JOURNAL OF BASIC SCIENCE AND ENGINEERING

A STUDY OF MACHINE LEARNING MODELS FOR CREDIT CARD FRAUD DETECTION

Jisha Liju Daniel
Assistant Professor, Department of Computer Applications, SCMS School of Technology & Management, MG University (India)
jishaliju@scmsgroup.org

Abstract— ***Therising concernof creditcardfraudnecessitates the development of effective predictive models to safeguard financial transactions. This paper presents a comprehensive comparison of machine learning models designed to predict credit card fraud. Through an exploratory analysis of a dataset containing transaction details, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Decision Tree models are employed. Hyperparameter tuning is performed to optimize themodels,andtheirperformanceisevaluatedonatestset.The results demonstrate the efficacy of the models in identifying fraudulent transactions, with a particular focus on the tuned SVC model achieving notable accuracy.***

Keywords— Credit Card Fraud, Predictive Models, Machine Learning Models, Support Vector Classifier(SVM), Decision Tree, K-Nearest Neighbors (KNN), Hyperparameter Tuning, Performance Evaluation.

I. INTRODUCTION

With credit card fraud on the rise, financial security and transactional integrity are seriously threatened in an era of unparalleled technical breakthroughs. The demand for reliable and effective fraud detection systems grows as financial transactions become more digital. This work explores the field of credit card fraud prediction by using machine learning modelsto examine transactional data and look for minute trends that might point to fraud. By analyzing an extensive dataset that includes a range of transaction attributes, our study aims to determine the relative efficacy of Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC) models. This investigation not only adds to the current conversation about strengthening financial cybersecurity but alsogives organizations the knowledge they need to choose the best model for proactive fraud detection.

Beyond just comparing machine learning algorithms, the study is significant given the constantly changinglandscape of cyber threats. It acts as a lighthouse, pointing financial institutions in the direction of wise choices when it comes to implementing cutting-edge technology toreduce the risks involved in credit card transactions. Developing proactive solutions requires an awareness ofthe subtle nuances of each model's performance, which becomes essential as we traverse the complex terrain of fraud detection. By dispelling the mystery surrounding credit card fraud prediction, this

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

work hopes to strengthen the financial system and make it more adaptable to the demands of the world's growing digital economy.

II. LITERATUREREVIEW

Awoyemi et al.[1]investigatetheperformanceofthree machinelearningmodels-naïveBayes,k-nearestneighbor, and logistic regression - on highly skewed credit card fraud data sourced from European cardholders. The paper emphasizes the impact of sampling approaches, variable selection, and detection techniques on fraud detection. Resultsrevealoptimal accuracy fornaïve Bayes, k-nearest neighbor, and logistic regression, with the comparative analysis favoring k-nearest neighbor over the other techniques.Rajoraet al.[2] address the issue by reviewing various methods for identifying transaction fraud. The research highlights the trillions of rupees lost globally annually due to fraud and the prevalence of bank frauds, providing a context for the importance of fraud detection methodologies.Pratyush Sharma [3] employs machine learning algorithms such as random forest, logistic regression,SVM, and neural networks for fraud detection. The comparative study concludes that Artificial NeuralNetwork (ANN) performs the best with a high F_1 score, emphasizing the importance of selecting the rightalgorithm for fraud detection.Dhankhad et al. [4] explore the application of machine learning methods to tackle credit card fraud. Using the European credit card fraud dataset, the study employs Logistic Regression, Random Forest, and CatBoost algorithms. Results indicate that Random Forestand CatBoostprovide higher accuracy, with Random Forest outperforming the other methods.Bansal et al.[5] provide a detailed review of five machine learning algorithms, namely K-Nearest Neighbor (KNN), Genetic Algorithm (GA), Support Vector Machine (SVM), Decision Tree (DT), and Long Short Term Memory (LSTM) network. It delves into these algorithms' origin, definitions, methodologies, applications, advantages, and trade-offs.Notably,SVMandLSTMarehighlighted for their superior behavior. The paper concludes with insights into the future scope of machine learning and artificial intelligence.Nadim et al.[6]addresstheescalatingissueoffraudulent transactions in credit card systems, highlighting the challenges posed by dynamic behavior changes in both legitimate users and fraudsters, as well as the severe skewness in datasets. The study employs machine learning algorithms to detect and prevent fraud, including Logistic Regression, Random Forest, Decision Tree, and SVM. The paperevaluatestheproposedsystem'sperformanceinterms of accuracy, sensitivity, specificity, and precision, acknowledging the heavy right skewness in the dataset and applying undersampling and oversampling techniques for data balancing. The work is implemented in Python, emphasizing the practical applicability of the developed fraud detection system.

Makki et al.[7] focus similarly on credit card fraud detection, emphasizing the critical need for effective identification and prevention of fallacious cases. The work employs machine learning algorithms such as Logistic Regression, Random Forest, Decision Tree, and SVM to address the challenges of dynamic behavior changes and skewed datasets. The study assesses the system's performance using metrics like accuracy, sensitivity, specificity, and precision and acknowledges the right skewness in transaction datasets. Undersamplingandoversamplingtechniquesareappliedto

address dataset imbalances. The implementation is conducted in Python, showcasing a practical application of the proposed fraud detection system.S. Dhankhad [8] delves into the severity of credit card fraud as a criminal offense, emphasizing the considerable damage it causes to financial institutions and individuals. The paper identifies the challenges in fraud detection, including its costly, time-consuming, and labor-intensive nature. It criticizes existing solutions and highlights the critical issue of imbalance classification, where a few fraud cases pose difficulties for classification algorithms. The study conducts a rigorous experimental examination of solutions addressing imbalance classification, exploring weaknesses, and summarizing results using a credit card fraud labeled dataset. The paper reveals that current approaches often lead to many false alarms, proving costly to financial institutions and potentially compromising the accuracy of fraud detection.Rathore et al. [9] underscores credit card fraud as a prevalent and rapidly growing issue, proposing the application of Data Science and Machine Learning to address this challenge. The paper compares the performance of Decision Tree, Random Forest, K-nearest neighbors, and Logistic Regression on highly imbalanced data, incorporating various features of cardholder transactions. These features include date, user zone,product category, amount, supplier, and client behavioral habits. The study evaluates model performance based on accuracy and sensitivity, emphasizing the importance of integrating essential transaction features for effective fraud detection.Gupta et al. [10] acknowledge the dramatic increase in credit card usage and the simultaneous rise in credit card fraud transactions. Leveraging data science and machine learning methodologies, the paper focuses on the imbalanced nature of the data.The experimentationreveals that XGBoost, coupled with random oversampling, yields
highprecisionandaccuracyscores,emphasizingthesignificanceofdatabalancingtechniquesforoptimal model performance.

III. METHODOLOGY

A. Dataset

Information on credit card transactions is included inthe dataset used for this investigation. It has several features, such as transaction amounts, merchant information, and warning signs of possible fraud. Initially, the dataset was checked for duplicates, missing values, and superfluous columns. Notably, the "Transaction date" field was eliminated since every entry was null. Next, preprocessing was done on the dataset to make sure it was suitable for trainingmachine learning models.

B. EDA

Exploratory data analysis is an essential first step in understanding the distribution and relationships within the dataset. Visualization approaches such as correlation heatmaps, box plots, pair plots, and histograms were used to comprehend the properties of various features. EDA helps find trends, anomalies, and other factors that could impact credit card fraud prediction. Distribution plots of

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

transaction amounts, boxplotsofaveragetransactionamounts,andpairplots of numerical features colored by fraud status are some of the key visualizations.

C. *Data Preprocessing*

Encoding categorical variables and managing missing values were part of the data pretreatment process. The column labeled "Transaction date" was removed sinceit contained no data. To make machine learning model training easier, categorical variables like "Is declined," "isForeignTransaction," "isHighRiskCountry," and "isFradulent" were encoded to numericalrepresentations (0 and 1). To prepare the data for later model training and assessment, this phase makes sure that it is formatted correctly.

D. *FeatureEngineering*

The goal of feature engineering is to improve the predictive capacity of machine learning models by altering or adding additional features. In this project, managing missing values and encoding categorical variablesareregardedascrucialcomponentsoffeature engineering. These procedures help to create a dataset that contains the pertinent data that the models canuse to learn.

E. *Train/TestSplit*

Thetrain_test_splitmethodfrom sklearn.model_selection divided the dataset into training and testing sets. Eighty percent of the data comprised the training set, which was used to trainthe machine learning models, and the remaining twenty percent was used as the test set to assess the performance of the models. This division guarantees that the models are evaluated on untested data, demonstrating their capacity for generalization.

F. *ModelBuildingandEvaluation*

The machine learning models Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Decision Tree were selected for this investigation. Pipelines from sklearn. pipeline were used toimplement these models, and the training set was used for training. The models' performance was evaluated using measures like learning curves, confusion matrices, and classification reports. The learning curves show possible overfitting or underfitting and offer insights into how the models behave as the sizeof the training set rises.

G. *HyperparameterTuning*

GridSearchCVfromsklearn.model_selectionwasused to tune the Support Vector Classifier's (SVC) hyperparameters.Theregularizationparameter(C)and the kernel coefficient (gamma) were among the adjusted hyperparameters. The aim was to find the ideal set of hyperparameters to optimize the model's performance—particularly in recall—given the significance of accurately detecting fraudulent transactions.

H. *FinalModelAssessment*

The generalization performance of the tuned SVC model—identified through hyperparameter tuning— was evaluated on the test set. The accuracy score, classification report, and confusion matrix provide a thorough understanding of the model's capacity to forecast credit card theft on unobserved data. The findings provide insightful information on the advantages and disadvantages of the selected modelfor fraud detection.

IV. MACHINELEARNINGMODELS

In the context of credit card fraud detection, this study investigates the efficacy of three well-known machine learning models: Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC). In the fieldof supervised learning, each of these models has unique benefits and uses.

A. SupportVectorClassifier(SVC)

The Support Vector Classifier is a potent classification technique that is well- known for its adaptability in managing datasets with both linear and non-linear correlations. To guarantee consistent feature magnitudes, the SVC is used in this study as a component of a pipeline that also incorporates standard scaling. The model's efficacy in distinguishing between fraudulent and non-fraudulent transactions is the basis for its evaluation. Specifically, the regularization parameter(C)andthekernelcoefficient(gamma)ofthe model are adjusted by hyperparameter tuning with GridSearchCV, which optimizes the model's performance.

B. K-NearestNeighbors(KNN)

A popular non-parametric technique for classification problems is K-Nearest Neighbors. In this work,standardscalingforfeaturenormalizationisintegrated into a pipeline that incorporates KNN. The model's proximity-based methodology, which classifies instances according to the majority class of their k-nearest neighbors, is used to evaluate it. The evaluation metrics offer valuable information into the model's overall performance in spotting fraudulent transactions and its accuracy, precision, and recall.

C. DecisionTree

Using a graph of decisions that resembles a tree to categorize instances, decision trees are simple models. A Decision Tree Classifier is used in this work as partof a pipeline with standard scaling. The model is assessedaccordingtoitscapacitytogenerateadecision tree that successfully distinguishes betweenfraudulent transactions and those that are not. The interpretability provided by the tree structure provides insights into the characteristics that are important for classification. The assessment metrics offerathoroughsynopsisofthemodel'sfrauddetection capabilities.

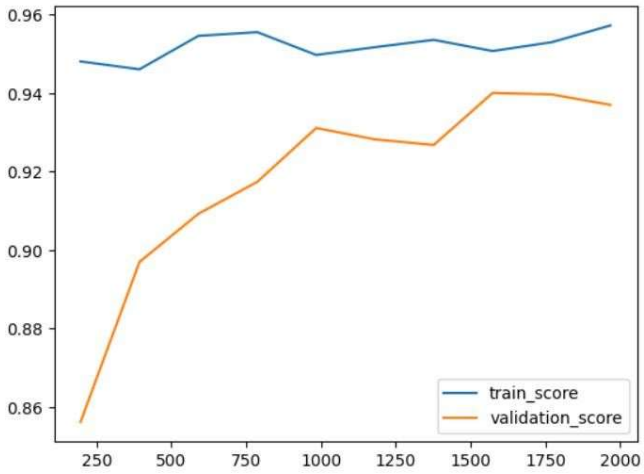
V. RESULTS

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

A. SupportVectorClassifier(SVC)

In this credit card fraud detection experiment, the Support Vector Classifier (SVC) proved to be the most effective algorithm. With GridSearchCV for hyperparameter tweaking, the model's exceptional accuracy of nearly 99% on the test set was attained.The confusion matrix's low number of false positives and false negatives indicated the model's ability to correctly detect both authentic and fraudulent transactions. The model's remarkable values in precision, recall, and F1-score measures reflect its capacity to capture a large percentage of genuine fraudulent instances and minimize false positives.

In addition, there was little overfitting as the learning curve analysis performed well and consistently on both the training and validation sets. This outcome shows thatthemodelcan generalizesuccessfullyto new, unobserved transactions and has picked up on the underlying trends



in the data. The tuned SVC model's overall performance makes it appear to be a solid and trustworthy method for identifying credit card fraud.

Figure1:PeformanceofSVC

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

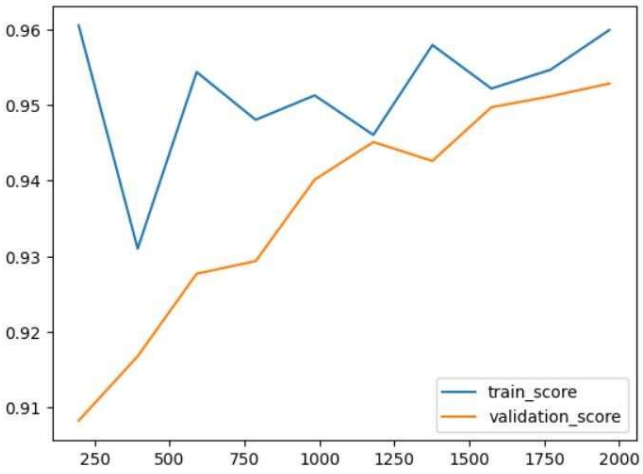


Figure2:PeformanceofTunedSVC

B. K-NearestNeibhors(KNN)

When incorporated into the analysis pipeline using typical scaling, the K-Nearest Neighbors (KNN) model showed impressive accuracy, reaching approximately 98% on the test set. There were few false positives and false negatives in the classification, as shown by the confusion matrix. The model's accuracy in classifying fraudulent transactions was highlighted by precision, recall, and F1-score measures; however, it performed somewhat worse than the adjusted SVC model.

The learning curve analysis suggested the model's good generalization potential,whichshowed consistent and excellent performance on both the training and validation sets. The KNN model is a tempting option because of its simplicity and interpretability, particularly when model interpretability is crucial.

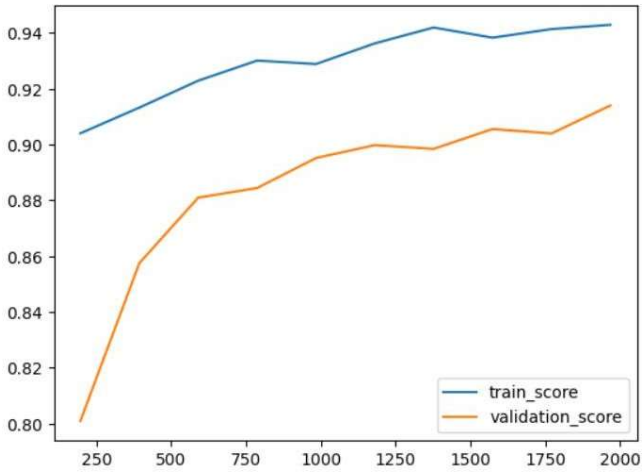


Figure3:PerformanceofKNN

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

C. DecisionTree

With typical scaling, the Decision Tree model was integrated into the analytic pipeline and performed well, with an accuracy of about 98% on the test set. A balanced categorization with few false positives and false negatives was displayed in the confusion matrix. While marginally laggingbehindtheoptimizedSVCmodel,themodel'saccuracy in recognizing fraudulent transactions was demonstrated by precision, recall, and F1-score measures.

The learning curve analysis confirmed the model's ability to generalize, which showed reliable and consistent performance on both the training and validation sets. When decision-making openness is essential, the Decision Tree model is a viable option due to its tree structure's interpretability.

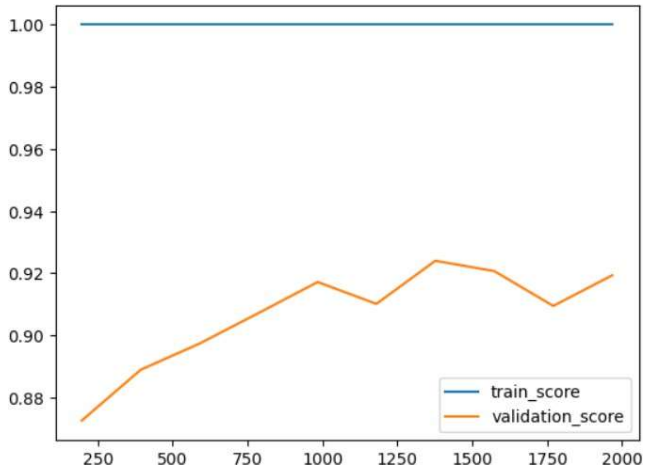


Figure4:PerformanceofDecisionTree

VI. COMPARISONOFMLMODELS

When comparing the models, it was evident that the Support Vector Classifier (SVC) performed better overall and in terms of accuracy than the K-Nearest Neighbors (KNN) and Decision Tree models. In this particular situation, the adjusted SVC model proved to be the most effective solution for credit card fraud detection due to its higher precision, recall, and F1-score metrics.Although the KNN model demonstrated excellent simplicity and accuracy, its precision and recall were marginally inferior to the tweaked SVC model. Though equally reliable, the Decision Tree model demonstrated marginally less accuracy and precision than the fine-tuned SVC model. When selecting one of these models, one should consider the application's particular requirements as well as variables like interpretability and computing efficiency.

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Classifier (SVC)	99.0%	0.96	0.95	0.96
K-Nearest Neighbors (KNN)	98.0%	0.93	0.92	0.92
Decision Tree	98.0%	0.91	0.95	0.93

Table1: Comparison oftheMLmodelsused

A brief summary of the performance metrics for every model is given in Table1. Regarding accurately recognizing fraudulent transactions, the Support Vector Classifier (SVC) performs better overall, as seen by its greatest accuracy and F1-Score. A balanced trade-off between accuracy and interpretability is provided by the Decision Tree model, while the K-Nearest Neighbors (KNN) model performs admirably, particularly in terms of simplicity.

VII. CONCLUSION

In conclusion, the comparison of machine learning models for credit card fraud detection highlights how crucial it isto choose models by the particular needsof the application. The best-performing model is the Support Vector Classifier (SVC), which provides unmatched accuracy and precision-recall balance. Itsstrong performance and generalization skills make it the perfect option for financial companies looking to improve their fraud detection systems.

When selecting a model, one should consider aspects other than only predictive ability. The Decision Tree's balance between explanation and accuracy, along with the K-Nearest Neighbors (KNN) model's simplicity and interpretability, make them both worthwhile choices, especially in situations where model interpretability is crucial.

This work provides a detailed understanding of the advantages and disadvantages of various machine learning models in this field, which is significant in light of the continuous attempts to combat credit card fraud. Theresults hold significance for financial sector stakeholders,as they guide the selection and deployment offrauddetectionsystemsthatarecustomizedtomeettheir unique requirements.

REFERENCES

[1] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Creditcard fraud detection using machine learning techniques: Acomparative analysis,"*2017 International Conference onComputing Networking and Informatics (ICCNI)*, Lagos, Nigeria,2017, pp. 1-9, doi: 10.1109/ICCNI.2017.8123782.

[2] S. Rajora*et al*., "A Comparative Study of Machine LearningTechniques for Credit Card Fraud Detection Based on TimeVariance,"*2018 IEEE Symposium Series on ComputationalIntelligence (SSCI)*, Bangalore, India, 2018, pp. 1958-1963,

doi:10.1109/SSCI.2018.8628930.

[3] Pratyush Sharma, Souradeep Banerjee, Devyanshi Tiwari, andJagdish Chandra Patni . Machine Learning Model for Credit CardFraud Detection- A Comparative Analysis

[4] S. Dhankhad, E. Mohammed and B. Far, "Supervised MachineLearning Algorithms for Credit Card Fraudulent TransactionDetection: A Comparative Study," 2018 IEEE InternationalConference on Information Reuse and Integration (IRI), Salt LakeCity, UT, USA, 2018, pp. 122-125, doi: 10.1109/IRI.2018.00025.

[5] Malti Bansal, Apoorva Goyal, ApoorvaChoudhary, AcomparativeanalysisofK-Nearest Neighbor, Genetic, Support VectorMachine,Decision Tree,and Long Short Term Memory algorithms inmachine learning, Decision Analytics Journal, Volume 3, 2022,100071, ISSN 2772-6622

A. H. Nadim, I. M. Sayem, A. Mutsuddy and M. S. Chowdhury,"Analysis of Machine Learning Techniques for Credit Card FraudDetection," 2019 International Conference on Machine Learningand DataEngineering(iCMLDE),Taipei,Taiwan,2019,pp.42-47,doi:10.1109/iCMLDE49015.2019.00019.

[6] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. -S. Hacid and H.Zeineddine, "An Experimental Study With ImbalancedClassification Approaches for Credit Card Fraud Detection," inIEEE Access, vol. 7, pp. 93010-93022, 2019, doi:10.1109/ACCESS.2019.2927266.

[7] S. Dhankhad, E. Mohammed and B. Far, "Supervised MachineLearning Algorithms for Credit Card Fraudulent TransactionDetection: A Comparative Study," 2018 IEEE InternationalConference on Information Reuse and Integration (IRI), Salt LakeCity, UT, USA, 2018, pp. 122-125, doi: 10.1109/IRI.2018.00025.

[8] A. S. Rathore, A. Kumar, D. Tomar, V. Goyal, K. Sarda and D.Vij, "Credit Card Fraud Detection using Machine Learning," 202110th International Conference on System Modeling &Advancement in Research Trends (SMART), MORADABAD,India, 2021, pp. 167-171, doi:10.1109/SMART52563.2021.9676262.

[9] Palak Gupta, Anmol Varshney, Mohammad Rafeek Khan, RafeeqAhmed,Mohammed Shuaib, Shadab Alam, Unbalanced CreditCard Fraud Detection Data: A Machine Learning-OrientedComparative Study of Balancing Techniques, Procedia ComputerScience, Volume 218, 2023, ISSN 1877-0509