Credit Card Fraud Detection

What patterns or behaviors differentiate fraudulent transactions from legitimate ones?

By D7

A TECHNICAL REPORT

submitted to

Professor Francis Pereira

Presented December 05, 2024

TABLE OF CONTENT

I. Introduction

1. **Overview of the Problem**:

In an increasingly cashless society driven by digital payment methods, credit card transactions are more common than ever. While being cashless is convenient, it also gives rise to many opportunities for credit card fraud. Credit card fraud is now one of the leading issues of fraud in today's society that occurs on a daily basis with 52 million Americans experiencing credit card fraud in 2023. According to *Security.org*, leading digital safety experts from across the five major U.S credit card companies have estimated the following statistics around credit card fraud:

- Around 60% of U.S credit card holders are victims of credit card fraud, with 45% of card holders being repeat victims.
- In 2023, over 52 million Americans had fraudulent charges on their credit cards, with amounts totaling to over $5 billion.
- The median fraud charge amount is around $100, having increased 26% over the past two years from $79.
- Most of the charges are from personal data and account information being accessed remotely, making up 93% of the charges involved.

2. **Objective:**

In order to find recommendations for reducing the number of fraud cases, we analyzed a sample dataset to identify patterns in credit card fraud. By isolating fraudulent transactions, we used frequency tables to examine common attributes like categories and locations. We then applied chi-square and two-sample tests to explore relationships between variables. These

insights aim to uncover patterns and help prevent future fraud cases. Ultimately, this will help us discover potential patterns in credit card fraud cases, with the goal of preventing such cases from occurring again in the future.

II. Data Description:

1. **Dataset Overview:** The dataset, sourced from Kaggle and updated 8 months ago, contains 555,719 credit card transactions recorded between July 1, 2020, and December 31, 2020. It is available for analysis at Kaggle's Credit Card Fraud Prediction.

Category	Variable	Description	Data Type
Transaction Information	trans_date_trans_time	Date and time of the transaction	Temporal/Date-Time
	trans_num	Unique transaction identifier	Categorical/Identifier
	unix_time	Unix timestamp	Numeric
	amt	Transaction amount	Numeric
	category	Transaction category	Categorical
Credit Card and Merchant Details	cc_num	Credit card number	Categorical/Identifier
	merchant	Merchant name or ID	Categorical
	merch_lat	Merchant latitude	Numeric
	merch_long	Merchant longitude	Numeric
Customer Information	first	First name	Categorical
	last	Last name	Categorical
	gender	Gender	Categorical
	dob	Date of birth	Temporal/Date-Time
	job	Customer's job	Categorical
	street	Street address	Categorical
	city	City	Categorical
	state	State	Categorical
	zip	Zip code	Categorical
	city_pop	Population of the city	Numeric
Geographical Information	lat	Customer's latitude	Numeric
	long	Customer's longitude	Numeric
Fraud Detection	is_fraud	Fraud status	Binary/Categorical

Table 1: Overview of Dataset Variables and Their Data Types

2. Summary Statistics & Data Profiling:

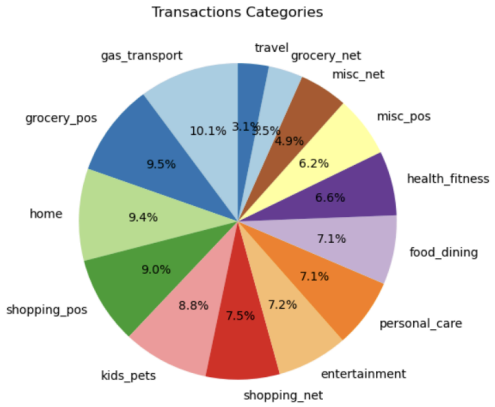


Figure 1**:** The percentage of transactions across 14 categories (category) illustrates that the database contains a total of 14 transaction types.

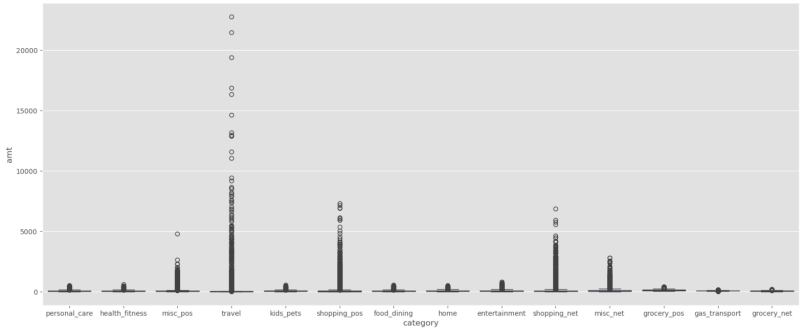


Figure 2: Transaction amount of different categories illustrates that travel expenses have the highest transaction amounts, with two transactions over $20,000 each.

Count:	555,719
Mean:	69.4
Std:	156.7
Min:	1
25%:	9.6
50%:	47.3
75%:	83
Max:	22,768.10

Table 2: Summary Statistics for Transaction Amount (amt) shows that the average transaction amount is $69.4 (SD: $156.7), ranging from $1 to $22,768.1.

Count:	555,719
Mean:	46.4
Std:	17.4
Min:	15
25%:	33
50%:	44
75%:	58
Max:	96.0

Table 3: Summary Statistics for Age shows that consumers range from 15 to 96 years old, with an average age of 46, and over 50% aged 44 or younger.

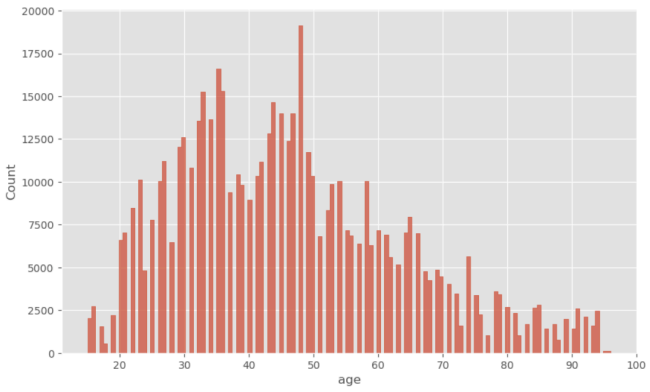


Figure 3: User age distribution illustrates that the age distribution peaks between 30 and 50, with a concentration around the late 40s, and tapers off significantly above 60, indicating fewer older users.

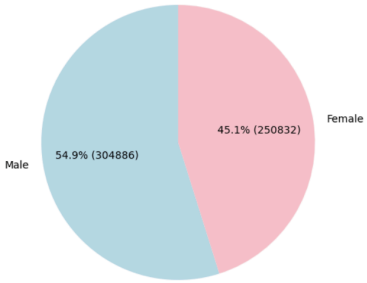


Figure 4: Gender distribution in transaction (gender) shows that Males account for 54.9% (304,886) of transactions, while females make up 45.1% (250,832).

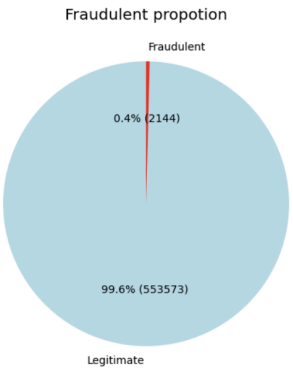


Figure 5: The distribution of Fraud and Non_Fraud transactions (is_fraud) shows that Fraudulent transactions make up 0.4% (2,144), compared to 99.6% (553,573) legitimate. To address the imbalance, we analyzed 10% of non-fraudulent transactions and included all fraudulent ones.

Overall, credit cards are primarily used for transportation, fuel, and grocery purchases. The average transaction is around $70, with higher amounts linked to travel. Males make up the majority of users, and most consumers are middle-aged to young adults, with half aged between 30 and 60.

III. Data Analysis

a. Exploratory Data Analysis (EDA):

Fraud trends show that online shopping (506 cases), in-person grocery shopping (485 cases), and miscellaneous online shopping (267 cases) are the most affected categories. Fraud peaks late at night, from 10 P.M. to midnight, with 1,088 cases. The average fraudulent transaction is $528.35, ranging from $1.78 to $1,320.92. New York leads with 175 cases, followed by Pennsylvania (114) and Texas (113).

To better visualize and understand fraudulent activity, we have created various charts (***Figure 6-Figure 11***):

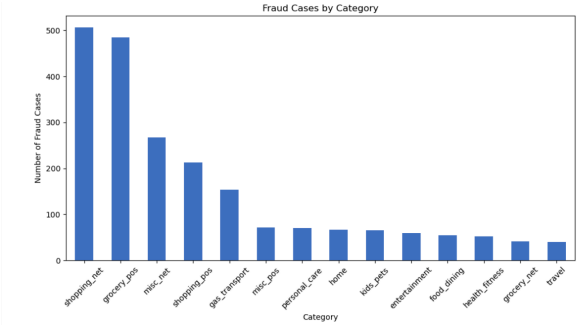


Figure 6: Fraud Cases by Category

The bar chart in ***Figure 6*** indicates that the top three categories are online shopping, in-person grocery shopping, and miscellaneous online shopping, possibly due to faulty credit card readers.

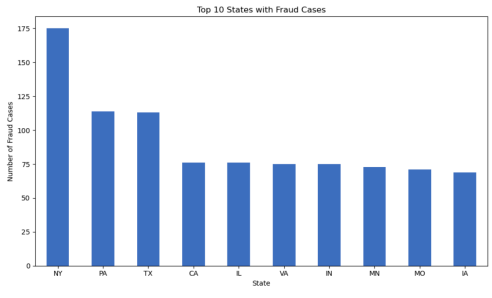


Figure 7: Top 10 States with Fraud cases

Figure 7 :The top 10 states with the most fraud cases are highlighted, with New York leading (~175 cases), followed by Pennsylvania (~120 cases) and Texas (~120 cases).

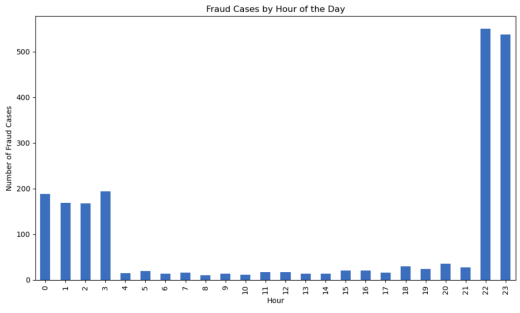


Figure 8: Fraud Cases by Hour of the Day highlights that the highest number of fraud cases occur around midnight, at hours 22 and 23, likely due to online-shopping scams as more people are at home.

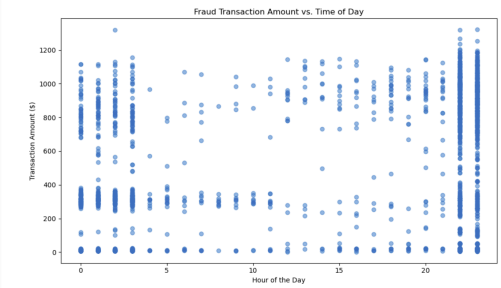


Figure 9: Fraud Transaction Amount vs.Time of Day shows that fraud varies throughout the day, with the majority of cases occurring around midnight.

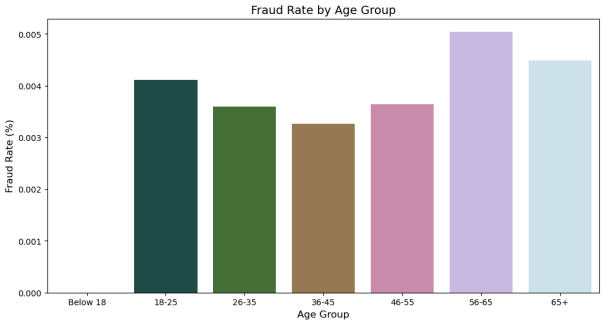


Figure 10: Fraud Rate vs Age group

Figure 10 shows higher fraud rates among younger (18-25) and older adults (56+), likely due to younger people being less aware of scams and older adults being less familiar with technology.

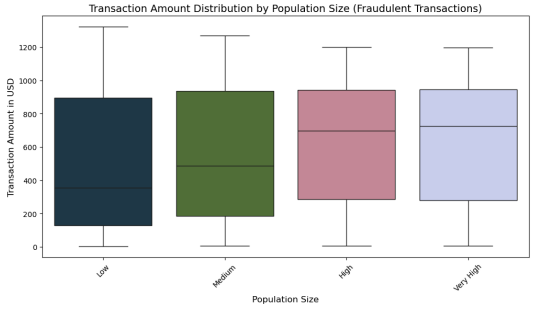


Figure 11: Transaction Amount Distribution by income group

Figure 11 shows transaction amounts across population groups (Low, Medium, High, Very High) with box plots. The median transaction amount increases as population size grows, with the Low group having the lowest median and the Very High group the highest.

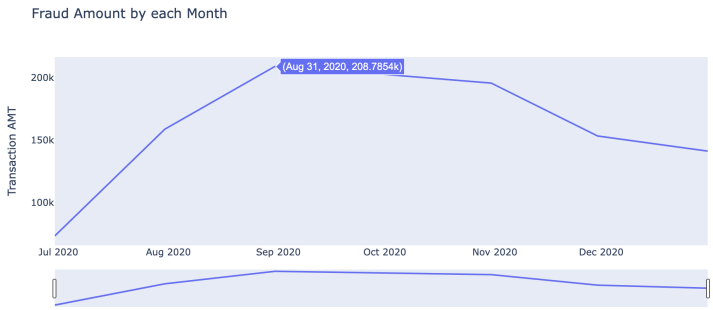
b. Statistical Analysis:

For a credit card transaction database, there are many factors that can help us identify potential fraudulent transactions. Next, we will explore these factors one by one and closely identify which potential transactions are worth analyzing and paying attention to.

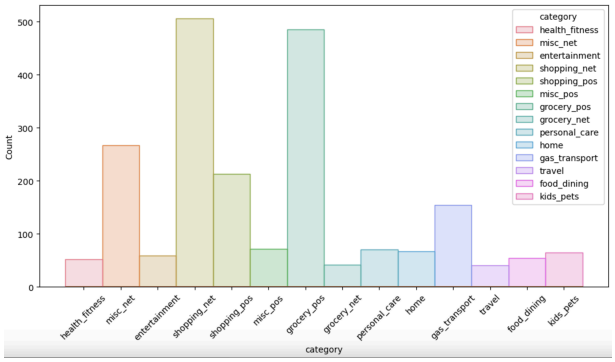
We all know that each transaction has a corresponding transaction time. Let's start by analyzing the time first. By generating a frequency table of fraud transactions by time and a time series chart based on the amount of fraudulent transactions per month, we can clearly observe that fraud transactions occur every day, and the data is relatively scattered with minimal patterns.

```
trans_date_trans_time
30/06/2020 23:16    3
26/07/2020 23:25    2
20/07/2020 03:05    2
13/11/2020 00:48    2
02/12/2020 22:27    2
                   ..
18/08/2020 22:07    1
18/08/2020 22:05    1
18/08/2020 18:35    1
18/08/2020 18:24    1
22/12/2020 23:13    1
Name: count, Length: 2108, dtype: int64
```

trans_date_trans_time	amt
2020-06-30	73274.93
2020-07-31	158669.49
2020-08-31	208785.43
2020-09-30	202700.99
2020-10-31	195572.97
2020-11-30	153182.19
2020-12-31	141138.68



This means that it becomes very difficult to pinpoint fraudulent transactions on a daily basis. Next, we consider transaction categories. By using a bar chart, we can clearly see that online shopping fraud transactions are the most frequent, but that's about it. There are various online shopping platforms, and if we cannot precisely identify specific platforms, it will cost us a significant amount of time and resources to address this issue.




```
category
shopping_net      506
grocery_pos       485
misc_net          267
shopping_pos      213
gas_transport     154
misc_pos           72
personal_care      70
home               67
kids_pets          65
entertainment      59
food_dining        54
health_fitness     52
grocery_net        41
travel             40
Name: count, dtype: int64
```

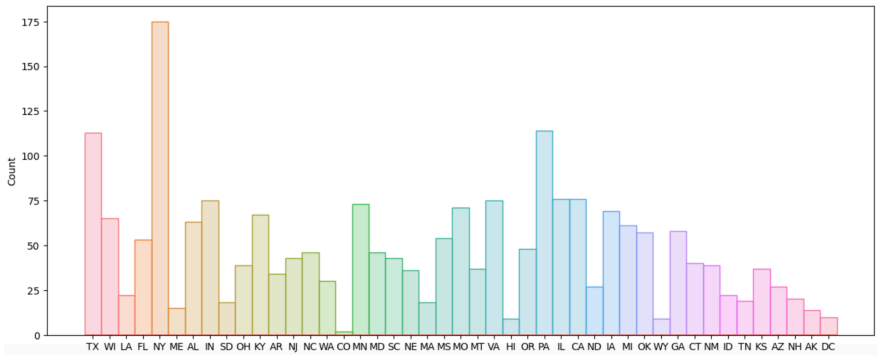
Finally, we want to identify more specific features through location. First, we can see that, overall, New York State has the highest number of fraudulent transactions. When we further break it down, we find that Camden is the city with the most fraud, and '00315 Ashley Valleys' is the area with the highest concentration of fraud.

```
credit_new.state.value_counts().head(10)
```

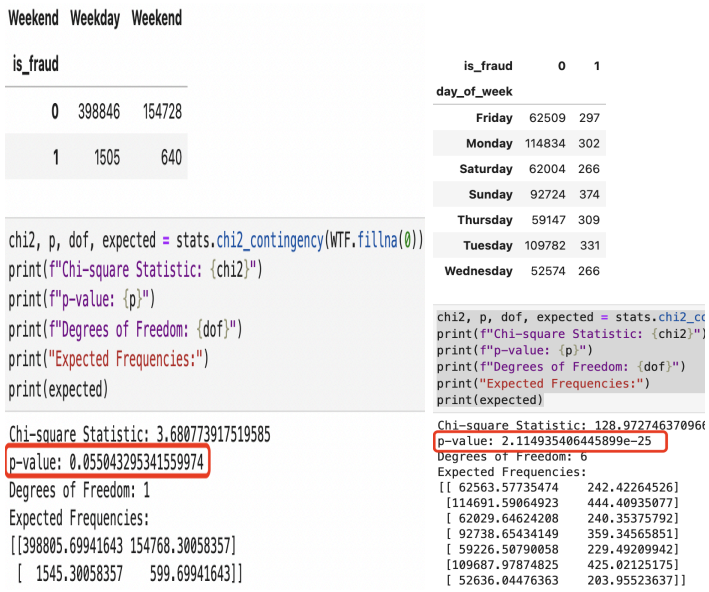
```
state
NY    175
PA    114
TX    113
CA     76
IL     76
VA     75
IN     75
MN     73
MO     71
IA     69
Name: count, dtype: int64
```

```
city
Camden          27
Birmingham      25
Burrton         19
Clarks Mills    18
Bradley         16
                ..
Deltona          3
Littleton        2
Sherman          2
Laramie          2
Notrees          2
Name: count, Length: 215, dtype: int64
```

```
street
00315 Ashley Valleys            19
444 Robert Mews                 18
4664 Sanchez Common Suite 930   16
9853 Flynn Hollow Apt. 602      16
358 Pruitt Square               16
                                ..
7337 Browning Falls Suite 027    3
3379 Williams Common             2
67210 Lori Creek                 2
9870 Collins Flat                2
63542 Luna Brook Apt. 012        2
Name: count, Length: 218, dtype: int64
```



Then, we can use the chi-square test for further analysis. We can determine whether there is an association between fraudulent transactions and whether they occur during weekdays or weekends, or if there is a correlation with specific days of the week.



is_fraud	Weekend	Weekday	Weekend
0	398846	154728	
1	1505	640	

```
chi2, p, dof, expected = stats.chi2_contingency(WTF.fillna(0))
print(f"Chi-square Statistic: {chi2}")
print(f"p-value: {p}")
print(f"Degrees of Freedom: {dof}")
print("Expected Frequencies:")
print(expected)
```

```
Chi-square Statistic: 3.680773917519585
p-value: 0.055043295341559974
Degrees of Freedom: 1
Expected Frequencies:
[[398805.69941643 154768.30058357]
 [  1545.30058357    599.69941643]]
```

is_fraud	0	1
day_of_week		
Friday	62509	297
Monday	114834	302
Saturday	62004	266
Sunday	92724	374
Thursday	59147	309
Tuesday	109782	331
Wednesday	52574	266

```
chi2, p, dof, expected = stats.chi2_co
print(f"Chi-square Statistic: {chi2}")
print(f"p-value: {p}")
print(f"Degrees of Freedom: {dof}")
print("Expected Frequencies:")
print(expected)
```

```
Chi-square Statistic: 128.97274637096
p-value: 2.1149354064458899e-25
Degrees of Freedom: 6
Expected Frequencies:
[[ 62563.57735474    242.42264526]
 [114691.59064923    444.40935077]
 [ 62029.64624208    240.35375792]
 [ 92738.65434149    359.34565851]
 [ 59226.50790058    229.49209942]
 [109687.97874825    425.02125175]
 [ 52636.04476363    203.95523637]]
```

From these two results, we can see that the first p-value is slightly greater than 0.05, so we can say that there is no significant relationship between fraudulent transactions and whether they occur on weekdays or weekends. However, from the second result, we can conclude that there is a relationship between fraudulent transactions and the specific day of the week, as the p-value is less than 0.05.

At the same time, we can also analyze the relationship between fraudulent transactions and gender. We used a t-test to examine whether there is a difference in the total amount of fraudulent transactions between males and females. From the result, we can see that there is a significant

relationship, as the p-value is much smaller than 0.05.

```
male=data_fraud[data_fraud['gender']=='M']['amt']
female=data_fraud[data_fraud['gender']=='F']['amt']
stats.ttest_ind(a=male, b=female,equal_var=True)
```

```
TtestResult(statistic=6.409727639782657, pvalue=1.7873917726692276e-10, df=2143.0)
```

Lastly, a t-test between distance and is_fraud show that there is no statistically significant relationship between distance and fraud status.

```
T-Statistic: 0.17720220514074408
P-Value: 0.8593661741316395
```

IV. Hypothesis Development:

Overall, this paper suggests that age, gender, transaction timing, and job category—specifically within the legal field—have a significant impact on the likelihood of being targeted by fraud. Specifically, it is proposed that older individuals, males, and those making late-night transactions are more likely to be targeted by fraud. In contrast, individuals employed in the legal field are less likely to experience fraud. The hypotheses are outlined below:

a. Age

H0: Age does not affect the likelihood of fraud.

Ha: Individuals aged 56+ are more likely to be targeted by fraud.

b. Gender

H0: Gender does not affect the probability of fraud.

Ha: Gender significantly affects the probability of fraud.

c. Time the Transaction Occurred

H0: Transaction time (day vs. night) does not affect the likelihood of fraud.

Ha: Transactions at night are more likely to be fraudulent than those during the day.

d. Job category-Legal

H0: Job category (legal) does not affect the likelihood of fraud.

Ha: Job category (legal) significantly affects the likelihood of fraud.

V. Methodology and Experimental Design

a. Data Cleaning:

We checked for missing or null values using the **.isnull().sum()** function and confirmed that the dataset had no missing values, ensuring data completeness for accurate analysis.

b. Data Enrichment:

We converted the trans_date_trans_time column to a datetime format and created a transaction_date column for trend analysis. User ages were calculated from DOB and grouped into age categories. Temporal features, like the hour of the day and weekend status, were added to track transaction patterns. Job titles were grouped into categories to analyze occupation impact, and a latenight feature was created to identify potential fraud risks. These transformations enabled deeper insights for better modeling and decision-making.

c. Experimental Design:

This study conducted a comprehensive Exploratory Data Analysis (EDA) to identify patterns in fraudulent transactions. We first analyzed fraud across different transaction types, such as grocery and gas purchases, highlighting high-risk areas. Geographical analysis revealed the top 10 states and top 20 cities with the highest fraud cases, aiding targeted interventions. Temporal analysis showed that fraud peaked during evening and midnight hours and exhibited distinct trends across weekdays and weekends. Demographic analysis uncovered fraud prevalence across various age groups and income levels, shedding light on potential vulnerabilities.

Hypothesis testing was employed to examine specific factors, including whether transaction amounts differed between male and female users, and the relationship between fraud occurrence and the day of the week. Time series analysis confirmed that fraudulent transactions occur daily, with a notable association to specific days of the week. The results suggest that targeting transactions in particular regions and during peak times can effectively reduce fraud.

VI. Logistic Regression Analysis

Hypothesis Testing

Preliminary hypothesis testing identified several variables with significant relationships to fraudulent transactions. Insignificant predictors included is_weekend and distance. Significant predictors of fraud included transaction amount, gender, latenight transactions (11 PM to 3 AM), high-risk transaction categories (online shopping, online miscellaneous, and in-person grocery), job category (legal professionals), and membership in a high-risk age group (56–65 and 65+). To further assess these relationships, a logistic regression analysis was performed.

Dataset and Model Setup

The dataset exhibited a high degree of imbalance, with fraud cases representing a small proportion of total transactions. To mitigate this imbalance and improve model performance, 10% of the majority class (is_fraud == 0) was randomly sampled and combined with all minority class cases (is_fraud == 1). This resulted in a balanced dataset containing 55,357 non-fraud cases and 2,143 fraud cases.

The logistic regression model included seven predictors:

- **Transaction characteristics**: amt, latenight, and is_high_risk_category.
- **Demographic variables**: gender_M, job_category_Legal, high_risk_age_group, and city_pop.

Logit Regression Results

Dep. Variable:	is_fraud	No. Observations:	57502
Model:	Logit	Df Residuals:	57494
Method:	MLE	Df Model:	7
Date:	Mon, 09 Dec 2024	Pseudo R-squ.:	0.4422
Time:	22:40:09	Log-Likelihood:	-5109.0
converged:	True	LL-Null:	-9158.7
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.8854	0.077	-76.786	0.000	-6.036	-5.735
amt	0.0050	9.88e-05	50.963	0.000	0.005	0.005
gender_M	-0.1351	0.057	-2.373	0.018	-0.247	-0.023
latenight	2.7510	0.072	38.378	0.000	2.611	2.892
is_high_risk_category	0.6827	0.058	11.802	0.000	0.569	0.796
job_category_Legal	-0.9723	0.297	-3.275	0.001	-1.554	-0.390
high_risk_age_group	0.1380	0.083	1.666	0.096	-0.024	0.300
city_pop	-5.479e-07	1.3e-07	-4.203	0.000	-8.03e-07	-2.92e-07

Key Findings

The logistic regression analysis revealed several key insights regarding predictors of fraudulent transactions:

1. **Transaction Characteristics**:

- **Late-night Transactions (11 PM to 3 AM):** Strongly linked to fraud ($z = 38.378$, $p < 0.001$, coefficient = 2.751)
- **High-Risk Categories**: Online shopping, miscellaneous net purchases, and grocery transactions were strongly correlated with fraud.
- **Transaction Amount**: Higher amounts increased fraud likelihood ($z = 50.963$, $p < 0.001$).

2. **Demographics**:
   - **Gender**: Males were slightly less likely to experience fraud ($z = -2.373$, $p = 0.018$).
   - **Job Category**: Legal professionals were less likely to experience fraud ($z = -3.275$, $p = 0.001$)..
   - **Age Group**: Legal professionals were less likely to experience fraud ($z = -3.275$, $p = 0.001$)..
   - **City Population**: sThe coefficient was too small (-5.479e-07) to be significant.

Model Performance

- **Pseudo R^2**: 0.4410, indicating a good model fit.
- **Misclassification Rate**: 2.49%, demonstrating the model's accuracy in correctly classifying transactions.
- **AUC-ROC**: 0.9179, reflecting strong discriminatory power between fraud and non-fraud cases.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	55357
1	0.84	0.41	0.55	2145
accuracy			0.98	57502
macro avg	0.91	0.70	0.77	57502
weighted avg	0.97	0.98	0.97	57502

Misclassification Rate: 0.0249
AUC-ROC: 0.9201882015835084

Given the imbalanced nature of the data, standard accuracy metrics were not the primary evaluation focus. Instead, the following metrics were emphasized:

- **Precision (Fraud)**: 84%, indicating that 84% of transactions predicted as fraud were truly fraudulent.
- **Recall (Fraud)**: 41%, indicating that the model identified 41% of all actual fraud cases.
- **F1-Score**: 55%, representing a balance between precision and recall.

Interpretation of Metrics

Precision measures the accuracy of predicted fraud cases, focusing on minimizing false positives. Recall captures the model's ability to detect all actual fraud cases. The F1-Score balances precision and recall, optimizing both fraud detection and false positive reduction.

VII. Findings and Conclusions

Summary of Insights

The analysis found that fraud is more common in late-night transactions (11 PM to 3 AM), high-risk categories (online shopping, miscellaneous purchases, grocery shopping), and larger amounts. While males and legal professionals are less likely to experience fraud, older users (56+) are more susceptible.

Implications

These findings have several implications for financial institutions seeking to reduce fraud risks effectively:

- **Improved Risk-Based Monitoring**: Focus on high-risk transactions, especially at night or in online shopping.
- **Targeted Fraud Prevention Strategies**: Customize fraud detection for demographics like older users or specific job categories.
- **Real-Time Detection**: Enhance systems with key predictors for quicker, more accurate fraud detection.
- **Cost Optimization**: Prioritize high-risk transactions to reduce false positive investigations and optimize budgets.

VI. Limitations

We sourced the dataset from Kaggle.com, a reliable platform for free data, but the lack of clarity on data collection methods poses limitations to the analysis. The dataset spans Q3 and Q4 of 2020, a period influenced by the early stages of the pandemic, which significantly impacted customer spending behavior.

The dataset contains 23 variables, categorized into customer, transaction, merchant, and geographical information. While suitable for this study, future research could include additional variables, such as merchant reputation and bank details, for deeper insights.

A key challenge was the dataset's imbalance, with 99.6% non-fraudulent and 0.4% fraudulent transactions. This skew risks biasing models toward the majority class, leading to poor fraud detection and misleading metrics. To address this, we sampled 10% of the majority class and combined it with all fraudulent transactions to create a more balanced dataset for analysis.

IX. Conclusion

This study examines the relationship between credit card fraud and external factors to identify patterns influencing fraud occurrences. Our findings show that older individuals, males, and late-night transactions are more likely to be targeted by fraud, while those in the legal field are less likely to experience fraud. This highlights the role of demographic factors and transaction timing in fraud. While fraud detection is essential, it poses significant challenges, especially with technological advances. This study offers a framework for banks to analyze and implement controls for fraud prevention, while encouraging further exploration by business analytics professionals.

Reference

Security.org. (2023). *Credit card fraud report: Statistics and facts for 2023*. Retrieved December 9, 2024, from https://www.security.org/digital-safety/credit-card-fraud-report/