Lauryne MOYSE
Farah JABRI
Rafael MOUROUVIN

MODS206 : Data analysis in economics 2

CREDIT CARD FRAUDS

SUMMARY

1. INTRODUCTION

1.1 About Credit card frauds

Credit card fraud is a significant concern for the financial industry. It is also an issue for both consumers and financial institutions. It occurs when an unauthorized person gains access to a person's bank information and uses it to make purchases or bank transfers. In the last ten years, the number of fraud cases has significantly increased. In fact, according to a report published by Nilson in 2021, one of the world's most reliable sources of information and statistics on the payments industry, global losses from bank fraud using credit cards were 23.97 billion dollars in 2015 and 28.65 billion dollars in 2019, which demonstrates the unprecedented growth of the problem of bank fraud.

Moreover, as the number of online payments is increasing with the digitalization of the economy, it is also leading to an increase in cyber fraud. In 2021, Juniper Research, an analysis company specialising in digital technology market research, estimated that online payment fraud could cost all companies approximately $205 billion from 2021 to 2025. This figure has been boosted by the COVID-19 pandemic, which has dramatically accelerated the digitalisation of the economy, between online shopping and various digital transactions.

The detection of bank fraud is becoming increasingly difficult. Indeed, counting only Visa transactions, there are on average 300 million transactions per day. The control of credit card fraud is therefore very challenging. As a result, financial institutions, banks and other payment platforms are investing in R&D to have the means to detect bank fraud. Indeed, the fraud detection market is booming and is estimated to be of a value of 68 billion dollars by 2027.

Thus, many researchers and economists have tried to find out what factors make it possible to detect fraud and what algorithms can automate this detection. High-performance tools using machine learning and artificial intelligence techniques have emerged as promising tools for combating bank fraud, particularly credit card fraud. Among other things, by analysing large data streams and identifying similar patterns that indicate fraudulent activity, these technologies can significantly increase fraud detection. These algorithms have the potential to achieve up to 90% accuracy in fraud detection.

1.2 State of the art

Many researchers have thoroughly studied this issue in order to deeply understand its aspects and develop some effective and efficient countermeasures to curb this phenomenon.

First of all, it is necessary to identify the different types of credit card fraud. The main bank card frauds are the following.

- **Counterfeit card fraud :** This type of fraud occurs when a cyber criminal creates a fake credit card using stolen data from an existing account. This data is obtained through methods such as skimming or data breaches.
- **CNP fraud :** The CNP fraud or card-not-present fraud happens when a cyber criminal does online purchases with stolen bank information.
- **Lost/ Stolen card fraud :** This type of credit card fraud occurs when a person makes transactions with a stolen or lost card
- **Account takeover :** This type of credit card fraud occurs when a cyber criminal gains access to a person's bank account through methods such as phishing and is able to make transfers and purchases with that bank account.

The increase with the digitalization of the economy has led governments to take strict measures to counter this phenomenon.

The European Union's Payment Services Directive 2 (PSD2) seeks to make payments more secure in Europe, boost innovation and help banking services adapt to new technologies. It seeks to increase the security of electronic transactions in the European Union. It was implemented in 2018 with a central security measure which was called Strong Customer Authentication for almost every electronic transaction. This process checks the user's identity through diverse elements such as : a password or a pin, the detection of the user's smartphone or hardware token, something unique that the user has (facial recognition or the user's fingerprint). These security measures make it more and more difficult for fraudsters to make illegal transactions.
Moreover, under PSD2, the risk of each transaction is assessed in real time. It takes into account factors such as transaction amount, location and historical data.

The US enacted the Fair Credit Billing Act in order to protect consumers from unfair credit billing practices. It allows people to dispute unauthorized transactions on their accounts.

Finally, the IGCI (Interpol global complex for Innovation) has created a unit called the Cyber Fusion Center specialized in cybercrime (including credit card fraud). This unit facilitates the sharing of information between governments/ law agencies and financial institutions.

All these measures are accompanied by research into detection algorithms to curb credit card fraud. The most effective methods use deep learning/machine learning.

- **A Fusion Approach Using Dempster-Shafer Theory and Bayesian Learning:** it uses the current and past behavior of the cardholder to create a profile and detects deviations from this profile using rules such as average daily/monthly spending profiles and different shipping and billing addresses.

- **Blast-Ssaha Hybridization**: This approach is a two-stage sequence alignment that uses a profile analyzer and a deviation analyzer to match incoming transactions with the cardholder database. If an unusual sequence is found, it is compared with the past fraud history database, and the final decision maker determines if the transaction is genuine or fraudulent.

- **Hidden Markov Model:** This approach uses a finite set of states with a probability distribution to determine if an incoming transaction is legitimate or fraudulent. Each state has a probability assigned to each cardholder, and incoming transactions are compared to a predefined threshold value. If the incoming transaction is unusual or does not have a sufficient probability, it is announced as fraudulent.

1.3 Our subject and motivations

As we have seen earlier, the increasing use of online transactions transformed the scope of financial transactions leading to more and more fraud behaviors. Online transactions increased the risks of credit card frauds. And with the flourishing of e-commerce online retailers are vulnerable. They are the most common type of identity theft and represent millions of dollars in losses.

Some characteristics of online transactions increase the risk of frauds. Anonymity is one of them: it is easy for fraudsters to hide their identities and locations. Moreover, the fastness of transactions and the global reach make it even more complicated: the Internet allows the purchase of things all over the world, it is complicating the track of fraudsters that are operating in a different country.

Therefore, credit card frauds are a key and current issue which needs to be tackled.That is why we believe it is interesting to understand the impact of online transactions on credit card fraud, as it could inform the development of detection strategies. We have chosen the following research question :

What is the impact of online transactions on credit card fraud ?

2. DATA DESCRIPTION

The dataset we chose is composed of 8 features. This dataset provides an overview of several transaction characteristics that can be used in order to analyze or detect an eventual credit card fraud.

Our dataset includes the following features, some floats expressed in an unknown unity :

- **distance_from_home :** this variable represents the distance between the location where the transaction took place and the location of the cardholder's residence.
- **distance_from_last_transaction :** this feature evaluates the distance between the previous and the current transaction of the cardholder.
- **ratio_to_median_purchase_price :** this variable presents the ratio of purchased price transaction to median purchase price.

The other variables of our dataset are dummy variables.

- **repeat_retailer :** This variable indicates if the transaction took place at a retailer where the cardholder already made previous purchases.
- **used_chip :** This variable indicates if the transaction was done through a chip (credit-card) or not. (chip transactions generally offer more security compared to other transactions as they are more difficult to clone).
- **used_pin_number :** This variable indicated if the transaction required a personal identification number, a PIN, for the authentication of the cardholder.
- **online_order :** This is the variable we chose to study more particularly. This variable specifies if the transaction was done through an online platform.
- **fraud :** This variable is our target variable. It indicates if the transaction is fraudulent or not.

The dataset presents 1.000.000 observations.

The dataset comes from Kaggle, an open-source dataset platform. The accuracy of the dataset is not verified by any qualified financial or governmental institution. We don’t know the country from which the dataset was retrieved.

3. DESCRIPTIVE ANALYSIS AND DATA VISUALIZATION

This part is aimed at analyzing our dataset to understand the representation of the different variables.

3.1 Data overview

```
Contains data
 Observations:     1,000,000
    Variables:             8
```

Variable name	Storage type	Display format	Value label	Variable label
distance_from~e	float	%9.0g		
distance_from~n	float	%9.0g		
ratio_to_medi~e	float	%9.0g		
repeat_retailer	byte	%8.0g		
used_chip	byte	%8.0g		
used_pin_number	byte	%8.0g		
online_order	byte	%8.0g		
fraud	byte	%8.0g		

Description of the dataset

As said before, we can observe that most of the variables are binary. Particularly, our variables of interest : “fraud” (explained variable) and “online_order” (explicative variable).

Variable	Obs	Mean	Std. dev.	Min	Max
fraud	1,000,000	.087403	.2824248	0	1
online_order	1,000,000	.650552	.4767959	0	1
distance_f~e	1,000,000	26.62879	65.39078	.0048744	10632.72
distance_f~n	1,000,000	5.036519	25.84309	.0001183	11851.1
ratio_to_m~e	1,000,000	1.824182	2.799589	.0043992	267.8029
repeat_ret~r	1,000,000	.881536	.3231569	0	1
used_chip	1,000,000	.350399	.4770951	0	1
used_pin_n~r	1,000,000	.100608	.3008091	0	1

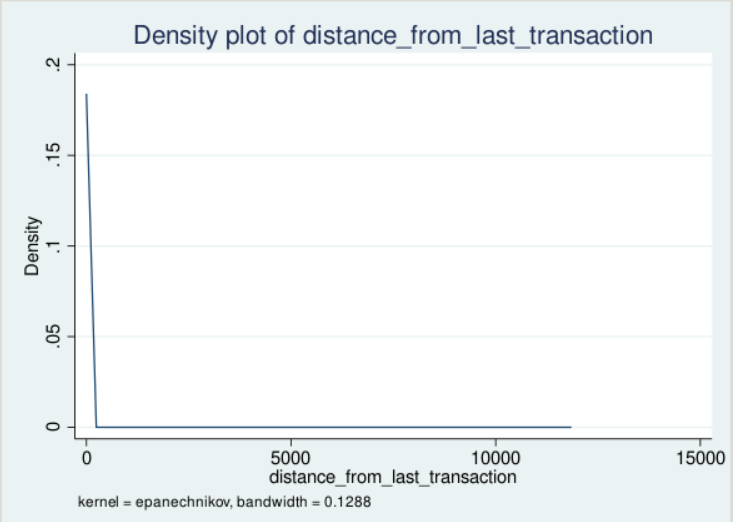
Summary of different variables from the dataset

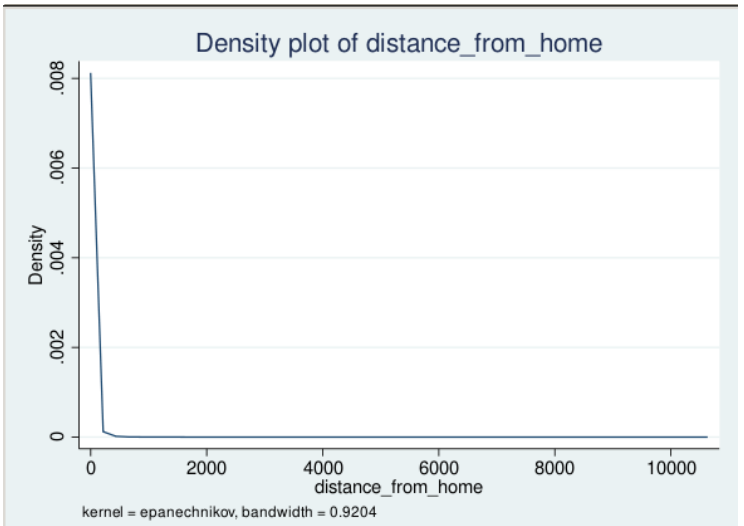
No data is missing, which makes our task easier since we won't have to do any cleaning. The means and standard deviation of continuous variables are hard to interpret since we do not know their unit. We will normalize them in order to avoid this concern.

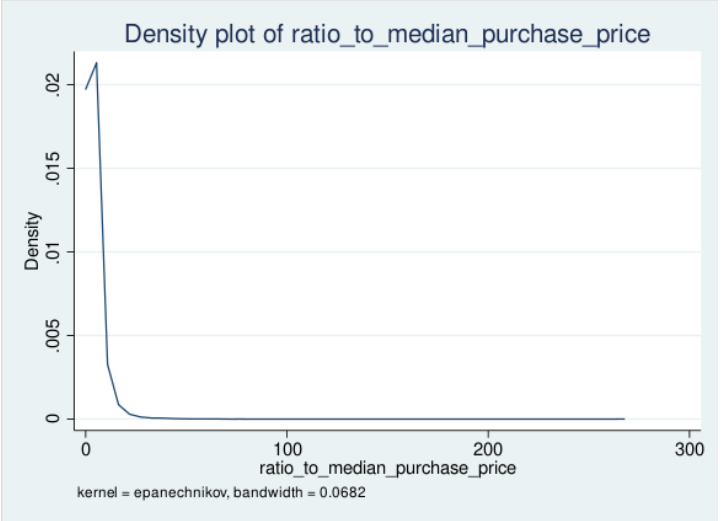
The size of our dataset is important which will allow us to have more significant results. However if we look at the mean of the variable “fraud”, we can see that only 8.7% of the registered transactions are fraud. Consequently, we have a significantly lower amount of data for fraudulent transactions specifically.

3.2 Distribution of data

We can take a look at the repartition of our continuous variables :







They are distributed quite similarly with an abundance of low values and then sag very quickly. This confirms our choice to normalize them and not to standardize them.

Now, let's take a look at our dummy variables :

fraud	Freq.	Percent	Cum.
0	912,597	91.26	91.26
1	87,403	8.74	100.00
Total	1,000,000	100.00	

used_pin_number	Freq.	Percent	Cum.
0	899,392	89.94	89.94
1	100,608	10.06	100.00
Total	1,000,000	100.00	

used_chip	Freq.	Percent	Cum.
0	649,601	64.96	64.96
1	350,399	35.04	100.00
Total	1,000,000	100.00	

repeat_retailer	Freq.	Percent	Cum.
0	118,464	11.85	11.85
1	881,536	88.15	100.00
Total	1,000,000	100.00	

online_order	Freq.	Percent	Cum.
0	349,448	34.94	34.94
1	650,552	65.06	100.00
Total	1,000,000	100.00	

As said before, a majority of transactions are not fraudulent. Similar trends can be observed : the majority of transactions are online, do not use pin number or chip code and have taken place at an already known retailer. We will try to study in more detail the correlations between these different characteristics.

3.3 Correlation between data

Knowing the correlation between our data will be an important asset for our study. We therefore display the correlation matrix :

	fraud	online~r	distan~e	distan~n	ratio_~e	repeat~r	used_c~p
fraud	1.0000						
online_order	0.1920	1.0000					
distance_f~e	0.1876	-0.0013	1.0000				
distance_f~n	0.0919	0.0001	0.0002	1.0000			
ratio_to_m~e	0.4623	-0.0003	-0.0014	0.0010	1.0000		
repeat_ret~r	-0.0014	-0.0005	0.1431	-0.0009	0.0014	1.0000	
used_chip	-0.0610	-0.0002	-0.0007	0.0021	0.0006	-0.0013	1.0000
used_pin_n~r	-0.1003	-0.0003	-0.0016	-0.0009	0.0009	-0.0004	-0.0014

Correlation Matrix

Our variable of interest “fraud” is highly correlated to “ratio_to_median_purchase_price”. It is also significatively correlated to our other variable of interest “online_order”, and to “distance_from_home” and “used_pin_number”. Finally, it is little or not correlated with the variables “distance_from_last_transaction”, “repeat_retailer” and “used_chip”.

“Online_order meanwhile, is very poorly correlated with the other variables, except “fraud”.

4. EMPIRICAL STRATEGY

1) **Make a first naive analysis :**

We will start by studying the distribution of “fraud” according to “online_order”. Since our variables of interest are binary, we can make frequency crosstabs and already make preliminary observations in order to intuit our conclusion.

2) **Identify the control variables of the model in order to limit the omitted variable bias:**

Control variables must be a determinant of the dependent variable and be correlated with the independent variable.

As seen previously with the correlation matrix, “ratio_to_median_purchase_price”, “distance_from_home” and “used_pin_number” are significantly correlated to “fraud”. But unfortunately none of the covariates is highly correlated to online order. We may have made a bad choice of database. But we will still try to see the effect of the presence of the previous variables in our model.

3) **Choose between the logit and probit model :**

Since our explained and explanatory variables are binary, we will have to choose one of these two models. But which one?

The main difference between these two types of regression is the link function used to model the error term distribution. In a probit regression, the link function is the standard cumulative normal distribution function. whereas in a logit regression, the link function is the logistic function.

A probit regression therefore assumes that the errors are normally distributed, while a logit regression assumes that they are distributed according to a logistic law. Both methods would produce quite similar results, given the size of our database.
As we have pointed out several times, we have little occurrence of the dependent variable (relatively to our number of entries), so logit regression may be preferable because it is less sensitive to extreme values. However, the question does not necessarily arise since we intend to use normalized variables.

In the end, as logit regression is more regularly used, we finally decided to settle on the use of the latter.

4) **Perform various logit regressions :**

Our Y variable will always be “fraud” but our set of covariates will vary. It will first be composed of “online_order” only. Then, we will add control variables one by one to minimize omitted variable bias. We will watch the evolution of the coefficient in front of “online_order” as well as the different test results and draw conclusions.

5. TESTS AND RESULTS

5.1 Qualitative approach

```
. tabulate online_order fraud, row
```

Key
frequency *row percentage*

online_order	fraud		Total
	0	1	
0	344,756 98.66	4,692 1.34	349,448 100.00
1	567,841 87.29	82,711 12.71	650,552 100.00
Total	912,597 91.26	87,403 8.74	1,000,000 100.00

fraud	online_order		Total
	0	1	
0	344,756 37.78	567,841 62.22	912,597 100.00
1	4,692 5.37	82,711 94.63	87,403 100.00
Total	349,448 34.94	650,552 65.06	1,000,000 100.00

This cross frequency tables provides interesting results.

First, 98.7% of the transactions that are not online are not fraudulent either. It is a significant number which invites not to be wary of this kind of operation but rather of those online.

Almost 95% of fraudulent transactions have been realized online. This observation can help us begin to respond to our research question. If a transaction is online, there is a higher risk that it is fraudulent compared to another type of transaction. It would now be a question of quantifying this risk.

5.2 Quantitative approach

We run our various logit regressions, adding control variables gradually.

• **Regression of “fraud” according to “online_order”**

```
Iteration 0:   log pseudolikelihood = -296487.78
Iteration 1:   log pseudolikelihood = -276045.76
Iteration 2:   log pseudolikelihood = -272768.07
Iteration 3:   log pseudolikelihood = -272688.98
Iteration 4:   log pseudolikelihood = -272688.94
Iteration 5:   log pseudolikelihood = -272688.94

Logistic regression                               Number of obs = 1,000,000
                                                  Wald chi2(1)  =  24444.04
                                                  Prob > chi2   =    0.0000
Log pseudolikelihood = -272688.94                 Pseudo R2     =    0.0803
```

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	2.370489	.0151618	156.35	0.000	2.340773	2.400206
_cons	-4.296978	.0146979	-292.35	0.000	-4.325785	-4.268171

Given the value of the log pseudolikelihood, the model has a relatively good fit to the data, but this is not enough to conclude definitively about the correctness of the fit. If we take a look at the coefficient of correlation, we can observe that it is quite weak. Only 8% of the variation of “fraud” is explained by the model.

A null p-value can be observed for “online_order”, therefore the variable is actually significant in determining fraud.

The variables are positively correlated as shown by the sign of the coefficient of “online_order”. This confirms our first observations : the online nature of a transaction is a factor that reinforces the probability of its fraudulency.

• Regression of “fraud” according to “online_order”, "ratio_to_median_purchase_price”

```
Logistic regression                                     Number of obs = 1,000,000
                                                        Wald chi2(2)  =  21101.33
                                                        Prob > chi2   =    0.0000
Log pseudolikelihood = -184017.62                       Pseudo R2     =    0.3793
```

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	4.526214	.0499215	90.67	0.000	4.42837	4.624059
norm_ratio_to_median_purchase	163.6307	1.161496	140.88	0.000	161.3543	165.9072
_cons	-7.939848	.0552018	-143.83	0.000	-8.048042	-7.831655

Note: 0 failures and 376 successes completely determined.

Naturally, the R2 has increased noticeably. Almost 38% of fraud variation is explained now. The coefficient in front of “online_order” went from 2.37 to 4.55.

• Regression of “fraud” according to “online_order”,"ratio_to_median_purchase_price” and “distance_from_home”

```
Logistic regression                                     Number of obs = 1,000,000
                                                        Wald chi2(3)  =  17019.80
                                                        Prob > chi2   =    0.0000
Log pseudolikelihood = -164240.94                       Pseudo R2     =    0.4460
```

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	5.49446	.066317	82.85	0.000	5.364481	5.624439
norm_distance_from_home	131.9205	1.92061	68.69	0.000	128.1562	135.6848
norm_ratio_to_median_purchase	191.3214	1.502591	127.33	0.000	188.3764	194.2664
_cons	-9.600645	.0771921	-124.37	0.000	-9.751939	-9.449351

Note: 0 failures and 598 successes completely determined.

R2 has once again increased significantly just like the coefficient in front of “online_order” which went from 4.55 to 5.49.

• Regression of “fraud” according to “online_order”, "ratio_to_median_purchase_price”, “distance_from_home” and “used_pin_number”

```
Logistic regression                                     Number of obs = 1,000,000
                                                        Wald chi2(4)  =  16449.00
                                                        Prob > chi2   =    0.0000
Log pseudolikelihood = -144742.23                       Pseudo R2     =    0.5118
```

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	6.05083	.0742545	81.49	0.000	5.905294	6.196366
norm_distance_from_home	146.1087	2.176056	67.14	0.000	141.8437	150.3737
norm_ratio_to_median_purchase	215.712	1.714815	125.79	0.000	212.351	219.0729
used_pin_number	-13.41294	.5593944	-23.98	0.000	-14.50933	-12.31654
_cons	-10.24491	.0868139	-118.01	0.000	-10.41506	-10.07475

```
Note: 33431 failures and 858 successes completely determined.
```

We observe the same effects as before. Depending on the value of R2, nearly half of the variations in fraud data are explained by the model. The coefficient of “online_order” is now worth 6.05.

Thus, with each addition of a control variable, the coefficient of “online_order” has increased. If the variables were continuous, this would mean that for each additional unit of “online_order”, the fraudulent aspect of a transaction would increase further when :

- the ratio of purchased price transaction to median purchase price

- the distance between the location where the transaction took place and the location of the cardholder's residence

- and the fact that the transaction required a personal identification number for the authentication of the cardholder

are taken into account.

We obtain a final coefficient of 6.05, which means that the online aspect of a transaction has a very significant effect on the probability of occurrence of a fraud, ceteris paribus.

6. CONCLUSION

Our analysis was aimed at providing an insight into the relationship between online transactions and credit card fraud. The cross frequency tables and the results of the logit regressions were helpful to understand the extent to which an online transaction could influence the fact that this transaction can be a fraud.

From the cross frequency tables, we have observed that 95% of fraudulent transactions were occurring online. On the contrary, 98.7% of non-fraudulent transactions were not online. It suggested that, as we thought, indeed, online transactions have a higher risk of fraud in comparison to other transactions that occur offline.

Our quantitative approach that uses logit regressions with control variables confirmed the previous result. With each control variable that we added, the coefficient of the “online_order” feature increased. This indicates an even stronger relationship between online transactions and the probability of fraud.

The limit is that the dataset may come from an unsure source. We have taken it from kaggle but we don’t know whether this dataset is reliable or not. Moreover, it does not necessarily include all the control variables that would be relevant in our model. Finally, the correlation between our control variables and our explanatory variable being very weak, their addition in the model is not perfectly justified.

In spite of this limitation, we were able to have a clear insight on the role that plays online transactions in credit card frauds and to draw logical and coherent conclusions.

BIBLIOGRAPHY

- Microsof Dynamics - Types of credit card fraud
- RANA, Priya J. et BARIA, Jwalant. A survey on fraud detection techniques in ecommerce. International Journal of Computer Applications, 2015, vol. 113, no 14.
- Delamaire, L., Abdou, H., & Pointon, J. (2009). Credit card fraud and detection techniques: a review. Banks and Bank systems, 4(2), 57-68.
- Bhatla, T. P., Prabhu, V., & Dua, A. (2003). Understanding credit card frauds. Cards business review, 1(6), 1-15.
- Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.

ORAL

Slides 4/5/6

A modifier si possible. Le tourner plus sous la forme état de l'art et hypothèses économiques.

Slide 9

Souligner que la moitié des transactions ne sont pas des fraudes

Slide 10

Dire que les variables continues ont à peu près la même distribution, donc pas de standardisation.

Mais unités inconnues et range différents donc normalisation.

Slide 22

Our analysis was aimed at providing an insight into the relationship between online transactions and credit card fraud. The cross frequency tables and the results of the logit regressions were helpful to understand the extent to which an online transaction could influence the fact that this transaction can be a fraud.

From the cross frequency tables, we have observed that 95% of fraudulent transactions were occurring online. On the contrary, 98.7% of non-fraudulent transactions were not online. It suggested that, as we thought, indeed, online transactions have a higher risk of fraud in comparison to other transactions that occur offline.

Our quantitative approach that uses logit regressions with control variables confirmed the previous result. With each control variable that we added, the coefficient of the "online_order" feature increased. This indicates an even stronger relationship between online transactions and the probability of fraud.

Limits :

- unsure source, we don't know whether this dataset is reliable or not

- does not necessarily include all the control or omitted variables that would be relevant in our model

- the correlation between our control variables and our explanatory variable being very weak, their addition in the model is not perfectly justified

To go further : add some interaction terms in our model in order to determine whether the impact of ‘online_order’ on ‘fraud’ depends on our control variables

In spite of this limitation, we were able to have a clear insight on the role that plays online transactions in credit card frauds and to draw logical and coherent conclusions.