**REPORT: CREDIT CARD FRAUD DETECTION SYSTEM**

**INTRODUCTION**

Credit card fraud poses a significant global challenge, leading to considerable financial losses for both individuals and businesses. Prompt detection and prevention of illicit activities are vital for managing risk in contemporary payment systems. To address this concern, our project develops a cutting-edge real-time fraud detection system using Python and Google Cloud Platform (GCP). Our principal objectives consist of leveraging machine learning algorithms, attaining scalability, delivering top-notch performance, and furnishing meaningful visualizations via a user-friendly dashboard.

**PROBLEM STATEMENT**

Project goals aim to forecast fraudulent credit card transactions using machine learning models. Early recognition of potentially suspicious transactions benefits both the banking institution and the end consumer. Our dataset originates from Kaggle, representing transactions spanning two days in September 2013 performed by European credit cardholders. Due to extreme class imbalance, the positive class (frauds) accounts for a mere 0.172% of total transactions. Our task lies in tackling data skewness while constructing an ideal model via careful selection and comparison of various algorithms.

**RELATED WORKS**

Past research on credit card fraud detection predominantly featured conventional statistical methods and rudimentary rule-based approaches. Recent advances favor sophisticated machine learning techniques complemented by big data frameworks, enhancing efficiency and effectiveness. Significant contributions encompass anomaly detection, deep learning implementations, and multi-algorithm ensemble hybrids. Although progress has been made, opportunities exist for innovations centered on heightened interpretability, expedited prediction times, and versatile applicability across dissimilar scenarios.

**MODEL DESCRIPTION**

Our chosen model employs an XGBoost-based machine learning architecture attributed to its remarkable precision and recall relative to alternatives cited in existing literature. Further enhancement comes from integrating an ensemble strategy involving Logistic Regression, Random Forest Classifiers, and Support Vector Machines. Individually, these constituent classifiers operate collaboratively—splitting input samples based on arbitrarily picked attributes aimed at minimizing residual errors iteratively.

**PROPOSED SOLUTION APPROACH**

Tackling credit card fraud detection necessitates a systematic approach, broken down into four core components:

1. Data collection and preprocessing: Collect expansive datasets inclusive of diverse transactional records, subjecting them to vigorous cleansing routines discarding inconsistencies and missing entries.

* Acquire data from Kaggle.
* Treat missing values in respective columns.
* Analyze class distributions and observe frequency distributions.

2. Hyperparameter tuning and ensemble methodologies: Refine individual classifiers forming our ensemble model through meticulous parameter adjustments, subsequently aggregating output to yield precise predictions.

3. Visualization and monitoring dashboard: Create a user-friendly display presenting essential metrics relating to identified fraudulent occurrences supported by customizable visual summaries promoting informed decision-making.

**THEORETICAL ANALYSIS**

Presently, we have undertaken exploratory data analysis elucidating latent tendencies inherent in authentic vs. fraudulent transactions. Upcoming stages entail crafting plausible hypotheses regarding boundary conditions separating legitimate and questionable conduct, contrasted with empirical evidence gleaned from curated databanks. Following establishment of theoretically grounded bedrock, select a fitting XGBoost machine learning approach harmonious with intricacies permeating this specialized domain.

**SIMULATION RESULTS AND DISCUSSION**

- Best Model Selection on Imbalanced Data

Among tested candidates (Logistic Regression, XGBoost, Decision Tree, and Random Forest), most demonstrated commendable performance. Particularly, Logistic Regression and XGBoost excelled based on ROC-AUC scores. Given tradeoffs, however, prefer XGBoost yielding a ROC score of approximately 1.0 on training data and 0.98 on testing data. Relinquishing marginal gains offered by heavier models could translate into tangible savings for financial institutions.

- Balancing Inequality in Datasets

Recognizing severe imbalance, we explored distinct techniques for restoring equilibrium:

- Undersampling: Reduce non-fraudulent transactions to match fraudulent transaction volume

- Oversampling: Duplicate fraudulent transactions until reaching equivalent non-fraudulent transaction numbers

- SMOTE: Apply Synthetic Minority Over-sampling Technique generating synthetic examples near minority instances

- ADASYN: Employ Adaptive Synthetic Sampling approach amplifying densities of lower-density regions

Following balance restoration, reassessed candidate models. Simpler models generally prevailed, particularly Logistic Regression post-SMOTE treatment yielding desirable ROC scores nearing 0.99 on training data and 0.97 on testing data.

**COST-BENEFIT ANALYSIS**

Experimentation revealed satisfactory performance across several models considering both balanced and imbalanced data. When comparing, prioritize ease of deployment, availability of requisite hardware, software, or computational prowess necessary to execute said models. Account for subtle fluctuations in return on investment given slight variations in outcome scores.

**SUMMARY FOR BUSINESSES**

- Smaller Transaction Volumes: Prioritize elevated precision to avoid misclassifications, thereby conserving human intervention expenditure needed for verifying suspected transactions. Low precision translates to unnecessary operational burdens stemming from excessive manual checks.

- Higher Transaction Values: Focus on maximizing recall, enabling prompt identification of sizable fraudulent transactions, thus preventing substantial losses.

Based on our experiments, recommend employing the Logistic Regression model trained on SMOTE-treated data for its simplicity, intelligibility, and modest resource allocation prerequisites. Deliver outstanding returns on investment by capitalizing on strong ROC scores and notably high recall rates, ultimately protecting both banking entities and valued clients from deleterious impacts of persistent credit card fraud.

**FUTURE DIRECTIONS**

Anticipate exploring novel avenues in autoencoder-based anomaly detection together with dynamic threat neutralization guided by reinforcement learning principles. These pursuits strive to fortify defenses shielding consumers and corporations alike from detrimental ramifications accompanying escalating incidents of credit card fraud.

Data Source:<https://www.kaggle.com/mlg-ulb/creditcardfraud>