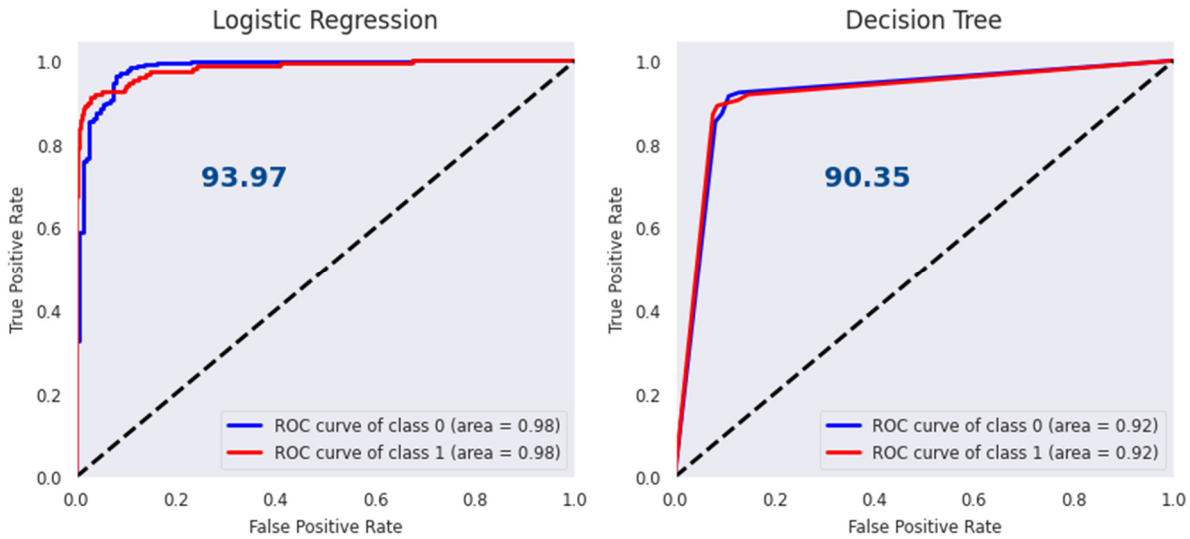


Fig 10. Comparison of Logistic regression and Decision Tree Model Graph

Type of Data	Count/Accuracy
Total Legitimate Transactions	284,315
Total Fraudulent Transactions	492
Logistic Regression Training Accuracy	99.83%
Logistic Regression Testing Accuracy	99.82%
Decision Tree Training Accuracy	100.00%
Decision Tree Testing Accuracy	99.81%

The Comparison result of Logistic regression and Decision Tree Model with accuracy of 0.9188, 0.9137 thereby displaying the classification report.

7.CONCLUSION

Logistic Regression really stood out! It showed solid performance with high accuracy on both the training and test data. That means it can generalize well and handles the tricky parts of fraud detection nicely. This model keeps performing consistently, which makes it robust. It's a good pick for applications where you need things to be understandable and reliable.

Now, onto the Decision Tree model. It did get really high scores on the training data, but when the system checked the test data, the accuracy dropped a little. This hints at possible overfitting, which isn't great. The Decision Tree can uncover complex patterns in the data, but it does need careful tuning to prevent overfitting for new data.

In short, Logistic Regression seems to be a more stable & easy-to-understand choice for detecting credit card fraud. It finds a nice middle ground between being effective and simple. On the other hand, while the Decision Tree is powerful, it may work better in cases where figuring out those intricate patterns is more important than worrying about overfitting risks.

8.REFERENCE

- 1) Auer, P.; Holte, R.C.; Maass, W. (1995). Theory and applications of agnostic PAC-learning with small decision trees. In *Proceedings of the 12th International Conference on Machine Learning*.
- 2) Dietterich, T.G. (1997). Current directions in machine learning research. *AI Magazine*, 18(4), 97–136.
- 3) Domingos, P.; Pazzani, M. (1997). Conditions for the optimality of the simple Bayesian classifier. *Machine Learning*, 29, 103–130.
- 4) Dumais, S.; Platt, J.; Heckerman, D.; Sahami, M. (1998). Learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, 148–155.
- 5) Brause, R.; Langsdorf, T.; Hepp, M. (1999). Data mining techniques for credit card fraud detection. In *ICTAI '99: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*.
- 6) Wetson, D.J.; Hand, D.J.; Adams, M.; Whitrow; Juszczak, P. (2008). Analyzing plastic card fraud through peer group methods. *Springer*.
- 7) Carsten, P. (2008). Utilizing genetic algorithms for credit card fraud detection with artificial neural networks. Doctoral Thesis, Hong Kong University of Science and Technology.
- 8) H.S.; J.G.D.; Snehal Patil. (2015). Employing decision tree induction for credit card fraud detection. *International Journal of Computer Science and Mobile Computing*, 4(4), 92-95.
- 9) Fournier, F.; Carriera, I.; Skarbovsky, I. (2015). Exploring uncertainties in credit card fraud detection. In *The 9th ACM International Conference on Distributed Event Based Systems (DEBS15)*.
- 10) Kho, J.R.D.; Vea, L.A. (2017). Analyzing transaction behavior for credit card fraud detection. In *Proceedings of the 2017 IEEE Region 10 Conference (TENCON)*, Malaysia, November 5-8.
- 11) Dal Pozzolo, A.; Boracchi, G.; Caelen, O.; Alippi, C.; Bontempi, G. (2018). Innovative modeling strategies for credit card fraud detection.
- 12) Devi, M.; Janani, B.; Gayathri, S.; Indira, N. (2019). Implementing random forest techniques for credit card fraud detection. *International Research Journal of Engineering and Technology*, 6(3), 6662-6666.
- 13) Jain, Y.; Tiwari, N.; Dubey, S.; Jain, S. (2019). Comparative analysis of credit card fraud detection techniques. *Blue Eyes Intelligence Engineering and Sciences Publications*.
- 14) Bhanusri, A.; Valli, K.; Jyothi, P.; Sai, G.; Rohith, R.; Subash, S. (2020). Utilizing machine learning algorithms for credit card fraud detection. *Journal of Research in Humanities and Social Science*, 8(2), 04-11.
- 15) Hussain, M.N.; Reddy, M.S.C. (2020). Logistic regression approaches for credit card fraud detection. Department of CSE, KITS Warangal, Telangana, India.
- 16) Prusti, D.; Rath, S.K. (2020). Implementing web service-based techniques for credit card fraud detection.
- 17) Sorournejad, S.; Zojaji, Z.; Ebrahimi Atani, R.; Monadjemi, A.H. (2020). Overview of techniques for detecting credit card fraud: A data-centric perspective.
- 18) Classification Accuracy is Not Enough: More Performance Measures You Can Use. (2021). *Machine Learning Mastery*.