Application of Machine Learning Models to improve Fraud Detection

Exploratory Data Analysis & Machine Learning using financial data.

Department of Business, Government & Economics

Seattle Pacific University

Seattle, USA

ABSTRACT

Credit Card fraud is a major concern and issue in banking. Hence, fraud detection becomes extraordinarily relevant for financial institutions to better serve their customers and reduce bank losses. In this paper, with the help of RStudio, we have analyzed a large credit card transaction dataset from a European bank, implemented some machine learning algorithms, to better identify and detect credit card fraud. We suggest implementing such machine learning algorithms to enable banks to reduce fraudulent transactions.

1. INTRODUCTION

It is vital for banks to develop cutting-edge techniques and risk management strategies to be able to quickly identify fraudulent transactions across diverse platforms. Applying adaptive and predictive analytics can create fraudulent transaction risk scores, enabling real-time monitoring and rejecting fraudulent transactions. The banking sector has prioritized cybersecurity highly to reinforce credibility and trust toward customers. There are several factors that demonstrate the necessity of cyber defenses; First, everyone seems to be using digital payment methods like credit or debit cards. Having reliable safeguards in place might protect confidential data. Second, data breaches may have devastating consequences for a bank. If the breach has been caused by an outdated security system, surely it will lead its customers to move their business base elsewhere. Third, when a bank's data is compromised, the customer loses time and money. Fourth, financial institutions retain valuable personal data, thus, there is a higher chance that it could be compromised if not safeguarded correctly.

2. ARTIFICIAL INTELLIGENCE APPLIED TO FRAUD DETECTION

Many financial institutions use rules-based systems to identify credit card transactions – pre-programmed rules to identify changes in behavior or predict outcomes - with manual evaluations to detect fraud. But with an ever-growing number of transactions and an increasing number of techniques used by credit card fraudsters, applying machine learning might be a solution. In fact, an ML model analyzes huge sets of data using algorithm complexity – a measure of how long algorithms take to complete a task given an input of size N - to identify patterns. It all stems from seeing fraudulent transactions showing patterns differently from genuine ones. Therefore, the algorithm detects fraudulent activity faster and more accurately than rigid rule-based systems, because it is more scalable. While humans unknowingly overlook valuable information, AI can be trained to analyze even the most seemingly unrelated piece of information to find a hidden pattern. It all starts by gathering and categorizing as much historical data as possible. Then, the training data is used to teach the model how to predict whether a certain customer or transaction is fraudulent or not. For any ML program to be successful, it requires to have as much fraudulent data as possible, so to feed the algorithm with a lot of references to learn from. Once the training is completed, the model becomes specific to the business and can be considered ready to use in a bank's fraud management framework. Unfortunately, the model must be updated frequently to keep meeting the institution's standards. Why should an institution implement an ML model? AI can offer a plethora of benefits; **Speed**, the model may evaluate vast amount of data in fraction of a second. Moreover, it can perpetually collect and analyze new data in real time. **Efficiency**, it can perform repetitive tasks and detect subtle changes in patterns across huge datasets. Also, it may analyze hundreds of thousands of payments per second, way more than several financial analysts can do. **Scalability**, as the number of transactions increases over time, the pressure goes along. Thus, additional costs and time are spent on analysis. With an algorithm, the more data the better. It improves its performance as more data comes in, enabling it to detect fraud faster and with more efficiency. **Accuracy**, a model can be trained to detect patterns across seemingly insignificant data or identify non-intuitive trends which would be hard, or perhaps impossible, for analysts to snatch. As a result, there will be fewer false positives and frauds that go undetected.

3. DATASET

For the purposes of this research, I used a synthetic dataset – information manufactured artificially rather than generated by rea-world event - by BankSim. An agent-based simulator of bank payments founded on a sample of aggregated transactional data provided by a bank in Europe. The goal would be to model relevant scenarios that combine normal payments and injected fraudulent signatures. It contains no personal information or disclosure of legal and private customer transactions. Thus, it is easy to be shared by academia to develop and reason about fraud detection methods. In addition, synthetic data has the benefit of being easier to collect, faster and cheap for experimentation.

To create the dataset, the simulator has been run for approximately six months, iterated, and calibrated to being reliable for testing. Consequently, log files have been scrutinized and selected. Virtual Credit Card thieves have been injected who aim to steal an average of three cards per step and perform about two fraudulent transactions per day. The results are 594643 records collected, where 587443 are normal payments and 7200 fraudulent transactions.

4. DATASET FEATURES

Each record in the dataset contains the following fields:

DAYs: it represents the day when the transaction happened. There are a total of 180 steps, so the data runs for 180 days – variable removed from the dataset.

CUSTOMER: the unique ID of the person who initialized the transaction. It is formed by the letter C, followed by a unique sequence of 10 numbers. There is a total of 4100 customers in the dataset.

AGE: the variable is split into intervals, starting from 0 to 6 and the letter U which stands for Unknown – only for gender equal to Enterprise. The intervals are:

- 0: younger than 18 years old.
- 1: between 19 and 25 years old.
- 2: between 26 and 35 years old.
- 3: between 36 and 45 years old.
- 4: between 46 and 55 years old.
- 5: between 56 and 65 years old.
- 6: older than 65 years old.

GENDER: F for female, M for male, E for enterprise, U for unknown. The last group has 170 customers aged from 19 and 45 years old.

MERCHANT: The unique ID of the party which receives the transaction. The sequence is formed by the letter M, followed by a series of 9 numbers. A total of 50 merchants are recorded in the dataset.

CATEGORY: There are 15 unique categories that label the general type of the transaction – *Transportation, Food, Health, Wellness & Beauty, Fashion, Bars & Restaurants, Hyper, Sports & Toys, Tech, Home, Hotel Services, Miscellaneous, Contents, Travel, and Leisure.*

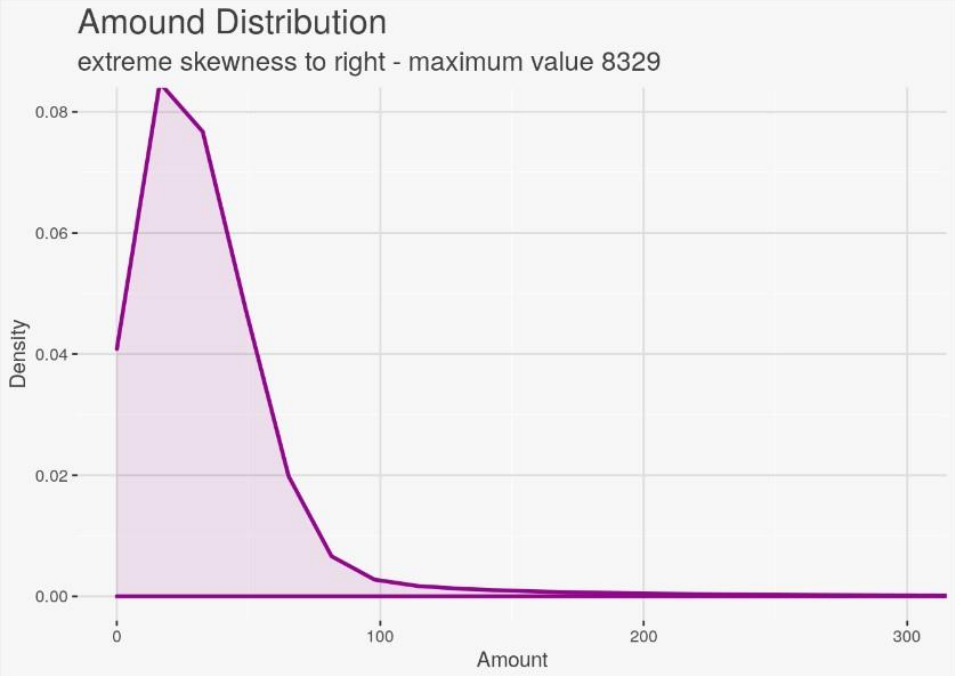
AMOUNT: The value of the transaction.

FRAUD: A flag column coded with 0 if the transaction is benign and with 1 if the transaction is fraudulent.

5. EXPLORATORY DATA ANALYSIS

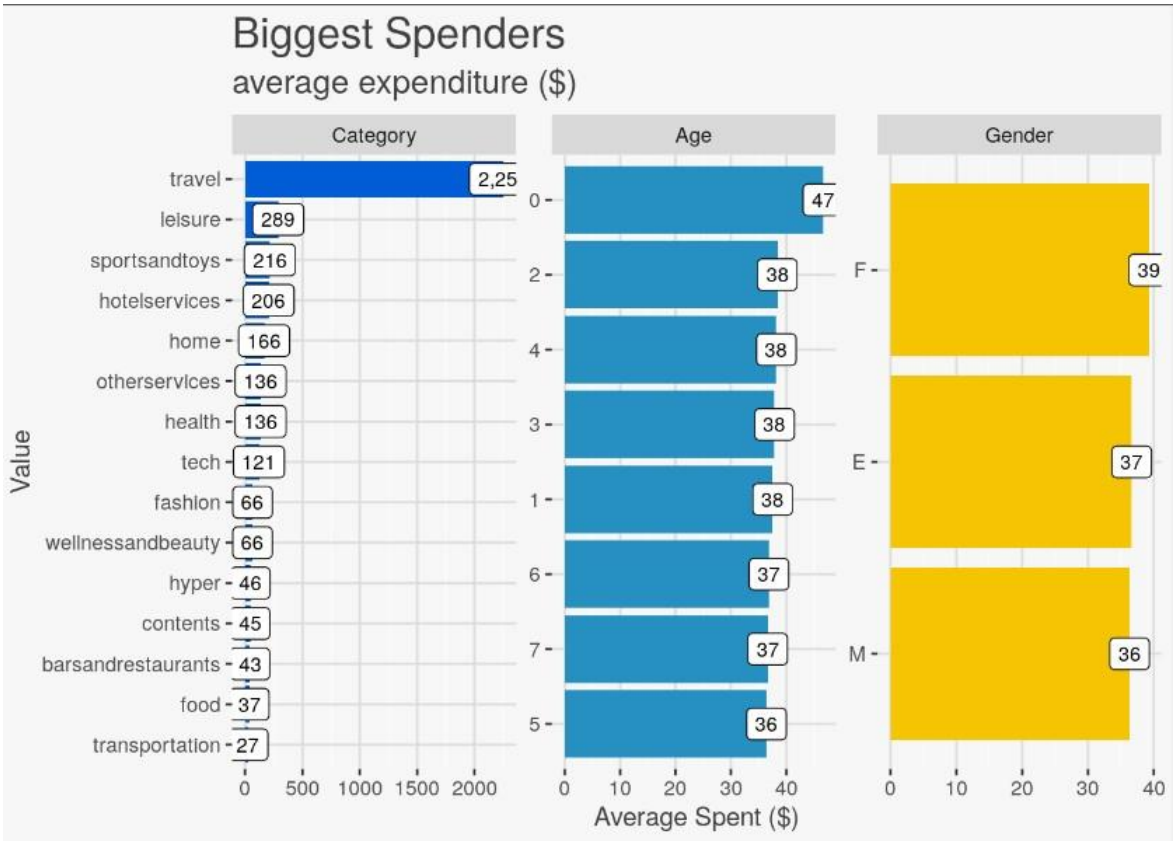
6.1.1 – AMOUNT DISTRIBUTION

We first looked at the amount field. The amount distribution of the transactions in the dataset is extremely skewed to the right, with more than 95% of the transactions having value between $1 and $100, but with the rest reaching extremely high values, with the maximum being more than $8000.



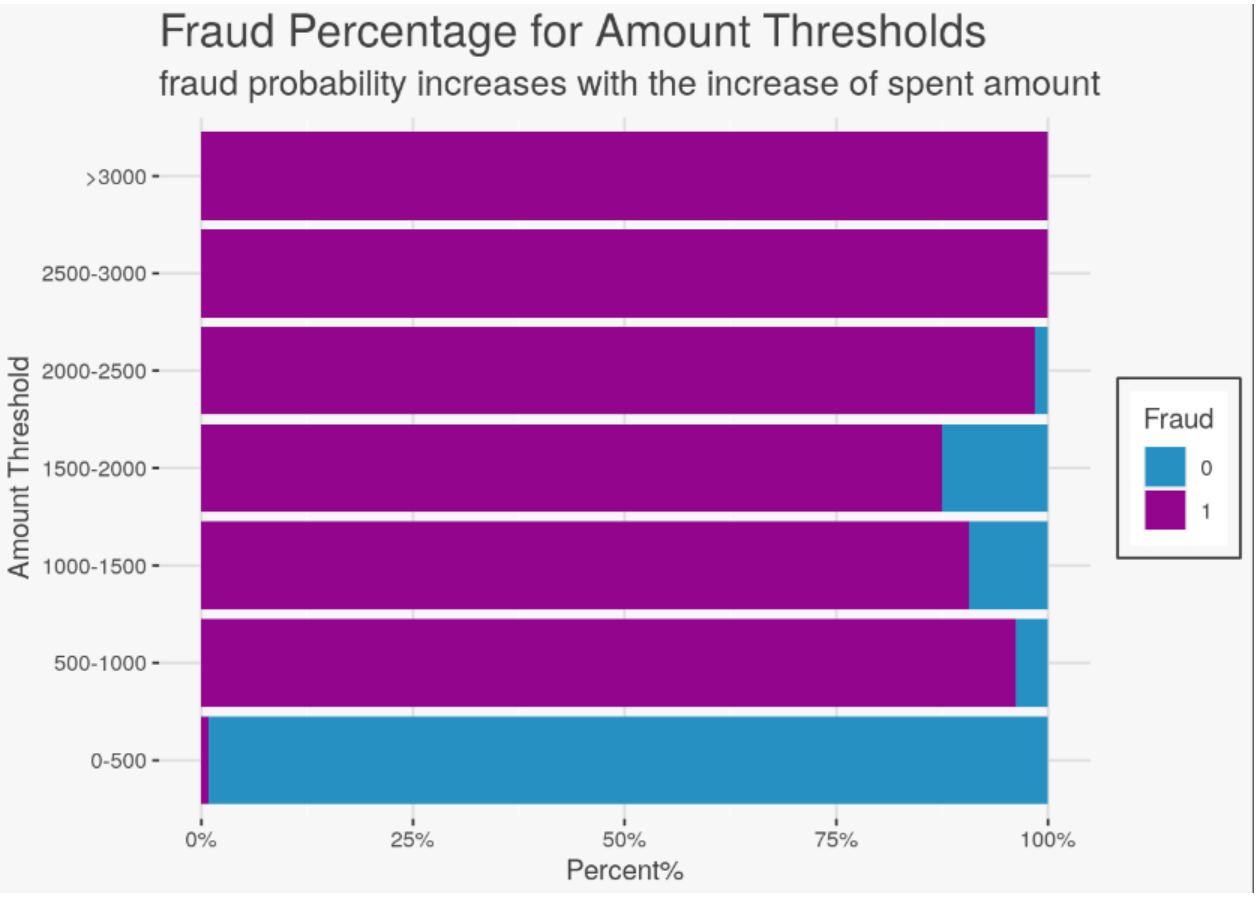
6.1.2 – BIGGER SPENDERS

Then we looked at the amount of field and compared it to the category, age, and gender fields to come up with the following results:

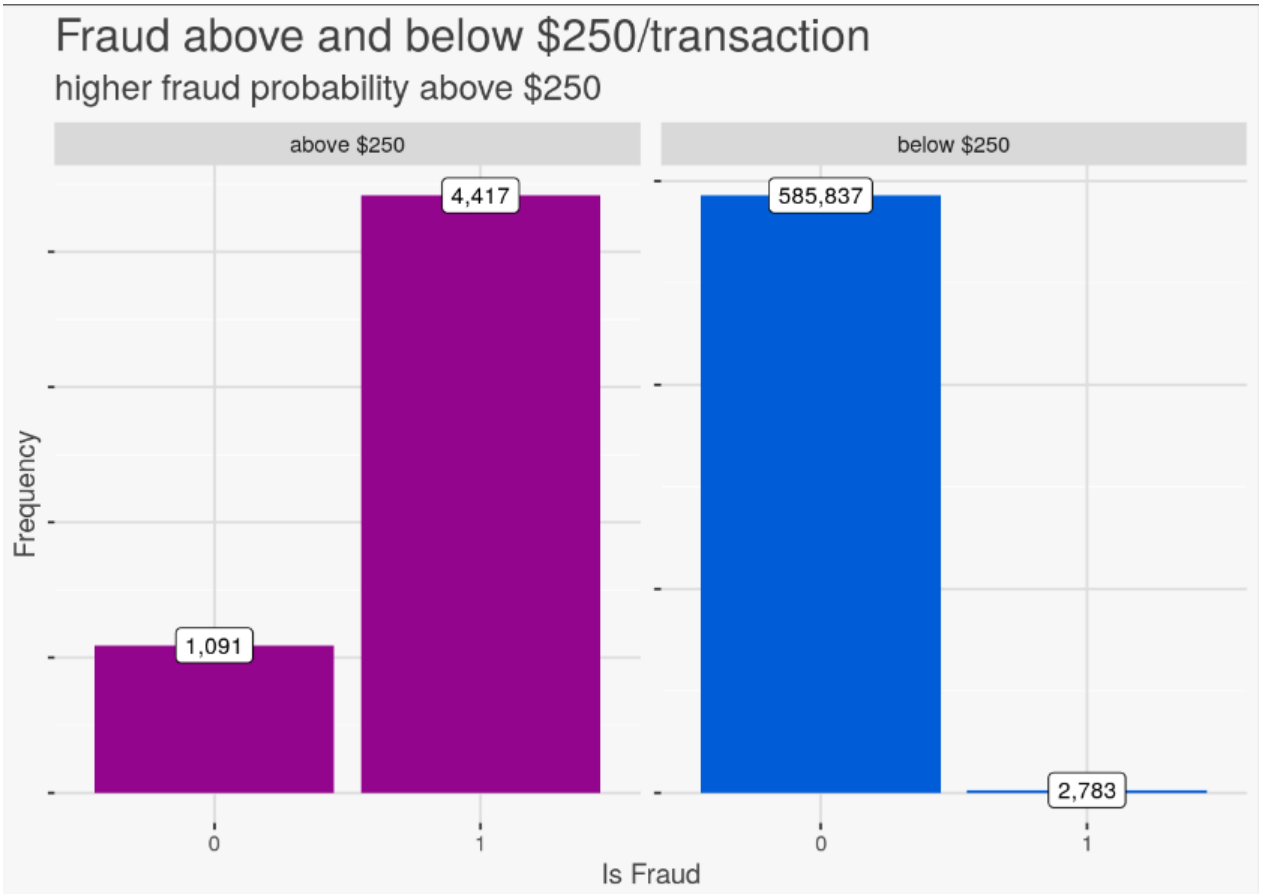


6.1.3 – FRAUD PERCENTAGE FOR SPENT AMOUNT THRESHOLDS

When comparing the valid transactions with the fraudulent transactions we found the following: There are close to no fraudulent transactions between $1 and $500, but the more the amount threshold increases, higher the chance to stumble into a fraudulent transaction.

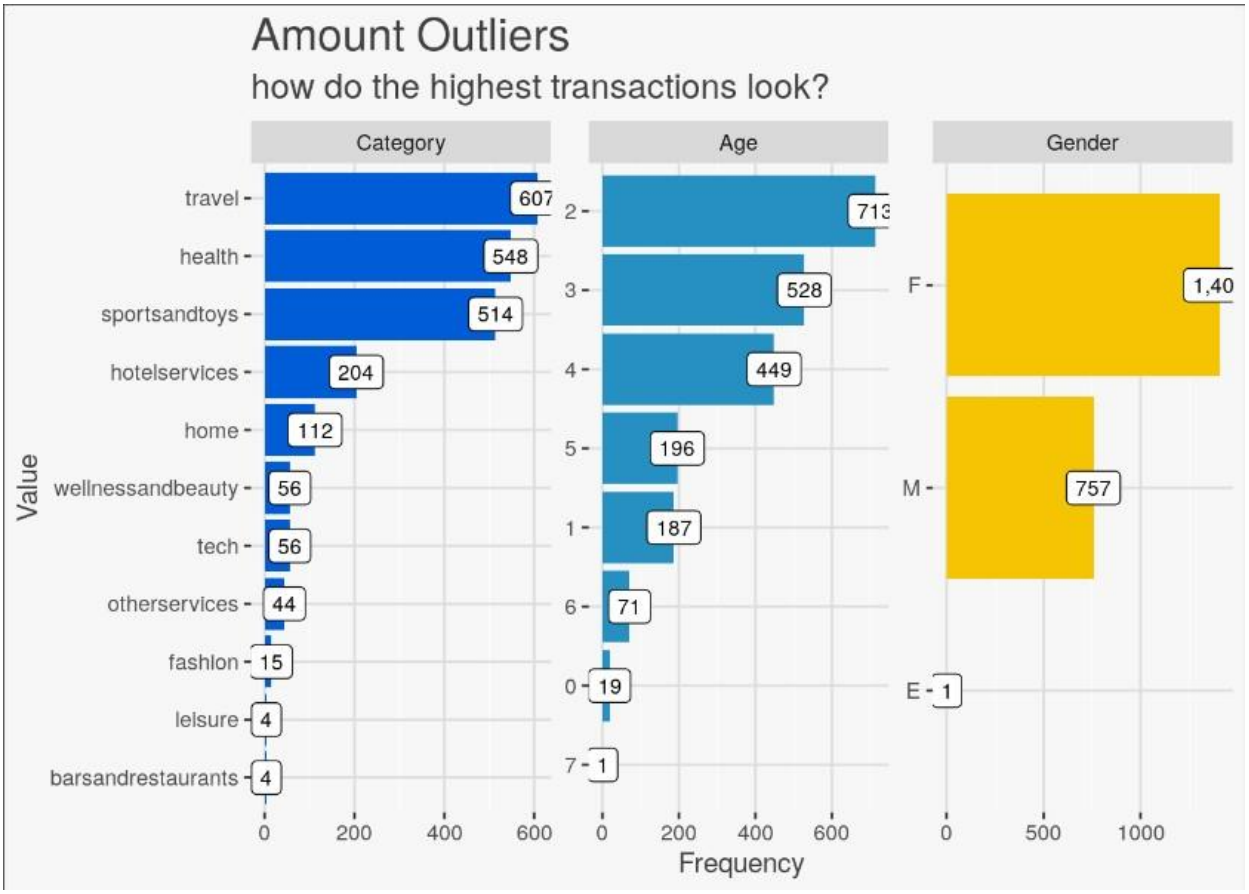


6.1.4 – AMOUNT DISTRIBUTION FOR VALUES ABOVE AND BELOW $250



6.1.5 – AMOUNT OUTLIERS

There was further investigation made into how these extremely high transactions look: the majority are expenses in the travel, health, sports & toys categories, usually done by people between 18 and 45 years old with 713 records. Also, Gender tends to lean toward women with 1400 cases compared to their male counterparts, 757.



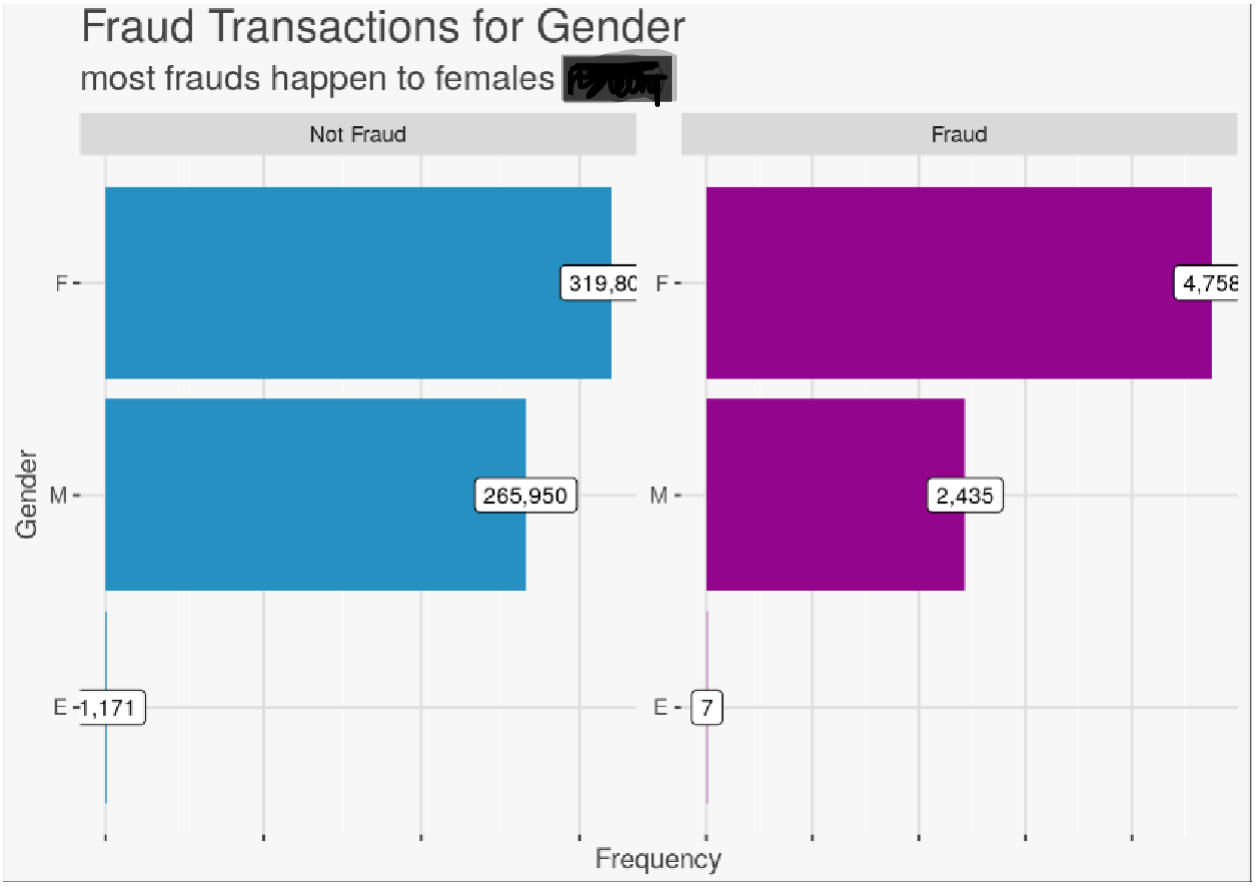
6.2. FRAUD

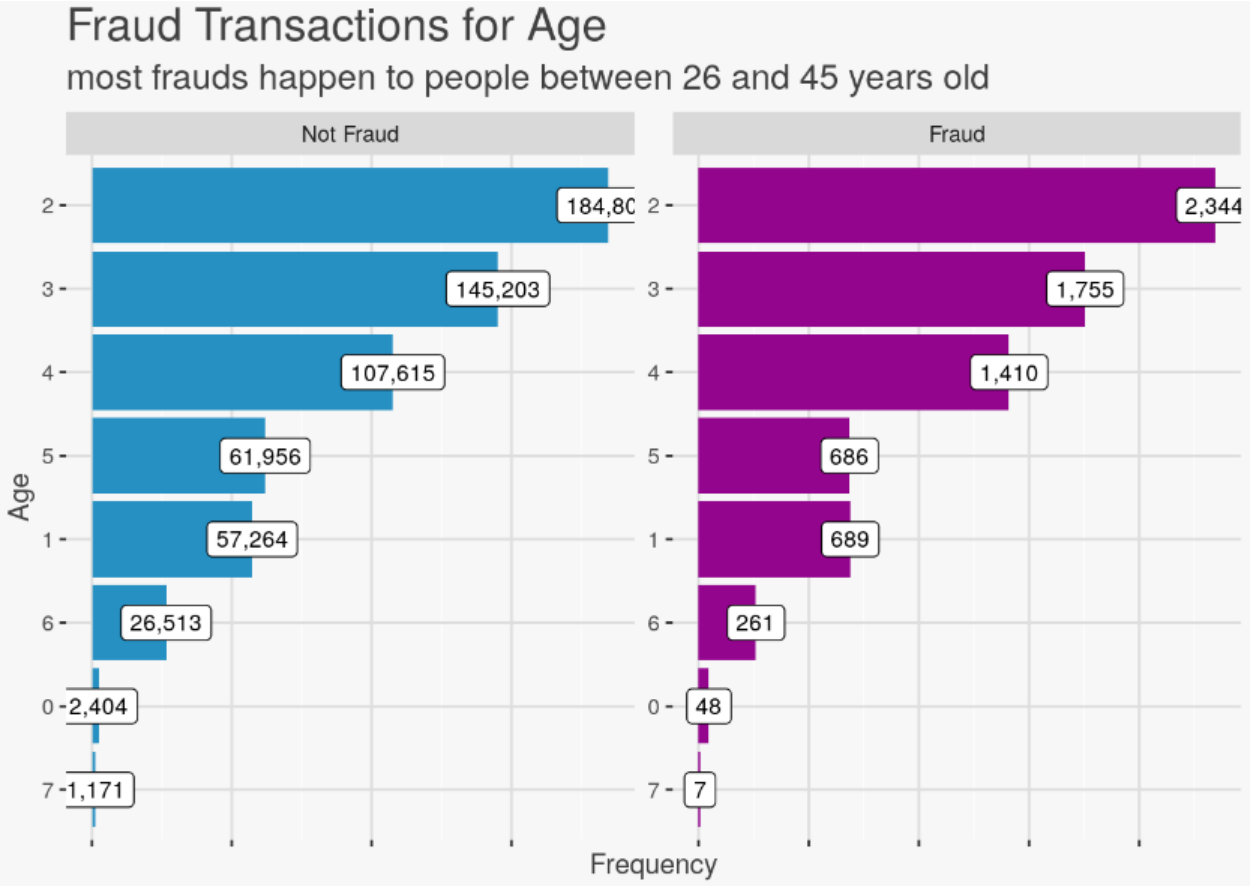
6.2.1 – FRAUD FREQUENCY

The percentage of fraudulent transactions is 1.2%, with only 7160 fraud cases compared to the nonfraudulent ones, which are 586928.

6.2.2 – GENDER AND AGE FRAUD FREQUENCY

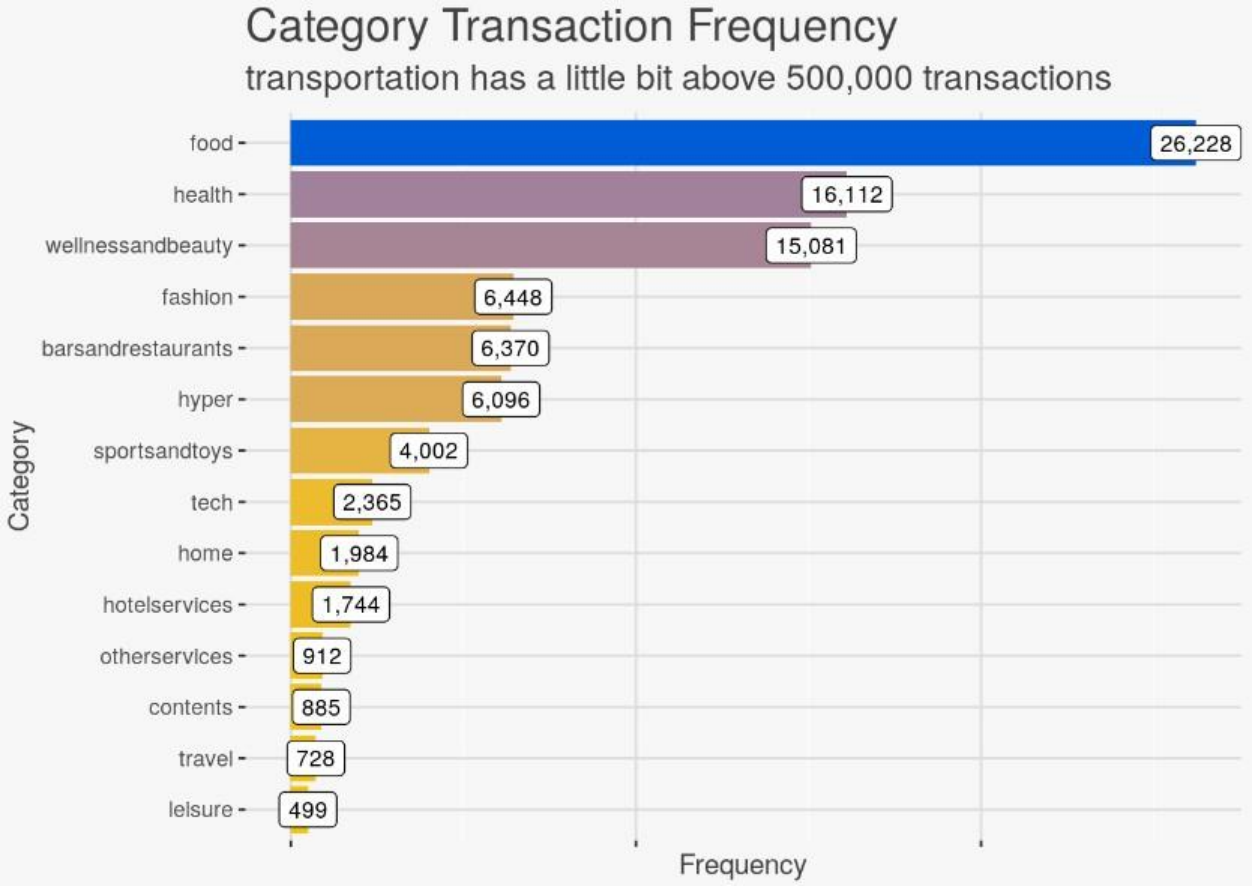
For both variables “gender” and “age”, there is no clear pattern between fraud and nonfraud cases. The distribution frequencies follow a similar pattern, so it cannot be said that one level influences fraud more than another. Regardless, an insight about the demography of the customers can be drawn from the following bar plots: most are aged between 19 and 45 years old, while females hold a bigger percentage of the total transactions than men. 55% of the total transactions are performed by women, while 44% are held by men, and only 1% are made by enterprises.

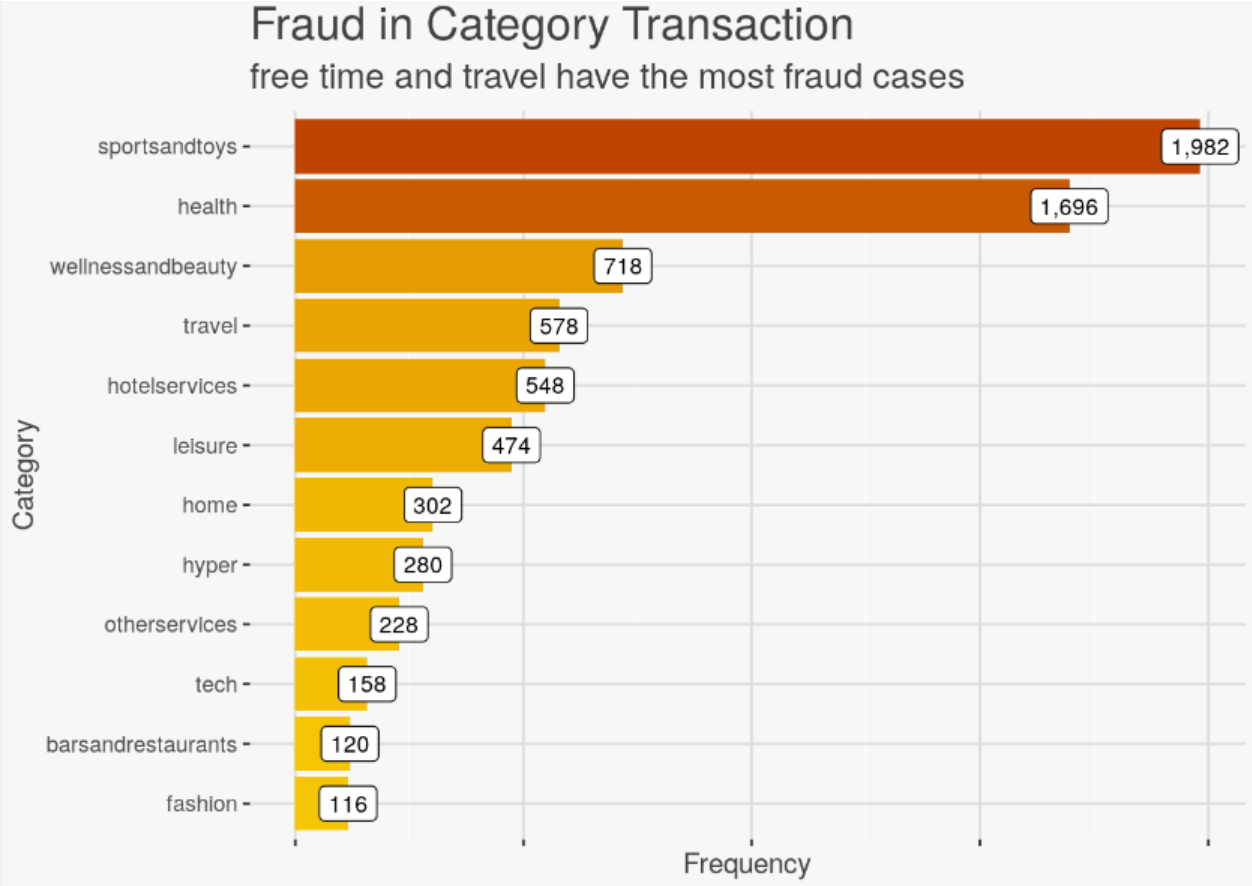


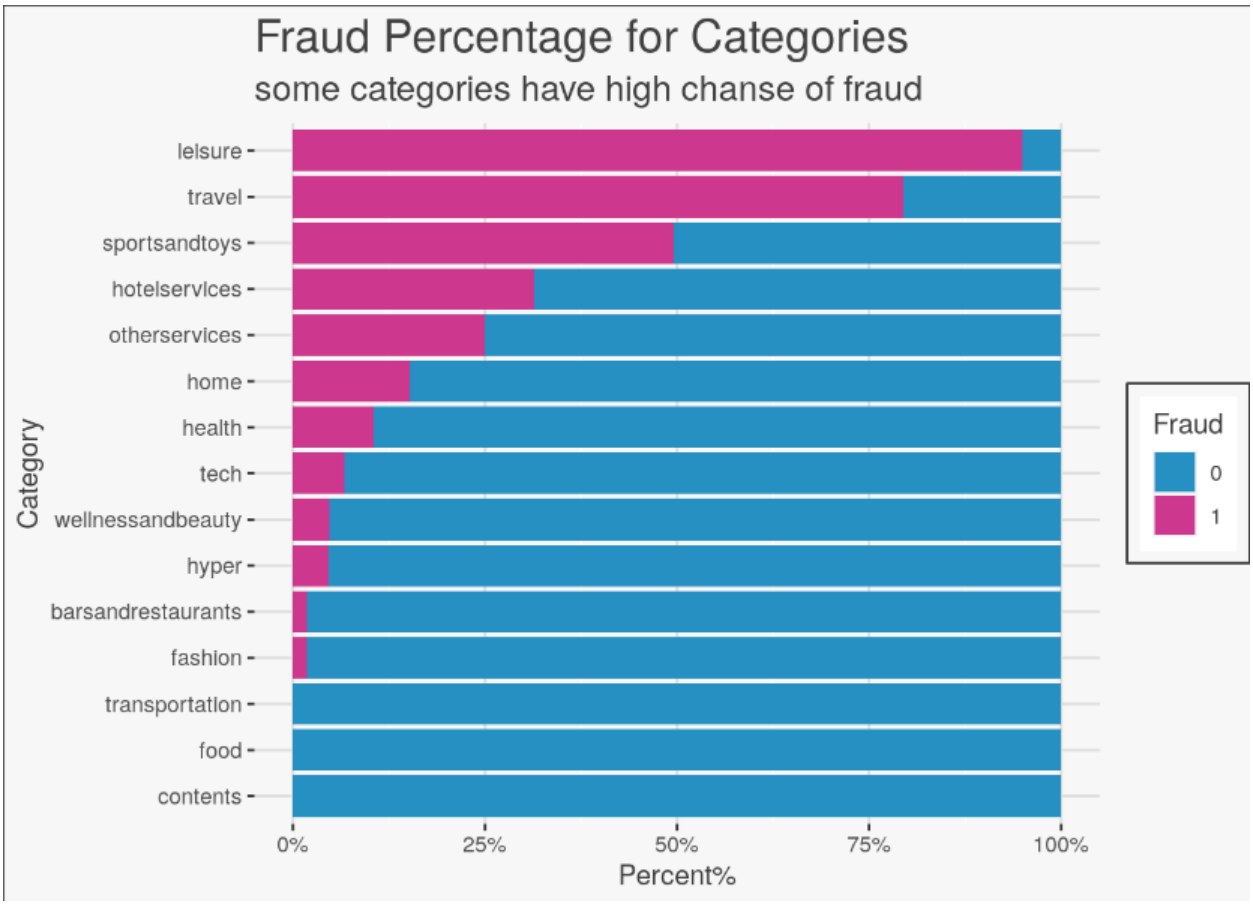

Fraud Transactions for Age
most frauds happen to people between 26 and 45 years old
Not Fraud
Fraud
Age
Frequency
2
3
4
5
1
6
0
7
184,80
145,203
107,615
61,956
57,264
26,513
2,404
1,171
2,344
1,755
1,410
686
689
261
48
7

6.2.3 – FRAUD IN SPENDING CATEGORIES

Categories also experience a high class-imbalance. There are 500000 transactions made alone in the transportation category, while the rest 90000 transactions are split unevenly between the other 14 categories left – Bar chart below. In absolute value, the categories that have the most fraud cases are in sports, toys, health, wellness, and beauty areas. Going even further and looking at the Fraud Percentage for Categories chart, there can be seen that leisure, travel, sports & toys, have more than half of the transactions flagged as fraudulent, while transportation, food and contents have 100% clean cases.




Fraud in Category Transaction
free time and travel have the most fraud cases
Category
sportsandtoys
health
wellnessandbeauty
travel
hotelservices
leisure
home
hyper
otherservices
tech
barsandrestaurants
fashion
1,982
1,696
718
578
548
474
302
280
228
158
120
116
Frequency


Fraud Percentage for Categories
some categories have high chanse of fraud
Category
lelsure
travel
sportsandtoys
hotelservices
otherservices
home
health
tech
wellnessandbeauty
hyper
barsandrestaurants
fashion
transportation
food
contents
0%
25%
50%
75%
100%
Percent%
Fraud
0
1

6. CLASSIFYING FRAUD (SUPERVISED LEARNING)

The purpose of building a fraud classification model is to assign to each new incoming transaction with a high certainty a probability of it being a fraud. Hence, any illegal attempt can be avoided.

6.1 – FEATURE ENGINEERING

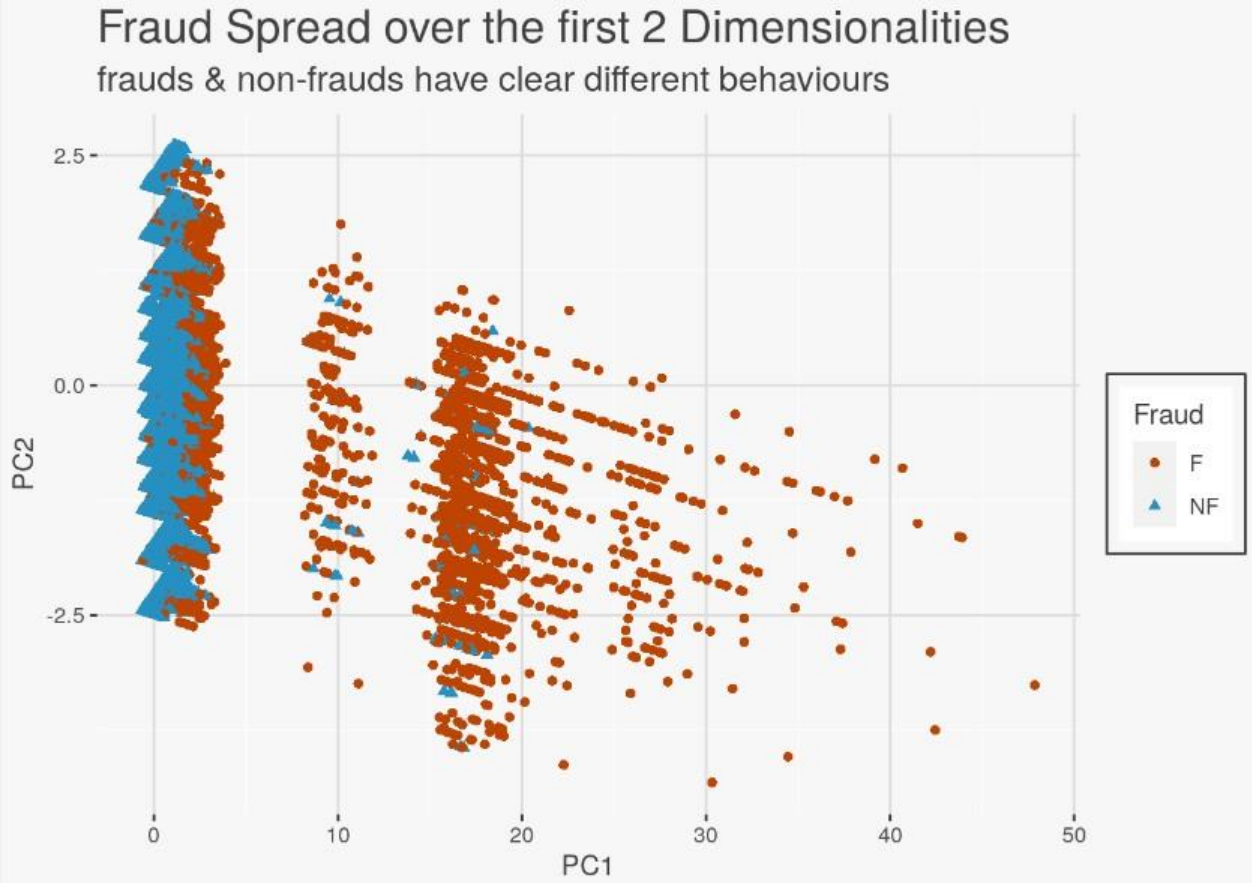
We then used Feature engineering - transformation of raw data into a feature vector or simply the merge of two attributes in a dataset. This could arguably be the utmost important step during the preprocessing part of a modeling problem. It has the purpose to transform raw data into dimensionalities that can be better understood by a predictive model, hence it remarkably improves the evaluation metrics.

- Recode character variables to numeric.
- Convert characters to double.
- Create IDs for “Merchant” and “Customer”.
- Create new variables “Total_transactions” for Customer, Merchant, Age, Gender, Category, and Amount Thresh.
- These new variables are to be transformed into percentages from Total.
- Remove “Step”
- Create dummy variables for Age, Gender, Category, and Amount Thresh.

6.2 – PRINCIPAL COMPONENT ANALYSIS

The purpose of creating a PCA before applying any classification technique is to visualize in 2D how the fraud and nonfraud transactions are grouping and if there is any clear separation between them.

- Using Age, Gender, Category, Amount, Amount Thresh.
- The first 2 PCs explain ~ 47% of variability.
- All nonfraud values are extremely concentrated over PC2, while fraud is more spread over the PC1 dimensionality.
- There are some cases of nonfraud that have strong fraud “behavior.”



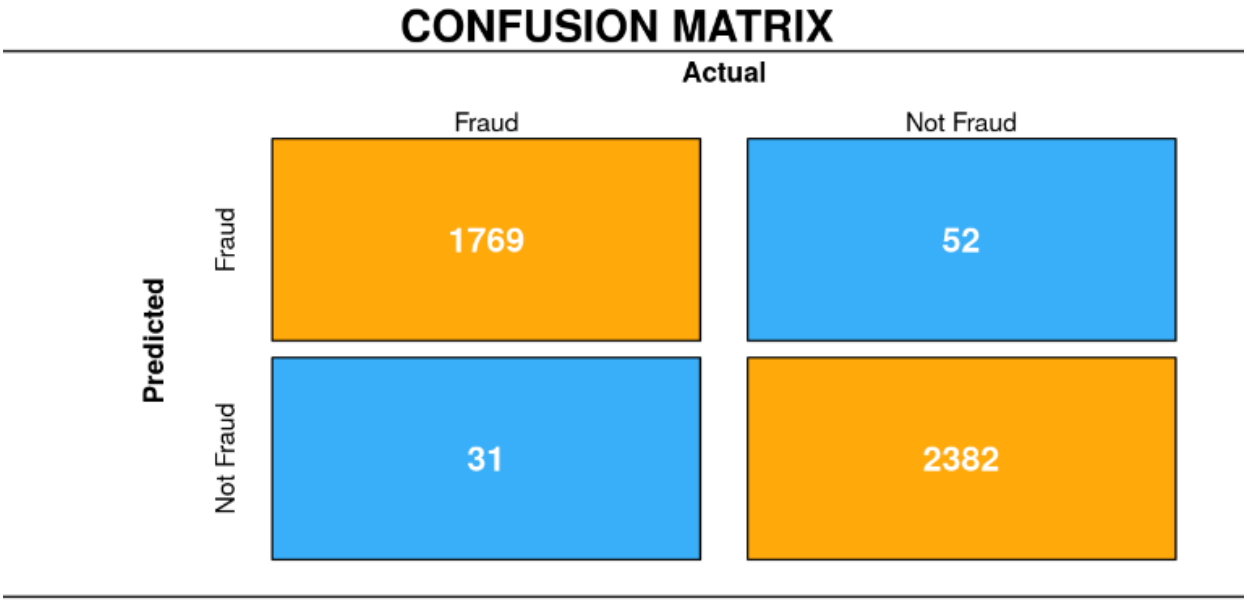
6 – CLASS IMBALANCE

The fraud and nonfraud cases available in the present dataset have extremely imbalanced weights, with only 1.2% out of the total cases being fraud. Thus, there are only 7000 observations labeled as fraud, while the rest 580000 observations are labeled as clean transactions. In a classification problem, this would create difficulties for a model to correctly identify the fraud label, because it is so scarce throughout the dataset. This data structure issue can be solved by using two different sampling techniques: **Under sampling or Over sampling**.

64 – UNDERSAMPLING “SPLITTING THE DATA 65% - 35%”

Under sampling method consists of keeping all available fraud transactions, while under sampling the nonfraud transactions to around the same number.

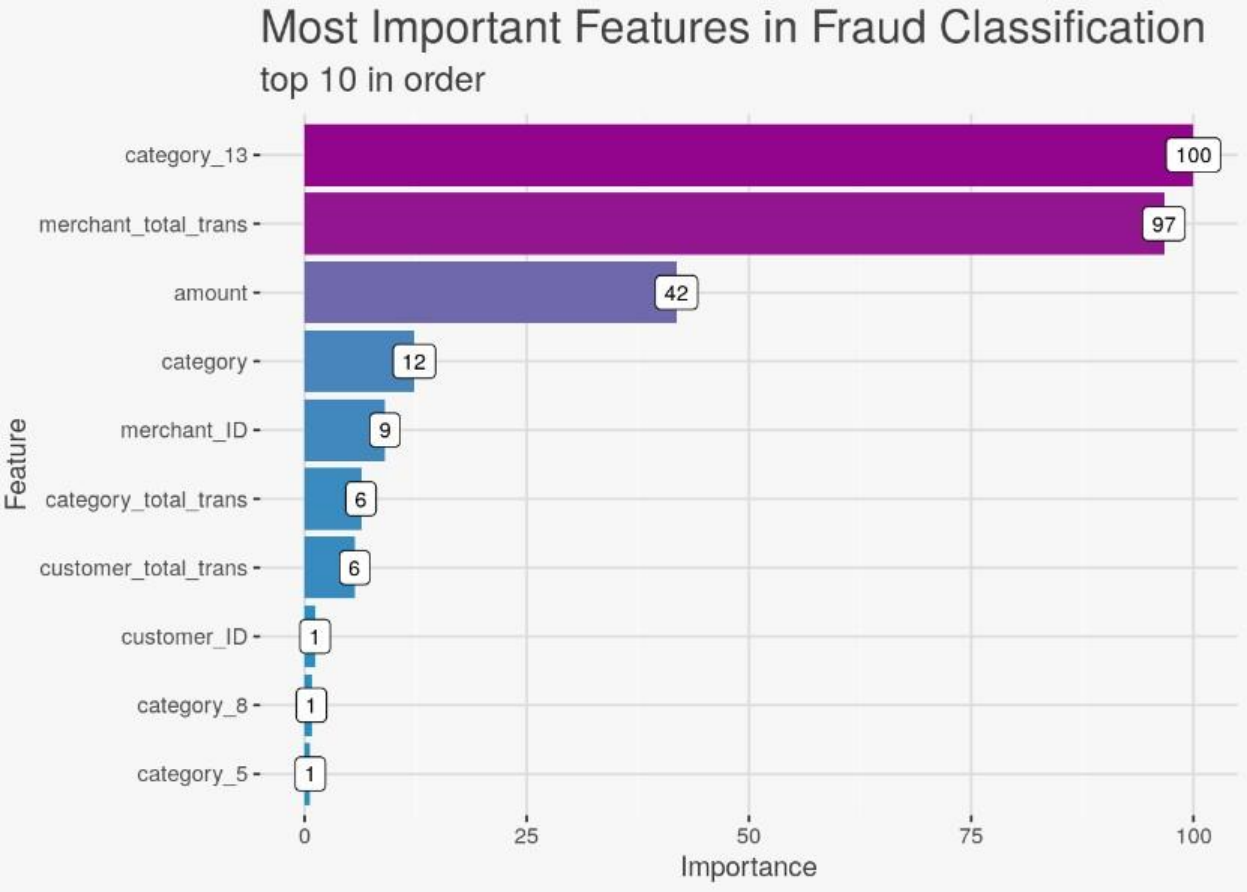
- The split is made so that proportions within the data for Age, Gender, Category, Amount Thresh, and Merchant remain the same.
- Final table dimension is fraud data: 7160 and nonfraud data: 9647.
- We split 65% - 35% to be sure that the model classifies nonfraud as correct as possible.



DETAILS

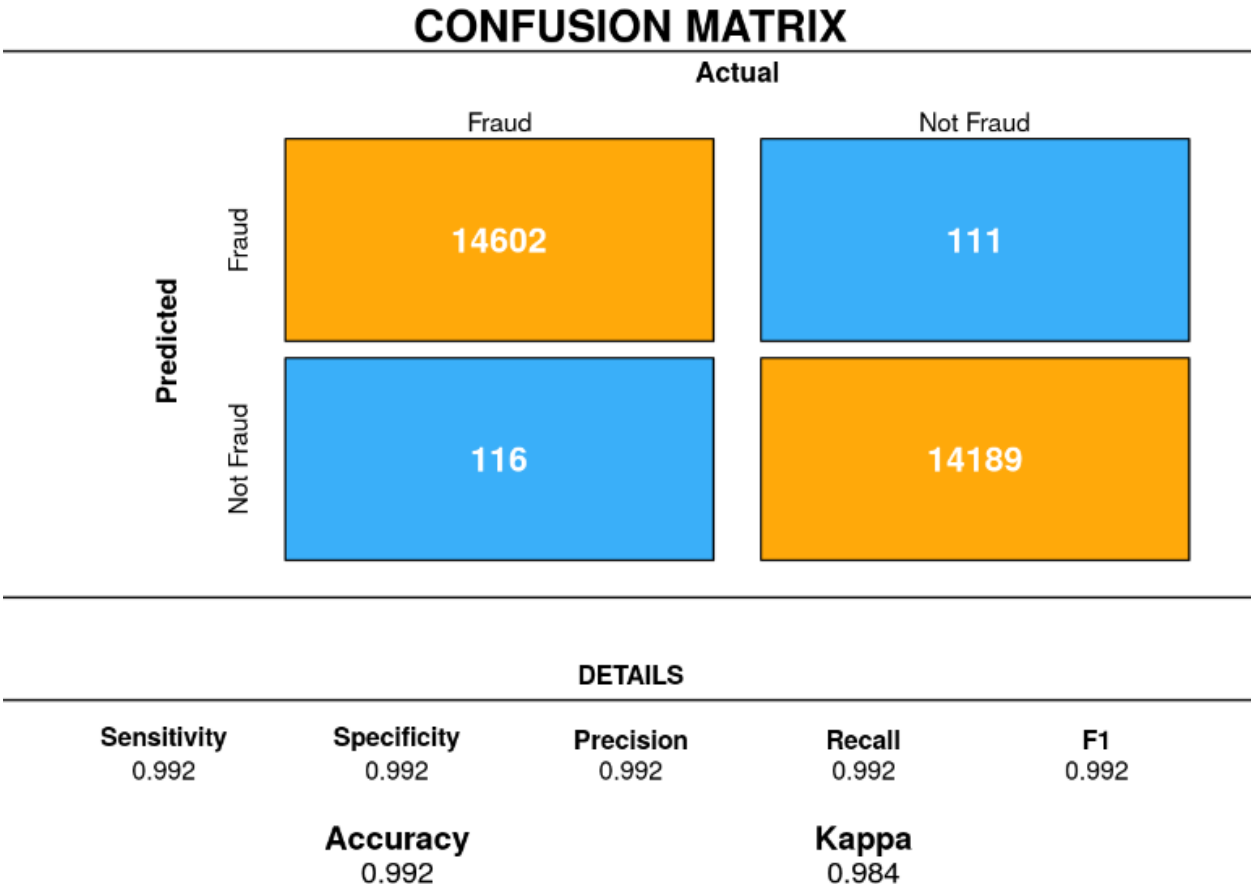
Sensitivity	Specificity	Precision	Recall	F1
0.983	0.979	0.971	0.983	0.977

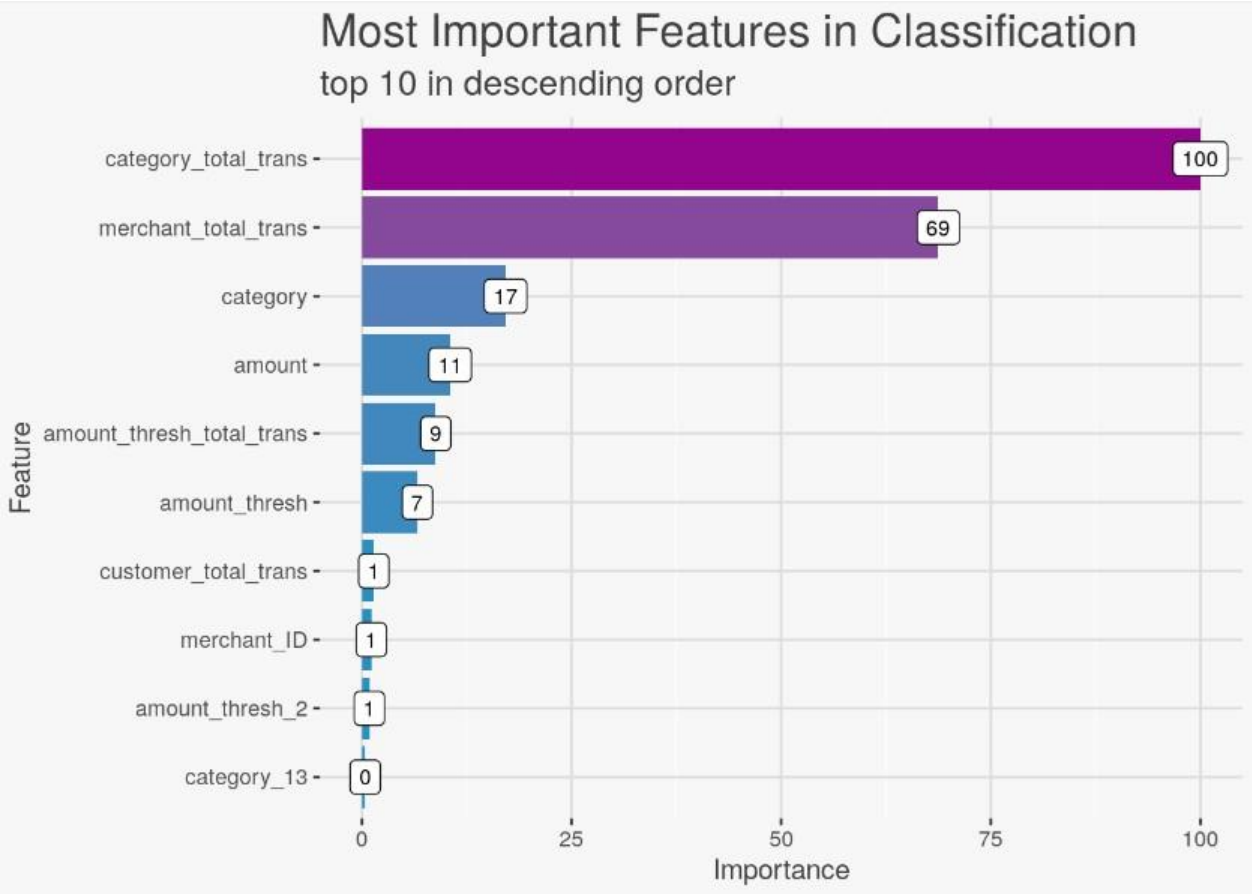
Accuracy	Kappa
0.98	0.96


Most Important Features in Fraud Classification
top 10 in order
category_13
merchant_total_trans
amount
category
merchant_ID
category_total_trans
customer_total_trans
customer_ID
category_8
category_5
Feature
100
97
42
12
9
6
6
1
1
1
0
25
50
75
100
Importance

6b – OVERSAMPLING

Another technique to deal with class imbalance is over sampling. Because using under sampling the model is training on just a few observations from the dataset (7000 + 9000 = 16000 observations, where the full data has almost 600000 observations), I chose to use over sampling because it deals with truer nonfraud cases while over sampling the fraud ones.




Most Important Features in Classification
top 10 in descending order
category_total_trans
merchant_total_trans
category
amount
amount_thresh_total_trans
amount_thresh
customer_total_trans
merchant_ID
amount_thresh_2
category_13
100
69
17
11
9
7
1
1
1
0
Feature
0
25
50
75
100
Importance

7. UNSUPERVISED LEARNING

For this specific training, we decided to create three personalized labels before training the model.

NORMAL BEHAVIOR

- Transaction amounts small (below $500).
- Payments for transportation and food transactions (do not have any fraud cases).
- Some merchants do not have any cases of fraud within their transactions reported.

ABNORMAL BEHAVIOUR

- Transactions with high amounts (above $500).
- Transactions made during travel or leisure activities.
- Some merchants reported all transactions were frauds.

CUSTOMER SEGMENTATION

- No need for customer segmentation.
- Fraud does not depend on the person who makes the transaction but on the nature of the transaction itself.

METHODOLOGY:

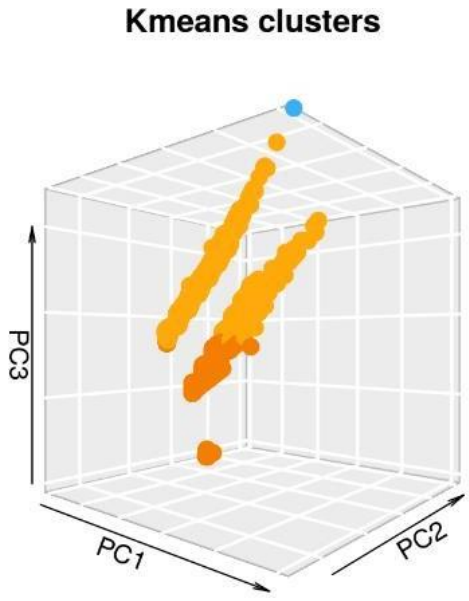
1) Find out which is the best number of clusters for the data.
2) Perform **KMEANS** on the data, using the best-found number.
3) Compute the distance between the points and the centroids.
4) The outliers for each cluster distance have abnormal behavior.
5) How many of these outliers are frauds? For which are not fraud, what makes them so abnormal?

1) CHOOSE BEST-NUMBER OF CLUSTERS

Because the data is too large, a sample of 6502 observation was extracted to obtain the best number of clusters. I found out that 6424 transactions are not fraudulent, whereas just 78 are. Thus, the best number of clusters is 3.

2) KMEANS CLUSTERS

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.1013093	27.53273	27.53273
Dim.2	1.0049000	25.12250	52.65523
Dim.3	0.9926223	24.81556	77.47079
Dim.4	0.9011684	22.52921	100.00000

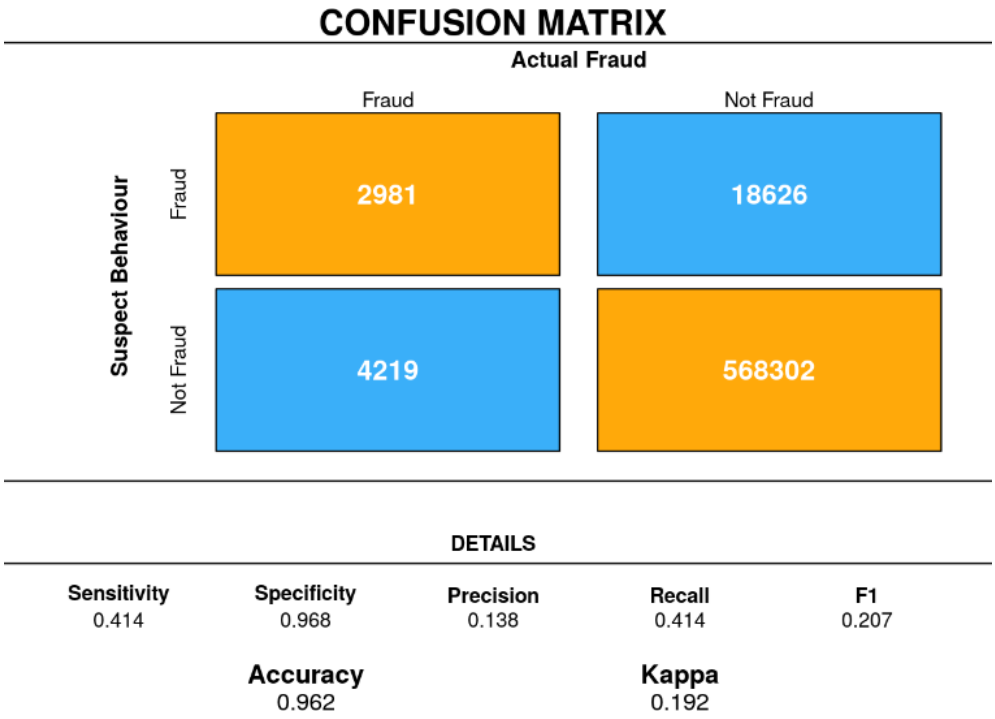


3) **DETERMINE THE OUTLIERS**

- Cluster 1 has the most observations, grouping 502245 of the total observations. 5 of them are frauds.
- Cluster 2 had 86838 observations from which 81745 are not frauds, and 5093 are frauds.
- Cluster 3 is the smallest one, with 2250 observations from which 2062 are frauds, and only 88 are not frauds.

How can a bank determine credit card fraud?

One possible way is to examine the clusters above. The first one is the "nonfraud" cluster, with usual behavior present. The second cluster has the most fraud cases (5k), but the other 80k nonfraud needs to be inspected. Finally, the third one is the fraud cluster, with the most suspect behavior.



The sensitivity is quite low, so the outliers' approach to identify **suspect behavior** did not work out well. Therefore, I have decided to analyze the 3 clusters again.

8. RESULTS

Based on the dataset analyzed, we have come up with the following three clusters:

CLUSTER 1

It holds most data with 502245 records of which 99.9% are not fraud. In fact, only 5 transactions are fraudulent. From the EDA, most of the behavior is tied to small amounts, payments for transportation, food, or health, and payments made to the top 3 merchants that have no fraud cases registered.

CLUSTER 2

It is a more homogeneous cluster. It is very small with 86838 observations which 5k transactions are fraud. The behavior is between the not suspect at all and extremely suspect behavior. The nonfraud cases are within fashion, restaurant, wellness, and beauty categories with an amount spent low, ~$60. Whereas the fraud cases are the ones that have a lower amount (less than $500, usually around ~$230) and there is no specific category to investigate, the transactions are spread throughout all categories.

CLUSTER 3

It is the smallest cluster, with 2150 records. Mostly composed of fraudulent transactions – 2062. These are classic frauds with most transactions being for hotels and services, travel, fashion, or restaurants. The amount spend is very high, with an average of ~$1300. In contrast, nonfraud transactions are made in the travel category (with 75% chances of being fraud) with an average amount of ~$1000.

9. FURTHER RESEARCH

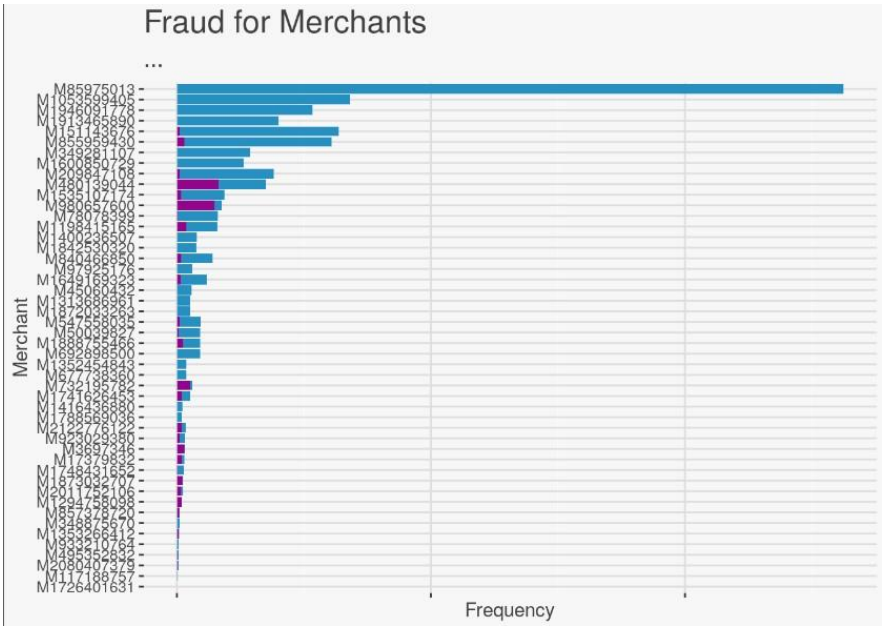
As discussed in the paper, Clustering is a technique in unsupervised machine learning which groups data points into clusters based on the similarity of information available for the data points in the dataset. Thus, the data points belonging to the same clusters are like each other in some ways while the data items belonging to different clusters are dissimilar. To complete the analysis, it would be necessary to run a DBSCAN clustering to extend the research further. Using the DBSCAN model, the number of clusters does not need to be predefined, it can be used to identify fraudulent transactions as very small clusters, and extremely reliable in detecting outliers and white noise – false positives & false negatives.

During the exploratory data analysis, not all the attributes have been used to gather pieces of information from the dataset. For example, there might be a correlation between the Merchants and fraudulent transactions as the below bar chart shows.

This analysis was done on one dataset produced by one Spanish bank. Machine learning algorithms should be tried to identify live fraudulent transactions.

It is not clear whether this dataset is representative for transactions by another bank…even a bank within the same country, let alone anywhere else in the world… We need to compare this dataset with datasets from other banks…

Further research is needed to analyze the acceptable number of false positives – perhaps depending on the nature and philosophy of the bank. Some banks may be more cautions in approving transactions that could be fraudulent than other banks…



10. REFERENCES

- K-means Cluster Analysis · UC Business Analytics R Programming Guide (uc-r.github.io)
- K-Means Clustering in R Programming - GeeksforGeeks
- Realtime Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm (researchgate.net)
- How to Detect Banking Fraud in a Constantly Evolving Cyberspace? | TIBCO Software
- How Banks Conduct Transaction Fraud Investigations (chargebackgurus.com)
- Paper Title (use style: paper title) (researchgate.net)
- Fraud Management in Banking: Detection, Prevention & More – Hitachi Solutions (hitachi-solutions.com)