BY

Md Nazmun Hasan Nafees

Background

Credit risk assessment is a critical process in the financial industry, determining whether an applicant is eligible for a loan based on financial stability and past credit behavior. Financial institutions rely on robust models and data-driven insights to minimize risk and make informed lending decisions.

Problem Statement

Loan default and misclassification of credit risk can lead to significant financial losses for lenders. Traditional credit evaluation methods are often manual, time-consuming, and prone to human bias. The challenge lies in developing an efficient, automated system that can accurately predict loan approvals while uncovering key insights into risk factors.

Objective

This project aims to analyze loan applicant data using the **German Credit Dataset** and build a predictive model for loan approval decisions. The key goals are:

- Conduct **exploratory data analysis (EDA)** to uncover insights into applicant demographics, financial behavior, and risk indicators.
- Identify the most significant factors affecting loan approval and creditworthiness.
- Develop and evaluate **machine learning models** to classify loan applications as approved or rejected.
- Provide a scalable, data-driven solution for financial institutions to enhance their credit rating and risk assessment processes.

By addressing these objectives, this project will offer valuable insights into credit risk patterns and provide a robust system for automated loan evaluation.

Part 1: Data Cleaning

Why Data Cleaning?

Data cleaning is essential to ensure accuracy, consistency, and reliability in our analysis and model predictions. The German Credit Dataset contains missing values, inconsistent formats, and unnecessary symbols that can negatively impact machine learning models and statistical analysis. Cleaning the data helps in standardizing formats, handling missing values, and improving model interpretability.

Data Cleaning Steps Performed

1. **Converting Time to Datetime Format:** Standardized date-related values by converting them to pandas datetime format for better analysis.

2. **Standardizing Currency Values:** Removed currency symbols and converted monetary amounts to a consistent currency format.

3. **Handling Capped Values:** Adjusted values like days (max 365), marks (out of 100), and percentages (out of 100) to ensure correctness.

4. **Handling Missing Data:** Dropped columns with excessive missing values to prevent bias and improve model performance.

5. **Fixing Data Types:** Converted categorical and numerical columns to appropriate data types for efficient processing.

6. **Removing Strings in Numeric Columns:** Cleaned numeric columns by removing non-numeric characters (e.g., converting "8$" to "8").

Part 3: Exploratory Data Analysis (EDA)

Why EDA?

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset, identifying patterns, detecting anomalies, and generating insights that inform decision-making. EDA helps uncover relationships between variables, detect fraud trends, and refine features for predictive modeling. It ensures that our assumptions are valid before applying machine learning models.

Objectives of EDA

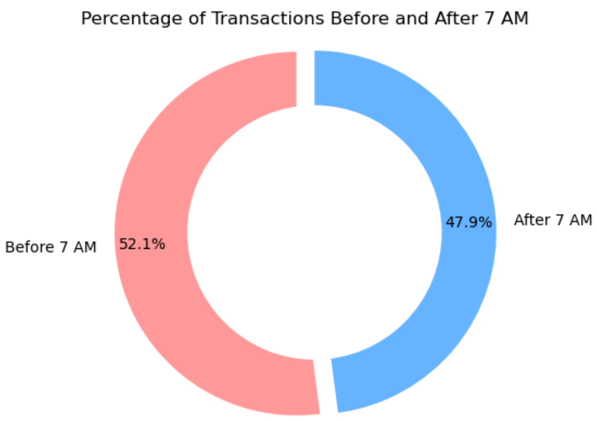
In this project, EDA is conducted to answer key questions about fraudulent transactions and customer behavior, including:

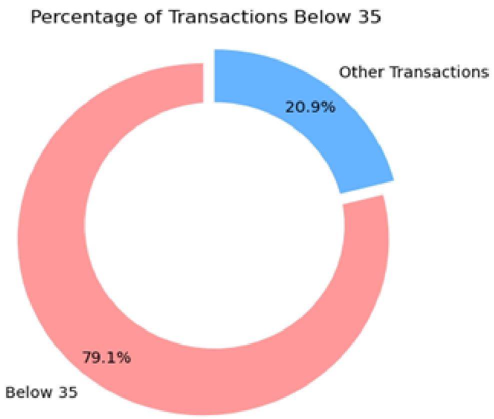
- **When did the highest percentage of fraudulent transactions occur?** (Time-based trends)
- **Which transaction amount range was most common?** (Distribution of transaction amounts)
- **Which transaction medium had the highest fraud rate?** (Entry mode analysis)
- **Which merchant types were most associated with fraud?** (Industry-based trends)
- **Which country had the highest number of fraudulent transactions?** (Geographical trends)
- **Did fraudsters target a specific gender?** (Demographic patterns)
- **Which age group was most susceptible to fraud?** (Age-based analysis)
- **Which bank experienced the highest fraud cases?** (Institutional vulnerability)

Part 4: Deep Dive into EDA: Insights.

Problem 1: Time and Amount of Fraudulent Transactions

- **Findings**:
  - **52.07%** of fraudulent transactions occur before or at 7 AM.
  - The transaction amounts are generally low, about 35.
  - **79.13%** of transactions are below 35.

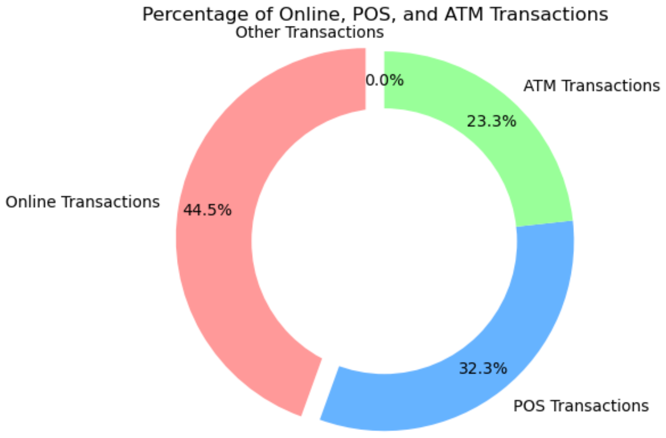


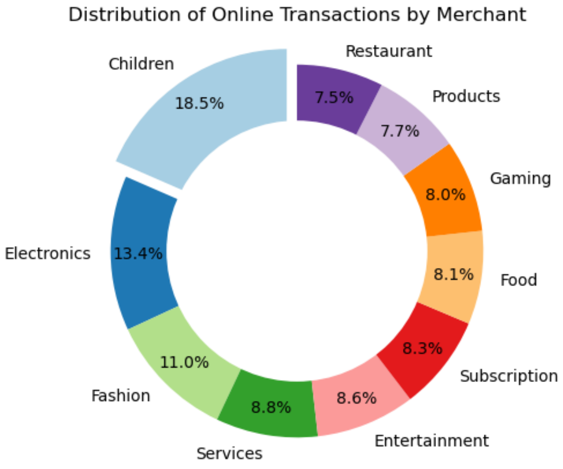


- **Solution**:
  - Implement a flagging system for transactions in the morning (before 7 AM) that are below 35. This can help in early detection and prevention of potential fraud.

Problem 2: Online Transactions and Merchants Contributing to Fraud

- **Findings**:
  - A significant portion of fraudulent transactions are online.
  - Top merchants involved:
    - **Children's Products**: 17.78%
    - **Electronics**: 13%





Distribution of Online Transactions by Merchant

- **Solution**:
    - Enhance monitoring and verification processes for online transactions, particularly those related to these merchant categories.
    - Implement stricter security measures, such as requiring additional verification steps for high-risk merchant categories.

Additional Insights:

- **CVC Code Usage**:
    - All online transactions are done by entering the CVC code. No CVC is used outside of online transactions.
- **Origin of Transactions**:
    - The origin of transactions is fairly distributed among four countries, suggesting that all four are suspicious.
    - 98% of the cardholders' origins are from the UK, indicating that hackers are primarily targeting residents of the UK.
- **Feature Importance Analysis (Using Random Forest Classifier)**:
    - **Time of Transaction**: The time of the transaction is a critical factor. Monitoring transactions occurring between midnight and 7 AM can help identify potential fraud.
    - **Transaction Amount**: The transaction amount is also a significant indicator, with amounts below 35 comprising 80% of fraudulent transactions.

Refined Solutions:

1. **Enhanced Flagging System**:
    - Develop a dynamic flagging system that not only flags transactions based on time and amount but also considers patterns in merchant categories and geographic origin.
2. **Stricter Verification for Online Transactions**:
    - Implement multi-factor authentication (MFA) for online transactions, especially for high-risk categories and during high-risk time frames.
3. **Geographic Monitoring**:
    - Increase scrutiny of transactions originating from the identified suspicious countries.
    - Implement additional verification for transactions where the origin does not match the typical behavior of the cardholder.
4. **Focus on UK Residents**:
    - Deploy targeted security measures to protect UK residents, such as anomaly detection and enhanced security checks.