Analysis of Loan Application: Fraud Detection

Name: Sai Harshit Maddila

College: Southern Alberta Institute of Technology

Date: June 18, 2025

Table of Content

Abstract

This study analyses loan application data to detect fraudulent activities, aiming to identify key patterns and risk factors associated with financial fraud. Leveraging a comprehensive dataset from Kaggle, the research focuses on understanding the demographics, loan characteristics, and behavioural indicators that correlate with fraud likelihood. A Power BI dashboard has been developed to visualize critical metrics such as Total Fraud Cases, Fraud Loss Rate, and Fraud Detection Rate, along with detailed breakdowns by fraud type, location, and credit score. The findings provide actionable insights for financial institutions to enhance their fraud detection capabilities, mitigate financial losses, and implement targeted preventative measures, ultimately strengthening the integrity of loan application processes.

Introduction

Loan fraud poses a significant threat to financial institutions worldwide, leading to substantial monetary losses and eroding public trust. As the volume of loan applications continues to grow, so does the sophistication of fraudulent schemes, making advanced detection mechanisms crucial. This study delves into a comprehensive loan application dataset to identify, analyse, and visualize the underlying patterns and indicators of fraud. By understanding the key drivers and characteristics of fraudulent applications, financial institutions can proactively develop robust strategies to minimize risk and protect their assets. This article outlines the data analysis process, highlights key findings through various visualizations, and presents an executive dashboard designed to provide actionable insights for effective fraud detection and prevention.

Dataset Overview

The dataset used for this analysis is sourced from Kaggle: "Loan Application and Transaction Fraud Detection"

https://www.kaggle.com/datasets/prajwaldongre/loan-application-and-transaction-fraud-detection.

This dataset was selected due to its direct relevance to the challenge of loan fraud detection and its comprehensive nature, offering a variety of features that can indicate fraudulent activity. It contains information related to customer data, loan application details, customer loan history, and a fact table to link these entities, alongside measures for fraud analysis.

Key entities and their attributes within the dataset include:

- **Customer_Data:** applicant_age, customer_id, gender, loan_status, number_of_dependents, residential_address.
- **Loan time history:** application_date, application_id.

- **CustomerLoanHistory:** cibil_score, customer_id, debt_to_income_ratio, employment_status, existing_emis_monthly, monthly_income, property_ownership_status.
- **loan application details:** application_id, interest_rate_offered, interest_rate_offered (bins), loan_amount_requested, loan_amount_requested (bins).
- **Fact Table:** application_id, customer_id, fraud_flag, fraud_type.
- **Measures_Table:** Fraud Cases Over Time, Fraud Cases YoY%, Fraud Loss Rate.

The fraud_flag attribute in the Fact Table is a crucial indicator, typically binary (0 for non-fraudulent, 1 for fraudulent), allowing for the identification and analysis of fraud cases. This rich dataset enables a multifaceted approach to understanding and detecting loan fraud.

Purpose and Importance

The purpose of this project is to analyze loan application and transaction data to identify and understand the patterns and characteristics associated with fraudulent activities. By leveraging Power BI for data visualization, the study aims to provide clear insights into various aspects of loan fraud, including its types, geographical distribution, and correlation with applicant demographics and financial history. This is crucial for financial institutions to:

- **Minimize Financial Losses:** Detect and prevent fraudulent loans, thereby reducing monetary write-offs.
- **Enhance Risk Management:** Develop more accurate risk assessment models for loan applications.
- **Improve Operational Efficiency:** Streamline fraud detection processes and reduce manual efforts.
- **Safeguard Reputation:** Protect the institution's credibility and maintain trust with legitimate customers.

The findings from this analysis will support data-driven decision-making, enabling the implementation of targeted interventions and more robust security measures to combat loan fraud effectively.

Purpose of Analysis

The primary objective of this analysis is to identify key risk factors and patterns that indicate probable fraud in loan applications. By examining various data points such as applicant demographics, credit scores, loan amounts, and transaction histories, the goal is to pinpoint correlations that can aid in early fraud detection and prevention. This analysis aims to:

- Uncover hidden trends and anomalies in fraudulent loan applications.
- Quantify the impact of different variables on fraud likelihood.
- Provide actionable insights for financial institutions to develop proactive fraud mitigation strategies.
- Build a comprehensive Power BI dashboard that serves as an executive tool for monitoring and understanding fraud.

Key Questions

This analysis seeks to answer the following key questions, which guided our investigative process:

1. **What patterns indicate probable fraud?**
2. **Which demographics correlate with fraud risk?**
3. **Do tenure and interest rate affect fraud likelihood?**
4. **How do key performance indicators (KPIs) like Total Fraud Cases, Fraud Loss Rate, and Fraud Detection Rate evolve over time?**
5. **What are the most prevalent types of loan fraud?**
6. **Where are fraud cases predominantly located?**
7. **Is there a relationship between credit score and the occurrence of fraud flags?**

Graphs & Analysis

Understanding the Data: Key Performance Indicators (Brief)

The executive dashboard provides an immediate snapshot of loan fraud through several key performance indicator (KPI) cards. These cards offer quick, high-level insights into the scale and impact of fraudulent activities. (Detailed explanations are provided in the "Dashboards" section.)

Entity Relationship Diagram (ERD)

The Entity Relationship Diagram (ERD) provides a clear visual representation of how different data entities are related within the loan application dataset. It is fundamental for understanding the database structure and how information flows between various tables.

Description: The ERD typically shows entities (tables) as boxes, with their attributes (columns) listed within. Lines connect these entities, indicating relationships (e.g., one-to-many, many-to-many). Key relationships observed in your ERD are:

- **Customer_Data** is linked to **Fact Table** (via customer_id).
- **Loan time history** is linked to **Fact Table** (via application_id).
- **CustomerLoanHistory** is linked to **Fact Table** (via customer_id) and potentially Customer_Data (also via customer_id).
- **loan application details** is linked to **Fact Table** (via application_id).
- **Measures_Table** A separate table for calculated metrics, not directly linked to the transactional data entities in the same way, but deriving its values from them.

What it shows: The ERD provides a schematic blueprint of the underlying database structure. It illustrates how different pieces of information (like customer details, loan specifics, and historical data) are organized and interconnected.

Why it's important: Understanding the ERD is crucial for:

- **Data Modelling and Integrity:** Ensuring that data is structured logically and relationships are correctly defined to prevent inconsistencies and maintain data quality.
- **Querying and Analysis:** Guiding analysts in writing accurate and efficient queries to join different tables and retrieve comprehensive data for reporting and analysis. For example, to analyze fraud type by customer demographics, one would join Fact Table with Customer_Data.
- **System Development:** Providing a clear roadmap for developers building applications that interact with the database.
- **Business Understanding:** Helping business stakeholders understand how various data points relate to each other, improving their comprehension of the data landscape and operational processes.

The Fact Table acts as a central hub, linking various dimensional tables (Customer_Data, Loan time history, CustomerLoanHistory, loan application details) to the core transactional events (e.g., loan applications) and their fraud status. This star schema-like design is efficient for analytical queries in Power BI, as it optimizes performance for aggregations and filtering.

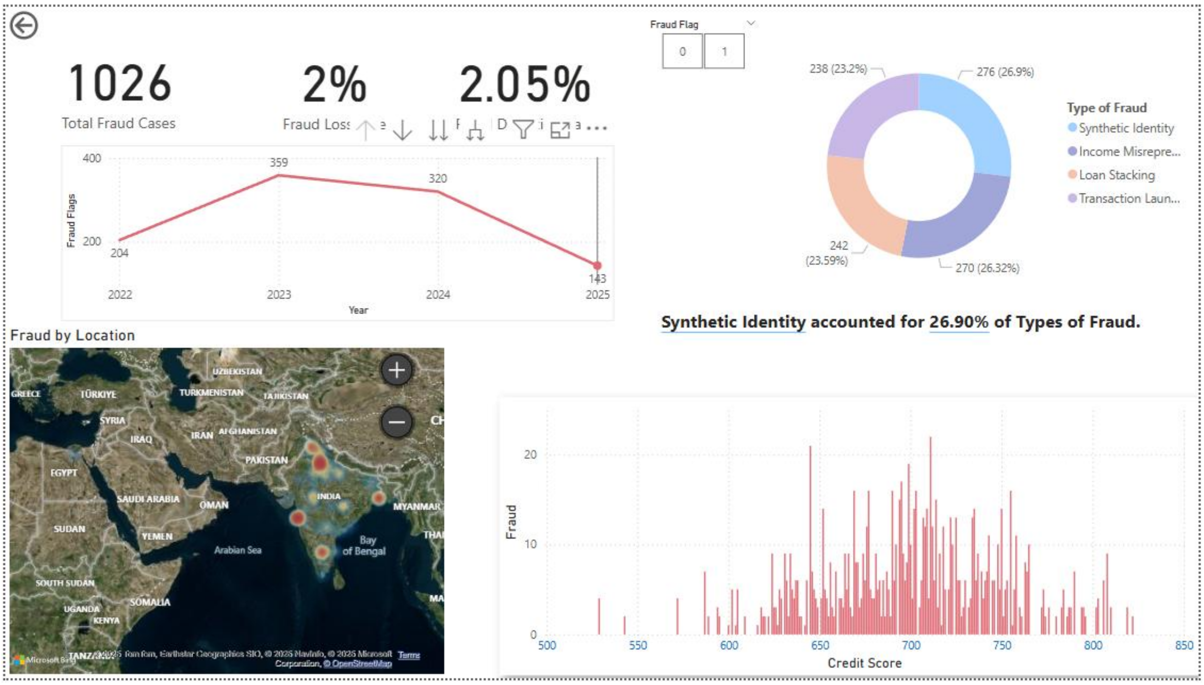
Dashboards

The Power BI dashboards serve as dynamic, interactive tools for monitoring and analyzing loan fraud. They consolidate various visualizations into a single, intuitive interface, enabling stakeholders to quickly grasp key insights and drill down into specific areas of interest.

Executive Dashboard Overview

This dashboard provides a high-level overview of the loan fraud landscape, designed for executive decision-makers. It integrates several key visuals to offer a comprehensive understanding of fraud patterns and impact.

Executive Dashboard



Synthetic Identity accounted for 26.90% of Types of Fraud.

Total Fraud Cases (KPI Card)

What it shows: This prominent card displays the total count of loan applications identified as fraudulent within the analyzed dataset. For example, Our dashboard shows 1026 Total Fraud Cases.

Why it's important to an executive: This metric provides an immediate and undeniable measure of the raw volume of fraud an organization is experiencing. For executives, it signals the sheer scale of the problem and helps in resource allocation, demonstrating whether the fraud challenge is a minor concern or a significant operational threat demanding urgent attention.

Fraud Loss Rate (KPI Card)

What it shows: This KPI card presents the percentage of financial losses incurred due to fraudulent loan applications, relative to a relevant baseline such as the total value of all loan applications or disbursed loans. Our dashboard shows 2% Fraud Loss Rate.

Why it's important to an executive: This is arguably the most critical financial metric for an executive, as it directly quantifies the monetary impact of fraud on the bottom line. A high or increasing fraud loss rate triggers immediate concerns about profitability and financial stability, prompting executives to invest in better fraud prevention or recovery mechanisms.

Fraud Detection Rate (KPI Card)

What it shows: This card illustrates the effectiveness of the current fraud detection mechanisms by displaying the percentage of actual fraudulent cases that were successfully identified and flagged by the system. Your dashboard shows 2.05% Fraud Detection Rate.

Why it's important to an executive: This metric directly reflects the efficiency and capability of the fraud prevention framework. For executives, a high detection rate instils confidence in the existing systems and processes, while a low rate signals potential weaknesses or blind spots that require immediate investigation and enhancement of detection technologies or strategies.

Trends it indicates: An increasing detection rate over time shows that the systems are improving in their ability to catch fraud. A stagnant or decreasing rate suggests that fraudsters might be evolving their methods faster than the detection systems, necessitating an upgrade in technology or analytical approaches.

Fraud Trends Over Time (Line Chart)

What it shows: This line chart visualizes the number of detected Fraud Flags (fraudulent applications) across different years (e.g., 2022, 2023, 2024, 2025). Each data point represents the total fraud cases for that year, connected to show historical progression. Your example shows values like 204 (2022), 359 (2023), 320 (2024), and 143 (2025).

Why it's important to an executive: This chart provides a crucial historical context, allowing executives to understand the dynamic nature of fraud. It helps in assessing whether fraud is a persistent, growing, or declining threat over time. This long-term perspective is vital for strategic planning, budgeting for fraud prevention, and evaluating the cumulative impact of past anti-fraud initiatives.

Trends it indicates: The chart can reveal upward trends (increasing fraud), downward trends (decreasing fraud, potentially due to effective interventions), or cyclical patterns.

For instance, the example shows an initial increase from 2022 to 2023, followed by a decline, suggesting that perhaps a new fraud scheme emerged and was then successfully mitigated.

Type of Fraud Distribution (Doughnut Chart)

What it shows: This doughnut chart breaks down the overall fraudulent cases by their specific categories, such as Synthetic Identity, Income Misrepresentation, Loan Stacking, and Transaction Laundering. Each slice represents a distinct fraud type, with its size indicating its proportion of the total fraud cases. Your dashboard indicates Synthetic Identity at 26.9%, Income Misrepresentation at 23.59%, Loan Stacking at 26.32%, and Transaction Laundering at 23.2%.

Why it's important to an executive: This visual is critical for tactical decision-making. Executives need to know which types of fraud are most prevalent to allocate resources effectively. Understanding the dominant fraud types allows for the development of targeted defences, training programs for staff, and specialized analytical models to counter the most significant threats.

Trends it indicates: If one type of fraud consistently represents a larger slice, it highlights a persistent vulnerability. Shifts in the distribution (e.g., one type growing rapidly while another shrinks) can indicate evolving fraudster tactics, prompting a re-evaluation of current prevention strategies.

Fraud by Location (Map Visual)

What it shows: This interactive map visualises the geographical distribution of fraudulent loan applications. It highlights regions or cities where fraud cases are more concentrated, using heatmap. The screenshot shows clusters in areas like New Delhi, Mumbai, Kolkata, Bengaluru and Ludhiana.

Why it's important to an executive: For executives managing operations across different regions, this map is invaluable for understanding geographical risk. It helps in identifying fraud hotspots, which can inform localized policy adjustments, targeted investigations, and resource deployment (e.g., deploying more fraud specialists to a high-risk region). It also aids in strategic decisions regarding market expansion or withdrawal based on fraud risk profiles.

Trends it indicates: Consistent high fraud activity in specific locations can indicate organized fraud rings or a pervasive local issue. Changes in hotspot locations might suggest that fraudsters are shifting their operations in response to detection efforts, or that new vulnerabilities have emerged in previously low-risk areas.

Fraud vs. Credit Score (Histogram)

What it shows: This histogram displays the frequency or count of fraudulent applications (Fraud Flags on the y-axis) across different ranges of Credit Score (x-axis, typically grouped into bins like 500-550, 550-600, etc.).

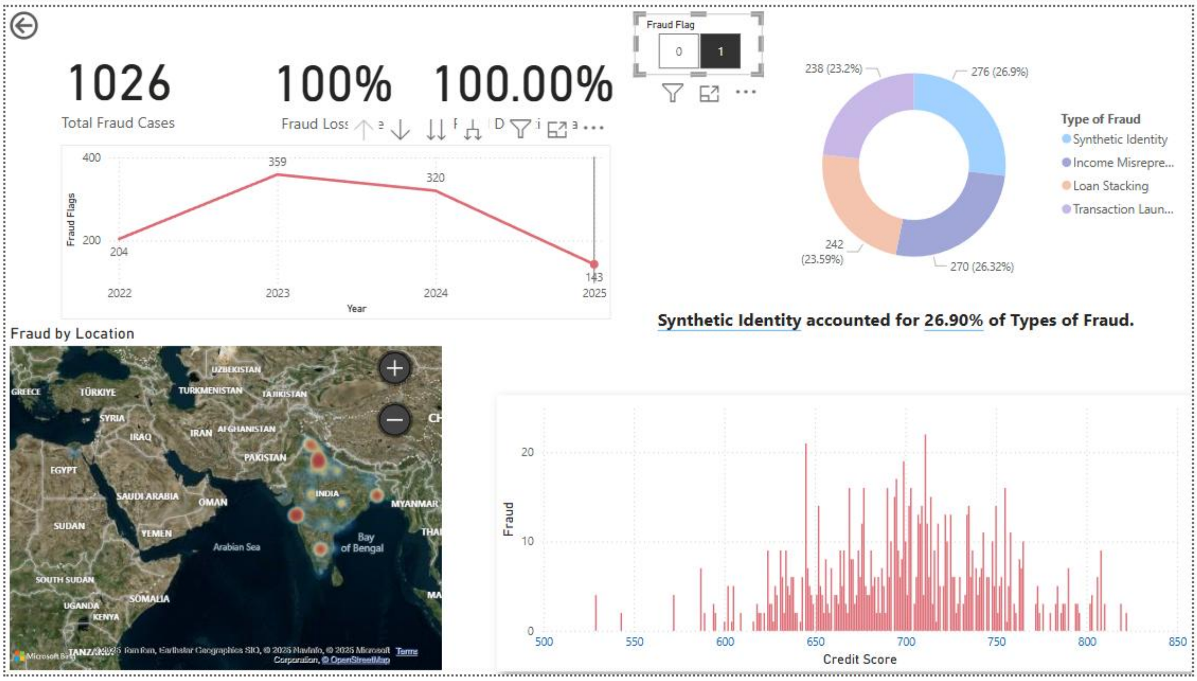
Why it's important to an executive: Understanding the credit profile of fraudsters is crucial for refining lending policies and risk assessment models. This visual helps executives determine if there's a particular credit score segment that disproportionately contributes to fraud. This insight can lead to adjustments in underwriting criteria, enabling faster approvals for low-risk applicants and stricter scrutiny for high-risk segments.

Trends it indicates: If the histogram shows a consistent peak in lower credit score ranges, it suggests a strong correlation between poor credit and fraudulent intent. A shift in these peaks (e.g., fraud appearing more frequently in higher credit score ranges) could indicate more sophisticated fraudsters who mimic good credit profiles, requiring more advanced behavioural analytics beyond just credit scores.

Filtered View Analysis

The filtered view of the dashboard demonstrates the power of interactivity in Power BI, allowing users to focus on specific segments of the data for deeper analysis.

Filtered View



What it shows: When a specific filter is applied (e.g., selecting Fraud Flag = 1 to view only fraudulent cases, and then further filtering by a particular Type of Fraud like "Income Misrepresentation"), this view dynamically updates all other visuals on the dashboard.

Why it's important to an executive: This capability allows executives and analysts to perform granular investigations into specific fraud scenarios. Instead of just seeing the aggregate picture, they can isolate and study the characteristics of a particular fraud type. This targeted insight is invaluable for understanding the nuances of how a specific fraud scheme operates, its geographical footprint, and the typical credit profile of perpetrators, which in turn informs precise counter-measures and policy adjustments. **Trends it indicates:** Filtering allows for trend analysis within specific fraud subsets. For example, observing the "Fraud Trends Over Time" in a filtered view for "Synthetic Identity" might reveal that this specific fraud type has a different temporal pattern compared to the overall fraud trend, helping to identify emerging or declining threats within particular categories.

References

- **Kaggle Dataset:** https://www.kaggle.com/datasets/prajwaldongre/loan-application-and-transaction-fraud-detection
- **GitHub Discussion:** https://github.com/kush777handa/DataScience/discussions/1

Conclusion

This analysis of loan application data has provided valuable insights into the key patterns and risk factors associated with financial fraud. The Power BI dashboard effectively visualises critical metrics, highlighting the significant influence of various factors such as fraud types, temporal trends, geographical locations, and credit scores. The findings emphasize that fraud is a multifaceted challenge, with varying prevalence across different categories and time periods.

Specifically, the ability to track Total Fraud Cases, Fraud Loss Rate, and Fraud Detection Rate offers financial institutions a clear understanding of the scale and impact of fraud. The breakdown by Type of Fraud allows for targeted prevention strategies, while the Fraud by Location map helps in identifying high-risk geographical areas. The correlation with Credit Score provides additional insights for risk assessment during the application process.

By leveraging these insights, financial institutions can enhance their fraud detection capabilities, reduce financial losses, and develop more robust, data-driven strategies for fraud prevention and management. This approach not only safeguards assets but also strengthens the integrity and trustworthiness of the loan application ecosystem.