Transactional Fraud Detection

Anukriti Manchanda
Maharaja Agrasen Institute of Technology,
Guru Gobind Singh Indraprastha University
Delhi, India
manchandaanukriti5@gmail.com

Anant Parashar
Maharaja Agrasen Institute of Technology,
Guru Gobind Singh Indraprastha University
Delhi, India
anantparashar005@gmail.com

Shubham Raturi
Maharaja Agrasen Institute of Technology,
Guru Gobind Singh Indraprastha University
Delhi, India
shubhamangwall@gmail.com

Guided By:
Mr. Ajay Kumar Tiwari
Assistant Professor
Maharaja Agrasen Institute of Technology,
Guru Gobind Singh Indraprastha University
Delhi, India

***Abstract*— The primary objective of data analytics is to unveil concealed patterns and leverage them to facilitate well-informed decision-making across diverse scenarios. The surge in credit card fraud, propelled by technological advancements, has rendered it a vulnerable target for fraudulent activities. This poses a significant challenge in the financial services sector, incurring substantial financial losses annually, amounting to billions of dollars.**

Developing an effective fraud detection algorithm is a complex undertaking, particularly due to the scarcity of real-world transaction datasets, attributed to confidentiality concerns and the inherent imbalance in publicly available datasets. In response to this challenge, our research addresses the issue by applying a spectrum of supervised machine learning algorithms to identify fraudulent credit card transactions, utilizing a real-world dataset.

Taking a step further, we harness these individual algorithms to construct a robust super classifier employing ensemble learning methods. Through our analysis, we discern the critical variables that contribute to heightened accuracy in the detection of fraudulent credit card transactions. This endeavor not only aids in enhancing the efficacy of fraud detection but also sheds light on pivotal factors influencing the success of such algorithms.

Moreover, our study extends beyond mere algorithmic application. We undertake a comprehensive comparative analysis, evaluating the performance of various supervised machine learning algorithms documented in the literature against the super classifier implemented in this paper.

Keywords*— *Credit Card Fraud Detection, Supervised Machine Learning, Classification, Imbalanced Dataset, Sampling Techniques, Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Dummy Classifier, Support Vector Machines, (SVM), XGBoost, LightGBM, Synthetic Dataset, Performance, Evaluation, Accuracy, Precision, F1 Score, Kappa, Recall, Feature Selection, Boruta, Hyperparameter Tuning, GridSearchCV, Confusion Matrix.

I. INTODUCTION

Today, globally, data is readily accessible, with organizations of all sizes storing information characterized by high volume, variety, speed, and significance. This data originates from diverse sources such as social media interactions, user purchases [1],[2]. It is utilized for analyzing and visualizing concealed patterns in the data [3]. Initial analyses of big data primarily focused on data volume, encompassing general public databases [4], biometrics, and financial analyses [5], [6].

In the realm of fraud, the realm of transactions proves to be a convenient and inviting target due to the substantial monetary gain achievable within a brief timeframe [7]. Perpetrators of transactional fraud aim to pilfer sensitive information, including transactional numbers, bank account details, and social security numbers [8], [6]. The endeavor to make each fraudulent transaction appear legitimate presents a formidable challenge in fraud detection [9]. The escalation in transactional dataset transactions indicates that around 70% of individuals in the US are susceptible to falling into the snares set by these fraudsters [10].

Transactional datasets often exhibit a significant imbalance, containing a higher volume of legitimate transactions compared to fraudulent ones [11]. Consequently, predictions may yield a notably high accuracy score without effectively identifying fraudulent transactions. Addressing this issue involves balancing class distribution through techniques like sampling minority classes [12]. In such sampling methods, training examples from the minority class are augmented in proportion to the majority class, enhancing the algorithm's capability to make accurate predictions [13].

This study employs seven machine learning models, evaluating their Accuracy, Recall, Precision, Kappa, and F1-Score. All machine learning algorithms undergo assessment using a synthetic dataset generated to mirror real-world scenarios and discern between fraudulent and non-fraudulent transactions. The primary objective of this study is to employ supervised learning methods on authentic datasets [14]

II. Related Study

Related study Utilizing logistic regression and artificial neural networks, the system identifies fraudulent and legitimate transactions based on their transaction scores. However, the overall performance of all machine learning models is adversely affected by the skewness present in the training dataset [15].

To address the issue of an unbalanced dataset, two distinct methods have been employed: intrinsic features and network-based features [16]. Intrinsic features involve a comparison of a customer's past transactions to identify any discernible patterns. On the other hand, network-based features leverage the connections among credit card holders and merchants, assigning a time-dependent suspiciousness score to each network object. These approaches yield a remarkably high accuracy score in Random Forest, achieving a mere 1% false positive rate, thereby creating an almost flawless model for detecting fraudulent transactions [11].

Comparative analyses were conducted across different modeling and algorithmic techniques using a real dataset, revealing that certain algorithms underperformed due to the dataset's unbalanced nature [11]. Addressing unbalanced datasets from both non-stream credit card and data streams, three distinct methods were employed: static, update, and DataStream. Additionally, two undersampling methods, namely SMOTE and Easy Ensemble, were applied to balance the dataset [17]. Notably, in Random Forest (RF) and Support Vector Machine (SVM), a decrease in the Area Under the Curve (AUC) was observed alongside an increase in F-measure [18].

The neural network architecture, employed in an unsupervised manner using real-time transaction entries [4], involves the utilization of a self-organizing map. Through optical classification, this map resolves issues associated with each specific group [19], achieving a 95% fraud detection rate with a ROC curve and without triggering false alarms.

Data Mining reports the development and implementation of a fraud detection system in a large e-tail merchant [20]. Using a cost-based performance, the algorithm is trained to obtain business outcomes, albeit requiring a longer training time [21]. A bank seller decision support system is utilized for banking fraud analysis and investigation. This system automatically detects fraud, assigns ranks, and comprehends user spending habits based on past transactions, employing mathematical and statistical techniques [22].

III. Our Approach and Methodology

1. Data Collection and Preprocessing:
- Gather relevant data: Obtain a comprehensive dataset of financial transactions, ensuring it includes information crucial for fraud detection.
- Handle missing values: Implement strategies such as imputation or removal to address missing data.
- Standardize data formats: Ensure consistency in data types and formats to facilitate subsequent analyses.

A. Exploratory Data Analysis (EDA):

- Conduct univariate, bivariate, and multivariate analyses: Explore relationships and patterns within the data to inform feature engineering.
- Identify outliers: Detect and address outliers that might skew the model's performance.

B. Feature Engineering:

- Create new features: Leverage domain knowledge to engineer features that might enhance the model's predictive capabilities.
- Transform variables: Apply transformations like log transformations to normalize skewed data.
- Use business assumptions: Integrate insights from business assumptions to guide feature creation.

C. Data Filtering:

- Remove unnecessary columns: Eliminate columns with no bearing on fraud detection, such as customer IDs or irrelevant timestamps.
- Filter rows: Exclude data points that do not align with the business problem, ensuring a focused dataset.

D. Data Preparation:

- Encode categorical variables: Convert categorical variables into numerical representations suitable for machine learning algorithms.
- Handle imbalanced data: Employ techniques like oversampling or undersampling to address class imbalances.
- Normalize or scale features: Enhance the model's performance by normalizing or scaling numerical features.

E. Feature Selection:

- Use algorithms like Boruta: Apply feature selection algorithms to identify the most relevant features for model training.
- Mitigate dimensionality: Reduce the number of features to prevent overfitting and enhance model interpretability.

F. Machine Learning Modeling:

- Select appropriate algorithms: Choose machine learning algorithms suitable for fraud detection, such as ensemble methods or anomaly detection techniques.
- Split data for training and testing: Allocate data for training and testing to evaluate the model's generalization capabilities.
- Train, validate, and test the model: Assess the model's performance through rigorous training-validation cycles and evaluate it on unseen data.

G. Hyperparameter Fine-Tuning:

- Use GridSearchCV: Systematically explore hyperparameter combinations to optimize the model's performance.
- Cross-validate results: Validate hyperparameter choices to ensure robustness across different subsets of the data.

H. Model Evaluation:

• Evaluate model metrics: Assess performance metrics like precision, recall, and F1 score to understand the model's effectiveness.

I. Continuous Improvement and Future Considerations:

• Monitor model performance: Regularly assess the model's performance using real-world data and update it as needed.
• Explore emerging technologies: Stay abreast of new technologies that could enhance fraud detection capabilities.
• Adapt to changing fraud patterns: Modify the model based on evolving fraud patterns and emerging threats.

IV. Results

Our research delved into transactional fraud detection, systematically evaluating machine learning models for their effectiveness in handling imbalanced datasets prevalent in fraud scenarios. In the initial cross-validation phase, XGBoost emerged as a standout performer, showcasing promising metrics in balanced accuracy, precision, recall, F1 score, and kappa.

Following this success, we fine-tuned the XGBoost model, resulting in exceptional performance metrics - an 88.1% balanced accuracy, 96.3% precision, 76.3% recall, and an 85.1% F1 score and kappa. This demonstrated the model's heightened ability to accurately identify and classify fraudulent transactions, suggesting its practical viability.

In validation on a test dataset, the final tuned XGBoost model excelled with a 91.5% balanced accuracy, 94.4% precision, 82.9% recall, and an 88.3% F1 score and kappa. These results underscore the model's consistent and resilient performance, highlighting its potential for real-world applications in transactional fraud detection.

Our study emphasizes the critical role of meticulous model selection and parameter tuning in developing effective fraud detection systems. XGBoost, with its adaptable ensemble learning approach, stands out as a potent tool, particularly adept at addressing challenges posed by imbalanced datasets in transactional fraud detection.

A. Discovered Insights

1) All the fraud amount is almost greater than 10.000. The values are greater than 10.000.

Fig. 1. Amount of frauds

2) The fraud transaction occurs in transfer and cash-out type. However they're almost the same value.

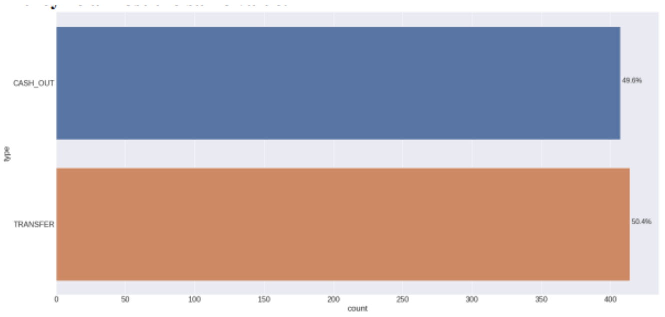


Fig. 2. Distribution Of Frauds

3) The majority transactions occurs in transfer-type, however transactions greater than 100.000 occur in cash-out and cash-in too.

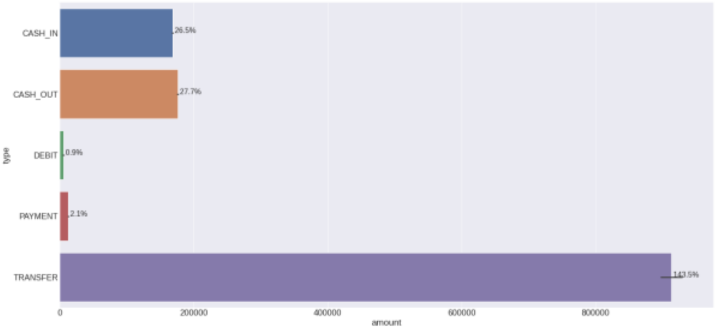
Fig. 3. Distribution Of Transactions

B. Cross Validation results

TABLE I. Dummy Model

Balanced Precision	Precision	Recall	F1	Kappa
4.99 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	-0.001 +/- 0.0

TABLE II. Logistic Regression



Balanced Precision	Precision	Recall	F1	Kappa
0.565 +/- 0.009	1.0 +/- 0.0	0.129 +/- 0.017	0.229 +/- 0.027	0.228 +/- 0.027

TABLE III. K-Nearest Neighbors

Balanced Precision	Precision	Recall	F1	Kappa
0.705 +/- 0.037	0.942 +/- 0.022	0.409 +/- 0.074	0.568 +/- 0.073	0.567 +/- 0.073

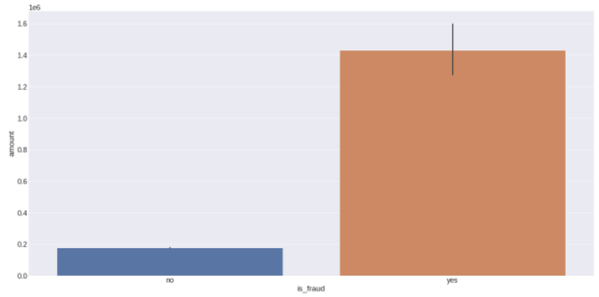
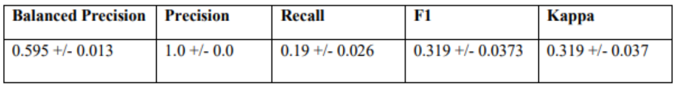
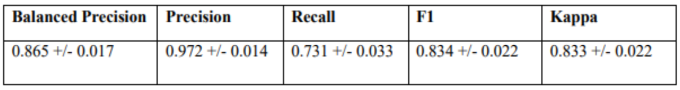


TABLE IV. SUPPORT VECTOR MACHINE (SVM)



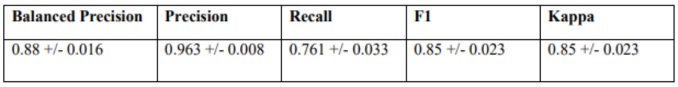
Balanced Precision	Precision	Recall	F1	Kappa
0.595 +/- 0.013	1.0 +/- 0.0	0.19 +/- 0.026	0.319 +/- 0.0373	0.319 +/- 0.037

TABLE V. RANDOM FOREST



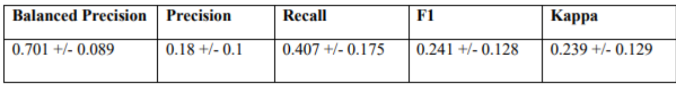
Balanced Precision	Precision	Recall	F1	Kappa
0.865 +/- 0.017	0.972 +/- 0.014	0.731 +/- 0.033	0.834 +/- 0.022	0.833 +/- 0.022

TABLE VI. XG BOOST



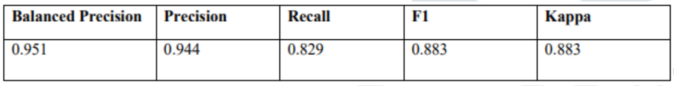
Balanced Precision	Precision	Recall	F1	Kappa
0.88 +/- 0.016	0.963 +/- 0.008	0.761 +/- 0.033	0.85 +/- 0.023	0.85 +/- 0.023

TABLE VII. LIGHT GBM



Balanced Precision	Precision	Recall	F1	Kappa
0.701 +/- 0.089	0.18 +/- 0.1	0.407 +/- 0.175	0.241 +/- 0.128	0.239 +/- 0.129

TABLE VIII. FINAL MODEL-XG BOOST WITH HYPERPARAMETER TUNING



Balanced Precision	Precision	Recall	F1	Kappa
0.951	0.944	0.829	0.883	0.883

V. CONCLUSIONS

In spearheading advancements in transactional fraud detection, this project champions a forward-thinking strategy, seamlessly integrating cutting-edge technologies while prioritizing an intuitive user experience. Departing from traditional methods, the platform unfolds as a transformative journey for institutions combating fraud. The incorporation of state-of-the-art technologies, prominently different machine learning algorithms to showcases the transformative capabilities of AI in conducting thorough analyses of transactional patterns and user behaviors, streamlining the fraud detection process effectively. At its essence, it's a modern, user-friendly interface carefully engineered for ease of use and interaction. With a lightweight and flexible backend, this interface provides a smooth user experience while positioning the system for the future. The platform's flexibility is a testament to its commitment to adapting to the ever-changing needs of institutions and the ever-changing transactional fraud landscape. Conclusively, this initiative will not only revolutionize transactional fraud detection, but also leave a lasting impression on strengthening financial security. Through a combination of technical ingenuity, user-centered design and ethical standards, the platform becomes a disruptive tool poised to redefine how institutions approach and protect against fraudulent activity. As financial entities continue to experience the benefits, the project and its impact are poised to reverberate, influencing the future of transactional fraud detection tools.

VI. REFERENCES

[1] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decis. Support Syst., vol. 50, no. 3, pp. 602–613, 2011.

[2] K. Chaudhary, J. Yadav, and B. Mallick, "A review of Fraud Detection Techniques: Credit Card," Int. J. Comput. Appl., vol. 45, no. 1, pp. 975–8887, 2012.

[3] "Mining of Massive Datasets Second Edition."

[4] F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System," vol. 6, no. 3, pp. 311–322, 2011.

[5] H. Nordberg, K. Bhatia, K. Wang, and Z.Wang, "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data," Bioinformatics, vol. 29, no. 23, pp. 3014–3019, Dec. 2013.

[6] M. Hegazy, A. Madian, and M. Ragaie, "Enhanced Fraud Miner: Credit Card Fraud Detection using Clustering Data Mining Techniques," Egypt. Comput. Sci., no. 03, pp. 72–81, 2016.

[7] M. Zareapoor and P. Shamsolmoali, "Application of credit card fraud detection:Based on bagging ensemble classifier," Procedia Comput. Sci., vol. 48, no. C, pp. 679–686, 2015.

[8] O. S. Yee, S. Sagadevan, N. Hashimah, and A. Hassain, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," vol. 10, no. 1, pp. 23–27.

[9] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," IEEE Access, vol. 6, pp. 14277–14284, 2018.

[10] N. Mahmoudi and E. Duman, "Detecting credit card fraud by Modified Fisher Discriminant Analysis," Expert Syst. Appl., vol. 42, no. 5, pp. 2510–2516, 2015.

[11] A. Dal Pozzolo, O. Caelen, Y. A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," Expert Syst. Appl., vol. 41, no. 10, pp. 4915–4928, 2014.

[12] M. A. Scholar, M. Ali, and P. Fellow, "Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets," vol. 13, no. 33, pp. 340–353, 2017.

[13] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: Adaption of different imbalance ratios," Expert Syst. Appl., vol. 98, pp. 105–117, May 2018.

[14] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification."

[15] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines," Int. Multiconference Eng. Comput. Sci., vol. I, pp. 442–447, 2011.

[16] V. Van Vlasselaer et al., "APATE: A novel approach for automated credit card transaction fraud detection using

networkbased extensions," Decis. Support Syst., vol. 75, pp. 38–48, 2015.

[17] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection," ACM SIGKDD Explor. Newsl., vol. 6, no. 1, p. 50, 2004.

[18] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decis. Support Syst., vol. 50, no. 3, pp. 559–569, 2011.

[19] C. C. Lin, A. A. Chiu, S. Y. Huang, and D. C. Yen, "Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments," Knowledge-Based Syst., vol. 89, pp. 459–470, 2015.

[20] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," Decis. Support Syst., vol. 95, pp. 91–101, 2017.

[21] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive logistic regression for credit scoring," Proc. – 2014 13th Int. Conf. Mach. Learn. Appl. ICMLA 2014, pp. 263–269, 2014.

[22] M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero, "BankSealer: A decision support system for online banking fraud analysis and investigation," Comput. Secur., vol. 53, pp. 175–186, 2015.