**Unlocking Insights: Fraud Detection and Customer Profiling**

Anushka A. Chougule

College of Information, University of Maryland

INST 737: Introduction to Data Science

Prof. Naeemul Hassan

December 9, 2024

**Unlocking Insights: Fraud Detection and Customer Profiling**

**Motivation**

Credit card transactions have become an integral part of the modern financial ecosystem, facilitating quick and seamless payments. However, this convenience comes with challenges: rising fraud cases and the growing need for personalized services. Fraudulent transactions incur significant financial losses and erode customer trust. Meanwhile, customer segmentation offers a way for businesses to tailor their services to meet customer needs and optimize resource allocation.

This project aims to tackle two critical challenges: (1) detecting fraudulent transactions effectively within an imbalanced dataset and (2) segmenting customers based on behavioral patterns to derive actionable business insights. By applying machine learning techniques, this project provides predictive models for fraud detection and prescriptive tools for customer segmentation.

**Problem Statement**

***Fraud Detection***. How can fraudulent transactions, which constitute less than 1% of the dataset, be effectively identified using machine learning models?

***Customer Segmentation***. How can customers be grouped meaningfully to improve business strategies, given overlapping behaviors and complex features in the dataset?

This project aims to tackle these challenges by employing machine learning techniques to achieve two main objectives:

**Fraud Detection**: Develop predictive models to identify fraudulent transactions with high accuracy, particularly in the context of an imbalanced dataset where fraud cases are rare but critical.

**Customer Segmentation**: Apply clustering methods to group customers based on their spending patterns, fraud risk, and demographics, offering actionable insights for personalized marketing and risk management strategies.

The project is significant as it integrates advanced analytics to solve real-world problems faced by financial institutions, highlighting the dual benefit of reducing financial losses and enhancing customer engagement.

The primary challenges addressed include managing the highly imbalanced dataset for fraud detection and clustering customers with overlapping transaction behaviors. These issues reflect the complexity of real-world data and the need for robust solutions.

To provide a comprehensive view of the project, the report is structured as follows:

**Introduction**: Motivation, objectives, and challenges of the project.

**Problem and Challenges**: An in-depth discussion of fraud detection and segmentation difficulties.

**Data Description**: Overview of the dataset, preprocessing, and key features.

**Methodology**: Explanation of techniques, models, and evaluation metrics used.

**Results and Analysis**: Detailed findings, including visualizations and insights.

**Discussion**: Implications, applications, and limitations of the results.

**Conclusion and Recommendations**: Key takeaways and actionable steps.

**References**: Sources and datasets used in the project.

**Research Questions**

This project was guided by the following research questions, which align with the dual objectives of fraud detection and customer segmentation. These questions were designed to address the underlying real-world challenges and ensure the analysis was focused on actionable insights.

**Fraud Detection**

1. What patterns or features are most indicative of fraudulent transactions?  
   Identifying specific transaction attributes such as amount, time, category, or merchant details that distinguish fraudulent activities from legitimate ones is critical to building reliable detection models.
2. How effectively can machine learning models handle imbalanced datasets for fraud detection?  
   Fraudulent transactions constitute less than 1% of the dataset, making it essential to evaluate the ability of models like Logistic Regression, Decision Trees, Random Forest, and XGBoost to perform well despite this imbalance.
3. What are the trade-offs between precision and recall in fraud detection models?  
   Exploring the balance between minimizing false positives and capturing a high proportion of fraudulent transactions to ensure practical applicability in financial systems.

**Customer Segmentation**

1. How can customers be grouped based on their transaction behavior and fraud risk?  
   Using clustering techniques to segment customers based on patterns like total transaction amount, average spending, fraud count, and geographic attributes.
2. What demographic or behavioral factors influence customer segmentation?  
   Understanding how variables such as age, city population, and transaction frequency contribute to the formation of meaningful customer clusters.
3. How can these customer segments be leveraged for business strategies?  
   Translating segmentation results into actionable insights for fraud prevention, personalized marketing, and operational efficiency.

**Alignment with Project Objectives**

The research questions are closely aligned with the project's overarching goals, emphasizing both fraud detection and customer segmentation. For fraud detection, the questions aim to identify critical features that signify fraudulent transactions and assess the performance of machine learning models in handling rare fraud cases within an imbalanced dataset. For customer segmentation, the focus is on uncovering meaningful patterns in customer behavior, such as spending habits and fraud risk, and leveraging these insights to inform targeted business strategies. Together, these questions provide a structured framework for achieving the project's dual objectives of enhancing fraud prevention and enabling data-driven decision-making in customer engagement.

**Methodology**

The methodology for this project was designed to address the dual objectives of fraud detection and customer segmentation. By leveraging machine learning techniques and exploratory data analysis, the approach ensures a systematic progression from raw data preprocessing to actionable insights.

**Overall Approach**

The project employed a structured pipeline comprising data preprocessing, feature engineering, model training, and evaluation. This approach allowed us to tackle both supervised learning for fraud detection and unsupervised learning for customer segmentation.

**Steps in the Methodology**

***Data Preprocessing***

**Handling Missing Values**: Missing data, particularly in features such as merchant zip codes, was analyzed and excluded when necessary to maintain data quality.

**Feature Transformation**: Temporal data (e.g., transaction timestamps, customer birthdates) were converted into meaningful variables, such as transaction hour and customer age.

**Scaling Features**: Numerical attributes were standardized using StandardScaler to ensure compatibility for clustering and machine learning models.

**Addressing Class Imbalance**: Fraudulent transactions accounted for less than 1% of the dataset. This imbalance was addressed through stratified sampling for validation and by focusing on evaluation metrics that account for imbalances (e.g., ROC-AUC, precision, recall).

***Fraud Detection***

**Algorithms Employed**

Four machine learning models were implemented to classify transactions as fraudulent or legitimate:

**Logistic Regression**: A baseline model for its simplicity and interpretability.

**Decision Tree**: To capture non-linear patterns in the data.

**Random Forest**: An ensemble model for improved precision and recall.

**XGBoost**: An advanced gradient boosting algorithm, offering high accuracy and the ability to handle imbalanced datasets effectively.

**Evaluation Metrics**

**ROC-AUC Score**: To assess the model’s ability to distinguish between fraudulent and legitimate transactions.

**Precision and Recall**: To evaluate performance in scenarios with rare fraudulent cases.

**Confusion Matrix**: To analyze classification errors and fine-tune model thresholds.

***Customer Segmentation***

**Clustering Technique**

K-Means clustering was employed to group customers based on their transaction patterns, spending habits, and fraud risk. This algorithm was selected for its simplicity and effectiveness in identifying distinct clusters.

**Feature Engineering**

Aggregated transaction data at the customer level, creating features such as:

**Total Spending (total\_amt)**

**Average Transaction Amount (avg\_amt)**

**Fraud Count (fraud\_count)**

**Demographic Factors**: Age, geographic attributes, etc.

**Determining the Optimal Number of Clusters**

The **Elbow Method** was used to identify the ideal number of clusters by analyzing the inertia curve.

**Cluster Validation Metrics**

**Silhouette Score**: To measure intra-cluster cohesion and inter-cluster separation.

**Davies-Bouldin Index**: To evaluate cluster compactness and separation.

***Techniques and Algorithms Summary***

**Supervised Learning for Fraud Detection**: Focused on classification models, addressing imbalanced data and prioritizing critical evaluation metrics.

**Unsupervised Learning for Customer Segmentation**: Leveraged clustering techniques to group customers into actionable segments.

**Evaluation and Visualization**: Comprehensive visualizations (e.g., heatmaps, scatter plots) were used to interpret results and communicate findings effectively.

**Data Preparation**

**Collection**

The dataset used in this project, titled "Credit Card Transactions Dataset," was sourced from Kaggle, an open platform for data sharing and machine learning challenges. It contains over 1.3 million transaction records, making it a robust resource for analyzing customer behavior and identifying fraudulent activities. Each record provides transaction-level details, including customer demographics, transaction amounts, timestamps, and merchant details. The dataset is pre-labeled, offering a clear binary classification of transactions as fraudulent (1) or legitimate (0).

**Curation**

The dataset underwent significant preprocessing to prepare it for analysis

**Data Transformation:**

Temporal features, such as trans\_date\_trans\_time and dob (date of birth), were converted into datetime formats to derive meaningful attributes, such as customer age and transaction hour.

Numerical features were scaled using StandardScaler to ensure consistency,particularly for clustering analysis.

**Feature Engineering:**

Aggregated attributes, such as total transaction amount (total\_amt) and fraud count (fraud\_count), were created to facilitate customer segmentation.

Ratios, such as amt\_to\_avg\_transaction\_amt\_ratio, were derived to identify outliers or unusual transaction patterns.

**Dataset Organization:**

For fraud detection, the dataset was retained in its transactional format.

For customer segmentation, data was aggregated at the customer level, ensuring the focus was on long-term spending habits and behaviors.

**Cleaning**

To ensure data quality and integrity, the following steps were implemented:

***Handling Missing Values***

Columns such as merch\_zipcode contained missing values. These were either excluded from the analysis or imputed based on the context.

***Duplicate Records***

Duplicate entries were identified and removed to avoid bias or redundancy in the analysis.

***Outlier Detection***

Extreme values in variables like amt (transaction amount) were reviewed and, where necessary, addressed to prevent skewed results.

***Class Imbalance***

The is\_fraud variable showed significant imbalance, with fraudulent transactions constituting less than 1% of the dataset. This was mitigated during the evaluation stage using metrics like ROC-AUC and precision-recall curves.

**Annotation**

The dataset came pre-labeled with a binary is\_fraud column, identifying transactions as either fraudulent (1) or legitimate (0). This labeling was essential for training supervised learning models in the fraud detection task. Additionally:

Categorical variables, such as category (transaction type) and gender, were encoded where necessary for model compatibility.

No additional manual labeling or annotation was required, as the dataset was already curated for machine learning purposes.

**Summary of Variables**

The dataset includes a mix of transactional, demographic, and derived variables. Key features are summarized below:

***Transactional Details***

* **trans\_date\_trans\_time**: Transaction timestamp.
* **amt**: Transaction amount.
* **category**: Transaction category (e.g., grocery, travel).
* **merchant**: Merchant details.

***Customer Information***

* **cc\_num**: Credit card number (anonymized).
* **gender**: Customer gender.
* **dob**: Customer date of birth.
* **city, state**: Geographic details of the customer.

***Fraud Label***

* **is\_fraud**: Binary label indicating whether the transaction is fraudulent (1) or legitimate (0).

***Derived Features***

* **Temporal Attributes**: transaction\_hour, transaction\_day, transaction\_month, and age.
* **Aggregates**:
  + **avg\_amt**: Average transaction amount per customer.
  + **total\_amt**: Total transaction amount per customer.
  + **transaction\_count**: Number of transactions per customer.
  + **fraud\_count**: Total number of fraudulent transactions per customer.
* **Ratios**: amt\_to\_avg\_transaction\_amt\_ratio and amt\_to\_median\_transaction\_amt\_ratio.

**Solution Framework**

To address the dual objectives of fraud detection and customer segmentation, a structured and methodical pipeline was designed. This framework integrates data preprocessing, feature engineering, model development, evaluation, and actionable insights, ensuring a comprehensive solution to the problems.

**High Level Framework**

The project framework follows the steps outlined below, ensuring a logical flow from data preparation to result interpretation:

**Data Preprocessing:**

Load the raw dataset and clean it by addressing missing values, removing duplicates, and handling outliers.

Transform temporal features (trans\_date\_trans\_time, dob) to generate meaningful attributes like transaction hour and age.

Scale numerical features using StandardScaler to ensure compatibility across clustering and supervised learning models.

**Exploratory Data Analysis (EDA)**

Perform EDA to uncover patterns in the dataset, including transaction trends, fraud risk by category, and geographic spending behavior.

Generate visualizations like correlation heatmaps, bar plots, and scatter plots to aid in data understanding.

**Feature Engineering:**

***Derive new attributes:***

Aggregates: total\_amt, fraud\_count, and transaction\_count.

Ratios: amt\_to\_avg\_transaction\_amt\_ratio to identify anomalies in spending.

***Prepare separate datasets:***

Fraud Detection: Transaction-level data to train supervised models.

Customer Segmentation: Customer-level aggregated data for clustering.

**Fraud Detection:**

***Machine Learning Models***

Train and evaluate models like Logistic Regression, Decision Tree, Random Forest, and XGBoost.

***Address Class Imbalance:***

Evaluate models using metrics such as ROC-AUC, precision, and recall ensuring performance despite the rare occurrence of fraudulent transactions.

***Feature Importance:***

Identify the most significant predictors of fraud, such as transaction amount, transaction category, and time of transaction.

**Customer Segmentation:**

***Clustering***

Apply K-Means clustering to segment customers based on their behavior, including spending habits, fraud risk, and demographic factors.

***Cluster Evaluation***

Determine the optimal number of clusters using the Elbow Method and validate using Silhouette Score and Davies-Bouldin Index.

***Visualization***

Create scatter plots and heatmaps to interpret and present cluster characteristics.

**Model Evaluation**

Validate fraud detection models using cross-validation and a range of metrics to ensure robustness.

Evaluate clustering quality using intra-cluster cohesion and inter-cluster separation metrics.

**Insights and Recommendations**

Extract actionable insights, such as identifying high-value customers and high-risk groups for fraud prevention.

Propose strategies for fraud detection and personalized marketing.

**Pipeline Diagram**

A diagram of a data analysis process

Description automatically generated

Figure 1: High-Level Solution Framework

**Flowchart Explanation**

**Data Input**

The pipeline begins with data ingestion, where the raw dataset is loaded and preprocessed to clean and transform it into a usable format. This step involves handling missing values, detecting outliers, and standardizing numerical features to ensure compatibility across different analyses.

**EDA and Feature Engineering**

Exploratory Data Analysis (EDA) uncovers key patterns and relationships in the dataset, such as correlations between transaction features and fraud occurrence. Feature engineering follows, enriching the dataset with derived attributes such as age, transaction hour, transaction ratios, and aggregate spending metrics to enhance analysis and model performance.

**Task-Specific Analysis**

***Fraud Detection***

Supervised learning models, including Logistic Regression, Decision Trees, Random Forest, and XGBoost, are implemented to identify fraudulent transactions. Class imbalance is addressed using appropriate metrics, ensuring the models effectively handle rare fraud cases.

***Customer Segmentation***

Unsupervised learning, specifically K-Means clustering, groups customers into actionable clusters based on transaction patterns, fraud risk, and demographic factors. Validation metrics like Silhouette Score and Davies-Bouldin Index assess the quality of the clusters.

**Insights and Recommendations**

The results from fraud detection and customer segmentation are synthesized into actionable strategies. Fraud detection insights guide the development of robust fraud prevention measures, while customer segmentation enables personalized marketing and resource allocation. The findings provide a holistic understanding of customer behavior and transaction risks, driving informed decision-making.

**Experiment Results**

**Experiment Setup**

To evaluate the effectiveness of machine learning techniques for fraud detection and customer segmentation, the following steps were undertaken:

***Dataset Splitting***

The dataset was split into 70% training and 30% testing sets to ensure robust evaluation of the models.

For fraud detection, stratified sampling was used to preserve the proportional representation of fraudulent and non-fraudulent transactions, addressing the class imbalance.

***Parameter Settings***

**Fraud Detection**

Four models were implemented: Logistic Regression, Decision Tree, Random Forest, and XGBoost.

Hyperparameter tuning was performed for Random Forest and XGBoost using grid search to optimize parameters like the number of estimators and maximum depth.

**Customer Segmentation:**

K-Means clustering was used for unsupervised learning. The optimal number of clusters was determined to be 5 based on the Elbow Method.

***Evaluation Metrics***

**Fraud Detection**

ROC-AUC was the primary metric to evaluate model performance.

Additional metrics such as precision, recall, F1-score, and confusion matrix analysis provided deeper insights into model accuracy and sensitivity.

**Customer Segmentation:**

The Silhouette Score and Davies-Bouldin Index were calculated to assess intra-cluster cohesion and inter-cluster separation, indicating the quality of the clusters.

**Results**

***Fraud Detection***

**Model Performance:**

Logistic Regression achieved a mean ROC-AUC score of 0.861, serving as a reliable baseline.

XGBoost outperformed other models, with an ROC-AUC score of 0.9971, demonstrating exceptional accuracy in distinguishing fraudulent transactions.

Random Forest achieved strong performance with a good balance of precision and recall, making it an effective model for fraud detection.A graph of different colored bars

Description automatically generated with medium confidence

Figure 2: Performance Metrics Comparison Across Models

A graph of a bar chart

Description automatically generated with medium confidence

Figure 3: ROC-AUC Score Comparison Across Models

**Feature Importance:**

Transaction amount (amt) was the most significant predictor of fraud, followed by transaction time (transaction\_hour) and category.A graph with blue and white text

Description automatically generated

Figure 4: Feature Importance in Fraud Detection (XGBoost)

**Visualization:**

A feature importance plot highlighted the relative contribution of features to the models.

The ROC curve for XGBoost underscored its high performance, with a near-perfect area under the curve.

***Customer Segmentation***

**Clustering Results:**

The Silhouette Score of 0.293 indicated moderate intra-cluster cohesion and inter-cluster separation.

The Davies-Bouldin Index of 1.07 reflected some overlap between clusters, influenced by the complexity and variability in customer behaviors.

**Cluster Insights:**

**Cluster 0**: Customers with high transaction amounts and moderate fraud counts, representing high-value, low-risk individuals ideal for retention strategies.

**Cluster 1**: Medium-to-high spending customers with the highest fraud counts, indicating a need for focused fraud prevention measures.

**Cluster 2**: Low spenders with relatively high fraud counts, emphasizing the importance of strict fraud detection in these cases.

**Cluster 3**: Low-risk, low-value customers with minimal spending and fraud, suitable for light engagement or upselling opportunities.

**Cluster 4**: Moderate spenders with balanced fraud risk, presenting a mix of marketing potential and fraud detection needs.

A chart of different colored circles

Description automatically generated

Figure 5: Customer Segments Based on Spending and Fraud Risk

**Visualization:**

Scatter plots demonstrated the clustering results based on features like total transaction amount and fraud count.

Heatmaps provided a comparative view of key metrics across clusters.

**Findings**

***Fraud Detection***

XGBoost emerged as the most effective model, offering near-perfect predictions, even with a highly imbalanced dataset.

The insights from feature importance underscored the need to monitor high-value transactions and specific transaction categories for fraud detection.

***Customer Segmentation***

The segmentation successfully grouped customers based on spending patterns, geographic factors, and fraud risk, providing actionable insights for marketing and fraud prevention.

High-value clusters highlighted opportunities for loyalty programs, while high-risk clusters indicated areas for targeted fraud detection.

**Limitations**

***Class Imbalance***

Fraudulent transactions constituted less than 1% of the dataset, posing significant challenges in model training. Although techniques like stratified sampling and performance evaluation using precision-recall curves were employed, additional methods such as Synthetic Minority Oversampling Technique (SMOTE) could further improve model robustness by creating synthetic samples for the minority class, enhancing model training on fraudulent transactions.

***Clustering Quality:***

The moderate Silhouette Score and Davies-Bouldin Index indicate some overlap between clusters, likely due to limited differentiation in customer behaviors. Incorporating external datasets, such as regional fraud trends or economic indicators, could provide richer contextual features to improve cluster separation and quality.

***Feature Engineering***

While derived features like transaction ratios and aggregated spending metrics added value, further exploration of advanced feature extraction techniques, such as embedding-based representations, could enhance the predictive power of the models.

***Real-Time Applications***

The models were designed and evaluated in a static environment, limiting their immediate applicability for real-time fraud detection. Future work could focus on integrating these models into dynamic systems with real-time data pipelines, enabling continuous monitoring and detection.

**Discussion**

**Interpretation of Results in the Context of Research Questions and Goals**

The results of this project provide compelling evidence for the efficacy of machine learning techniques in addressing the dual challenges of fraud detection and customer segmentation. In the context of the research questions, the findings validate the hypothesis that certain transaction patterns, demographic features, and spending behaviors are strong indicatorsof fraud and customer segmentation

**Fraud Detection**

The XGBoost model, with its high ROC-AUC score of 0.9971, demonstrated exceptional predictive accuracy in identifying fraudulent transactions. This highlights the reliability of machine learning models in extracting meaningful patterns from imbalanced datasets, directly answering the research question about the accuracy of predictive models.

Key features, such as transaction amount, category, and transaction hour, were identified as significant predictors, emphasizing the role of transaction-level data in fraud detection.

The ability to detect fraud in a highly imbalanced dataset validates the robustness of the models and aligns with the project’s goal of minimizing financial losses for businesses.

**Customer Segmentation**

The clustering analysis revealed actionable insights into customer behavior, with distinct segments reflecting varying levels of spending and fraud risk.

High-value, low-risk clusters (e.g., Cluster 0) represent opportunities for loyalty programs, while high-risk, high-fraud clusters (e.g., Cluster 2) underscore the need for targeted fraud prevention strategies.

The moderate clustering quality metrics, including a Silhouette Score of 0.293, highlight the complexity of customer behaviors but demonstrate the potential for meaningful segmentation.

**Practical Significance and Applications**

The findings from this project have significant real-world implications for businesses, particularly in the financial sector. The practical applications include:

**Fraud Prevention:**

By implementing high-performing models like XGBoost, businesses can proactively flag suspicious transactions, reducing financial losses and protecting customer trust.

The insights into significant fraud predictors can guide the design of transaction monitoring systems, focusing on high-value transactions, specific time windows, and transaction categories prone to fraud.

**Customer Relationship Management**

Understanding customer segments allows businesses to tailor marketing strategies. High-value customers in low-risk clusters can be targeted with personalized offers and loyalty programs, enhancing retention and engagement.

For low-value customers in low-risk clusters, upselling campaigns can be designed to encourage higher spending, optimizing revenue potential.

**Resource Optimization**

Fraud detection and monitoring efforts can be prioritized for high-risk clusters, ensuring efficient allocation of resources.

Marketing investments can be focused on high-value clusters, maximizing return on investment.

**Broader Implications for the Field**

This project demonstrates the power of machine learning in addressing complex financial challenges, such as fraud detection and customer segmentation, and offers a roadmap for its application across industries. The broader implications include:

**Scalability Across Industries:**

The methodologies employed in this project can be adapted for other domains, such as insurance, healthcare, and e-commerce, where fraud detection and customer segmentation are equally critical.

**Addressing Real-World Challenges:**

The project highlights the importance of handling class imbalance and clustering complexities, common challenges in real-world datasets. Advanced techniques, such as ensemble methods and hybrid clustering algorithms, could further improve outcomes.

**Future Directions:**

Incorporating external data sources, such as economic indicators or regional fraud trends, could enhance the predictive power and relevance of the models.

Real-time model deployment could significantly improve operational efficiency, enabling dynamic fraud detection and segmentation strategies.

**Conclusion**

This project effectively addressed two key challenges in the financial sector: fraud detection and customer segmentation. Using machine learning, models like XGBoost and Random Forest demonstrated high accuracy in identifying fraudulent transactions, highlighting critical features such as transaction amount and category. These insights enable proactive fraud prevention while balancing operational efficiency and customer trust.

K-Means clustering successfully segmented customers based on spending behavior and fraud risk, offering actionable strategies for personalized marketing and fraud mitigation. High-value customers can be targeted with loyalty programs, while high-risk segments require enhanced monitoring.

While challenges such as class imbalance and moderate cluster separation were identified, they present opportunities for improvement through advanced techniques and external data integration. This project underscores the value of data analytics in reducing financial risks and enhancing customer engagement, providing a foundation for future advancements across industries.

**References**

Choksi, P. (2023). *Credit Card Transactions Dataset*. Retrieved from Kaggle: <https://www.kaggle.com/datasets/priyamchoksi/credit-card-transactions-dataset>

Bachmann, J. (n.d.). *Credit fraud: Dealing with imbalanced datasets* [Notebook]. Kaggle. July 3, 2019, from <https://www.kaggle.com/code/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets>

GeeksforGeeks. (n.d.). *How to handle imbalanced classes in machine learning?* Retrieved from <https://www.geeksforgeeks.org/how-to-handle-imbalanced-classes-in-machine-learning/>

GeeksforGeeks. (n.d.). *XGBoost.* Retrieved from <https://www.geeksforgeeks.org/xgboost/>

GeeksforGeeks. (n.d.). *K-Means Clustering – Introduction.* Retrieved from [https://www.geeksforgeeks.org/k-means-clustering-introduction/\](https://www.geeksforgeeks.org/k-means-clustering-introduction/%5C)

Hassan, N. (2024). *Course Materials for INST 737: Introduction to Data Science*. Department of Information Studies, University of Maryland. Unpublished manuscript.