BAN 620 Data Mining

Project Report

Team -2

NAME	NET ID:
Veera Venkata Sai Kalyan Kaparaju	ss4785
Adrian Padilla	bu75 69
ARPIT GHATORIYA	op9580
Praneeth Reddy Gaddam	id8973

Instructor: Balaraman rajan

OVERVIEW:

The project focuses on analyzing credit card fraud patterns using a dataset obtained from Kaggle, spanning January 1st, 2019, to December 31st, 2020. With 1,048,575 records and 22 variables, the dataset covers merchant details, transaction specifics, and customer demographics. The objective is to understand factors influencing credit card spending, detect fraud patterns, propose preventive measures, and assess the impact of demographic factors on spending and fraud rates.

AIM OF THE PROJECT: This project aims to carefully examine data from credit card transactions to figure out how fraud happens, what factors affect how people use credit cards, and how demographics influence spending habits. Using advanced analysis methods, the goal is to improve the security of credit cards. The study aims to discover patterns, factors that predict fraud, and ways to prevent it. This is important because credit card fraud can cause financial problems for both banks and people who use credit cards.

INTRODUCTION:

Our project is focused on analyzing fraud patterns in credit card data. In recent years, the pervasive threat of credit card fraud has inflicted financial harm on millions of individuals worldwide. Our project centers on the meticulous analysis of fraud patterns embedded within credit card transaction data. As a critical issue affecting both financial institutions and consumers, the urgency to address and mitigate credit card fraud has intensified. Alarming projections estimate that the cumulative impact of such fraud is poised to surpass $165 billion in the next decade. This project aims to delve into the intricacies of credit card transactions, leveraging advanced analytical techniques to identify and combat fraudulent activities, thereby contributing to the enhancement of credit card security in the face of an evolving threat landscape.

PROBLEM STATEMENT:

The increasing prevalence of credit card fraud poses a significant threat to financial institutions and consumers alike. To address this challenge, we aim to conduct a comprehensive analysis that explores the predictors and influencers of credit card spending, common trends in credit card fraud, and effective preventive measures to enhance credit card security.

KEY QUESTIONS TO BE ANSWERED:

1. Predictors/Influencers of Credit Card Spending: What are the primary factors influencing credit card spending, and how can we identify key predictors in the dataset?
2. Trends/Patterns in Credit Card Fraud: What are the prevalent trends and patterns associated with credit card fraud? Can we uncover distinctive characteristics or behaviors that indicate potential fraudulent transactions?
3. Fraud Prevention and Security Improvement: What strategies and measures can be implemented to prevent credit card fraud? How can credit card security be enhanced to mitigate the risk of unauthorized transactions?
4. Demographic Factors and Credit Card Spending: Are there significant demographic factors, such as age and gender, that predict higher credit card spending? How do these factors contribute to spending patterns?

5. Fraud Rate Variation with Demographic Factors: How does the rate of credit card fraud vary across different demographic factors, such as gender and age? Are certain groups more susceptible to fraudulent activities?
6. Occupation and Credit Card Spending Habits: What is the relationship between the occupation or job title of the cardholder and their credit card spending habits? Can we identify patterns based on occupation?
7. Regional Analysis of Credit Card Fraud: Which states exhibit the highest incidence of credit card fraud based on the provided dataset? Are there geographical factors contributing to the concentration of fraudulent activities?

Objective:

The main objective of this analysis is to gain insights into the factors influencing credit card spending, detect patterns indicative of credit card fraud, propose effective preventive measures, and explore the impact of demographic and occupational factors on spending behavior and fraud rates. The findings will contribute to the development of targeted strategies for fraud prevention and improved credit card security.

DATA SOURCE:

We took our dataset from Kaggle for Credit Card Fraud Prevention
Source Link: Kaggle
Our dataset consists of 1,048,575 total records & 22 variables.
Our dataset contains credit card transactions from Jan 1st, 2019 to Dec 31st, 2020.
This means we have data of around 1 million credit card users for a time period of 2 years.

DATA DESCRIPTION:

The dataset in question encompasses a comprehensive set of credit card transactions, providing detailed information on both legitimate and fraudulent activities. We have got imbalanced dataset. The data includes:

1. Merchant Details:

- Name
- Category
- Latitude
- Longitude
- Transaction Number

2. Transaction Details:

- Credit Card Number
- Transaction Date and Time
- Amount
- ID
- Place

3. Customer Details:

- Name
- Date of Birth
- Gender
- Address
- Job
- City Population

This structured dataset lays the foundation for a thorough analysis of credit card transactions, offering insights into merchant characteristics, transaction specifics, and customer demographics. The inclusion of fraudulent transactions allows for a targeted exploration of patterns and factors associated with potential fraud, enhancing the dataset's utility for comprehensive research and analysis in the realm of credit card security.

Data Dictionary:

1. **trans_date_trans_time :** Transaction time stamp
2. **cc_num :** Credit card number
3. **merchant :** merchant name
4. **category :** transaction category
5. **amt :** Transaction amount
6. **trans_num :** transaction number of transaction
7. **unix_time :** time in unix format
8. **merch_lat :** lattitude of the merchant
9. **merch_long :** longitude of merchant
10. **is_fraud :** nature of transaction (fraud or not fraud)
11. **dob :** date of birth of card holder
12. **first :** First name of card holder
13. **last :** Last name of card holder
14. **gender :** Sex of card holder
15. **street :** transaction address
16. **city :** transaction city
17. **state :** transaction state
18. **zip :** transaction zipcode
19. **lat :** transaction lattitude
20. **long :** transaction longitude
21. **city_pop :** Population of the city
22. **job :** job of the card holder

Here, **is_fraud** is our target variable

Missing Values – There were no missing values in the whole dataset.

```
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 22 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   cc_num            1048575 non-null  float64
 1   merchant          1048575 non-null  object
 2   category          1048575 non-null  object
 3   amt               1048575 non-null  float64
 4   gender            1048575 non-null  object
 5   street            1048575 non-null  object
 6   city              1048575 non-null  object
 7   state             1048575 non-null  object
 8   zip               1048575 non-null  int64
 9   lat               1048575 non-null  float64
 10  long              1048575 non-null  float64
 11  city_pop          1048575 non-null  int64
 12  job               1048575 non-null  object
 13  trans_num         1048575 non-null  object
 14  unix_time         1048575 non-null  int64
 15  merch_lat         1048575 non-null  float64
 16  merch_long        1048575 non-null  float64
 17  is_fraud          1048575 non-null  int64
 18  trans_hour        1048575 non-null  int64
 19  trans_day_of_week 1048575 non-null  int32
 20  trans_year_month  1048575 non-null  period[M]
 21  age               1048575 non-null  int32
dtypes: float64(6), int32(2), int64(5), object(8), period[M](1)
```

Unique Values – The number of unique values in dataset is stated below

trans_date_trans_time	476595
cc_num	943
merchant	693
category	14
amt	48602
first	348
last	479
gender	2
street	965
city	879
state	51
zip	952
lat	950
long	951
city_pop	865
job	493
dob	950
trans_num	1048575
unix_time	1030650
merch_lat	1016437
merch_long	1034825
is_fraud	2

DATA PRE-PROCESSING:

Data preprocessing is an important step in data analysis, as it helps to ensure that the data is clean and consistent, and that it is in a format that can be easily analyzed by machines. The specific steps involved in data preprocessing will vary depending on the type of data and the intended analysis.

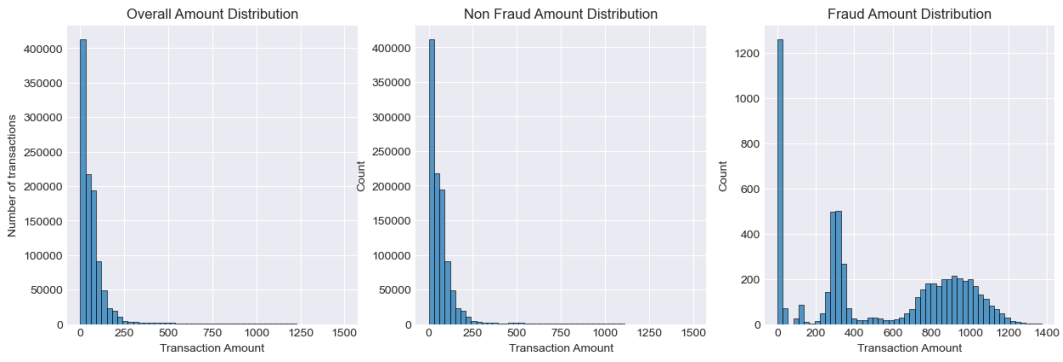
1) One common step in data preprocessing is to drop irrelevant columns. These are columns that are not needed for the analysis, and that can actually add noise and make it more difficult to analyze the data. We initiated the process by removing index columns to streamline data structure and optimize analysis efficiency.
2) Second step is to convert the trans_date_trans_time column into a datetime format. This will make it easier to analyze the data by date and time.

3) Three new columns are derived from the trans_date_trans_time column: hour, day of week, and month-year. These new features can be used to analyze the data by time of day, day of week and month.
4) We have also derived age from date of birth. This will help us to know about fraud patterns in different age groups.
5) Then we drop all the irrelevant columns for our analysis. We do this by dropping trans_date_trans_time, first name, last name, and date of birth columns as they are no longer needed.
6) Finally, we converted category, gender and is_fraud columns into categorical variables.

```
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 22 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   cc_num              1048575 non-null  float64
 1   merchant            1048575 non-null  object
 2   category            1048575 non-null  category
 3   amt                 1048575 non-null  float64
 4   gender              1048575 non-null  category
 5   street              1048575 non-null  object
 6   city                1048575 non-null  object
 7   state               1048575 non-null  object
 8   zip                 1048575 non-null  int64
 9   lat                 1048575 non-null  float64
 10  long                1048575 non-null  float64
 11  city_pop            1048575 non-null  int64
 12  job                 1048575 non-null  object
 13  trans_num           1048575 non-null  object
 14  unix_time           1048575 non-null  int64
 15  merch_lat           1048575 non-null  float64
 16  merch_long          1048575 non-null  float64
 17  is_fraud            1048575 non-null  category
 18  trans_hour          1048575 non-null  int64
 19  trans_day_of_week   1048575 non-null  int32
 20  trans_year_month    1048575 non-null  period[M]
 21  age                 1048575 non-null  int32
```

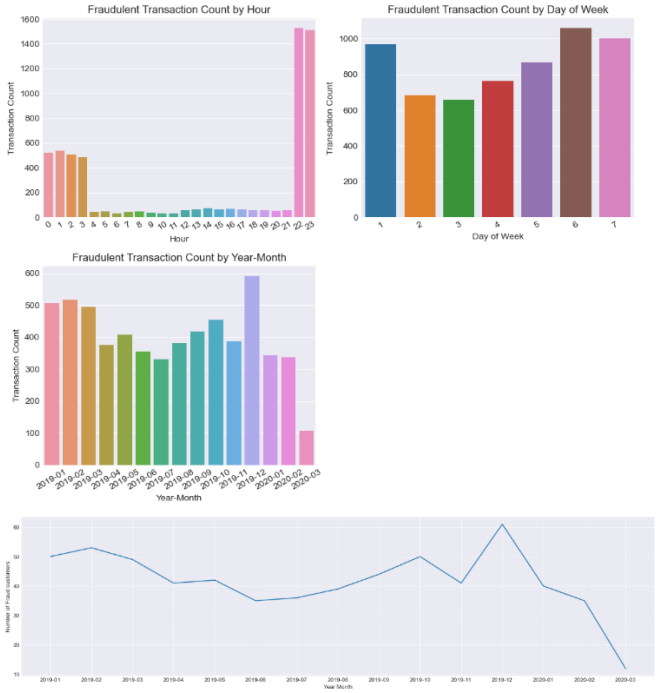
Data pre-processing is an important step in the data analysis process, and it can make a big difference in the quality of your results. By taking the time to cleaning & preparing data, we can ensure that we are getting most accurate and reliable results possible.

Amount ($)



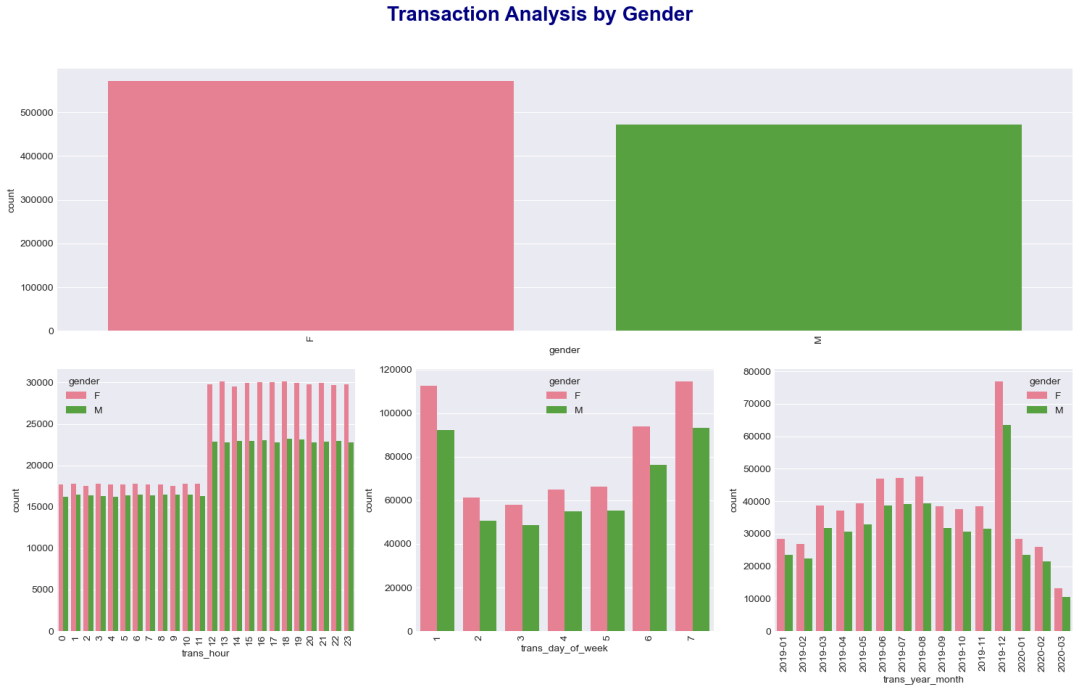
The charts above represent the distribution of dollar amounts spent for the total distribution, non-fraud distribution and fraud distribution. In our total distribution, we had 1,048,575 total records, with roughly 99.43% being non-fraud transactions and .57% being fraud transactions (6,006 records). In the total distribution chart, we can notice that a huge majority of the transactions are roughly $250 or less. The same applies for the distribution of non-fraud transactions. As for the distribution of fraud transactions, the chart illustrates a great majority of transactions that are roughly $250 or above. From this, we were able to conclude that if someone were to see a transaction for $250 or more on their credit card statement, then it should call for more attention as there is a higher risk for fraud, as opposed to if the transaction was below $250.

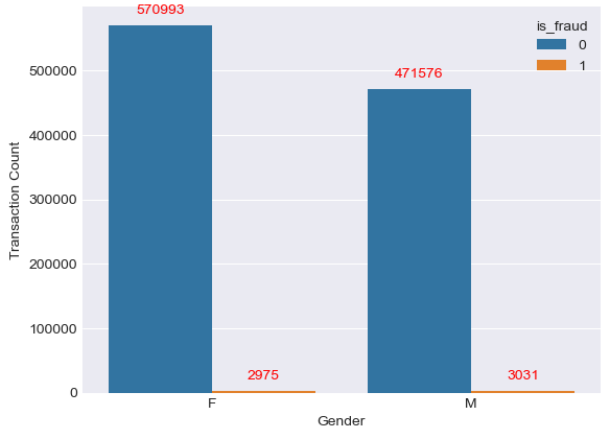
Time Increments (Hour, Day, Month)



The following charts illustrate the relationship between fraudulent transactions and different increments of time (hour, day, month) for the previous two years. From our charts, we were able to notice that out of our 6,006 records of fraud many of them occurred overnight (10pm-4am). In addition, we also found that there was slightly higher fraud counts over the weekend, including Friday, Saturday, and Sunday. In terms of which months had the most fraud, the month of December 2019 had the highest fraud count, but it did not have the highest rate of fraud. The month of February 2019 had the highest rate of fraud with 1.03%, followed by January 2019 with .96% rate of fraud. Based on our findings, it seems that people need to be more cautious when making purchases overnight and over the weekend. Also, people seem to be victims of fraud at the beginning of the year, compared to other months.

Gender



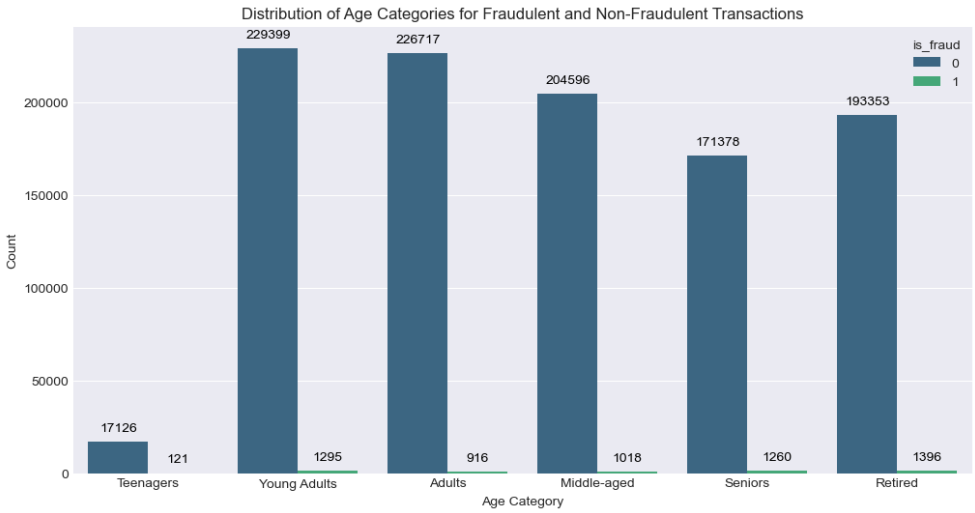


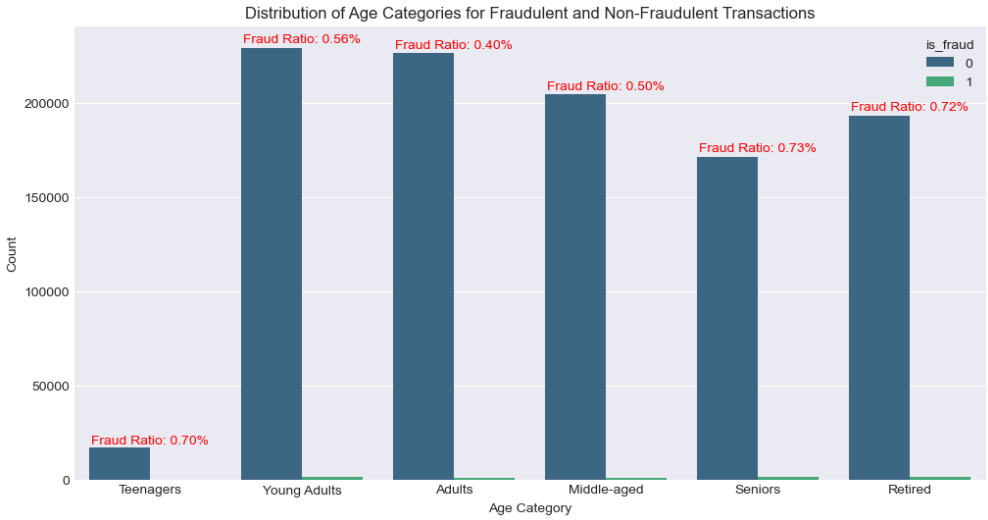
The next series of charts display the distribution of our transactions across gender, along with the distribution of our transaction across gender and increments of time (hour, day, month). For our transaction distribution, females made up 54.45% of our total transactions and males made up the remaining 45.55%. For our fraud transaction distribution across gender, we found that females made up 49.53% of our total fraud count, while males made up a higher percentage at 50.47%, regardless of having roughly 100,000 fewer transactions than female. With further investigation, we found that males had a .64% rate of being a victim of fraud, compared to females with a rate of .52%. In other words, this means that males were roughly 23% more likely to be a victim of fraud compared to females.

For our distribution breakdown of gender and our selected time increments, we noticed that they followed a similar pattern in transaction activity. Both females and males participated in more credit card

transactions in the second half of the day (12pm -12am), have more transactions over the weekend (Fri-Sun) and December was both their highest transaction count month.

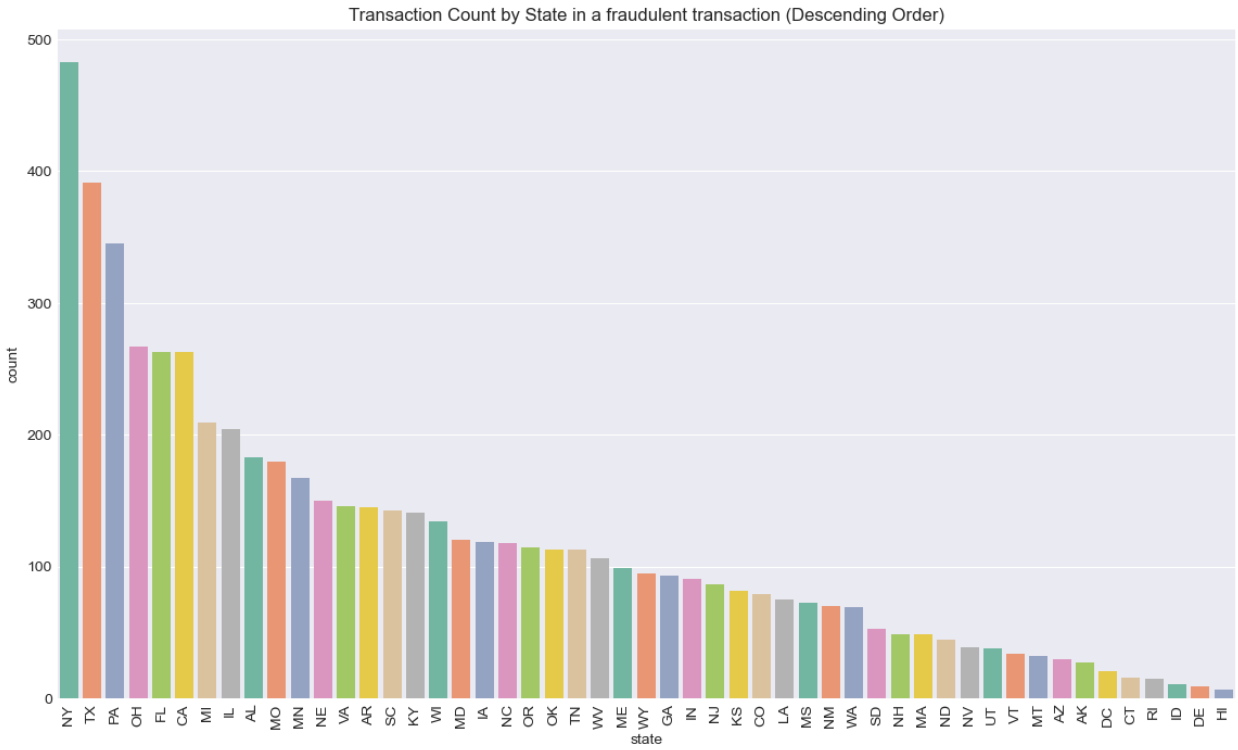
Age





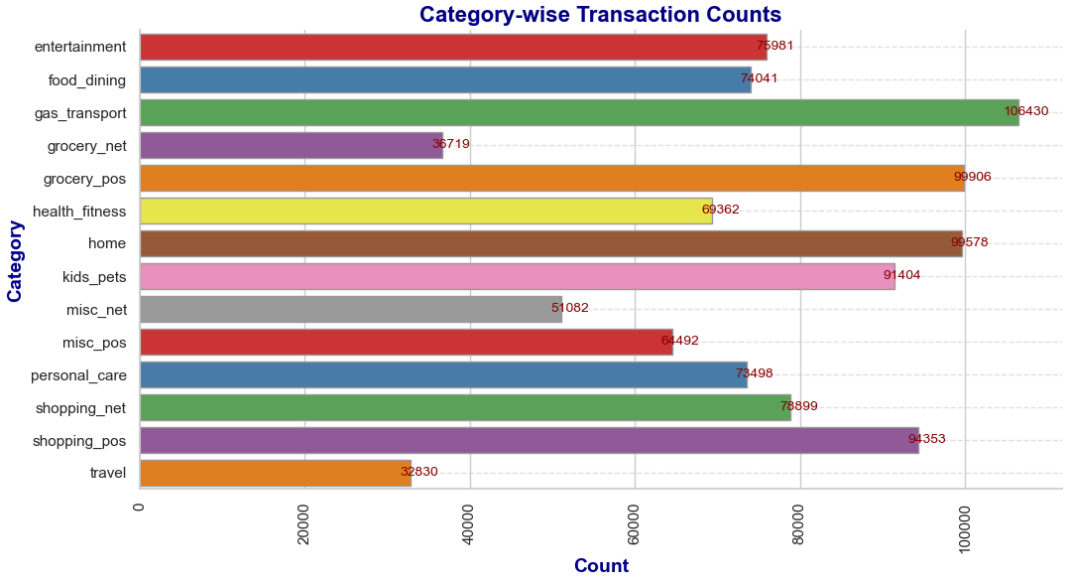
The next series of visuals displays our distribution of fraud and non-fraudulent transactions across different age groups. We have broken up the age groups into teenagers (13-18), young adults (19-31), adults (32-41), middle-aged (42-49), seniors (50-61) and retired (62 and up). Based off our data, we found that the retired group had the highest count for fraudulent transaction and teenagers had the lowest count. When calculating the fraud rates, we found that the age groups with the highest rate of fraud were the senior group (.73%) and the retired group (.72%). Although teenagers made up the smallest count of fraud transactions, they were in a close third with a rate of .70%. The group with the lowest rate of fraud was the adult group (.40%). When comparing the senior group to the adult group, we found that seniors were approximately 82% more likely to be victims of fraud compared to adults.

State



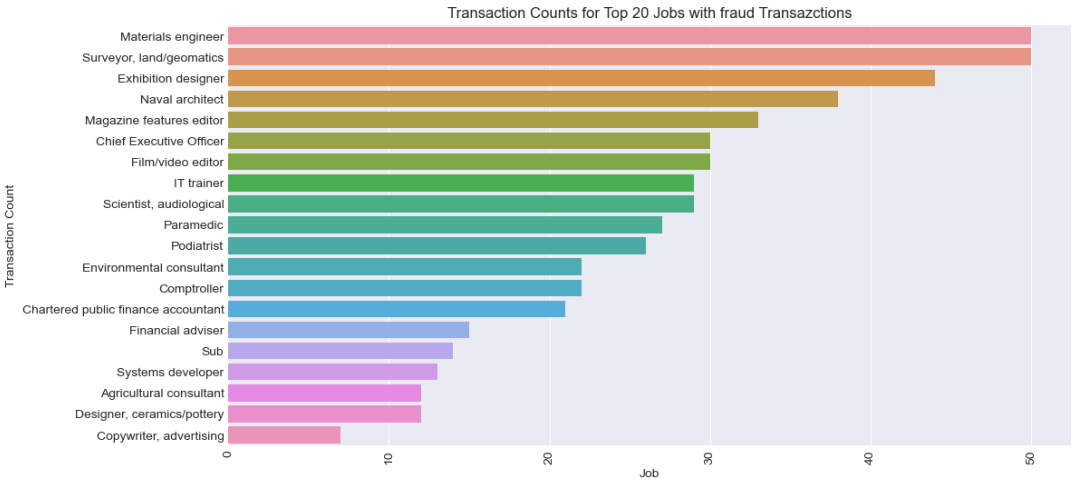
In the following chart we can see the distribution count of fraud records across all fifty states in the US. We can see that New York has the highest count of fraud transaction out of the fifty states and Hawaii has the lowest count. When conducting our ratios, we were able to find that the states with the highest rate of fraud were Delaware (100%), Rhode Island (33.63%), Alaska (1.58%), Nevada (.86%) and Tennessee (.80%). Although Delaware and Island had the highest rates of fraud, the sample size was relatively low compared to the other states, so it is difficult to conclude if these ratios are accurate. The states that had the lowest rate of fraud were Arizona (.34%), Hawaii (.34%), Montana (.34%), Connecticut (.26%) and Idaho (.24%). When comparing the fraud rate of Alaska and Idaho, we found that people in Alaska were roughly 6.5 times more likely to be victims of fraud compared to Idaho.

Item Category



The following charts display the distribution of our total transactions across different item categories. From our chart, we found that gas_transport had the highest count of total transactions in our dataset. In terms of the item categories that had the highest fraud count, grocery_pos (1396) and shopping_net (1375) had the highest counts. Grocery_pos is represented as grocery shopping offline or in-person and shopping_net is represented by online shopping. When conducting our ratios, we were able to find that the item category with the highest rates of fraud were shopping_net (1.74%) and misc_net (1.45%). Misc_net is represented by items that were shopped online and miscellaneous from our other categories. The item categories with the lowest rate of fraud were health_fitness (.15%) and home (.15%). In addition, we compared the fraud rates of online shopping (1.74%) and in-person (.70%) and found that online shopping was approximately 2.5 times riskier than in-person shopping. Based on our ratios, we concluded that online shopping is significantly riskier than shopping in person.

Job Occupation



	job	is_fraud	Transaction count	job_count	Transaction percentage
403	Forest/woodland manager	1	9	9	100.000000
123	Careers adviser	1	15	15	100.000000
843	Ship broker	1	7	7	100.000000
853	Solicitor	1	11	11	100.000000
41	Air traffic controller	1	8	8	100.000000
811	Sales promotion account executive	1	14	14	100.000000
235	Dancer	1	19	19	100.000000
455	Homeopath	1	11	11	100.000000
535	Legal secretary	1	12	12	100.000000
71	Armed forces technical officer	1	8	8	100.000000
345	Engineer, site	1	12	12	100.000000
103	Broadcast journalist	1	9	9	100.000000
213	Contracting civil engineer	1	7	7	100.000000
491	Information officer	1	8	8	100.000000
487	Industrial buyer	1	10	10	100.000000

The chart and table above display the fraud tranaction count across all the job occupations listed in our dataset. The chart represents the job occupations with the highest fraud transaction count, which include materials engineer and surveyor with the highest count, at 50 a piece. The table represents the job occupations that had a 100% fraud rate from their transaction record. We found 15 total job occupations with a 100% fraud rate, indicating that these occupations may have an underlying issue with their credit card information and job.

ANALYSIS & RESULTS –

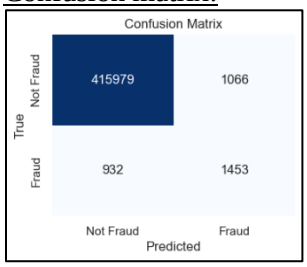
Decision Tree:

We chose to use the Classification Tree model to predict the fraud transactions in the given dataset.

```
Decision Tree Accuracy: 0.9952363922466204

Decision Tree Classifier:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    417045
           1       0.58      0.61      0.59      2385

    accuracy                           1.00    419430
   macro avg       0.79      0.80      0.80    419430
weighted avg       1.00      1.00      1.00    419430
```

The Decision Tree model did really well with an overall accuracy of 99.52%, showing that it's good at getting things right. But when it comes to the minority class (class 1), the model struggles a bit. It has lower precision, recall, and F1-score for this class, meaning it has some difficulty accurately finding positive cases. While the model performs strongly on the majority class (class 0), it needs some work on handling the minority class better to make its predictions even more reliable overall.

Confusion matrix:



Predicting True Positives in %: **60.92%**

Since we got an imbalance Dataset, we need to train our model using Under-Sampling technique.

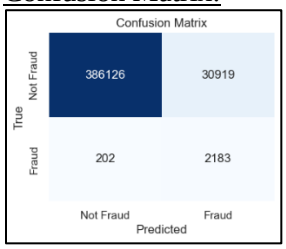
Decision Tree Classifier (Using the Under Sampling Method):

We choose to use the Classification Tree model to predict the Fraud Transactions with under sampling method.

```
Decision Tree Accuracy: 0.9258016832367737

Decision Tree Classifier:
              precision    recall  f1-score   support

           0       1.00      0.93      0.96    417045
           1       0.07      0.92      0.12      2385

    accuracy                           0.93    419430
   macro avg       0.53      0.92      0.54    419430
weighted avg       0.99      0.93      0.96    419430
```

The Decision Tree model performs well overall with a high accuracy of 92.58%, indicating good general performance. However, there's a noticeable trade-off between precision and recall, particularly for the minority class (class 1). While the model effectively captures positive instances with high recall for class 1, it sacrifices precision, leading to a lower F1-score. To improve the model's overall performance, it might be beneficial to explore further optimization strategies that balance this trade-off.

Confusion Matrix:



Predicting True Positives in %: **91.53%**

The first Decision Tree model had a high overall accuracy of 99.52% but struggled to detect fraud (class 1). After applying Under-Sampling, the second model achieved 92.58% accuracy with a notable improvement in identifying fraud cases (91.53% true positives). Although overall accuracy slightly decreased, the trade-off resulted in a more effective fraud detection model. Further optimization is recommended for a well-balanced performance.

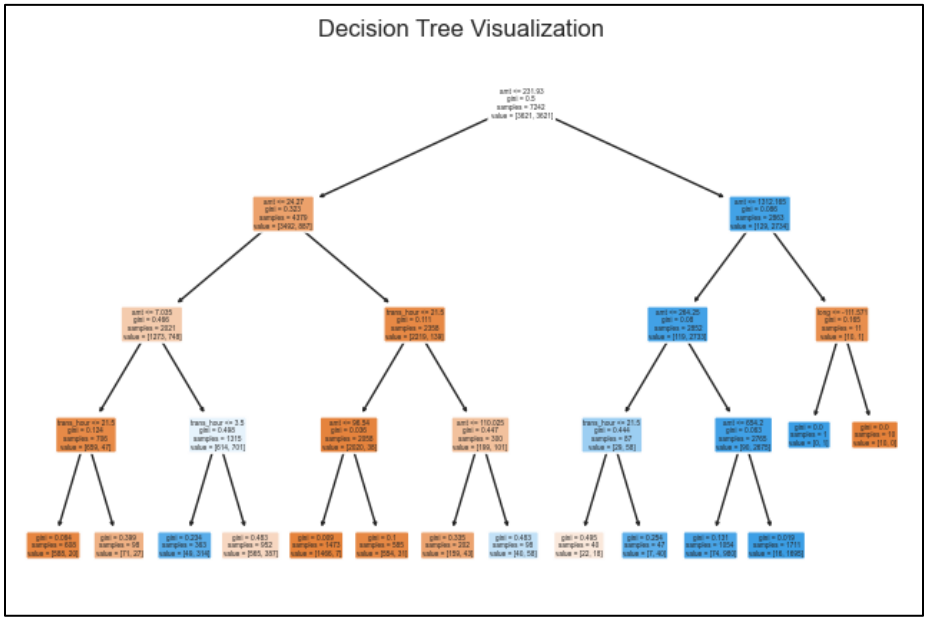
Feature importance analysis -

```
Feature Importance:
             Feature  Importance
0                amt    0.737863
2         trans_hour    0.125705
3                age    0.043285
1           city_pop    0.034438
5                lat    0.022222
6               long    0.021386
7  trans_day_of_week    0.010559
4           gender_M    0.004543

Accuracy: 0.9258016832367737
```

The analysis of feature importance reveals that the 'amt' (transaction amount) significantly influences the Decision Tree model, contributing a substantial 73.79% to its decision-making process. Other impactful

features include 'trans_hour,' 'age,' and 'city_pop.' With an overall accuracy of 92.58%, the model demonstrates its capability to predict the target variable. Exploring additional insights into feature importance can offer guidance for potential optimizations to enhance the model's performance.

Data Visualization for the decision tree:



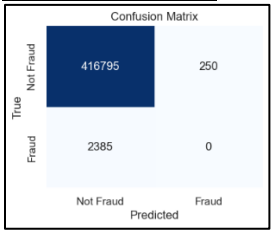
Logistic Regression:

Then we used Logistic Regression model to predict the Fraud Transactions.

```
Logistic Regression Accuracy: 0.9937176644493718

Logistic Regression Classifier:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00    417045
           1       0.00      0.00      0.00      2385

    accuracy                           0.99    419430
   macro avg       0.50      0.50      0.50    419430
weighted avg       0.99      0.99      0.99    419430
```

The Logistic Regression model demonstrates robust predictive performance with a high overall accuracy of 99.37%. Despite this, challenges arise in accurately identifying the minority class (class 1), evident from the low precision, recall, and F1-score for this class. These findings point towards a potential imbalance issue, underscoring the importance of conducting further evaluation and considering adjustments to improve the model's capacity to correctly capTrue Positive instances.

Confusion matrix:



Predicting True Positives in %: 0%

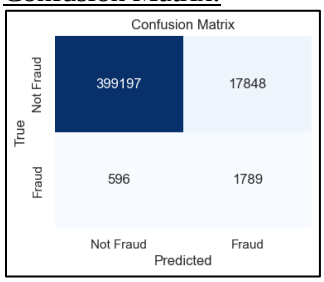
Logistic Regression (Using the Under Sampling Method):

We choose to use the Logistic Regression model to predict the Fraud Transactions with under sampling method.

```
Logistic Regression Accuracy: 0.9560260353336671

Logistic Regression Classifier:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98    417045
           1       0.09      0.75      0.16      2385

    accuracy                           0.96    419430
   macro avg       0.54      0.85      0.57    419430
weighted avg       0.99      0.96      0.97    419430
```

The Logistic Regression model demonstrates a commendable overall accuracy of 95.60%, indicating effective general predictive performance. However, the model faces challenges in achieving a balance between precision and recall for the minority class (class 1), resulting in a lower F1-score. Despite the imbalanced nature of the dataset, the model shows potential, with notable improvements in correctly identifying positive instances compared to the previous scenario.

Confusion Matrix:



Predicting True Positives in %: **75.01%**

The initial Logistic Regression model had a high accuracy of 99.37% but struggled with the minority class (class 1), indicating an imbalance issue. After applying Under-Sampling, the model achieved 95.60% accuracy with a significant boost in correctly identifying fraud cases (75.01% true positives). Despite challenges in balancing precision and recall, the model showed potential for improved fraud detection, highlighting progress from the initial scenario. Further refinements are suggested for optimization.

```
Coefficients:
             Feature  Coefficient
6               long     0.010344
0                amt     0.006554
1           city_pop    -0.000000
4           gender_M    -0.000052
7  trans_day_of_week    -0.000463
2         trans_hour    -0.001576
5                lat    -0.004466
3                age    -0.004940

Intercept:
[-0.00011527]

Accuracy: 0.9560260353336671
```

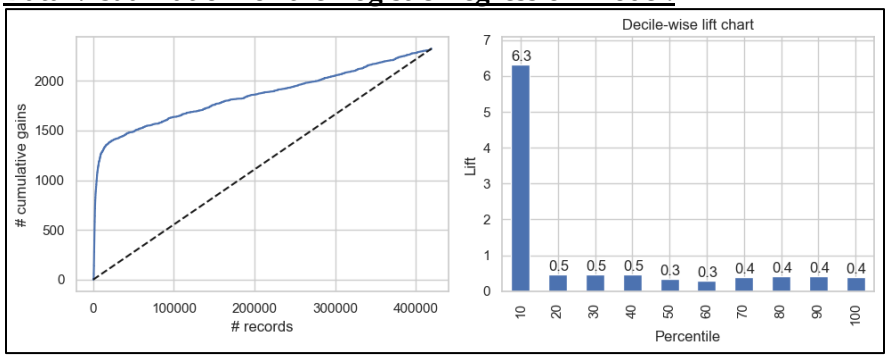
The coefficients of the Logistic Regression model indicate that 'long' and 'amt' positively impact the target variable, whereas 'city_pop,' 'gender_M,' 'trans_day_of_week,' 'trans_hour,' 'lat,' and 'age' exert negative

influences. With an accuracy of 95.60%, the model demonstrates its proficiency in making accurate predictions. The non-zero coefficients underscore the significance of 'long' and 'amt' in shaping the model's decision-making process.

Probability:

```
cutoff = 0.7
        actual     p(0)     p(1)  predicted  predicted_class
785452       0 0.995496 0.004504          0                0
709108       0 0.992586 0.007414          0                0
466450       0 0.994897 0.005103          0                0
657287       0 0.992503 0.007497          0                0
15941        0 0.994347 0.005653          0                0
```

Data Visualization for the Logistic Regression Model:



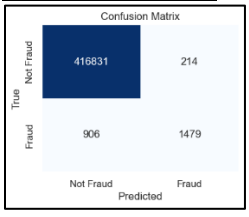
XG Boost Model:

The last model we used to predict fraud transactions is XG Boost model

```
Accuracy: 0.9973297093674749

XGBoost Classifier:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    417045
           1       0.87      0.62      0.73      2385

    accuracy                           1.00    419430
   macro avg       0.94      0.81      0.86    419430
weighted avg       1.00      1.00      1.00    419430
```

The XGBoost Classifier displays outstanding overall accuracy of 99.73%, showcasing its robust predictive capabilities. It excels in achieving perfect precision and recall for the majority class (class 0). Additionally, the model performs well in identifying positive instances (class 1) with a high precision of 87% and a moderate recall of 62%, resulting in a balanced F1-score. The model's strong classification performance across both classes highlights its effectiveness, especially in handling imbalanced datasets.

Confusion matrix:



Predicting True Positives in %: **62.01%**

Since we got an imbalance Dataset, we need to train our model using **Under-Sampling** technique.

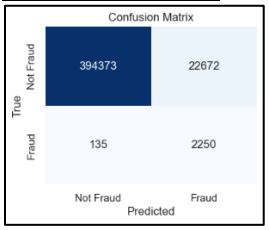
XG Boost (Using the Under Sampling Method):

We choose to use the XG Boost model to predict the Fraud Transactions with under sampling method.

```
Accuracy: 0.9456238228071431

XGBoost Classifier:
              precision    recall  f1-score   support

           0       1.00      0.95      0.97    417045
           1       0.09      0.94      0.16      2385

    accuracy                           0.95    419430
   macro avg       0.54      0.94      0.57    419430
weighted avg       0.99      0.95      0.97    419430
```

The XGBoost Classifier attains a noteworthy accuracy of 94.56%, showcasing commendable overall predictive performance. Nevertheless, difficulties arise in maintaining a balance between precision and recall for the minority class (class 1), leading to a reduced F1-score. While the model excels in correctly identifying positive instances with a high recall, it comes at the cost of precision. Fine-tuning is recommended to enhance the model's capacity to effectively classify both classes.

Confusion Matrix:



Predicting True Positives in %: **94.56%.**

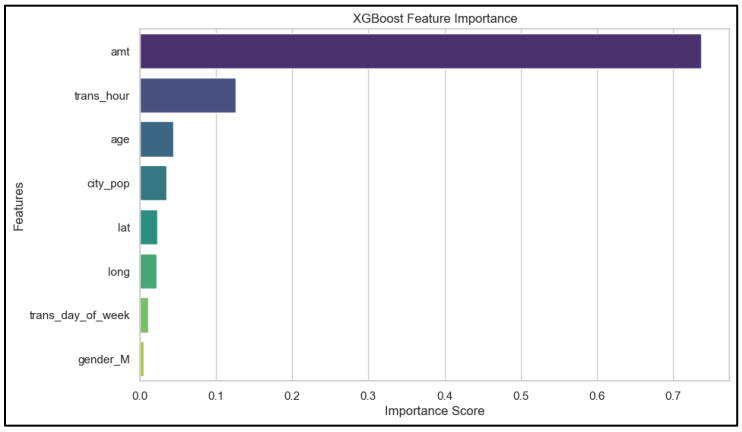
The first XGBoost model performed exceptionally well with an accuracy of 99.73%, showing it can predict both normal and fraudulent transactions accurately. After applying under-sampling to address imbalances, the model still achieved a strong accuracy of 94.56%. It excelled in catching fraud (94.33% accuracy), but there's room to improve how it identifies all instances of fraud while keeping mistakes to a minimum. Fine-tuning the model could enhance its effectiveness in handling both normal and fraudulent cases.

```
Feature Importance:
             Feature  Importance
0                amt    0.536954
2         trans_hour    0.211833
3                age    0.055626
4          gender_M    0.047788
1          city_pop    0.042167
5                lat    0.038829
7  trans_day_of_week    0.033734
6               long    0.033070

Accuracy: 0.9456238228071431
```

The analysis of the XGBoost Classifier emphasizes that the most impactful feature is 'amt' (transaction amount), contributing more than 53.7% to the model's decision-making process. Other notable features include 'trans_hour,' 'age,' 'gender_M,' and 'city_pop.' Despite achieving a high overall accuracy of

94.56%, delving into the importance of features can provide valuable insights for guiding additional optimization efforts aimed at enhancing predictive performance.

Data Visualization for the XG Boost:



Initial Model Regression accuracy on validation data:

	Models	ACC	True Positives (%)
0	Logistic Regression	99.371766	0.000000
1	Decision Tree classifier	99.523639	60.922432
2	XG boost	99.732971	62.012579

The Logistic Regression model achieved a high overall accuracy of 99.37%, but it didn't predict any True Positives. The Decision Tree classifier demonstrated an accuracy of 99.52%, with a True Positives rate of 60.92%. The XG Boost model performed well with a 99.73% accuracy, and it predicted True Positives at a rate of 62.01%.

	Models	ACC	True Positives (%)
0	Logistic Regression	95.602604	75.010482
1	Decision Tree classifier	92.580168	91.530398
2	XG boost	94.562382	94.339623

The Logistic Regression model achieved an accuracy of 95.60% with a True Positives rate of 75.01%. The Decision Tree classifier displayed an accuracy of 92.58% and excelled in predicting True Positives at a rate of 91.53%. The XG Boost model performed well with a 94.56% accuracy, demonstrating a high True Positives rate of 94.34%. Given the particular goals and priorities of our aim to detect credit card fraud, opting for XG Boost as the preferred model appears justified. This underscores the significance of assessing models using various metrics rather than relying solely on overall accuracy.

What did we learn?

Our analysis focused on identifying fraudulent patterns within the dataset. By employing various analytical techniques, we gained valuable insights into the diverse strategies employed by individuals engaging in fraudulent activities.

Challenges of Imbalanced Datasets:

One key challenge we encountered was the imbalance in the dataset, were instances of regular transactions significantly outnumbered instances of fraud. This imbalance posed a risk of our model being overly optimistic in its accuracy assessment, potentially overlooking instances of fraud due to the dominance of regular transactions.

Addressing Imbalance with Under Sampling:

To counteract the dataset imbalance, we implemented an under-sampling technique. This involved deliberately reducing the number of instances of regular transactions in our analysis. While this approach did impact the overall accuracy metric, it substantially improved our model's ability to detect and understand patterns associated with fraud. This trade-off between accuracy and improved class balance was a strategic decision to enhance the model's fraud detection capabilities.

Performance Evaluation of Models:

We experimented with three different models to evaluate their effectiveness in detecting fraud patterns. Notably, XG Boost emerged as the most successful among them, achieving a commendable accuracy rate of 94.56%. This high accuracy suggests that XG Boost is a robust model for identifying and predicting fraudulent transactions within the dataset.

Recommendations for Future Work:

Moving forward, it is imperative to continue refining and expanding our models to handle evolving fraud patterns. Additionally, exploring other techniques for addressing imbalanced datasets, such as oversampling or more advanced algorithms, may contribute to further improvements in both accuracy and the ability to capture nuanced fraud behaviors.

In conclusion, our findings underscore the significance of addressing imbalances in fraud detection datasets and highlight the effectiveness of XG Boost as a promising model for enhancing accuracy in identifying fraudulent transactions. Continued research and refinement in this area will be essential to stay ahead of emerging fraud patterns in the dynamic landscape of financial transactions.

Conclusion and Recommendations

After conducting our research and analysis, we created the following list of recommendations to combat the ongoing issue of credit card fraud. Our recommendations include:

1. Caution with Credit Card Info:

- People over 50, be careful sharing credit card details. Only share with trusted sources to avoid scams.

2. Regular Statement Checks:

- Check credit card statements often. Look for any big or strange charges to catch problems early.

3. Transaction Alerts:

- Sign up for alerts. Get messages for every transaction. Quick notifications help stop any unauthorized purchases.

4. Set Payment Limits:

- Talk to your bank. Set limits on how much money can be spent in one go. It adds extra protection.

5. Secure Online Shopping:

- Only shop on secure websites. Check for "https://" to make sure your info is safe.

6. Avoid Public Wi-Fi:

- Don’t use public Wi-Fi for transactions. It's safer to shop or check your accounts at home.

7. Use Strong Passwords:

- Create strong, unique passwords. Mix letters, numbers, and symbols for better online security.

8. Keep Cards Safe:

- Keep your cards in a safe place. Avoid leaving them where others can easily access them.

9. Report Lost Cards Immediately:

- If a card is lost, report it right away. This helps prevent unauthorized use.

10. Educate Yourself:

- Stay informed about common scams. Knowledge is a strong defense against fraud.