Syracuse
University

IST718: Big Data Analytics

Group -7: Saikumarreddy Pochireddygari, Sravan Kumar M, Indraneel Somayajula, Shiva Kumar

XENTE FRAUD PREDICTION

Introduction and Project Overview:

Background

In an increasingly digital world, e-commerce and financial services are at the forefront of technological advancement and customer convenience. However, these advancements also bring forth significant challenges, particularly in the realm of security and fraud detection. The need for robust, efficient, and intelligent fraud detection systems has never been more critical.

Project Objective

The primary objective of this project is to develop a sophisticated big data analytics solution for fraud detection in financial transactions. This solution is tailored for Xente, a leading e-commerce and financial services provider in Uganda. The project aims to leverage the power of big data and machine learning to identify and prevent fraudulent transactions, thereby enhancing the security and reliability of Xente's services. This project addresses this pressing need by utilizing advanced analytics to detect and prevent fraud in real-time, thereby protecting both the company and its customers from financial losses and reputational damage.

Scope of the Project

This project encompasses a comprehensive approach to fraud detection, including:

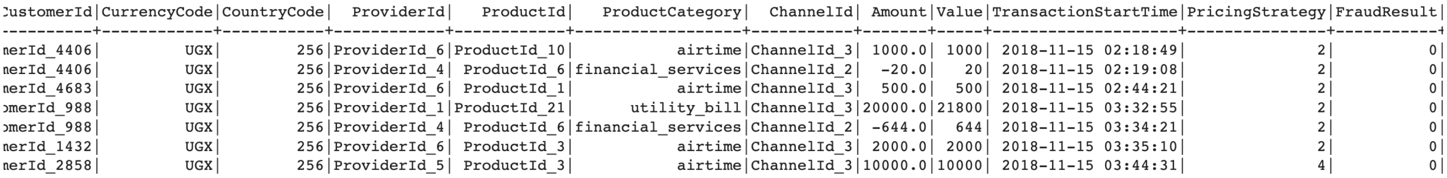
- Detailed data analysis of transactional data.
- Application of various machine learning models to identify potential fraud.
- Development of a efficient system that can accurately predict fraudulent transactions.

Using a dataset that includes a wide range of transactional attributes, such as transaction IDs, account IDs, product categories, and more, the project team employed a technology stack comprising pyspark, python, seaborn, bokeh, and matplotlib for data processing and analysis.

Data Analysis and Methodology:

Data Overview

This project utilized a comprehensive dataset from Xente, encompassing transaction IDs, account IDs, product categories, amounts, and timestamps. This dataset provided a foundational base for our fraud detection analysis.



```
----------+------------+-----------+------------+------------+------------------+-----------+-------+-----+--------------------+---------------+-----------+
CustomerId|CurrencyCode|CountryCode|  ProviderId|   ProductId|   ProductCategory|  ChannelId| Amount|Value|TransactionStartTime|PricingStrategy|FraudResult|
----------+------------+-----------+------------+------------+------------------+-----------+-------+-----+--------------------+---------------+-----------+
merId_4406|         UGX|        256|ProviderId_6|ProductId_10|           airtime|ChannelId_3| 1000.0| 1000| 2018-11-15 02:18:49|              2|          0|
merId_4406|         UGX|        256|ProviderId_4| ProductId_6|financial_services|ChannelId_2|  -20.0|   20| 2018-11-15 02:19:08|              2|          0|
merId_4683|         UGX|        256|ProviderId_6| ProductId_1|           airtime|ChannelId_3|  500.0|  500| 2018-11-15 02:44:21|              2|          0|
omerId_988|         UGX|        256|ProviderId_1|ProductId_21|      utility_bill|ChannelId_3|20000.0|21800| 2018-11-15 03:32:55|              2|          0|
omerId_988|         UGX|        256|ProviderId_4| ProductId_6|financial_services|ChannelId_2| -644.0|  644| 2018-11-15 03:34:21|              2|          0|
merId_1432|         UGX|        256|ProviderId_6| ProductId_3|           airtime|ChannelId_3| 2000.0| 2000| 2018-11-15 03:35:10|              2|          0|
merId_2858|         UGX|        256|ProviderId_5| ProductId_3|           airtime|ChannelId_3|10000.0|10000| 2018-11-15 03:44:31|              4|          0|
```

Methodology

The methodology for this project involved several key stages:

- Data Collection and Preprocessing:
  - Data was sourced from Xente's databases and underwent preprocessing to ensure quality, including cleaning and normalization.
- Exploratory Data Analysis (EDA):
  - An extensive EDA was conducted to understand transaction patterns and key variables within the dataset. Analysis was done on Debit Data and Credit Data Separately.
- Technology Stack:
  - The project leveraged PySpark for big data processing, and machine learning, Python for data analysis and visualization tools like Seaborn, Bokeh, and Matplotlib.
- Feature Selection and Engineering:
  - We focused on identifying and engineering features pivotal in detecting fraudulent transactions, such as transaction frequency and average amounts per customer Id and many more.
- Modelling and Testing:
  - Once the features were Finalized we built models and tested the model performance using F-1 Metric.

Data Analysis

- We have decided to break the data analysis in to two parts one is Credit Data Analysis and Debit Data Analysis, Also, we did compare why some transactions were fraud and not fraud in either of the analysis.
- We observed class imbalance issue challenge.

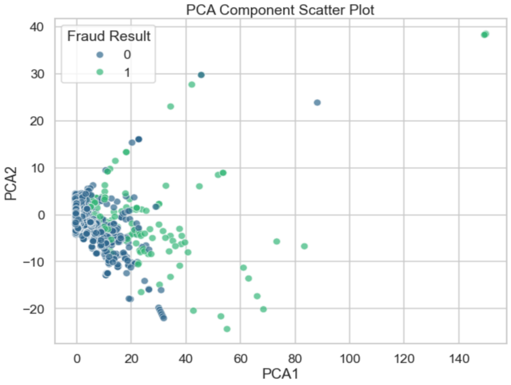
Feature Engineering and Data Preparation

Feature Engineering:-

Feature engineering was a critical step in our methodology, involving the creation of new features from the existing data that could more effectively identify potential fraud. Key features engineered included:

- **Transaction Flags:** Differentiating between weekdays and weekends and identifying debit or credit transactions.
- **Behavioral Patterns:** Analyzing spending habits, such as average transaction amounts and frequency, to identify deviations from the norm.
- **Categorical Analysis:** Evaluating transaction data based on product categories, providers, and payment channels for patterns that might indicate fraud.

We also evaluated how better these Feature engineered features were in identifying Fraud vs Non fraud just by using PCA plot shown below. We observed they were useful.



Data Preparation for Modelling:-

Data preparation was another crucial aspect of our project, ensuring the data was suitable for modeling. This stage involved:

- **Data Cleaning and Normalization:** Standardizing the data to a uniform format, dealing with missing values, and removing outliers.
- **Handling Imbalanced Data:** Our dataset was initially skewed with a higher proportion of legitimate transactions compared to fraudulent ones. We used techniques like **Synthetic Minority Over-sampling Technique (SMOTE) to address this imbalance**, which is critical for training effective machine learning models.

Model Development

Overview:-

In the core phase of our project, we focused on developing and training machine learning models capable of effectively detecting fraudulent transactions. This process involved selecting appropriate models, training them with our prepared dataset, and tuning them for optimal performance.

We utilized a variety of supervised machine learning models, each offering unique strengths in pattern recognition and anomaly detection. These included:

- **Decision Trees:** Useful for their interpretability and ease of visualization, helping in understanding the decision logic for fraud detection.
- **K-Nearest Neighbors (K-NN):** Employed for its simplicity and effectiveness in classification tasks.
- **Gradient Boosted Decision Trees (GBDT):** Chosen for their ability to handle a variety of data types and for their robustness against overfitting.
- **Random Forest:** A preferred choice due to its high accuracy and ability to run in parallel, speeding up the training process.
- **Logistic Regression:** Utilized for its efficiency in binary classification problems, particularly useful in distinguishing between fraudulent and non-fraudulent transactions.
- **Support Vector Classifier (SVC):** Implemented for its effectiveness in high-dimensional spaces and its versatility in handling various types of data.

Model Training and Evaluation:-

Each model was trained using the features engineered from our dataset. The training process involved:

- **Data Splitting:** Dividing the dataset into training and testing sets to ensure unbiased evaluation of model performance.
- **Parameter Tuning:** Using techniques like grid to find the optimal parameters for each model.
- **Performance Metrics:** Evaluating models based on metrics such as accuracy, precision, recall, and the F1 score. Special attention was given to the F1 score due to the imbalanced nature of our dataset.

Results and Analysis

1. Data Analysis Findings:-

   - Sum of Fraud transactions mean amount is greater than its counterpart Non-Fraud transactions for credit and debit data.
   - Product_15 was the most fraud-prone product in Debit Data transactions, While it was product_3 for Credit Data transactions
   - Financial Services Line had most fraud activity for Debit Data transactions, While it was Airtime category line for Credit Data
   - Pricing Strategy 4 was leading to more fraud in credit data transactions
   - Most of the fraud was happing during 12 pm – Mid Night for Debit transactions

2. Model Performance:-

Upon completion of the model training and evaluation phase, we analyzed the performance of each machine learning model based on various metrics. The key findings included:

- **Decision Trees and Random Forest:** These models demonstrated high accuracy but varied in their precision and recall, indicating a strong ability to identify fraudulent transactions but with some limitations in differentiating them from legitimate ones.
- **K-Nearest Neighbors (K-NN):** K-NN showed moderate performance, being more effective in certain scenarios but less so in others, depending on the dataset's complexity.
- **Gradient Boosted Decision Trees (GBDT):** GBDT emerged as one of the top performers, balancing accuracy with computational efficiency.
- **Logistic Regression:** This model provided a strong baseline, with good performance in terms of both speed and accuracy, making it a viable option for real-time fraud detection.
- **Support Vector Classifier (SVC):** SVC was particularly effective in high-dimensional spaces, although it required more computational resources.

3. Analysis of Results: -

The analysis revealed several insights:

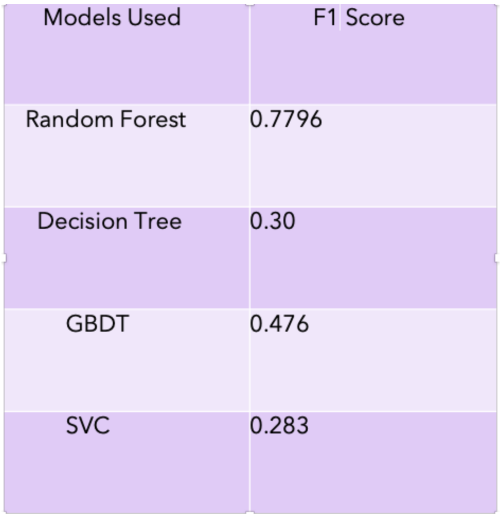
- **Strengths and Weaknesses:** Each model had its strengths and weaknesses, suggesting that a hybrid or ensemble approach might be beneficial for improving overall performance, For simplicity we decided to go with RF model.
- **Impact of Feature Engineering:** The engineered features played a significant role in model performance, underscoring the importance of thorough feature selection and engineering.
- **Model Adaptability:** Models like GBDT and Random Forest showed adaptability to the dataset, indicating their potential for application in dynamic, real-world scenarios.

4. F1 Score Emphasis:-

Given the imbalanced nature of our dataset, the F1 score, which balances precision and recall, was a critical metric. Models with the highest F1 scores were considered more effective in this context.

5. Conclusion from Results:-

The results indicated that while no single model outperformed others across all metrics, some models showed particular strengths in certain areas. This suggests the potential for using a combination of models in a real-world application to enhance overall fraud detection efficacy.



Models Used	F1 Score
Random Forest	0.7796
Decision Tree	0.30
GBDT	0.476
SVC	0.283

Recommendations and Future Work:

Based on our findings, we recommend the following enhancements to Xente's fraud detection systems:-

- **Hybrid Model Approach:** Implementing a combination of machine learning models, such as an ensemble of Random Forest and GBDT, could improve detection accuracy and robustness against evolving fraud patterns.
- **User Authentication Enhancements:** Exploring advanced authentication mechanisms, like biometric verification, to add an additional layer of security, when certain products are involved like product_15 &3.
- **Transaction Amount Vigilance**: When a customer's transaction amount exceeds their average account transaction, prompt for additional authentication to verify the transaction's legitimacy.

For future work, we propose the following areas of exploration:-

- **Deep Learning Techniques:** Investigating deep learning models, such as neural networks, for their potential to capture complex patterns and relationships in the data that traditional models might miss.
- **Geographical and Behavioral Analysis:** Expanding the analysis to include geographical data and user behavior patterns, which could uncover new fraud indicators.
- **Risk Scoring Algorithm:** Developing a risk scoring algorithm that quantifies the likelihood of a transaction being fraudulent, providing a more nuanced approach to fraud detection.
- **User Authentication Enhancements:** Exploring advanced authentication mechanisms, like biometric verification, to add an additional layer of security.

- **Cross-Platform Fraud Analysis:** Analyzing data across different platforms to identify cross-platform fraud schemes.

Conclusion

1. Summary of Findings:-

This project embarked on the ambitious task of enhancing fraud detection mechanisms for Xente, a leading e-commerce and financial service provider. Through rigorous data analysis, feature engineering, and the application of various machine learning models, we have made significant strides in identifying fraudulent transactions. Our findings revealed that no single model was universally superior; however, models like Gradient Boosted Decision Trees (GBDT) and Random Forest exhibited promising results in specific contexts.

2. Impact of the Project:-

The project's impact extends beyond the technical realm. By improving fraud detection capabilities, we contribute not only to safeguarding Xente's financial assets but also to maintaining the trust and security of its customer base. This is crucial in an era where digital transactions are prevalent, and the threat of fraud is ever-present.

3. Reflections and Learnings:-

Throughout this project, we learned the importance of nuanced data analysis and the critical role of feature engineering in predictive modeling. We also gained insights into the complexities of balancing data which could have imposed biased results in model.

4. Looking Forward:-

Looking ahead, the project lays a foundation for more advanced fraud detection systems. The recommendations and future work outlined offer a roadmap for continuous improvement and adaptation to the evolving landscape of digital fraud.

5. Final Thoughts:-

In conclusion, this project not only addressed the immediate need for effective fraud detection at Xente but also provided a blueprint for future advancements in this vital area of financial technology. The lessons learned and the methodologies developed have the potential to make a lasting impact on the field of big data analytics in fraud detection.

Citations:-

1. **Challenge:- https://zindi.africa/competitions/xente-fraud-detection-challenge**

2. **Synthetic Minority Over-sampling Technique (SMOTE):**

Citation: Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

3. **Machine Learning Models (e.g., Decision Trees, K-Nearest Neighbors, Gradient Boosted Decision Trees, Random Forest, Logistic Regression, Support Vector Classifier):**

Citation: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.