e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

FINANCIAL FRAUD DETECTION USING MACHINE LEARNING TECHNIQUES

Matvalam Sumanth*1, Y. Guravaiah*2, Vemplli Subhasini*3

*1PG Scholar, Department Of Computer Science And Engineering, AIML, India.

*2,3Associate Professor, Department Of Computer Science And Engineering, AIML, India.

ABSTRACT

Digital payments of all kinds are increasing all over the world. For instance, in 2018, payments totaling $578 billion were processed by PayPal. It is of utmost importance for financial institutions like banks and credit card companies to find fraudulent transactions in real time to withhold any suspicious transaction as majority of traditional approaches are manual, which is not only inefficient, expensive, and imprecise but also impractical. By analyzing a large amount of financial data, machine-learning-based methods can intelligently detect fraudulent transactions. Most banks and other financial institutions have dedicated teams of dozens of analysts working on automated systems to identify potentially fraudulent transactions through their products. In this research, publicly available data was used on different payment transactions, and solved the issue of fraud detection using different machine learning techniques. Machine learning and Deep Learning techniques was implemented for fraud detection and demonstrate that fraudulent and non-fraudulent transactions can be distinguished through exploratory analysis. General Terms Financial Fraud, Machine Learning, Deep Learning, Algorithms.

Keywords: Financial Fraud Detection, Logistic Regression, Random Forest, Deep Learning.

I. INTRODUCTION

Financial fraud, considered a tricky strategy for acquiring monetary advantages, has become a broad threat to organizations. The volume of transactions handled by payment companies is expanding rapidly. Alongside this change, a fast expansion in monetary misrepresentation occurs in these installment frameworks. To be better prepared to deal with cases of cybercrime, it is essential to investigate a method for resolving the issue of identifying fraudulent entries and transactions in large amounts of data. Obtaining financial benefits through illegal and fraudulent means is known as financial fraud. Insurance, banking, taxation, and the corporate sector are just a few examples of areas where financial fraud can occur. There is abundance of literature on financial fraud detection because it is vital for reducing cybercrimes and businesses. As stated by Mehbodniya, companies and industries are facing an increasing challenge from financial transaction fraud, money laundering, and other forms of financial fraud [2]. The economy and society suffer from persistent fraud, which results in daily losses of a significant amount of money despite numerous efforts to curb it. Predicting fraud activities was done using both unsupervised and supervised methods and the most widely used method for identifying financial fraud transactions has been classification. Maurya and Kumar gave a comprehensive review of their work on detecting credit card fraud. The authors comprehensively analyze various ML classification techniques, including their approaches and difficulties [1]. Krasic and Celar analyzed a decade's studies on data mining-based fraud detection in the financial sector [6]. However, despite several reviews in the field, most studies primarily focused on specific aspects of finance, such as the detection of fraudulent activities on credit cards, fraud in online banking, fraud in bank credit administration, and fraud in payment cards [5]. First and foremost, their primary objective is to integrate accounting, information systems, and analytics research. On the other hand, our goal is to use a financial dataset and to identify financial fraud transactions using supervised learning classification techniques to classify between legitimate and fraudulent transactions and to classify the type of transactions that are most prone to fraud.

II. METHODOLOGY

DATA

The dataset used in the project was obtained from Kaggle [https://www.kaggle.com/] which is an online platform and community for data science and machine learning enthusiasts and provides a platform where users can access a wide range of datasets. The dataset has approximately 6 million rows of data spread across 11 columns. The key columns that are important are “step” which maps a unit of time in the real world, 1 step is

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

1 hour of time. The column "type" includes transactions which are cash-in, cash-out, debit, payment or transfer, "Amount" transacted which is amount of the transaction in local currency, "Old and New balance of Customer Recipient" which is the initial balance of the originator before the transaction and balance after the transaction respectively, "nameDest" which is the identifier of the recipient who received the transaction, "oldbalanceDest and newbalanceDest" which is the Old and New balance of originator which is the initial balance of the originator before the transaction and balance after the transaction respectively and "IsFraud" which indicates whether the transaction is actually fraudulent or not. The value 1 indicates fraud and 0 indicates non-fraud.

Table 1: Description of the dataset's columns

Name of the variable	Description
step	Maps a unit of time in the real world. 1 step is 1 hour of time.
type	Indicates the type of transaction. This can be cash-in, Cash-out, Debit, Payment or Transfer
amount	amount of the transaction in local currency
nameOrig	Identifier of the customer who started the transaction
oldbalanceOrg	Initial balance of the originator before the transaction
newbalanceOrg	Originator's balance after the transaction
nameDest	Identifier of the recipient who received the transaction
oldbalanceDest	Initial balance of the recipient before the transaction
newbalanceDest	Recipient's balance after the transaction
isFraud	Indicates whether the transaction is fraudulent or not. The value 1 indicates fraud and 0 indicates non-fraud

2.1 DATA PREPROCESSING AND VISUALIZATION

Data preprocessing is a crucial step in the data analysis pipeline that involves preparing and transforming raw data into a format suitable for further analysis and modeling. It is the process of cleaning, organizing, and manipulating data to enhance its quality, consistency, and relevance. Data preprocessing plays a vital role in ensuring the accuracy, efficiency, and effectiveness of data analysis and machine learning tasks. Data standardization, also known as data normalization, is a preprocessing technique that transforms numerical data into a standardized scale. It involves rescaling the data to have zero mean and unit variance. In other words, it transforms the data distribution to have a mean of 0 and a standard deviation of 1. Data was standardized by converting all the data's columns to have the same range. Python's standard scaler function is used to accomplish this. The dataset was checked for any missing and there were no missing values. The scaled dataset was divided into training and testing datasets. 70% of the original data will be used for training and the remaining 30% for testing. Data visualization is the process of presenting data in a graphical or visual format to uncover patterns, trends, and insights that may not be apparent from raw data alone. It involves creating visual representations, such as charts, graphs, maps, or infographics, to communicate information effectively and intuitively. Data was visualized using the matplotlib library and identified patterns in the data. From figure 1, it was observed that Cash-out and Payment are the two types of fraud transactions that occur most frequently, and debit transaction is the type of fraud that occurs least frequently.

The time step variable is very important for analysis of fraud transactions. The number of transactions in each time step according to fraud status was measured to determine whether there are any particular time steps in which fraudulent transactions are more prevalent than others. Each time step is one hour from the data description. Figure 2 shows that the number of fraudulent and nonfraudulent transactions by time step is almost evenly distributed across time steps. The training of the classification models may benefit from this because it may serve as a distinguishing factor between the two categories.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

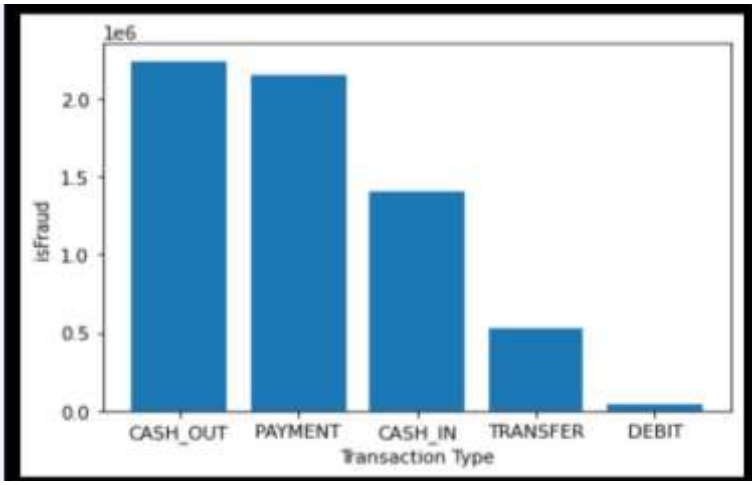

Figure 1: Frequency of transaction type that is most prone to fraud

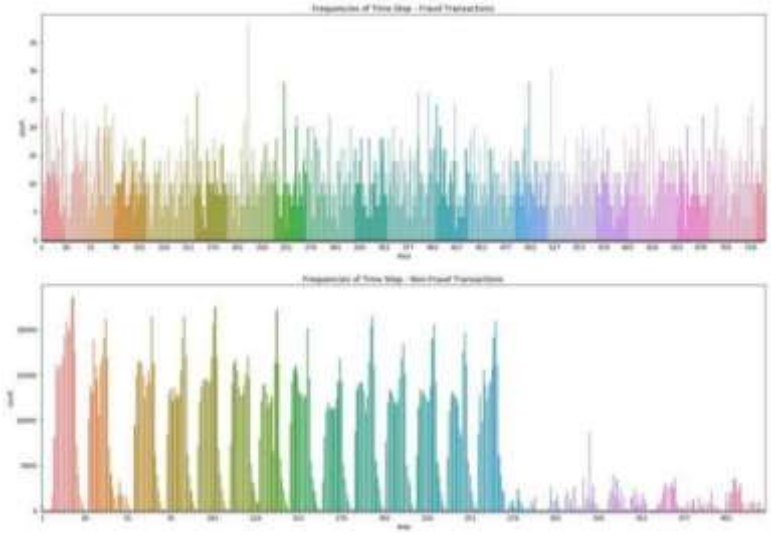

Figure 2: Frequencies of Time Step of Fraud and Non- Fraud Transactions

2.2 METHODOLOGY

This project took the usual approach to machine learning. The labeled class variable in the identified dataset served as the prediction variable in machine learning models. Firstly, data was processed, such as missing values, outliers and data standardization. Then, Logistic Regression (LR), Random Forest (RF) and Deep Learning models was implemented to select the best features after preprocessing and building models. In addition, the frequency of these indicators was counted when during extraction. When the frequency is higher than 4 times, the indicator will be chosen [2]. A comprehensive exploratory analysis was conducted of the data set and discovered potential fraud predictors and were able to distinguish between legitimate and fraudulent transactions by employing a variety of visualization tools. Figure 3 shows the steps carried out in our research. A well-structured workflow is crucial to derive meaningful insights and build effective models. Our first step is exploratory data analysis (EDA), where the raw dataset is examined to understand its properties, identify patterns, and uncover potential issues or anomalies. During this phase, summary statistics, data distributions, and correlations are analyzed to gain valuable insights. After EDA, the data cleaning and transformation process begins, involving the handling of missing values, outlier detection, and normalization or scaling of features to ensure the dataset is suitable for modeling. Once the data is cleaned and prepared, data visualization techniques were employed to visually represent the relationships and trends discovered during EDA. Visualizations help in gaining a deeper understanding of the data and can aid in feature selection for the subsequent modeling phase. At this point, the data might be split into training and testing sets to avoid overfitting during model evaluation. From our data visualization, imbalance in the dataset was detected and were able to find out which type of transactions are more prone to fraud.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

The next step involves training the machine learning and deep learning model on the prepared dataset. Logistic Regression, Random Forests and a Deep Learning model were implemented in this research and performed hyperparameter tuning on the best model to optimize the model's performance. After training, model evaluation is conducted to assess the model's accuracy and generalization ability. Evaluation metrics, such as accuracy, sensitivity, specificity, balanced accuracy as it is an imbalanced dataset, precision and F1-score were used to gauge the model's performance on the test data. Building models is an essential step in predicting the fraud or anomaly in the data sets. It can be determined how to make that prediction based on previous examples of input and output data [4]. Logistic Regression and Random Forest, both supervised machine learning methods, and a Deep Learning model was implemented to find a solution to the fraud detection issue.

Logistic regression is a statistical method used for binary classification, which means it's used to predict the probability of an event falling into one of two categories. Despite its name, logistic regression is a type of generalized linear model rather than a regression model. In logistic regression, the dependent variable or outcome variable is binary, taking on values such as 0 and 1, or "yes" and "no" and in this case, the model predicted whether a transaction is fraud or not fraud. The next model is a Random Forest model which is a popular machine learning algorithm that is used for both classification and regression tasks. It is an ensemble learning method that combines the predictions of multiple decision trees to make more accurate predictions. The Random Forest algorithm creates an ensemble of decision trees, where each tree is trained on a random subset of the training data and a random subset of the features. This randomness helps to introduce diversity among the individual trees and reduce overfitting.

The third model is a Deep Learning model which is a subfield of machine learning that focuses on training artificial neural networks with multiple layers (deep neural networks) to learn and make predictions or decisions. Deep learning models are inspired by the structure and function of the human brain, particularly its interconnected network of neurons. Models were created with cross-approval to avoid over fitting and acquire comprise of execution. The performance of the models was compared using performance measures like the Confusion Matrix and Area Under Curve (AUC). The amount transacted, the balance between the originator and recipient, and the transaction's time are all considered in this study. Some financial transactions, like credit card fraud, may not be affected by these factors that helped identify fraud. Python was used for this analysis and machine learning and deep learning models were run with the help of built-in libraries and techniques. Functions were defined when necessary to simplify detailed analyses or visualizations.

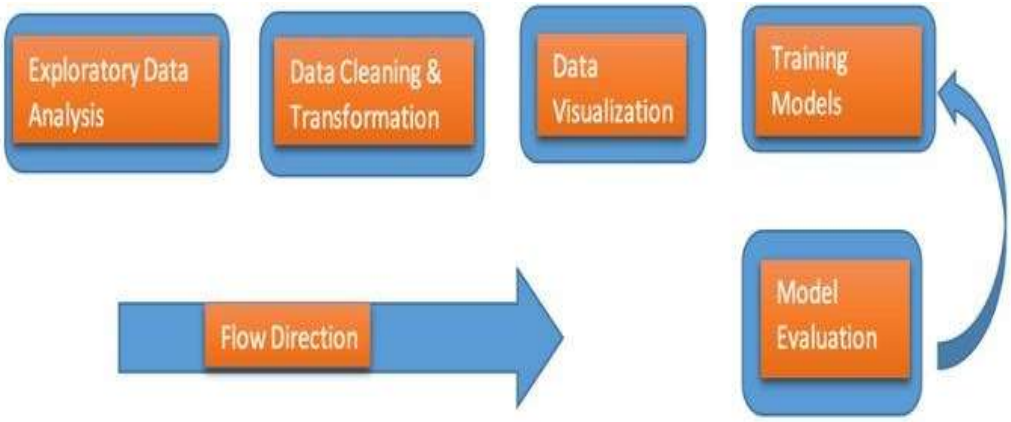

Figure 3: Project Process

III. MODELS AND METHODS

3.1 Logistic Regression

In this section, the logistic regression model is trained, and the mean recall score is calculated. Logistic regression is a statistical model that is widely used for binary classification tasks. A useful method for calculating the probability of binary classes based on one or more characteristics is logistic regression [2]. As part of a bigger fraud strategy, it can add value by evaluating the prediction ability of specific variables or combinations of variables [4]. When a transaction is active, logistic regression analyzes the values of its attributes and determines whether the transaction should move forward or not which is used for clustering [10].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

Logistic Regression model was employed from the scikit-learn library to predict financial outcomes based on a given dataset. The training datasets X train and Y train were used to train the Logistic Regression classifier, which was then used to predict results for the test dataset X test. By comparing the predicted results with the ground truth labels Y test using the accuracy score function from the sklearn metrics module, the accuracy of the Logistic Regression model was assessed which showed exceptional accuracy of 94%. Different metrics such as the true positive, true negative, false positive, and false negative values were tabulated by the confusion matrix function. Additional metrics such as sensitivity (tp / (tp + fn)), specificity (tn / (tn + fp)), balanced accuracy score, precision score, and f1_score was computed from these values to access the model's performance. Given input features *X* and parameters *w* and *b*, the logistic function calculates the probability of the positive class.

Logistic regression is a widely used statistical model for binary classification tasks, estimating the probability of binary classes based on input characteristics. It is valuable for evaluating specific variables or combinations in fraud detection and clustering. Our Logistic Regression model, implemented with scikit-learn, achieved exceptional accuracy of 94%. The model predicts the probability of the positive class using the logistic function (sigmoid), as defined in Equation 1:

The random forest models achieved an outstanding accuracy of 98% on the test dataset, highlighting their effectiveness in.

$$P(y|X) = \frac{1}{1+e^{-(\omega X+b)}} \quad \ldots\ldots \text{eq (1)}$$

P(y=1 | X): Probability of the positive class.

- X: Input features.
- ω (omega): Weights.
- b (beta): Bias term.

Training involves finding the optimal *w* and *b* that maximize the likelihood of the observed data. The model's output is then compared to a threshold (usually 0.5) to make predictions.

3.2 Random Forests

The random forest model is a powerful ensemble learning method widely used for both classification and regression tasks. It combines the concepts of decision trees and bootstrap aggregating which is also known as bagging to create a robust and accurate predictive model. At each split in the decision tree, only a random subset of predictor variables (features) is considered [7]. The algorithm typically picks a random selection of features once for training. Randomness is also incorporated into the algorithm, which helps to effectively prevent the over-fitting phenomena [9]. The two randomization processes, known as bagging and random feature selection, help reduce correlation between the trees and promote diversity in the ensemble. In this research study, the Random Forest classifier was employed from the scikit-learn library to predict outcomes on a financial dataset. The classifier was trained using the training dataset X train and Y train and subsequently used to predict outcomes for the test dataset X test. To further evaluate the performance of the Random Forest model, various metrics were calculated. The confusion matrix function provided a tabular representation of true positive, true negative, false positive, and false negative values. From these values, additional metrics were computed, including sensitivity (tp / (tp

+ fn)), specificity (tn / (tn + fp)), balanced accuracy score, precision score, and f1 score. These metrics offer insights into the model's performance in terms of correctly identifying positive and negative instances, as well as the overall balance between the classes.

Random forests, a robust ensemble learning method, combine decision trees and bootstrap aggregating to create an accurate model. To prevent overfitting, random forests introduce randomness via bagging and random feature selection. In this study, the Random Forest classifier from scikit-learn predicted financial outcomes with outstanding 98% accuracy. While avoiding overfitting, it produces a prediction for a sample x by aggregating predictions from all trees, as per Equation 2:

$$RandomForestPrediction(x) = 1/N(N@i = 1)(TreePrediction)i(x) \quad \ldots. \text{eq (2)}$$

N: Number of trees.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

The Random Forest model identifies significant features, as shown in Figure 4, which displays feature importance.

The prediction of a single decision tree for a sample x can be represented as:

$TreePrediction(x) = LeafValue_i \ if \ x \ reaches \ leaf \ i$

The final prediction of the random forest for sample x is obtained by aggregating predictions across all trees:

$RandomForestPrediction(x)$ predicting financial outcomes. The figure 4 depicts the random forest model's relative importance in terms of features. The variables that have the most significant impact on fraud prediction are depicted in the following plot.

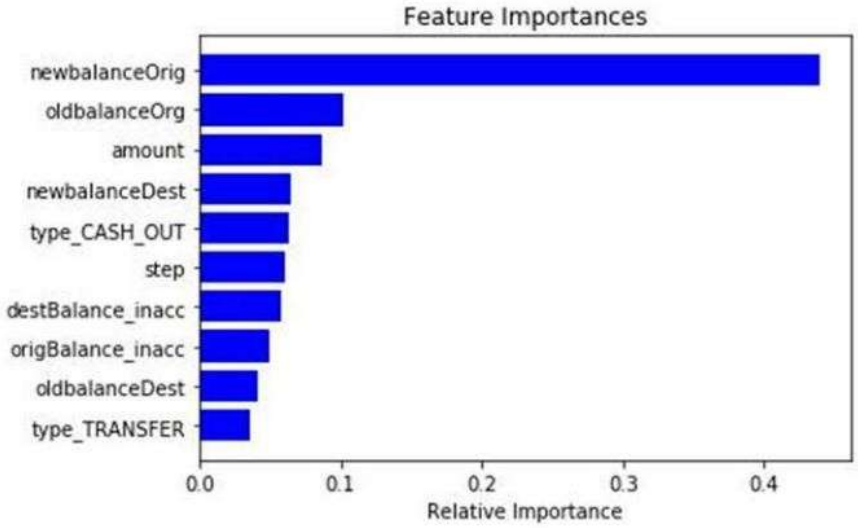

Figure 4: Random Forest Model Feature Importance

3.3 Deep Learning Model

Deep learning is a subfield of machine learning that focuses on training artificial neural networks with multiple layers, allowing them to learn complex patterns and representations from data. It is inspired by the structure and function of the human brain, particularly the interconnectedness of neurons in deep neural networks [10]. At the core of deep learning are artificial neural networks (ANNs), which consist of interconnected nodes, or "neurons," organized into layers. Each neuron performs a simple mathematical operation on its inputs and passes the result through an activation function. This study explores the utilization of deep learning models to predict financial outcomes, with a particular emphasis on achieving high accuracy. Sequential deep learning was used on our financial dataset to compare its performance with machine learning models. Sequential deep learning models refer to a class of deep learning models that are specifically designed to handle sequential data, such as time series or natural language sequences. The deep learning model was created using the Keras library. The model consists of two dense layers with 64 units and followed by an output layer with a sigmoid activation function. The model is compiled with the Adam optimizer and binary cross-entropy loss function. It is then trained on the training set for 10 epochs with a batch size of 32. The model is used to make predictions on the test set, and evaluation metrics such as accuracy was calculated. Deep learning models consist of interconnected layers. For sequential data like financial time series, a simple architecture can be as follows:

1. Input layer: Receives input data X with dimensions

$n \times m$ (n samples, m features).

2. Hidden layers: Composed of multiple neurons with activation functions, for example, $a[l] = g(W^{[l]}a^{[l-1]} + b^{[l]})$ where $a^{[l]}$ is the activation of layer l.

3. Output layer: Produces a prediction based on the task. For binary classification, a sigmoid activation is used: $y_{pred} = \sigma(W^{[l]}a^{[l-1]} + b^{[l]})$

The model's parameters ($W \ and \ b$) are learned by minimizing a loss function (e.g., binary cross-entropy) using

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

optimization methods like gradient descent.

The deep learning models offered the advantage of leveraging

$$= \frac{1}{N} \sum_{i=1}^{N} TreePrediction_i(x)$$

the temporal structure of the data, enabling them to model long- term dependencies and handle sequential information effectively. Thus, deep learning models demonstrated their efficacy in financial prediction tasks, providing valuable insights into the interplay of variables and temporal dynamics.

Deep learning, inspired by neural networks, excels in learning complex patterns from data. Sequential deep learning model was applied to the financial dataset, comparing it with machine learning models. Sequential deep learning models are tailored for sequential data, such as financial time series. The model, implemented with Keras, consists of two dense layers with 64 units and an output layer with a sigmoid activation function. The model learns parameters (W and b) by minimizing loss functions using optimization methods like gradient descent. It leverages the temporal structure of data, making it effective for financial predictions.

IV. RESULT

In this study, machine learning models and a deep learning model was implemented to identify fraud after analyzing the data on financial transactions. Data cleaning, exploratory analysis and predictive modeling were all part of the investigation. Data types were converted, checked for missing values, and summarized the data's variables during data cleaning. Exploratory analysis was conducted to investigate the class imbalance and examined each variable, particularly the type of transaction, amount, balance, and time step. It was discovered that derived variables that can assist in the detection of fraud. Various graphs were plotted to understand the data better and draw conclusions. In our research, Random Forest, Logistic Regression algorithms and a Deep Learning algorithm were implemented for predictive modeling.

In this task, accuracy was not a good metric for assessing model performance. The metric that should be maximized is recall/sensitivity, which is the ability of the model to find all the relevant cases within our dataset. Recall is the number of true positives divided by the number of true positives plus the number of false negatives, but there should be a tradeoff between recall and precision. This is because if recall is maximized, then precision is reduced, which is the ability of a model to identify only the relevant data points. In this case we wanted to find an optimal blend of precision and recall. The two metrics could be combined using F1 score, which is the harmonic mean of precision and recall. The best model with the optimum F1 Score is random forest classifier hence the best model to be used.

Table 2: Comparing the results of each algorithm

Before Balancing

Algorithm	Logistic Regression	Random Forest	Deep Learning
Accuracy	91%	93%	90%
Precision	0.83	0.92	0.88
Balanced Accuracy	0.73	0.86	0.82
F1 Score	0.61	0.81	0.78
Sensitivity/Recall	0.46	0.72	0.68

After Balancing

Algorithm	Logistic Regression	Random Forest	Deep Learning
Accuracy	94%	98%	92%
Precision	0.89	0.94	0.91

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

Balanced Accuracy	0.93	0.95	0.89
F1 Score	0.7	0.87	0.82
Sensitivity/Recall	0.89	0.97	*0.76*

After Balancing the data, it improved the performance of Logistic Regression, enhancing its ability to detect fraudulent transactions while maintaining high precision. Random Forest and Deep Learning continued to perform well, with Random Forest remaining the top-performing model. These results underscore the importance of addressing class imbalance in fraud detection, as it can substantially impact model performance. Future research should consider more advanced techniques for data balancing and feature engineering to further enhance model accuracy and fraud detection capabilities.

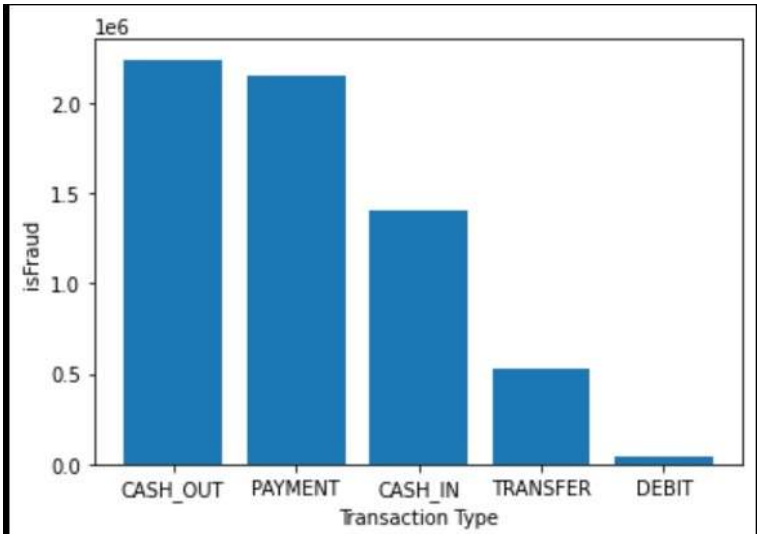

Figure 5: Transaction Type vs is Fraud

V. CONCLUSION

The framework for identifying fraudulent transactions in financial data was developed successfully. The creation of derived variables that may assist in class separation, addressing the class imbalance, and selecting the appropriate machine learning algorithm are all aspects of fraud detection that can be better understood with the help of this framework. Logistic Regression, Random Forest and Deep Learning were the three models that was implemented. Tree-based algorithms are practical for transaction data with well-differentiated classes, as evidenced by the Random Forest algorithm outperforming Logistic Regression and Deep Learning. Balanced Accuracy was also considered as it is an imbalanced dataset and Random Forest model outperforms in terms of both accuracy and balanced accuracy. This also emphasizes the value of conducting an in-depth, exploratory data analysis before creating machine learning models. Few features were identified from this exploratory analysis that distinguished the classes more effectively than the raw data. This research also distinguished the type of transactions most prone to fraud along with the implementation of machine learning in the finance domain. Future research work in the field of machine learning and data analysis holds promising avenues for further investigation and improvement. One crucial aspect is the exploration of alternative machine learning models, which can provide valuable insights into their performance and applicability in various domains. Additionally, investigating a wider range of balancing techniques, including under-sampling and oversampling methods, will help in addressing class imbalance issues more effectively. The utilization of ensemble methods like stacking or boosting is also an exciting area to explore, as these techniques can potentially enhance predictive accuracy by combining predictions from multiple models. Furthermore, delving into advanced deep learning architectures can unlock new possibilities for tackling complex data problems. Finally, it is essential to evaluate model performance using diverse metrics to gain a comprehensive understanding of their effectiveness and suitability for different applications, thus contributing to the continuous advancement of machine learning and artificial intelligence.

ACKNOWLEDGMENTS

Our sincere thanks to the experts and specialists who have helped and contributed towards development of the template.

VI. REFERENCES

[1] Maurya and A. Kumar, "Credit Card Fraud Detection System using machine learning technique," 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), 2022.

[2] Mehbodniya, I. Alam, S. Pande, R. Neware, K. P. Rane, M. Shabaz, and M. V. Madhavan, "Financial fraud detection in healthcare using machine learning and Deep Learning Techniques," Security and Communication Networks, vol. 2021, pp. 1–8, 2021.

[3] Sardeshmukh, S. Reddy, B. P. Gautham, and A. Joshi, "Bayesian networks for inverse inference in manufacturing Bayesian networks," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017.

[4] H. Shu, "Bayesian inference in Census-House Dataset," 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), 2021.

[5] Krasic and S. Celar, "Telecom fraud detection with machine learning on imbalanced dataset," 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2022.

[6] J. B, J. A. R, and D. P. Ganesh, "Credit card fraud detection with unbalanced real and synthetic dataset using Machine Learning Models," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), 2022.

[7] M. S., "Survey paper on fraud detection in Medicare using machine learning," International Journal of Psychosocial Rehabilitation, vol. 24, no. 5, pp. 4170–4174, 2020.

[8] P. Singh, V. Chauhan, S. Singh, P. Agarwal, and S. Agrawal, "Model for credit card fraud detection using machine learning algorithm," 2021 International Conference on Technological Advancements and Innovations (ICTAI), 2021.

[9] S. Dandotia and S. K. Tiwari, "Detection of credit card fraud transactions using machine learning algorithms techniques with data driven approaches: A comparative study," International Journal of Innovative Research and Growth, vol. 10, no. 11, 2021.

[10] S. khan, S. Kumar, and M. H. Kumar, "Credit card fraud detection using machine learning," International Journal of Scientific and Research Publications (IJSRP), vol. 11, no. 6, pp. 60–67, 2021.

[11] Z. Zhao and T. Bai, "Financial fraud detection and prediction in listed companies using smote and machine learning algorithms," Entropy, vol. 24, no. 8, p. 1157, 2022.