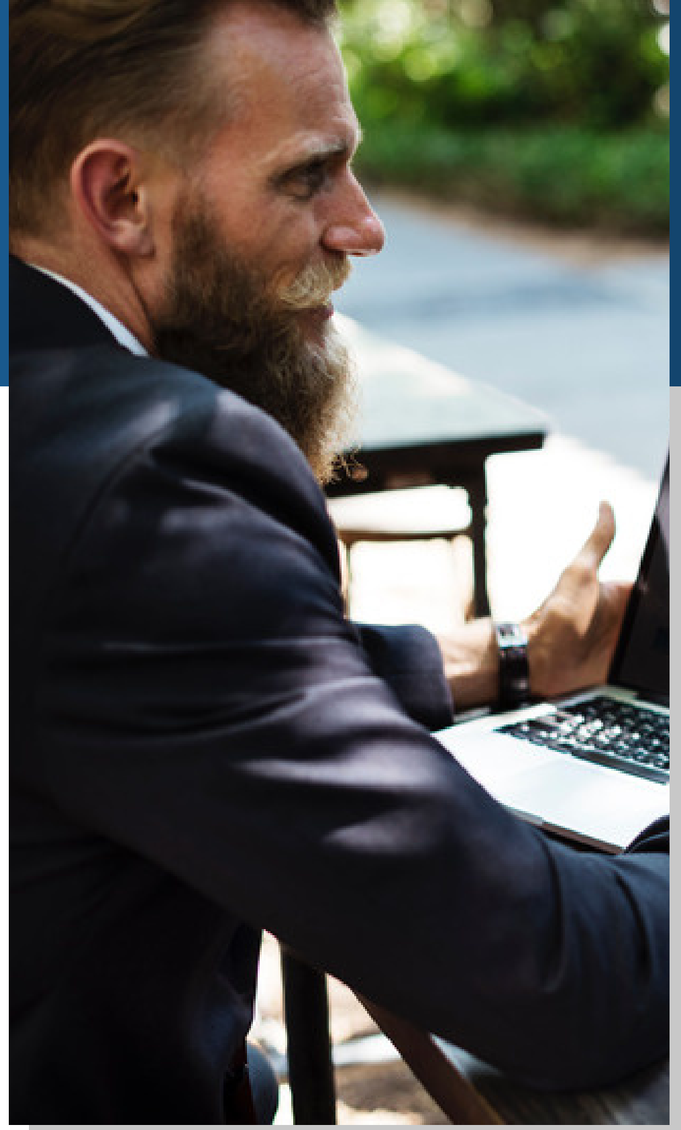
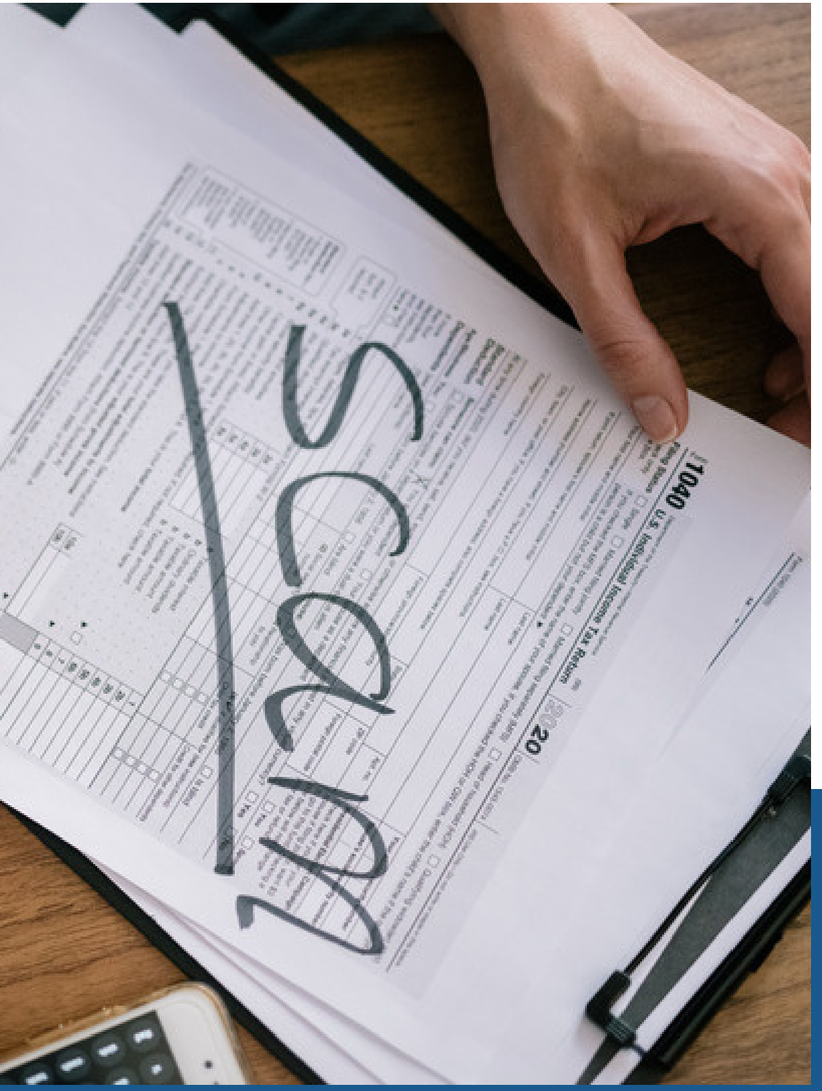
Modern Data Architectures for Big Data II

Fraud Detection Problem

Group 2
Federico Cañadas
Lia Dollison
Paulino Herrera
Sergio Reyes
Andrea Palomino
Jesus Fuster
Lutho Dabula
Johnny Naime

Project Index





1. Description

- Using the data sourced from Kaggle, our goal is to predict whether or not a Merchant will be a "fraudster"
- What does it mean to be a fraudster?
- We will use tree based machine learning algorithms in order to predict this binary classification problem

Merchant Fraud [noun]/Mər • CHənt • frôd/

Merchant fraud can refer to any situation in which a bad actor pretends to be a merchant, with the intent of committing fraud against either consumers or a financial institution.

Why is it important to uncover fraudulent merchants?

- Fraudulent merchants can harm consumers and e-commerce sites that host these merchants
- Fraudulent merchants might be stealing financial information, and selling unsafe or ineffective products, etc.
- They also harm the reputation of e-commerce.

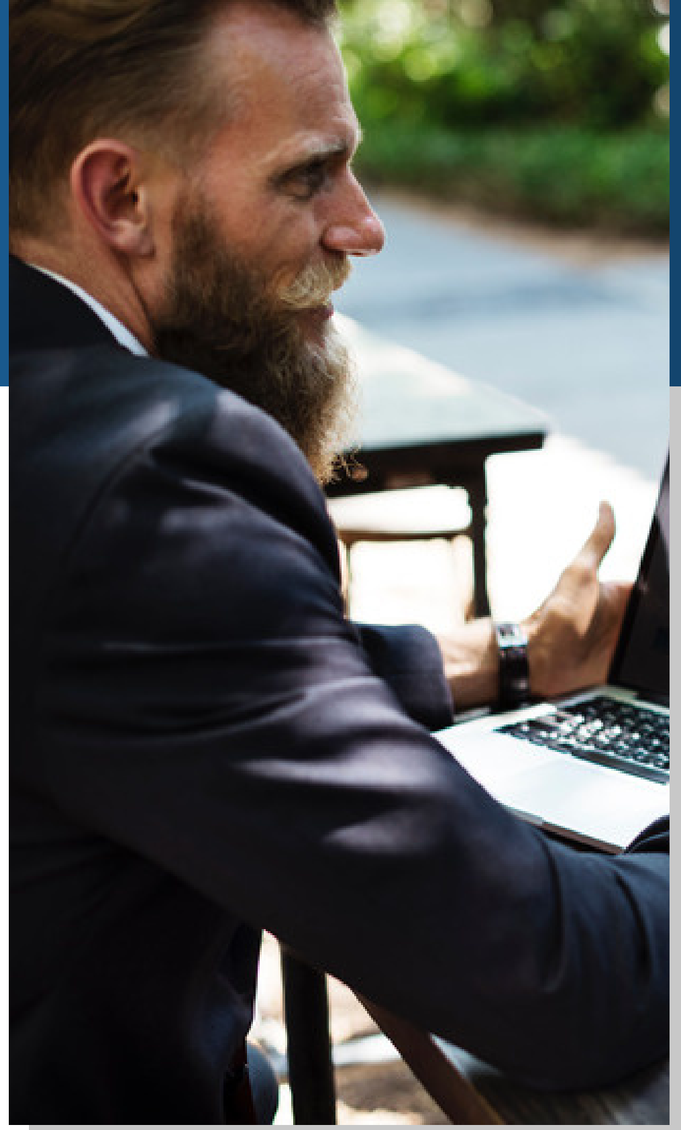
Data Flow

Sources → Ingestion → Storage → Processing

2. Data Sources

- We used Kaggle to download this dataset containing Payment Fraud activities in .csv format

- This download resulted in three datasets
  a. Order Data
  b. Merchant Data
  c. Label Data



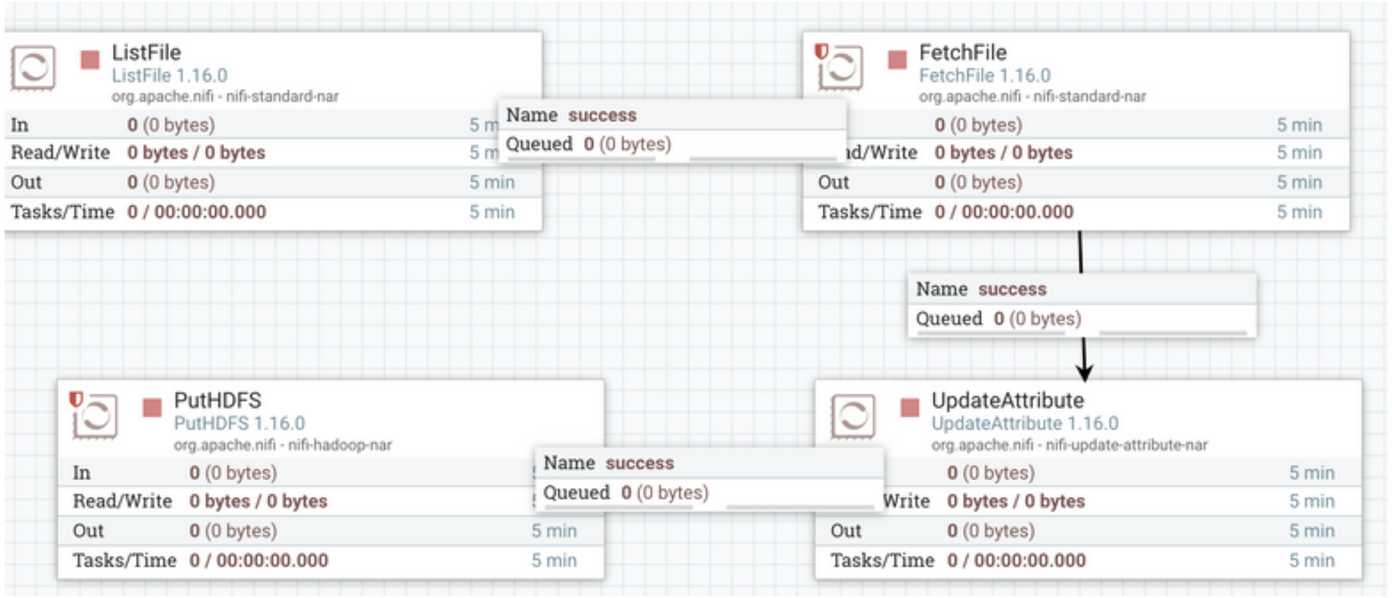
Data Flow

Sources → Ingestion → Storage → Processing

3. Big Data Pipeline

- Data is downloaded and then loaded into **HDFS** using **Nifi**
- **Spark Processing** is then used to begin the **Machine Learning Processing** and **Graph Processing** of our data

3. Big Data Pipeline



Processing Group:

1. **List File** – Creating an empty FlowFile with only the path to the Landing File. Staging area checked every 30 seconds.
2. **Fetch File** – Reading contents of files in staging area and stream it to incoming FlowFile
3. **UpdateAttribute** – change attributes of FlowFile by adding 2 properties
4. **PutHDFS** – store FlowFile in HDFS

Data Flow

Sources → Ingestion → Storage → Processing

4. Graph Processing

- Help to identify the **relationships** and **patterns** in fraudulent activites
- Detection of **Anomalies** and **outliers**
- Identification of the most **influential nodes** or clusters related to fraud

GraphFrames Spark

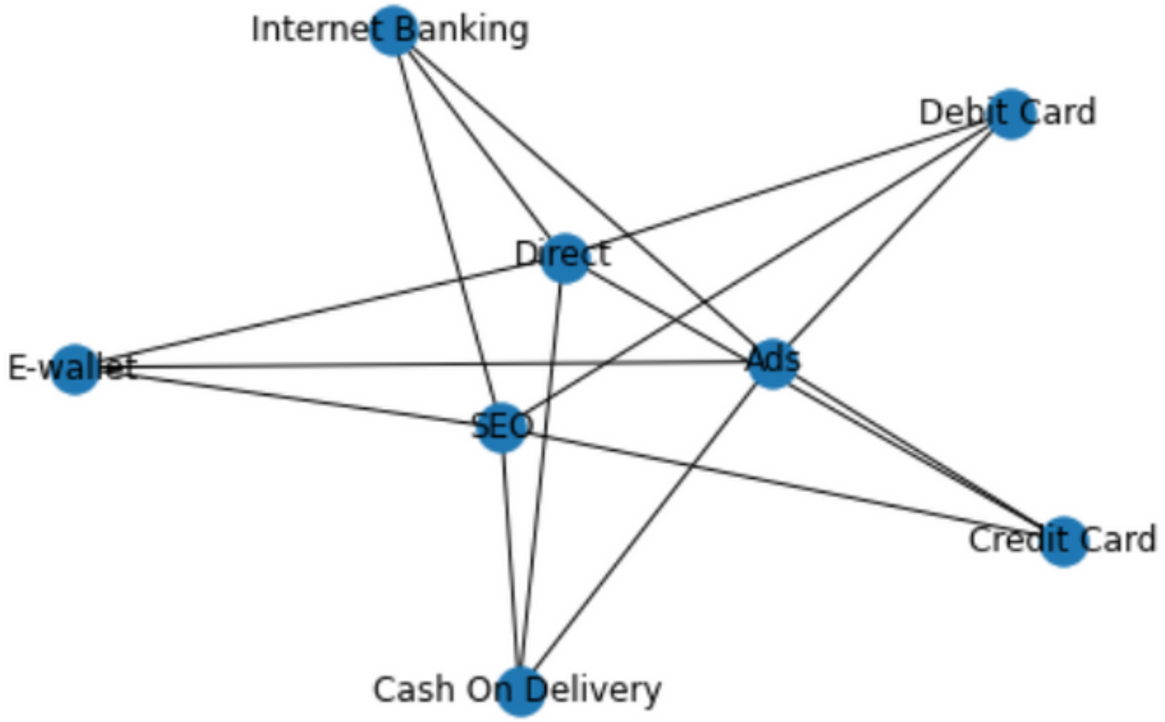
- Creating a GraphFrame with only the fraudsters

- Vertices:
  - Order_Payment_Method (ID)

- Edges:
  - Order_Source (src)
  - Order_Payment_Method (dst)
  - Gender (gnd)

GraphFrames Spark

- Graph 1:



GraphFrames Spark

- Most common Fraudster Profil by Order_Value_USD:

- SEO, E-wallet, Male and Age 43

	Order_Value_USD	Order_Source	id	Gender	Age
0	270	SEO	E-wallet	M	43
1	252	SEO	Internet Banking	M	24
2	250	Direct	Credit Card	F	46
3	250	Direct	Internet Banking	M	34
4	245	SEO	Credit Card	F	23
5	245	Ads	Debit Card	M	53
6	240	Ads	E-wallet	M	43
7	238	SEO	E-wallet	M	42
8	238	SEO	E-wallet	F	29
9	238	Ads	E-wallet	F	38
10	235	Ads	Internet Banking	F	39
11	230	Ads	Internet Banking	F	35
12	230	SEO	Internet Banking	F	28
13	230	Direct	Internet Banking	F	49
14	230	Ads	E-wallet	F	18

GraphFrames Spark

- Number of Edges per Order Source and Payment Method:

- Ads and Credit Card

	src	dst	count
0	Ads	Credit Card	853
1	SEO	Credit Card	757
2	Direct	Credit Card	552
3	Ads	Internet Banking	445
4	SEO	Internet Banking	406
5	SEO	E-wallet	340
6	SEO	Debit Card	337
7	Ads	E-wallet	325
8	Ads	Debit Card	307
9	Direct	Internet Banking	283
10	Direct	Debit Card	172

GraphFrames Spark

- From the total amount of fraudulents merchants the majority of this is done by Males.

	Gender	count
0	F	2035
1	M	2990

- From any of the methods employed to commit the fraud have a similar amount of average order value

	id	avg_order_value
0	Credit Card	93.610083
1	Internet Banking	92.004409
2	Debit Card	91.719363
3	E-wallet	90.696629
4	Cash On Delivery	89.732143

- The same analysis can be done for the order source, as it has almost the same average order value.

	Order_Source	avg_order_value
0	Ads	91.705762
1	Direct	92.405636
2	SEO	93.088748

5. ML Processing

- Accessible using Sparks Machine Learning library, **MLlib**
- Well suited for large scale machine learning applications due to its **scalability**
- Known for its high processing speed so it is ideal for training machine learning models on large datasets

EDA

After we set up the data in the Spark Eviroment, we analyzed the dataset.

Age

People in their 30's are more likely to commit fraud

Source

"Ads" are the order source with more fraudsters in comparison to SEO and Direct Orders

Payment Method

Credit Card is the prefered method for commit fraud.

Gender

Gender does not play a large role in fraudster description

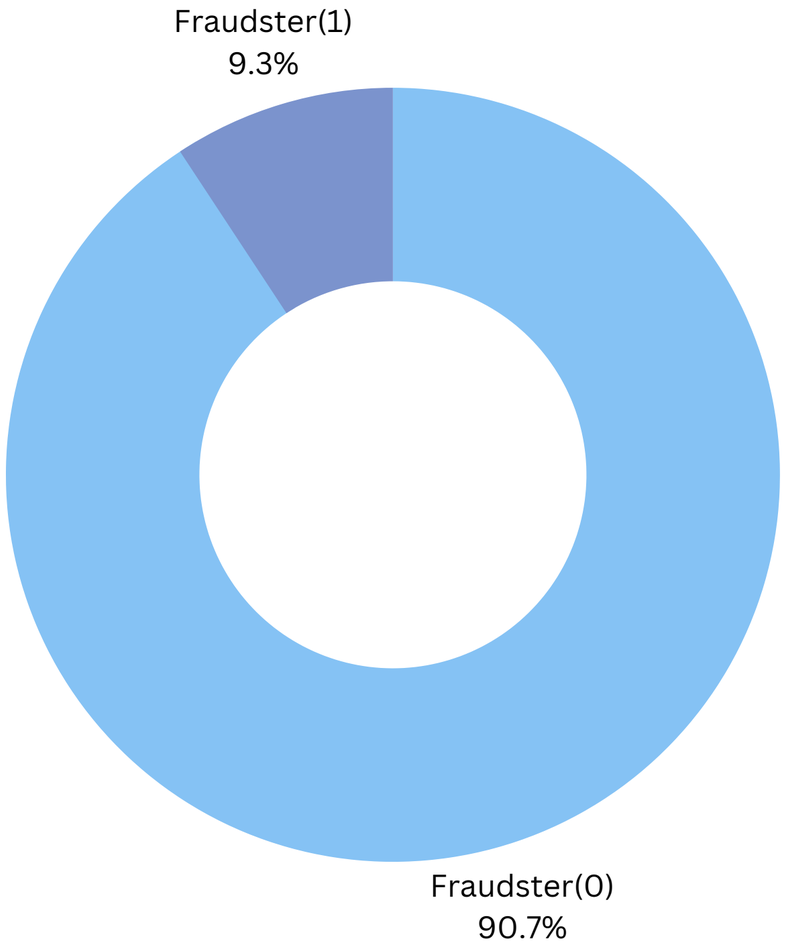
Feature Engineering

Quartly: We discover that in the first three months of the year there are more fraudsters.

We use the column "Date_of_Order" to extract the month of the year when the order was placed using regular expressions.

After that we replace those months with the values Q1 till Q4.

OVERSAMPLING



- Based on the data, we can extract the percentage of merchants who are "fraudsters". We see that only 9.3% of the data has Fraudster=1.
- We use the explode function to create additional data points for the minority class.
- After we create this new data points for the minority class we join it with the mojority one

ONE HOT ENCODING

We instantiate the One Hot Encoder and define "inputCols" as the names of the categorical columns we want to transform and "outputCols" as desired name of the result of the OHE transformation

Index	Color	Color_Array
0	Red	[Red]
1	Blue	[Blue]
2	Green	[Green]

Color_OneHotEncoded
(3,[0],[1.0])
(3,[1],[1.0])
(3,[2],[1.0])

each vector has a 1 in the position corresponding to the original value and a 0 everywhere else

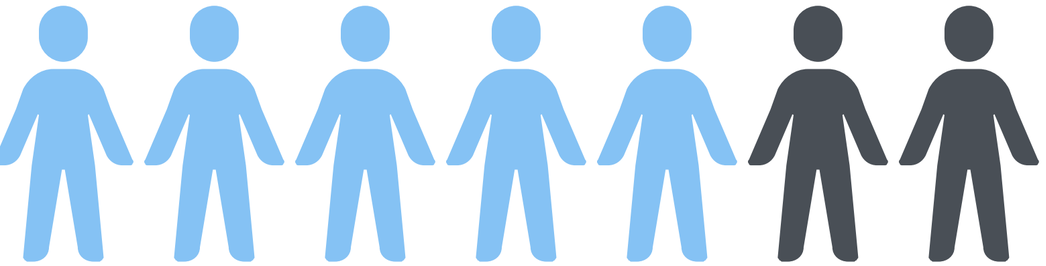
After the fitting the encoder and transforming the data using the "transform" method with only the columns of interest, the output is a dataframe with one hot encoded vectors as columns

on the next slide, you can see our columns of interest

ONE HOT ENCODING

Order_Source	This colunm has SEO, Direct and Ads
Quarter	Q1, Q2, Q3 and Q4
Order_Payment_Method	Credit Card, Internet Banking, Cash On Delivery, E-wallet, Debit Card
Gender	M: Male , F: Female

OUTLIERS



We have ages ranging between 18 to 72 y/o

We have Orders between $22 and $385

Models

DecisionTree

Basic decision tree model that has leaves and nodes to set-up conditions for the classification

AUC ROC: 0.28

RandomForest

Tree model that applies bagging to optimize the classification

AUC ROC: 0.83

Gradient-Boosted Trees

Tree model that applies boosting to optimize the classification

AUC ROC: 0.76

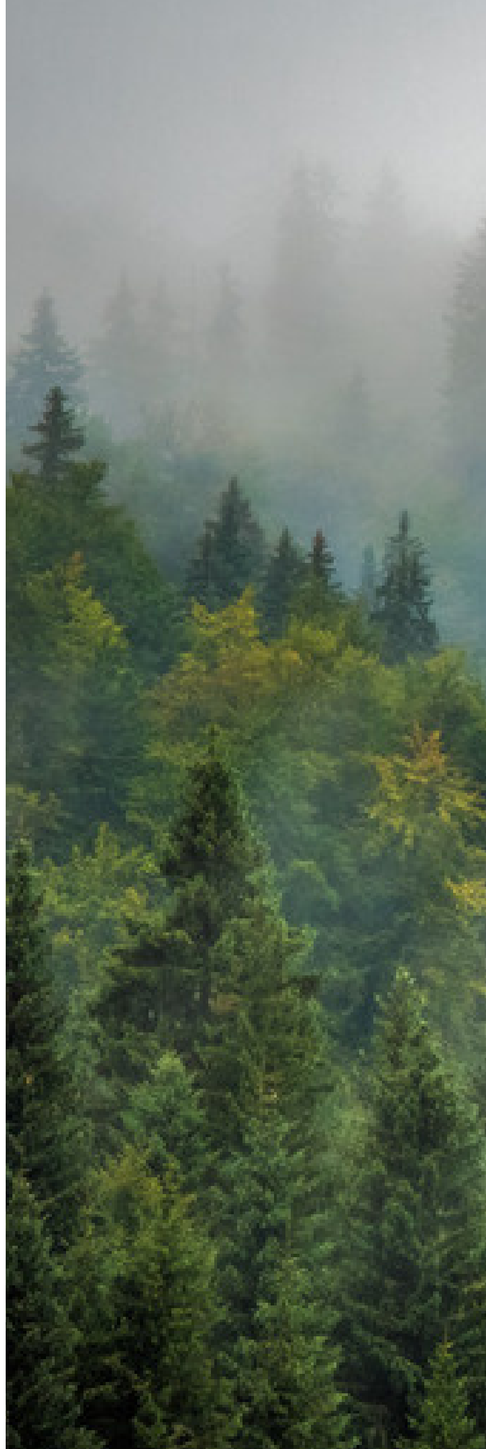
6. INSIGHTS

The most important features of our best model were the Order Source and the Order Payment Method to determinate if the merchant were a fraudster or not.

During the process of developing our Machine Learning Project we attempted to analyze the following hypothesis:

1. **How the number of years registered in the company affected the probability of being a "fraudster"?**

This variable was calculated by subtracting "Merchant_Registration_Date" from the current year.
We saw that every merchant had been registered for the same number of years (registered in 2018) so the new column was dropped due to lack of importance



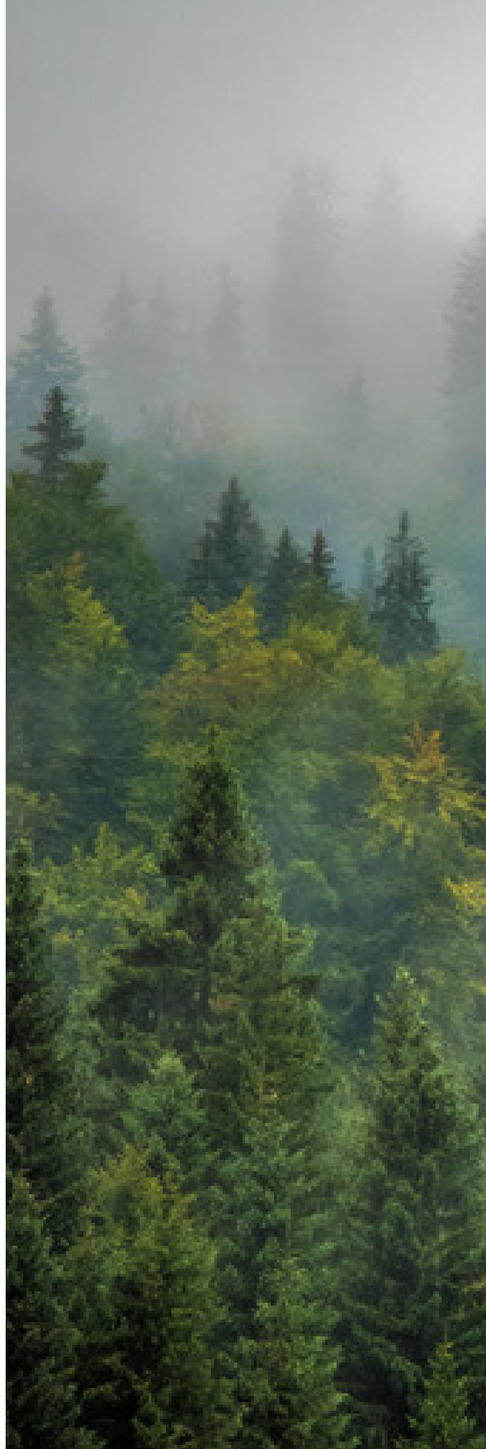
6. INSIGHTS

The most important features of our best model were the Order Source and the Order Payment Method to determinate if the merchant were a fraudster or not.

During the process of developing our Machine Learning Project we attempted to analyze the following hypothesis:

2. How many merchants have the same devices?

We thought that the fraudsters can operate with a single device.
After we apply the function to get the number of devices we discover that the devices in the whole dataset were unique.



7. CONCLUSIONS

- Merchants are more willing to commit fraud at the very beginning of the year. The fraudsters are young people around 30 years old.

- RandomForest was the best model to predict fraudster merchants with main features as Order Source and the Order Payment Method. We also see this relationship in the Graph with a common values Order Source "Ads" and Payment Method "Credict Card".

- Finally, Spark MLib was a good tool to deal with a big amount of observations and process the models faster than common Python.

Andrea Palomino

Federico Cañadas

Lia Dollison

Lutho Dabula

The Team

Paulino Herrera

Sergio Reyes

Johnny Naime

Jesus Fuster