Title of Project :- Developing a Machine Learning Model to Predict Payment Fraud.

Name :- Divyanshu Yadav

Email id :- divyanshuydv0002@gmail.com

Contact Number :- +91 85888-89331

Google Drive Link :- Drive Folder

Github Link :- https://github.com/DivZyzz

YouTube Link :- Full Project Explaniation Video

Tableau Dashboard:- Tableau Dashboard Link

Abstract

Developing machine learning models for payment fraud detection in financial transactions. Dataset includes transaction type, amount, and balances. Objective: robust models to distinguish fraud. Extensive data exploration, feature engineering, and model selection. Metrics: precision, recall, accuracy. Solutions for combating fraud and safeguarding financial institutions. Future work: advanced algorithms and real-time fraud detection.

Keywords

Payment Fraud, Machine Learning, Financial Transactions, Fraud Detection, Data Exploration, Feature Engineering, Model Selection, Precision, Recall, Accuracy.

Introduction

Payment fraud poses a significant challenge for financial institutions. Machine learning models offer real-time detection and prevention. Project aims to develop an efficient fraud prediction model. Data exploration, feature engineering, model selection for accuracy. Protect customers and minimise losses. Future work includes advanced algorithms and real-time detection capabilities.

Problem Statement

Build a machine learning model to predict payment fraud by distinguishing between legitimate and fraudulent transactions using features like transaction amount, type, and accounts. The model aims for high accuracy to prevent financial losses and protect customers' assets in real-time.

Methodology

1. Explore data for distribution and missing values.
2. Perform feature engineering using one-hot encoding and remove irrelevant columns.
3. Split data into training and testing sets.
4. Train and evaluate machine learning algorithms (Logistic Regression, Decision Trees, K-Nearest Neighbours, and Random Forests) using accuracy and recall metrics.
5. Validate models using cross-validation and optimise performance through hyper-parameter tuning.
6. Select the best model for real-time payment fraud prediction.

Summary

To build the model, the data was first preprocessed, including handling missing values and performing one-hot encoding on categorical variables. Then, the dataset was split into training and testing sets, with 80% for training and 20% for testing.

Four different classification algorithms were used: Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors. Each model was trained and evaluated using accuracy and classification reports, which provided insights into the model's performance.

Additionally, cross-validation was performed to assess the models' generalisation ability. The Random Forest model demonstrated the highest recall score during cross-validation, indicating its effectiveness in identifying fraudulent transactions.

The insights gained from the model can be used by financial institutions to implement real-time fraud detection systems, minimising financial losses and protecting customers' assets. Future work may involve exploring other advanced algorithms and feature engineering techniques to further enhance the model's performance and accuracy.

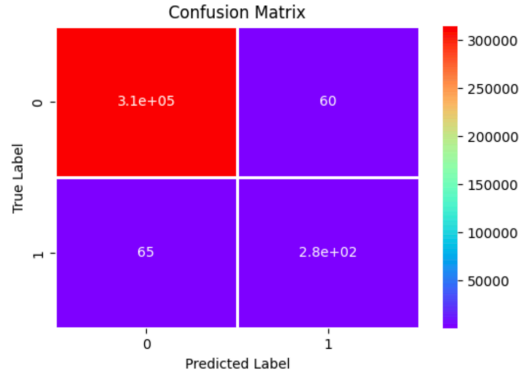
Model Evaluation

Precision Score: This means that 82% of all the things we predicted came true. that is 82% of clients transactions were detected to be fraudulent.

Recall Score: In all the actual positives, we only predicted 81% of it to be true.

For DecisionTreeClassifier, Accuracy score is 0.9996026359541347

	precision	recall	f1-score	support
0	1.00	1.00	1.00	314233
1	0.82	0.81	0.81	340
accuracy			1.00	314573
macro avg	0.91	0.90	0.91	314573
weighted avg	1.00	1.00	1.00	314573

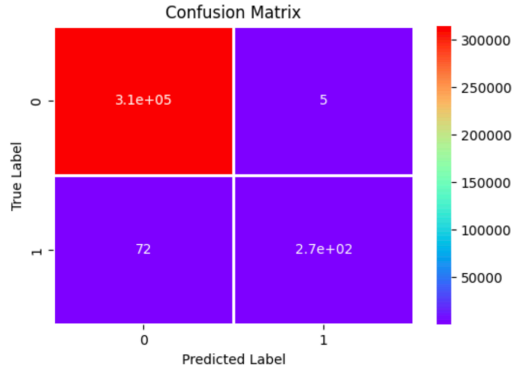


Precision Score: This means that 98% of all the things we predicted came true. that is 98% of clients transactions was detected to be a fraudulent transaction.

Recall Score: In all the actual positives, we only predicted 79% of it to be true.

For RandomForestClassifier, Accuracy score is 0.999755223747747

	precision	recall	f1-score	support
0	1.00	1.00	1.00	314233
1	0.98	0.79	0.87	340
accuracy			1.00	314573
macro avg	0.99	0.89	0.94	314573
weighted avg	1.00	1.00	1.00	314573



Result

Both the Decision Tree and Random Forest models outperform the Logistic Regression and K-Nearest Neighbors model by a wide margin. Since they both have similar recall scores, we should perform a cross-validation of the two models so we may declare which is the best performer with more certainty.

Upon training and evaluating our classification models, we found that the Random Forest and Decision Tree performed the best.

Conclusion

In conclusion, the project aimed to develop a machine learning model for predicting payment fraud in financial transactions. The dataset was explored, visualized, and preprocessed to prepare it for model training. Various machine learning algorithms, including Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors, were utilized to build the models.

The models were evaluated using accuracy and classification report metrics, and their performance was visualized using confusion matrices. Cross-validation was performed to assess the robustness of the models. The results showed that **[0.8763066972668836]** and **[0.8720556411470802]** for the Decision Tree and Random Forest models, respectively.

Future Work

1. Feature Engineering: Explore additional features or engineering techniques to enhance the model's predictive power. Domain knowledge or external data sources can be leveraged to extract more meaningful features.

2. Ensemble Methods: Investigate ensemble methods such as stacking or boosting to combine the predictions of multiple models and improve overall performance.

3. Hyperparameter Tuning: Perform a more extensive hyperparameter tuning process to optimize the model's parameters, leading to better generalization and higher accuracy.

4. Imbalanced Data: Address the issue of imbalanced data by employing techniques such as oversampling, undersampling, or using advanced algorithms like SMOTE (Synthetic Minority Over-sampling Technique).

5. Real-Time Implementation: Deploy the model in real-time to monitor transactions and detect fraud as it occurs, enabling immediate action to prevent financial losses.

6. Explainability: Investigate interpretability techniques to understand how the model makes predictions and provide explanations to stakeholders.

7. Model Updates: Continuously update and retrain the model with new data to ensure its effectiveness against evolving fraud patterns.

By focusing on these future work aspects, the payment fraud prediction model can be enhanced, providing more reliable and accurate predictions, and ultimately contributing to a safer financial ecosystem.

My Journey at Pickl.AI

During my internship at Pickl.AI, I had the opportunity to participate in a comprehensive internship course. The course covered essential topics such as data mindset, Python programming, machine learning, statistics, data visualization with Tableau, and SQL for database management. These foundational skills laid the groundwork for my data science journey.

Throughout the internship, we had weekly sessions to monitor our progress and provide feedback. These sessions were instrumental in keeping us on track and ensuring that we were grasping the concepts effectively.

Additionally, we had access to a doubt tracker, which allowed us to ask questions and seek clarification on any challenging topics. The doubt tracker was a valuable resource, enabling us to receive timely responses and deepen our understanding of complex concepts.

Overall, the internship course provided a structured and supportive learning environment, fostering both theoretical knowledge and practical skills. It played a crucial role in preparing me for the project at hand and equipped me with the tools needed to succeed in the data science field. I am grateful for the well-structured curriculum and the continuous support from the Pickl.AI team throughout my journey.