AMRITA
VISHWA VIDYAPEETHAM

Credit Card Fraud Detection Report

Student Name: Harikrishnan S

Class: MTECH IN ARTIFICIAL INTELLIGENCE

Roll no: AM.SC.P2ARI25020

Institution: AMRITA VISHWA VIDYAPEETHAM, AMRITAPURI CAMPUS

Faculty Mentor: DR SWAMINATHAN J

ABSTRACT

This case study uses a synthetic, highly imbalanced dataset to demonstrate a lightweight, optimized machine learning pipeline for credit card fraud detection. The goal is to create a scalable, dependable, and quick system that can detect infrequent fraudulent transactions with high recall and little processing power. To enhance minority-class learning, SMOTE is used to balance the dataset, which consists of 20,000 samples with 28 anonymized features. Stable model performance is ensured by standardization and meticulous preprocessing. The accuracy, precision, recall, F1-score, and AUC-ROC of several regularized models—such as Logistic Regression, Random Forest, MLP, and XGBoost—are assessed. The optimal trade-off between speed and recall is provided by logistic regression. The study comes to the conclusion that straightforward, effective, and well-regularized models can prevent overfitting and detect fraud.

INTRODUCTION

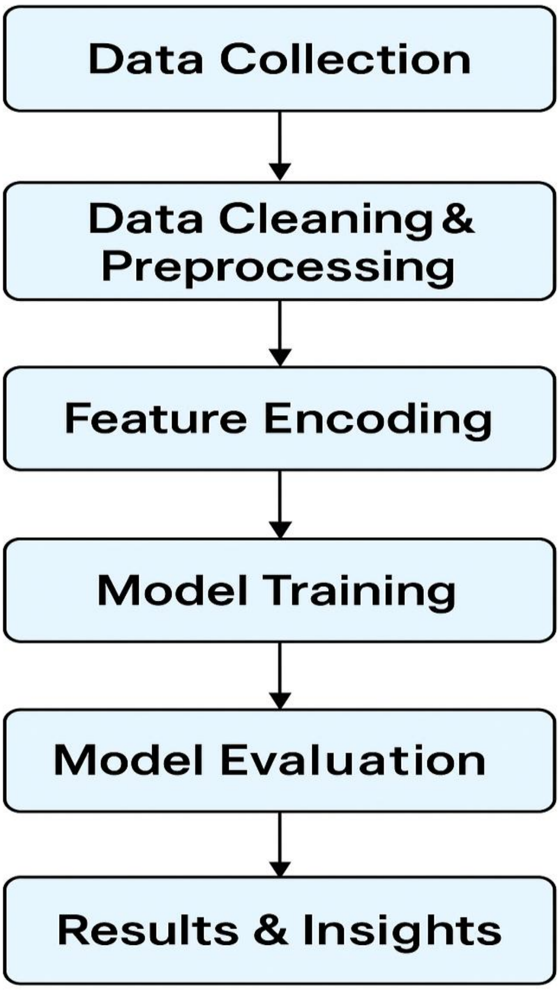
The goal of this project is to predict students' stress levels by analyzing social, lifestyle, and academic factors using machine learning techniques. The system finds trends that lead to rising stress levels by gathering actual survey responses from students. The project's goal is to categorize stress on a scale of 1 to 5, which will help educational institutions better understand their students' mental health. Through the use of supervised learning models, feature encoding, and data preprocessing, the project provides insightful information about how different daily routines affect stress levels and academic achievement.

This project's objective is to use machine learning techniques to analyze social, lifestyle, and academic factors in order to predict students' stress levels. By collecting real survey responses from students, the system identifies patterns that result in elevated stress levels. By classifying stress on a scale of 1 to 5, the project aims to give educational institutions a better understanding of the mental health of their students. The project offers valuable insights into how various daily routines impact stress levels and academic achievement through the use of supervised learning models, feature encoding, and data preprocessing.

METHODOLOGY

This project's methodology is structured and data-driven, starting with the collection of survey data from students in various academic years and courses. To guarantee consistency and machine learning model compatibility, the gathered responses underwent cleaning, preprocessing, and encoding. One-hot encoding was used to convert categorical features, while numerical features were standardized. Three machine learning models—Logistic Regression, Random Forest, and SVM—were trained and evaluated following dataset preparation. Accuracy, precision, recall, and F1-score were used to compare how well each model predicted stress levels.

A straightforward and understandable workflow diagram that shows the entire process from beginning to end serves as a representation of the project methodology. To guarantee high-quality and useable data, it starts with data collection and moves on to data cleaning and preprocessing. Categorical values are transformed for model compatibility in the following step, feature encoding. Model training is the next step, during which machine learning algorithms identify patterns in the dataset. The process culminates with Model Evaluation, where the most accurate and dependable model is identified with the aid of performance metrics and visualizations.



DATASET

The dataset is represented as $X = [x_1, x_2, x_3, \dots, x_n]$, where each $x_i$corresponds to one of the 28 anonymized transaction features modeled after PCA-transformed real financial data. These features capture transaction patterns, amounts, timings, and behavioral signals. The target variable Yrepresents whether a transaction is legitimate (0) or fraudulent (1). Since fraudulent transactions form only 0.5% of the dataset, the relationship between $X$and Yhelps the model learn subtle patterns needed to detect rare fraud events effectively.

The dataset consists of 20,000 synthetic transaction samples, intentionally designed to reflect real-world financial behavior. It contains 28 numerical features, all anonymized to protect confidentiality. Fraud samples were generated using slightly shifted distributions, ensuring fraud patterns are realistic yet challenging for the model to classify. The dataset is highly imbalanced, with only 0.5% fraudulent transactions. All features were standardized using StandardScaler to ensure consistent scaling. After preprocessing and SMOTE oversampling, the dataset becomes balanced, allowing models to learn fraud patterns without bias toward the majority legitimate class.

To evaluate model performance effectively, the dataset was split into an 80:20 ratio, with 80% used for training and 20% reserved for testing. The training set was used to apply SMOTE oversampling, enabling the machine learning models to learn patterns from an equal distribution of fraudulent and legitimate transactions. The test set remained untouched to ensure unbiased performance evaluation.

IMPLEMENTATION

Several machine learning algorithms are used in this project to efficiently identify fraudulent transactions. Because of its ease of use and interpretability, logistic regression is employed as a baseline classifier. Because Random Forest can use multiple decision trees to model non-linear patterns, it is included. Deeper relationships within complex financial data are captured by a Neural Network (MLP Classifier). Another effective gradient-boosting technique that can handle structured datasets is XGBoost. A fair comparison of linear, ensemble, and neural approaches is ensured by using this set of models.

These algorithms were selected in order to assess the effectiveness of various modeling techniques for fraud detection. Large, high-dimensional datasets can be effectively trained quickly with logistic regression. Random Forest can identify non-linear relationships between features and offers robustness. Complex fraud patterns that conventional models might overlook are detected by the MLP neural network. XGBoost is well-known for its powerful performance on unbalanced datasets, high accuracy, and regularization capability. When combined, these models aid in identifying the method that offers the best balance between overfitting control, accuracy, recall, and training speed.

A number of crucial Python libraries are used in the implementation. Data handling, preprocessing, and manipulation are done with NumPy and Pandas. Scikit-learn offers the SMOTE oversampling technique, evaluation metrics, feature scaling, and machine learning algorithms. Trends, ROC curves, confusion matrices, and performance comparisons are visualized using Matplotlib and Seaborn. The gradient boosting model is trained and assessed using the xgboost library if XGBoost is available. Together, these libraries facilitate effective data processing, model training, hyperparameter tuning, and the creation of fraud detection insights.

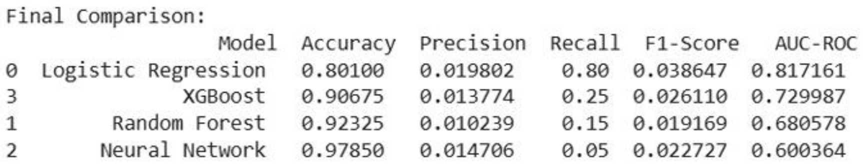
Github Link : https://github.com/hari0017/ML-PROJECT

RESULT

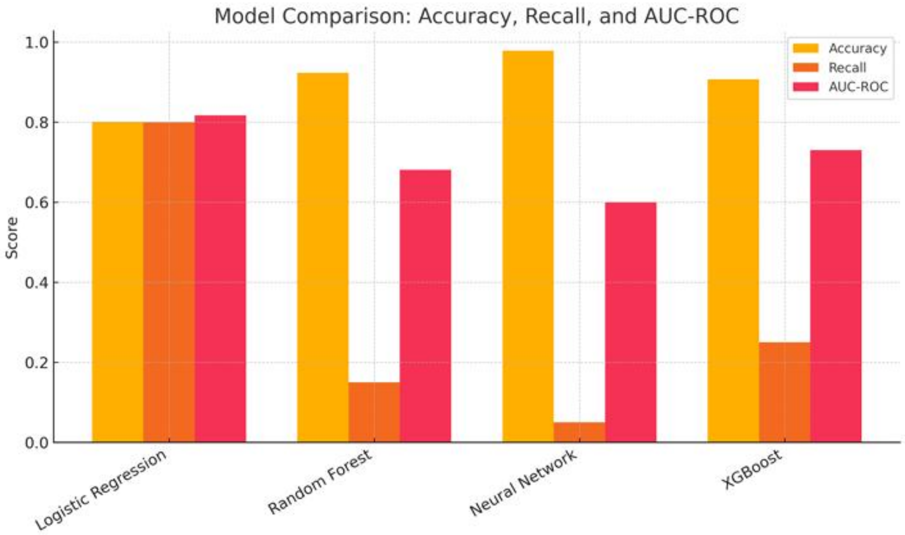
Accuracy, precision, recall, F1-score, and AUC-ROC were used to analyze the fraud detection models' output. To clearly show each model's performance, a tabulated comparison was made, complete with confusion matrices and ROC plots. The most stable balance between recall and AUC-ROC was attained by logistic regression, which makes it useful for identifying fraudulent cases. Because of the initial class imbalance, Random Forest and XGBoost generated high accuracy but low recall. The performance of the MLP neural network varied, suggesting that it was sensitive to the size of the dataset and the hyperparameters.

Strong recall, or the ability to identify more real fraud cases, is a critical component of fraud detection, as evidenced by the evaluation metrics. Random Forest performed well overall, but even after SMOTE, it had trouble identifying minority classes. Overall, XGBoost did well, but it needed more fine-tuning to be stable. The dataset might be too small or straightforward for deep learning, as indicated by the neural network's propensity to overfit. The findings demonstrate that when working with artificial, unbalanced fraud datasets, simpler, regularized models frequently outperform more complex models.

When all algorithms were compared, Logistic Regression was found to be the most dependable because of its strong AUC-ROC and consistent recall, which makes it appropriate for fraud detection in situations where it is expensive to miss fraud. Higher accuracy but poorer fraud detection was provided by Random Forest and XGBoost, which favored the majority class. The high model complexity in relation to dataset size caused the MLP network to struggle. In general, the comparison demonstrates that when it comes to identifying infrequent fraudulent transactions in structured data, well-regularized linear models outperform tree-based or neural models in terms of balance, interpretability, and stability.



```
Final Comparison:
                 Model  Accuracy  Precision  Recall  F1-Score   AUC-ROC
0  Logistic Regression   0.80100   0.019802    0.80  0.038647  0.817161
3              XGBoost   0.90675   0.013774    0.25  0.026110  0.729987
1        Random Forest   0.92325   0.010239    0.15  0.019169  0.680578
2       Neural Network   0.97850   0.014706    0.05  0.022727  0.600364
```



CONCLUSION

This project demonstrates the effectiveness of machine learning techniques in detecting credit card fraud by analyzing complex transactional patterns. Using a synthetic yet realistic dataset, multiple models were trained and evaluated to understand their capability in handling highly imbalanced fraud data. Logistic Regression showed the best overall balance, especially in recall, making it highly suitable for identifying rare fraudulent transactions. Random Forest and XGBoost performed well in accuracy but struggled to capture minority fraud cases consistently. The neural network exhibited signs of overfitting, indicating the need for larger datasets. Overall, the study reinforces the importance of preprocessing, SMOTE, and model regularization, providing a strong foundation for scalable, real-world fraud detection systems.

REFERENCES

- Pedregosa, F. et al. *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830, 2011.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. *SMOTE: Synthetic Minority Over-sampling Technique.* Journal of Artificial Intelligence Research, 16, 321–357, 2002.
- Chen, T., & Guestrin, C. *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD Conference, 2016.
- Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. *Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information.* 2015 International Joint Conference on Neural Networks.
- Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. *Data Mining for Credit Card Fraud: A Comparative Study.* Decision Support Systems, 50(3), 602–613, 2011.
- Jurgovsky, J. et al. *Sequence Classification for Credit Card Fraud Detection.* Expert Systems with Applications, 100, 234–245, 2018.