TheDataOpen

The Data Open

European Datathon - Team 13

Contents

1 Non-Technical Executive Summary

In this project, we analyzed BigSupply Co's supply chain management data sets, a multinational retailer that specializes in clothing, fitness, and electronics.

More precisely, the company makes a profit between the purchase price of the products and their resale price for each order submitted by a customer. This gain can be positive, in which case the firm makes a profit from the sale, or it can be negative (e.g., due to shipping costs, product discounts), in which case the firm makes a loss. After an exploratory analysis, we find that the firm's profits follow a sort of normal distribution, with a majority of profitable orders making it profitable. However, we find a large number of unprofitable orders. In addition, during the last months of 2018, the company has suffered large variations in profit, as well as a sharp drop in the last months of this year.

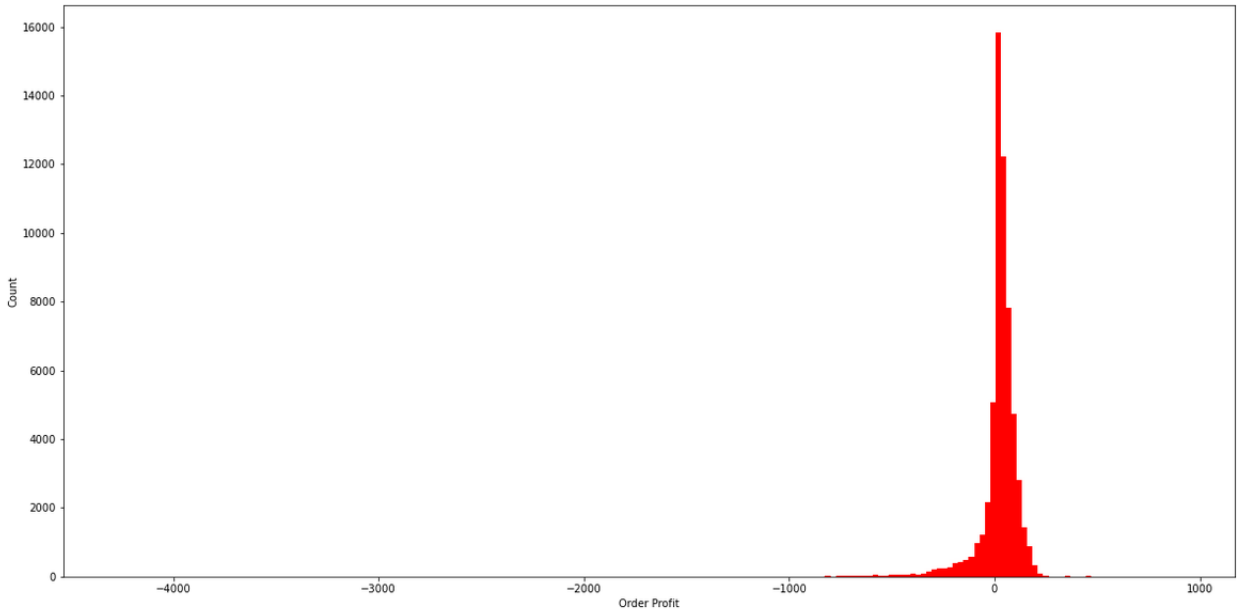

Figure 1: Orders Profit Distribution (2015-2018)

Our first goal was to understand which components were the most important in explaining the company's profit. In other words, we wanted to understand which customers are the most and least profitable over the long term, and what their behavioral characteristics are (type of customer, type of product ordered, country of shipment etc.).

Following this, we focused on the discount rate of each order, which against all intuition

does not seem to be correlated to the item quantity or to the customer profitability. We then built a new discount rate function which relies more on the long term profitability of each customer. By performing a backtesting, this new function leads to a 17% increase in revenue between 2015 and 2018. Such an analysis is obviously not complete, as the model has its limitations : the market is a dynamic system in which participants would adapt to a new discount rate. However, our study highlights the deficiencies in the way discount rates are assigned, and that a more consistent function could lead to a potentially significant increase in firm profits.

In addition, we also conducted an exploratory analysis of orders classified as potentially fraudulent. Specifically, we attempted to understand which client characteristics are present in orders that are considered fraudulent, and which features would help predict whether or not an order is fraudulent.

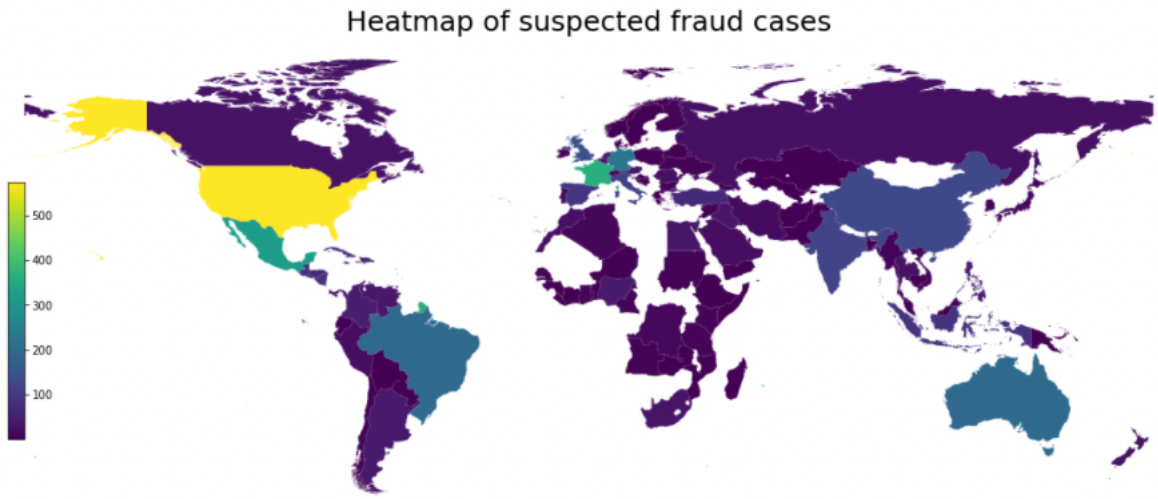

Figure 2: Fraudulent transactions distribution by country (2015-2018)

Problem Statement :

- ⋆ Which components explain the temporal evolution of the firm's profit?
- ⋆ How can we classify customers according to their long-term profitability?
- ⋆ How can we optimize the order discount rate function to make it more consistent with customer profitability?
- ⋆ What are the most common features of suspected fraud cases?

2 Modeling

Hypothesis & Reductions : In all this part, we filtered the dataset to only keep `COMPLETE` and `CLOSED` shippings. In fact, for all the other orders (e.g. `PENDING PAYMENT` or `CANCELED ORDERS`), the order profit cannot be included in the total revenue of the company until the shipping has been completed.

2.1 Client Sharpe Ratio

Let $\{1, ..., T\}$ be our time period (here we consider every day from January 2015 to December 2018). Denote by $\mathcal{C} := \{c_1, ..., c_n\}$ the set of clients who submitted an order during the time period.

Assume that each client c_i submits shipping orders, and that they bring a daily profit $o_i^{(t)} \in \mathbb{R}$ to the company, with $t \in \{1, ..., T\}$ (and $o_i^{(t)} = 0$ is no order is submitted by the client during the day). The i-th client brings an average profit of :

$$\overline{p}_i := \frac{1}{T} \sum_{t=1}^{T} o_i^{(t)} \in \mathbb{R}.$$

Notice that an analogy can be made with financial markets. Each client can be interpreted as a financial security, and the company as a trading firm. For each security (= each client), the company has a day trading strategy so that the inventory is empty at the beginning and the end of each day. Every day, the trading strategy of the i-th security brings a profit $o_i^{(t)} \in \mathbb{R}$. The combination of all these securities is a portfolio, and our goal is to optimize its performance.

In the modern portfolio management theory, a good porfolio maximizes the expected return of its assets with a low risk. Here, the shipping company is profitable is the orders bring high profits with a low 'profit volatility'. Hence, for each client, we can built a profit/risk metric of each client which is similar to a shape ratio :

$$\forall i \in \{1, ..., n\}, \quad \xi(c_i) := \frac{\overline{p}_i}{\sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(o_i^{(t)} - \overline{p}_i \right)^2}}$$

We define this function $\xi : \mathcal{C} \to \mathbb{R}$ as the *client sharpe ratio*.

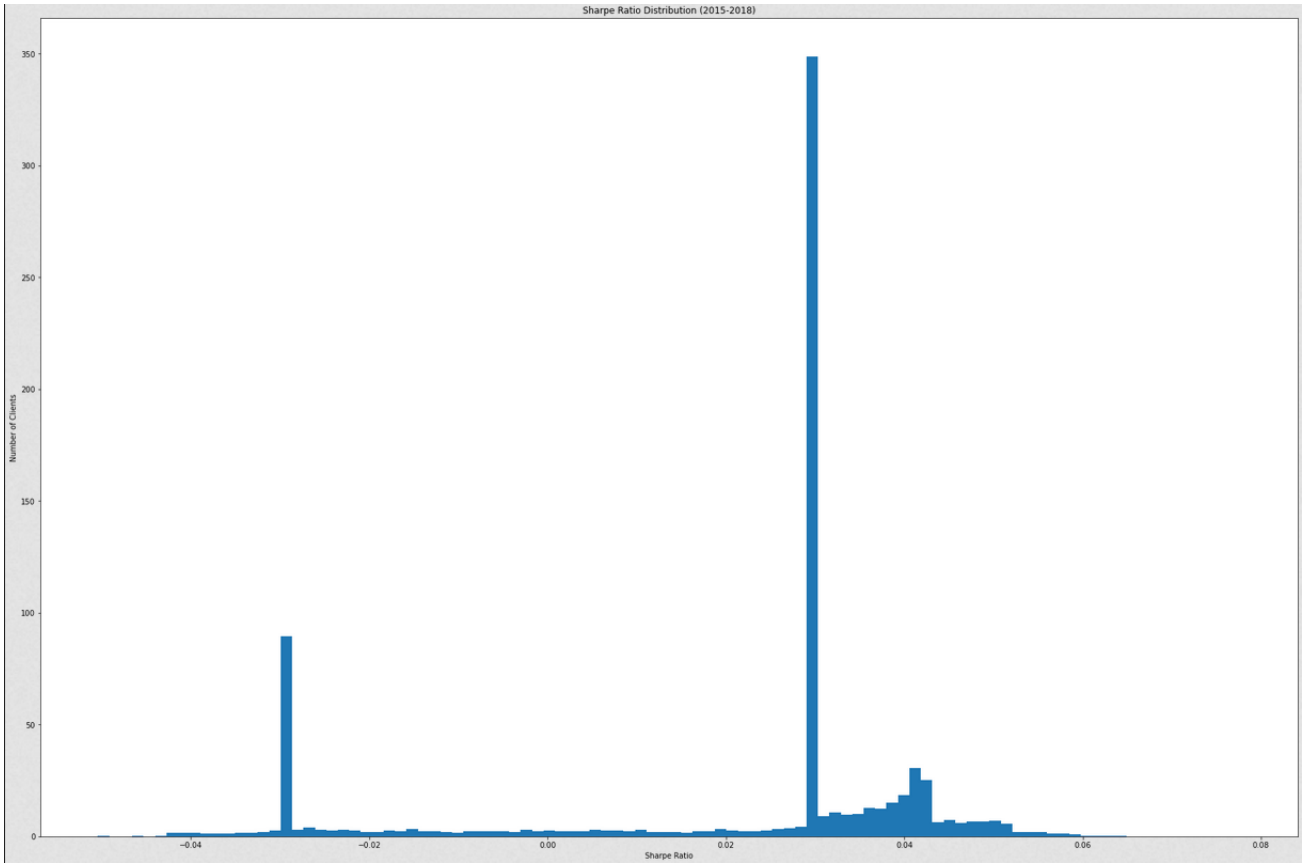

Figure 3: Clients Sharpe Ratios distribution (2015-2018)

The clients with the highest sharpe ratios will be those who bring high profits with a low variance (e.g. 200€ everyday for 3 years), whereas those with low and negative ones will generate high losses, with a lot of volatility.

2.2 Clients Segmentation

The sharpe ratio metric measures the *profit-risk profile* of each client. Based on that, we will propose a segmentation of $\mathcal{C}$. More precisely, let $\mathcal{D}$ be the distribution of $\xi(C)$, with $C \sim \text{Uniform}(\mathcal{C})$, and q_α the α quantile of this distribution ($\alpha \in [0,1]$). For every $(\alpha, \beta) \in [0,1]^2$ with $\alpha < \beta$, define the (α, β)-*profit risk set* by:

$$\mathcal{R}_{[\alpha,\beta[} := \left\{ c \in \mathcal{C} \,\middle|\, \xi(c) \in [q_\alpha, q_\beta[\right\}.$$

Here, we proceed to the following partition of the clients set :

$$\mathcal{C} := \mathcal{R}_{\left[0,\frac{1}{5}\right]} \cup \mathcal{R}_{\left[\frac{1}{5},\frac{2}{5}\right]} \cup \mathcal{R}_{\left[\frac{2}{5},\frac{3}{5}\right]} \cup \mathcal{R}_{\left[\frac{3}{5},\frac{4}{5}\right]} \cup \mathcal{R}_{\left[\frac{4}{5},1\right]}.$$

Clients in the set $\mathcal{R}_{\left[\frac{3}{5},\frac{4}{5}\right]}$ are those with a high sharpe ratio (high profit and low risk), and clients in the set $\mathcal{R}_{\left[0,\frac{1}{5}\right]}$ are those with the highest sharpe ratios (losses with a huge

volatility). These pattern can be obeserved when we plot the daily profit lines per profit-risk set :

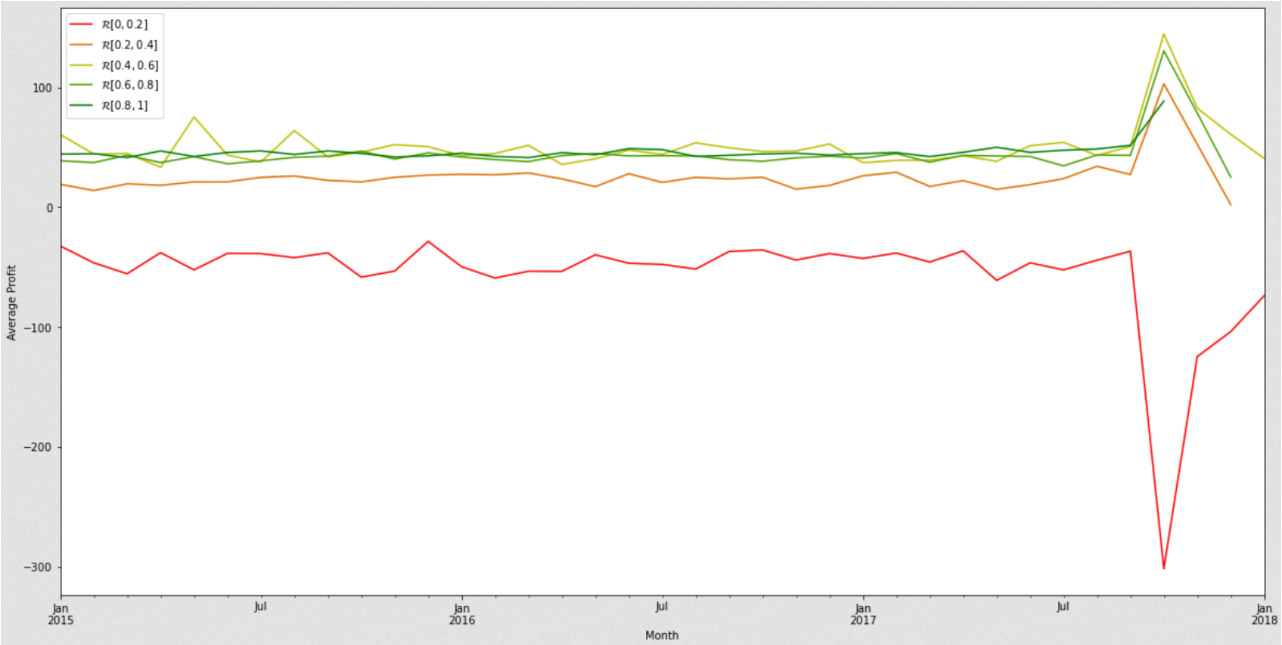

Figure 4: Monthly average profit per profit-risk set (2015-2018)

In figure 2, we see that clients with the lowest sharpe ratio are unprofitable for the company, and we observe huge profit variations in the last months of the year 2018. On the opposite side, profitable clients with a high volatility generate a lot of profit, but not enough to offset the losses. This uncontrolled PL volatility can explain the strong profit drop of the company in 2018 :

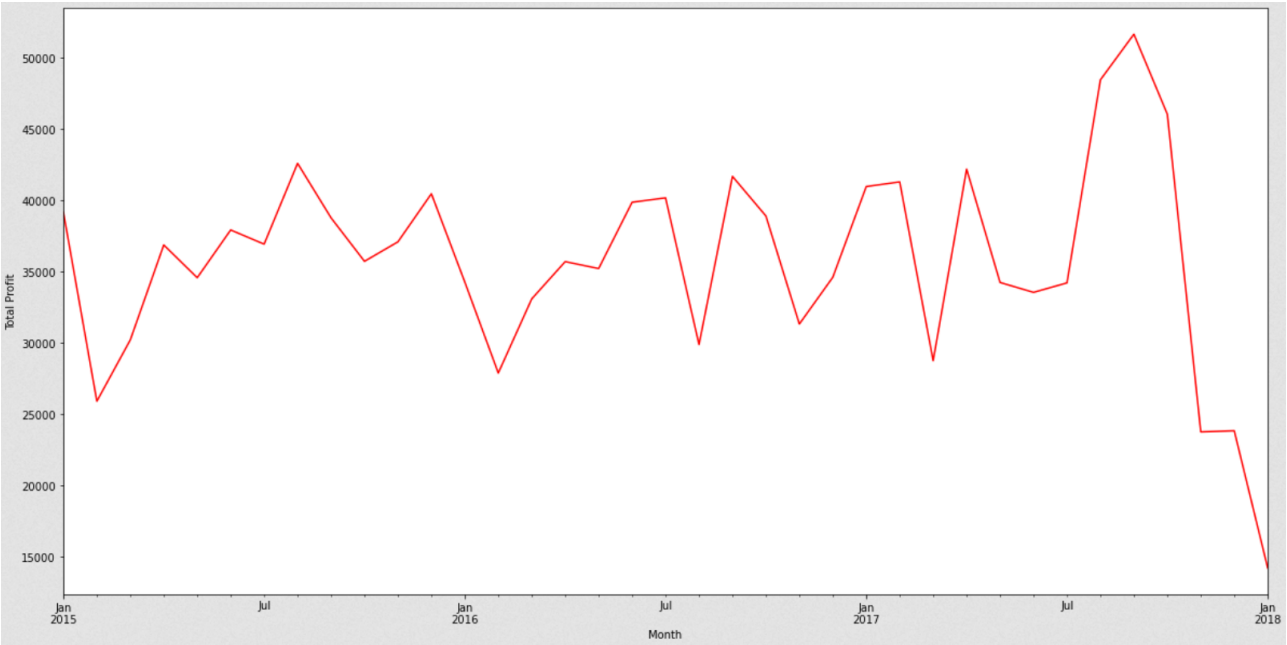

Figure 5: Total profit per month (complete & closed orders)

2.3 Order Discounting Rate : Statistical Analysis

We notice that the total price of each order can be discounted. Hence, each client only pays the total order price minus the discount amount

Sales	Order Item Discount	Order Item Total	Order Profit
327.75	13.10999966	314.64001460000003	91.25
327.75	16.38999939	311.35998539999997	-249.0899963
327.75	18.03000069	309.7200012	-247.7799988
327.75	22.94000053	304.80999760000003	22.86000061
327.75	29.5	298.25	134.2100067

Figure 6: `Order Item Total` = `Sales` - `Order Item Discount`

To continue the analogy with financial markets, the discount amount can be interpreted as the broker's order execution fee. If the broker charges high execution fees, then traders will have to substract them from their trading strategy profit.

The difference here is that the shipping company can choose its own discount rate for each order, which is equivalent to say that the trading company can choose its broker among an infinite pool of candidates with an infinite number of discount rates.

In practice, the discount rate is proportional to the ordered quantity, i.e. clients shipping large quantities will benefit from volume-dependent shipping discounts.

However, we observe that in our data set, there is no obvious correlation between the order quantity and the discount rate :

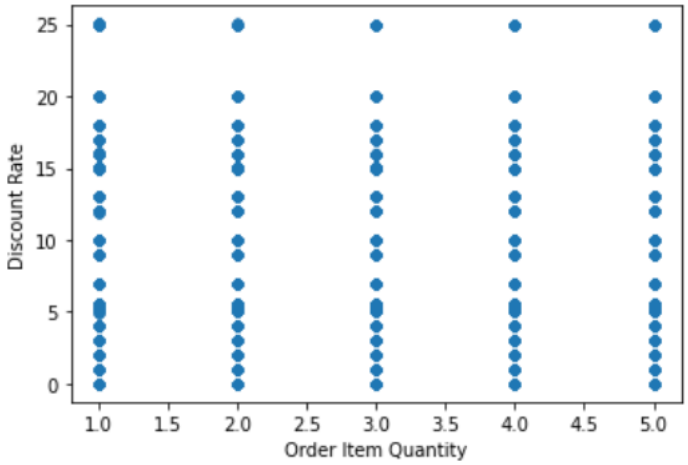

Figure 7: Discount Rate - Order Quantity scatter plot for every order between 2015 and 2018

Moreover, another intuition would be that 'bad clients', i.e. unprofitable clients with a negative sharpe ratio, should get low discount rates, to reduce the losses of the company. But again, there is no correlation between the order profit and the discount rate :

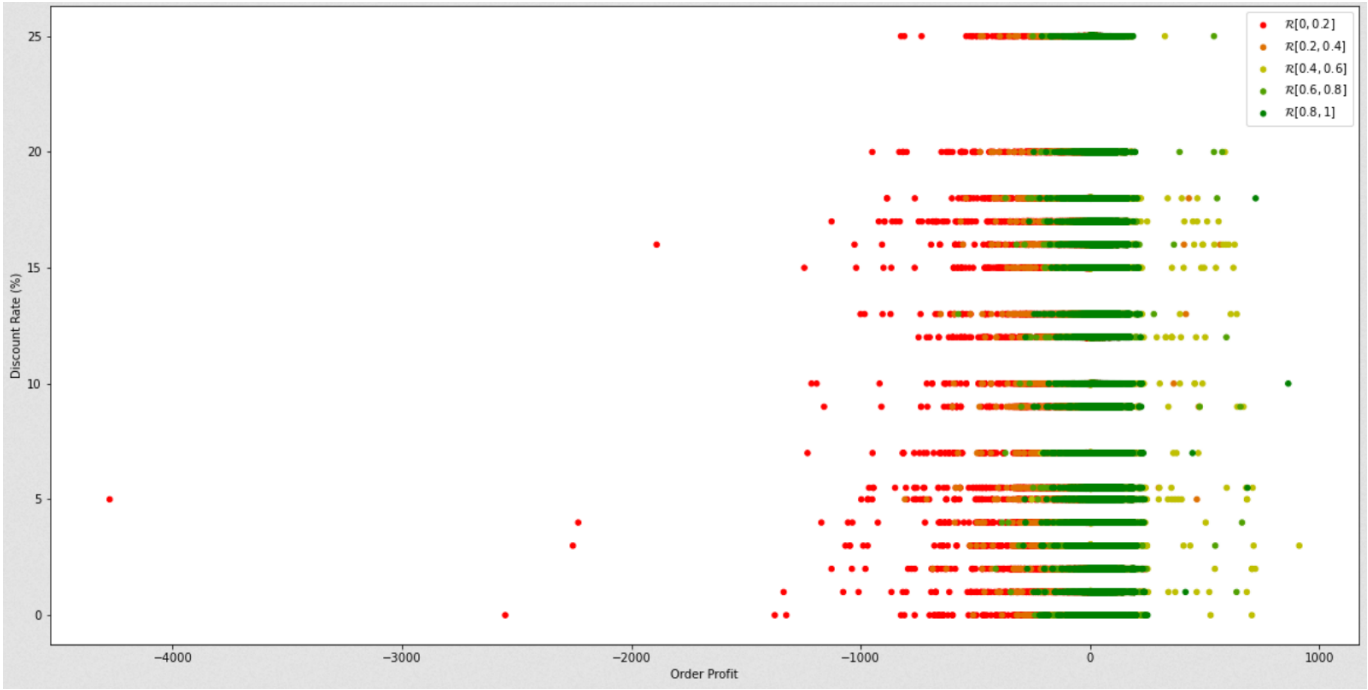

Figure 8: Order Profit - Discount Rate scatter plot for every order between 2015 and 2018

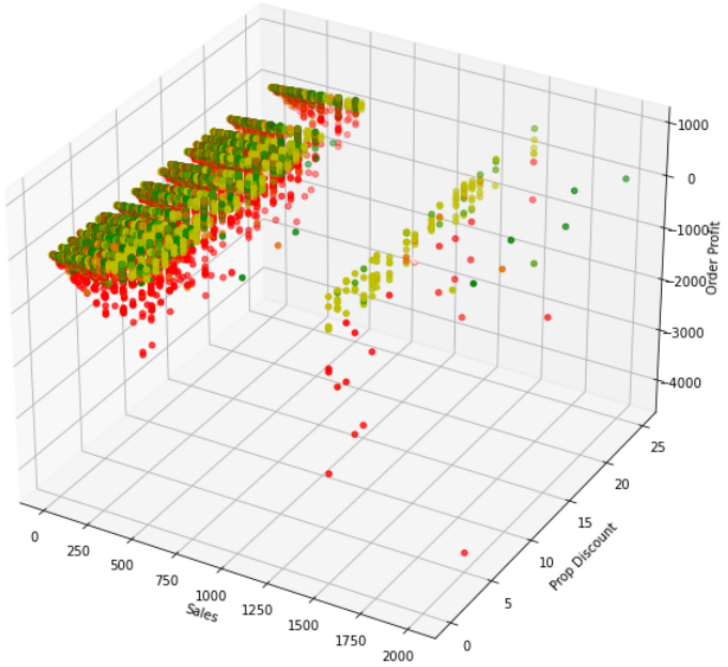

Figure 9: Order Profit - Discount Rate - Initial Order price scatter plot for every order between 2015 and 2018

We notice that profitable and unprofitable clients are treated the same in terms of discount rate, which is not intuitive. Hence, an unprofitable client can receive the maximum discount rate of 25%, which is not optimal. In the next parts, we will propose an optimization strategy to propose discount rates that are more consistent with the profit-risk profile of each client.

2.4 Order Discounting Rate : Client profile

Before considering our discount rate 'optimization' strategy, let us try to understand how the order discounting rate 'algorithm' of the company works, and what is the profile of the clients with the highest & lowest discount rates.

The order discounting rate follows a sort of gaussian distribution with regular peaks, centered around 10%, and all the values are contained between 0 and 25% :

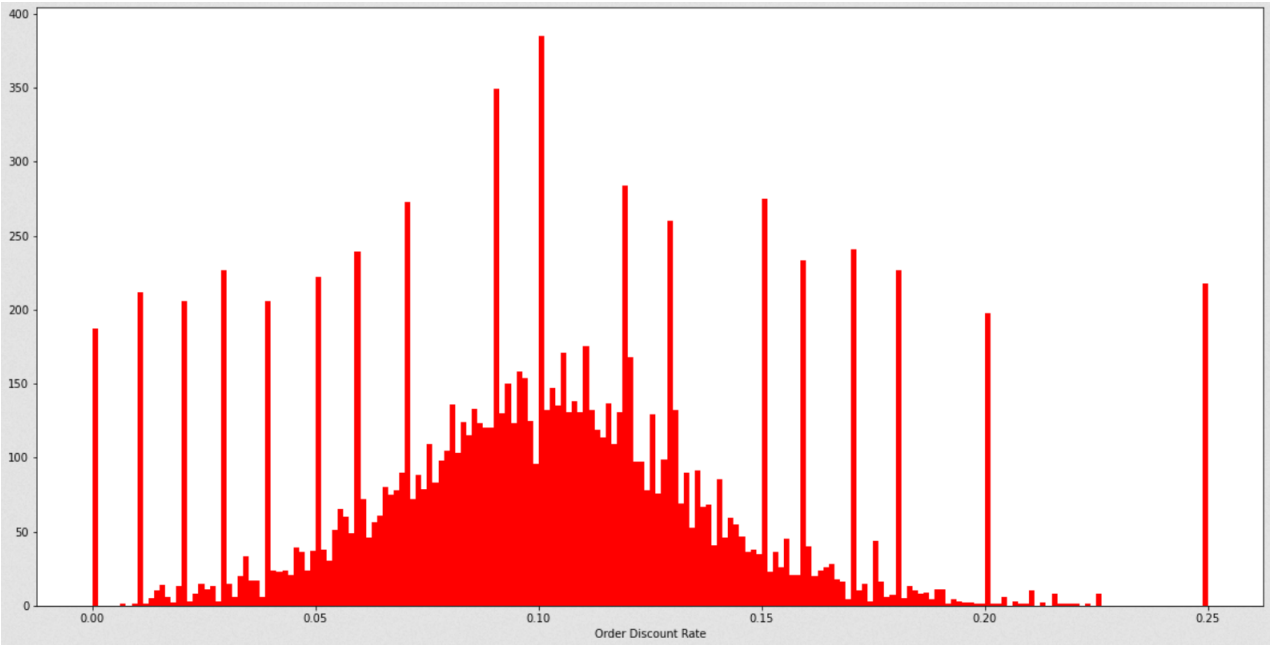

Figure 10: Average discount rate per client distribution (2015-2018)

Now, we want to determine who are the clients with the highest and the lowest discount rates, to understand the current 'pricing dynamic'.

To do that, we need to define our the characteristics of our clients. Assume that our clients are described by $g \geq 2$ categorical features (e.g. order country, customer segment, payment channel etc.), and that these features can take $f_1, ..., f_g \geq 1$ values respectively. By doing this, we obtain a combination of $\prod_{i=1}^{g} f_i$ combinations, and each client belongs to one of them.

Here, assume that the order of each client is considered through the prism of $g = 3$ categorical feaures:

- f_1 : The customer segment of the client (Consumer, Home Office etc.)
- f_2 : The market (Africa, Pacific Asia, Europe etc.)
- f_3 : The product class (Accessories, Camping & Hiking etc.)

For each client, we compute the average order discount rate of each client for each possible behaviour (= each combination of theses features) between 2015 and 2018, and we obtain the following table with 485 columns :

Order Customer Id	mean_Consumer_Africa_Accessories	mean_Consumer_Africa_Baseball & Softball	mean_Consumer_Africa_Boxing & MMA	mean_Consumer_Africa_Camping & Hiking	mean_Cons
1	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	
...	...	...	...	...	
20743	0.0	0.0	0.0	0.0	
20749	0.0	0.0	0.0	0.0	
20750	0.0	0.0	0.0	0.0	
20754	0.0	0.0	0.0	0.0	
20755	0.0	0.0	0.0	0.0	

2523 rows × 485 columns

Figure 11: Table describing the average order discount rate of each client for every possible behaviour

The average value of column is the average discount rate of all the clients with a specific behaviour. If we rank the 10 behaviours with the highest & lowest averages, we obtain the following bar chart :

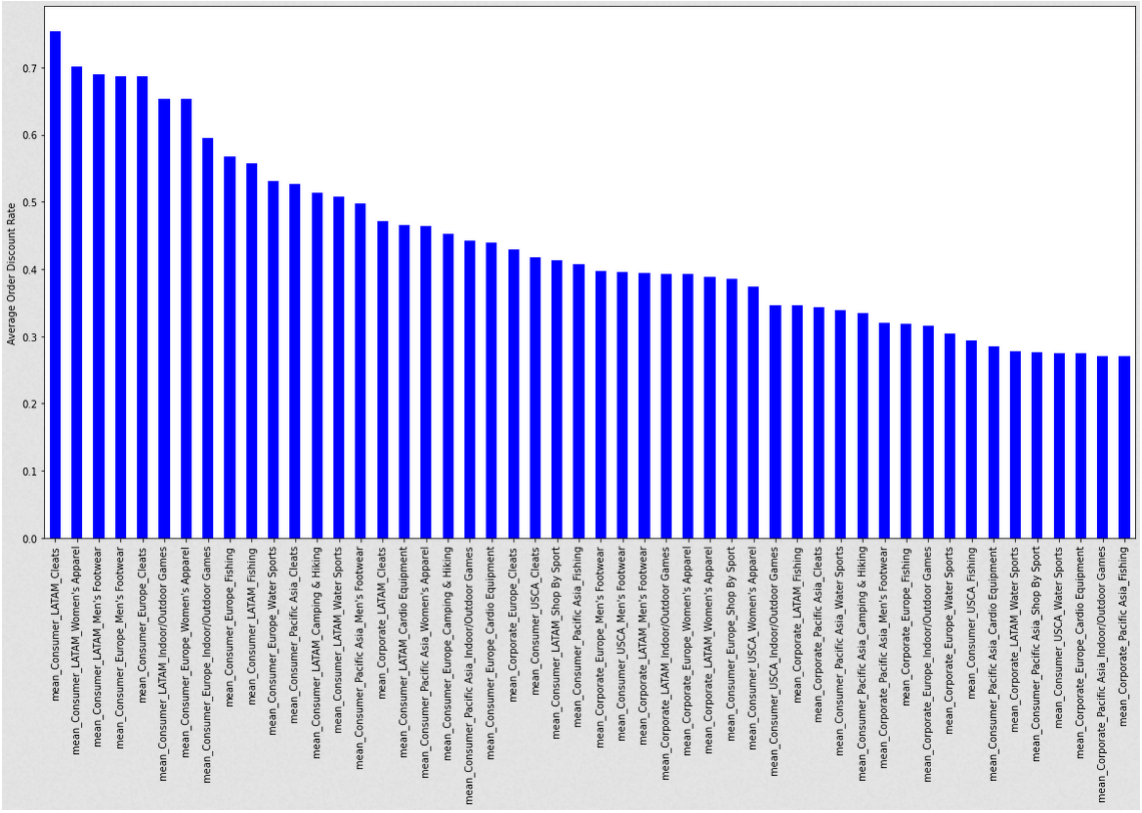

Figure 12: Profile of the clients with the highest average discount rate

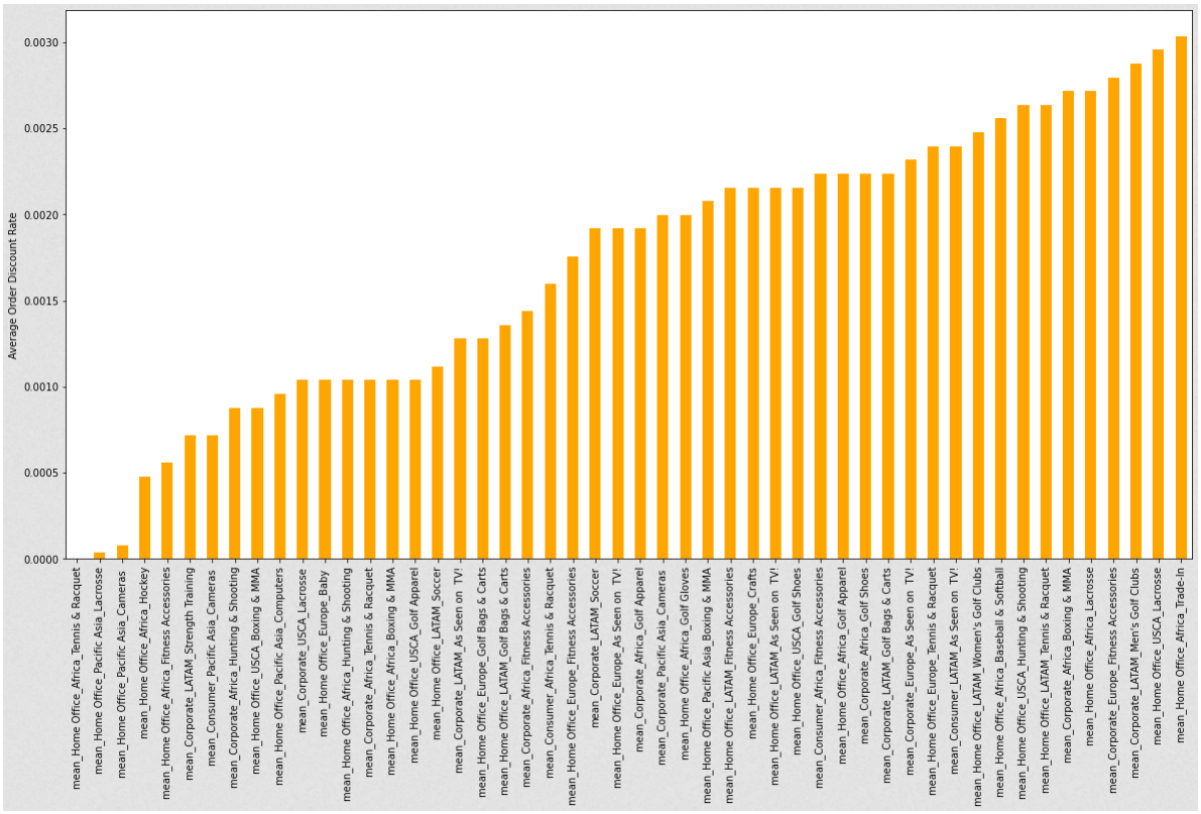

Figure 13: Profile of the clients with the lowest average discount rate

We observe that clients with the highest client with the highest discounting rates are consumers from Europe and Latin America. Those with the lowest discounting rates are home office clients from Pacific Asia and Africa. Hence, it seems that the discount rate is determined based on the segment type and the location of the clients. To confirm this intuition, we can perform a clustering of our clients set based on our table, and determine the most common charateristics of each clients in each cluster.

To perform this clustering, we first reduced the dimensionality of our dataset using the Uniform Manifold Approximation and Projection (UMAP), which is a competitive reduction algorithm with t-SNE to preserve the global structure of our initial dataset. Then we perform a hierarchical density-based spatial clustering of applications with noise algorithm (HDBSCAN) on the low dimensional projection (here 2D). We obtain 4 main clusters, and some noise :

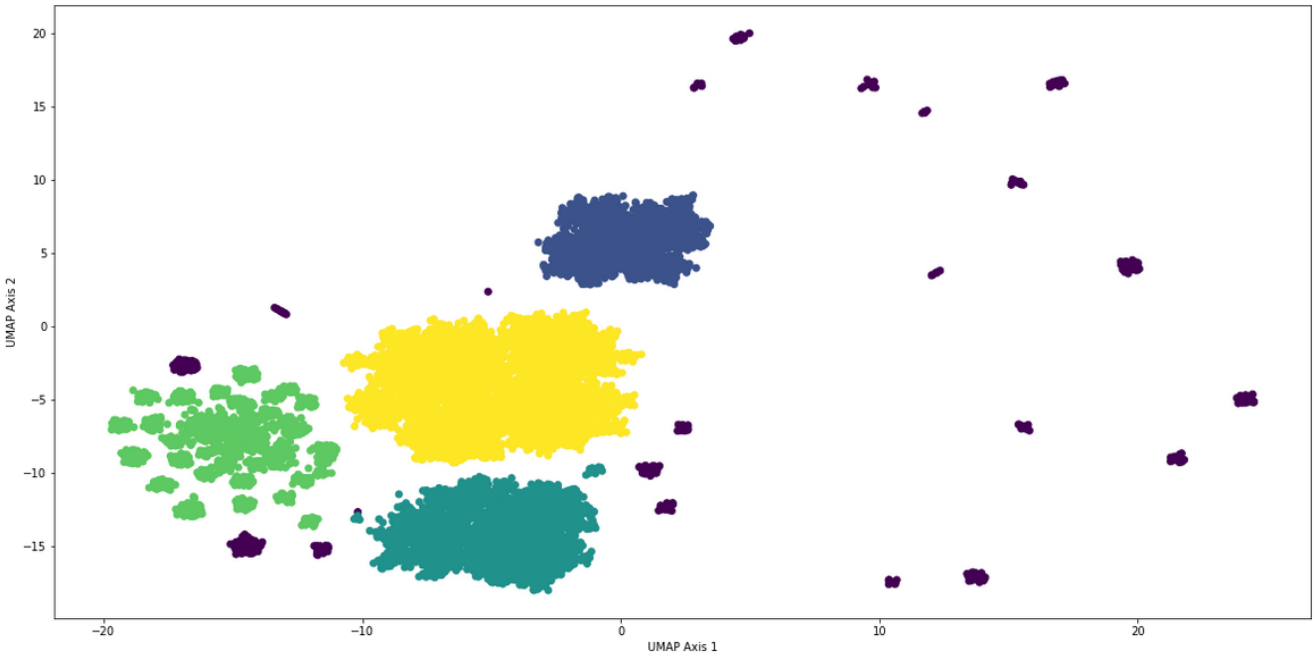

Figure 14: Clients clustering using UMAP and HDBSCAN.

By performing the same analysis, we conclude that each big cluster corresponds to one consumer segment (Corporate, Home Office, Consumer in Pacific Asia and Consumer in Europe and Latin America), cf. the Jupyter Notebook for more detail about the clusters compositions.

2.5 Order Discounting Rate : Optimization Strategy

2.5.1 Discount Rate Pricing Function

The intuition behind the optimization strategy is that unprofitable clients should have low discounting rates, and profitable clients should be allowed to get high discounting rates, to retain them. Instead of computing discount rates based on customer segment and location, we derive it from the profit-risk profile of the client.

Denote by $d_0 \in [0, 1]$ the maximum order discount rate for a client (here $d_0 = 0.25$). If a client submits an order with a total order item quantity $v > 0$ at time t, denote by $\Pi(c, t, v) \in [0, d_0]$ the current order discount rate proposed by the company. By doing this, we define a discount pricing function $\Pi : \mathcal{C} \times [0, T] \times \mathbb{N}_{>0} \to [0, d_0]$, which gives us the values in the current data set. After the order discount, the client c only has to pay $p\,(1 - \Pi(c, t, p))$, and it leads to a profit loss of $p\Pi(c, t, p)$ for the company. Given our previous analysis, we can assume that the function Π is independent of p.

Now, we want to built a new discount pricing function $\tilde{\Pi}$ that takes into account the total amount paid and the profit-risk profile of the client. Hence, it would be a function $\tilde{\Pi} : \mathcal{C} \times [0, T] \times \mathbb{N}_{>0} \times \to [0, d_0]$.

2.5.2 Discount Rate Pricing Function : Manual Construction

First, we assume that the order discount is not volume-based, not time dependent and that it only depends on the profit-risk profile of the client (i.e. $\tilde{\Pi} : \mathcal{C} \to [0, d_0]$).

A intuitive idea would be to consider a discount rate as a step function which takes high values for clients with a high sharpe ratio and low values for clients with a low sharpe ratio. If it leads to a lower discount rate, we keep it; otherwise, we keep the discount rate given by Π :

$$\forall c \in \mathcal{C}, \quad \tilde{\Pi}(c) := \min \left(\frac{d_0}{5} \mathbf{1}_{\mathcal{R}_{\left[\frac{1}{5}, \frac{2}{5}\right]}} (\xi(c)) + \frac{2d_0}{5} \mathbf{1}_{\mathcal{R}_{\left[\frac{2}{5}, \frac{3}{5}\right]}} (\xi(c)) + \frac{3d_0}{5} \mathbf{1}_{\mathcal{R}_{\left[\frac{3}{5}, \frac{4}{5}\right]}} (\xi(c)) + \frac{4d_0}{5} \mathbf{1}_{\mathcal{R}_{\left[\frac{4}{5}, 1\right]}} (\xi(c)), \Pi(c) \right)$$

The first term in the minimum has the following shape of step function (using the time period $[0, T]$ = [January 2015 - December 2018]

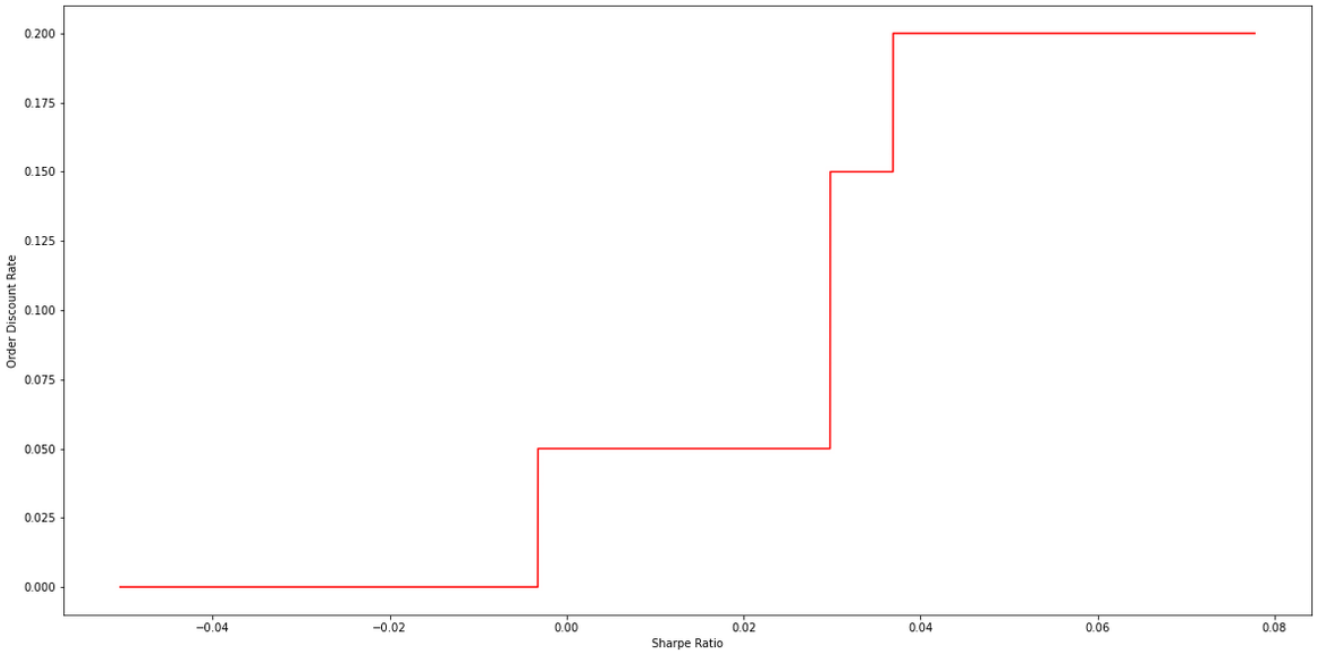

Figure 15: Fist term in the minimum in the definition of $\tilde{\Pi}$

Notice that this new discount rate pricing function necessarily improves the profit of each order, because $\tilde{\Pi}(c) \leq \Pi(c)$.

Now, we introduce the time dependency. When a client submits a new order at time t, we compute the sharpe ratio of the client based on its historic on the time period $[0, t]$,

and we compare it to the sharpe ratios of the other clients to determine the profit-risk profile of the client. From that, we apply $\tilde{\Pi}$ to determine the discount rate of the client. If the client has no order historic, the order discount is given by the current discount rate pricing function Π.

To keep the model simple, we decided to not include the total order price in the discount rate pricing function (i.e. $\tilde{\Pi}$ is not a volume-based discounting rate function).

2.5.3 Discount Rate Pricing Function : Backtesting

Here is a description of our backtesting methodology for the first three months :

- **January 2015** : We compute the sharpe ratio of all the active clients in January 2015, and we determine their profit-risk sets.
- **February 2015** :
  - If a client who was active in January 2015 submits a new order, we determine its order discount rate with $\tilde{\Pi}$, and we update the order profit value. Otherwise we keep discount rate original value (given by Π).
  - We compute the sharpe ratio of all active clients during January & February period (**with the new updated profits of February 2015**), and we determine their profit-risk sets.
- **March 2015** :
  - If a client who was active in January or February 2015 submits a new order, we determine its order discount rate with $\tilde{\Pi}$, and we update the order profit value. Otherwise we keep discount rate original value (given by Π).
  - We compute the sharpe ratio of all active clients during January & February & March period (**with the new updated profits of February & March 2015**), and we determine their profit-risk sets.

2.5.4 Discount Rate Pricing Function : Backtesting Results

We obtain the following results :

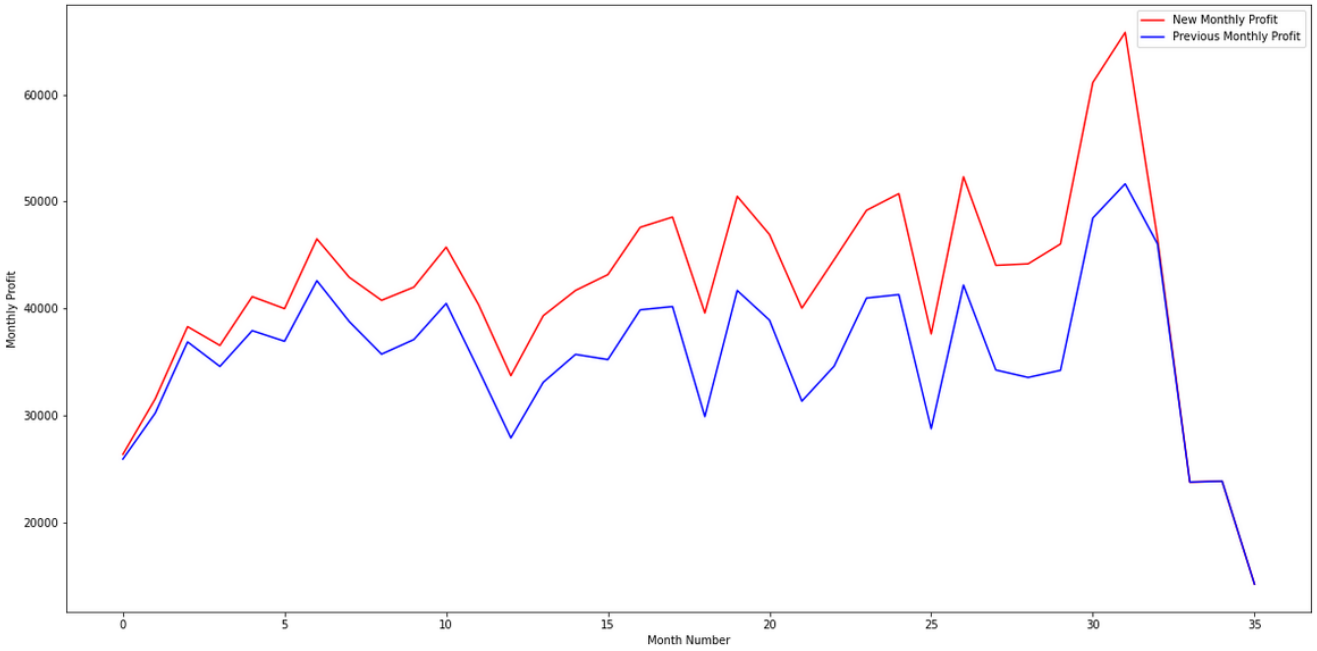

Figure 16: Comparison between previous and new monthly profits

Month	Previous Monthly Profit	New Monthly Profit	Profit Increase (%)
2015-02	25904	26351	1.7
2015-03	30207	31542	4.4
2015-04	36863	38290	3.9
2015-05	34566	36533	5.7
2015-06	37915	41095	8.4
2015-07	36920	39972	8.3
2015-08	42588	46495	9.2
2015-09	38768	42907	10.7
2015-10	35716	40750	14.1
2015-11	37083	41984	13.2
2015-12	40456	45724	13.0
2016-01	34255	40347	17.8
2016-02	27880	33705	20.9
2016-03	33081	39307	18.8
2016-04	35694	41675	16.8
2016-05	35208	43155	22.6
2016-06	39860	47579	19.4
2016-07	40167	48540	20.8
2016-08	29879	39569	32.4
2016-09	41673	50480	21.1
2016-10	38896	46905	20.6
2016-11	31318	40017	27.8
2016-12	34598	44558	28.8
2017-01	40961	49170	20.0
2017-02	41286	50728	22.9
2017-03	28745	37611	30.8
2017-04	42180	52306	24.0
2017-05	34233	44020	28.6
2017-06	33538	44163	31.7
2017-07	34204	46021	34.5
2017-08	48447	61121	26.2
2017-09	51646	65806	27.4
2017-10	46034	46561	1.1
2017-11	23752	23752	0.0
2017-12	23832	23832	0.0
2018-01	14210	14210	0.0
Total	**1,282,581**	**1,506,799**	**17.5**

Table 1: Backtesting Results

As expected, the higher the customer's purchase record, the more information we have on his profit-risk profile. As a result, the discount rate is more adapted to the profile of each customer, and allows us to reduce losses and therefore increase the company's profit. We notice that in the last months of 2018 (where the company looses a lot of money), the discount rates are very low for all the clients to stop losses.

3 Fraud Detection

3.1 Exploratory Data Analysis and Review of Suspected Fraud Orders

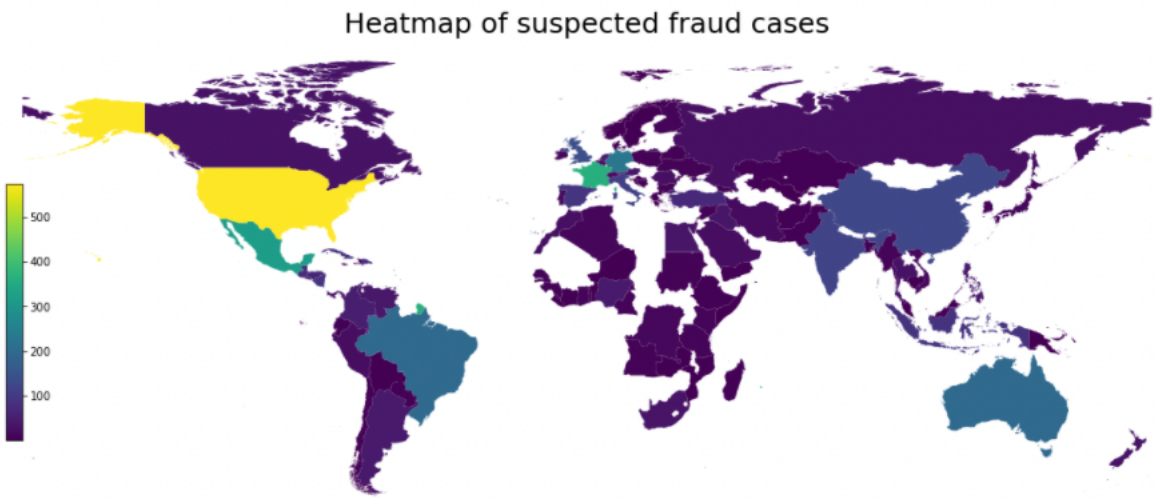

Figure 17

Analysing fraudulent orders were also deemed necessary to optimise supply chain operation. Key sales figures comparing overall and suspected fraud orders are tabulated as below:

Type of Order	Total Number of Orders	Sum of Sales	Sum of Profits
Total	180,519	36,784,735	3,966,903
Suspected Fraud	4062	825,935	85,137
% Of Total	2.25%	2.24%	2.15%

Table 2: Key Sales Figures of Overall and Fraudulent Orders

It is seen that 2.25% of all orders were flagged as fraudulent with suspected fraud orders accounting for a similar proportion of Sales and Profits: suspected fraud did not have a disproportionate effect on the key sales figures. However, these orders will not only represent a loss in revenue and profit but also disruptions to the distribution of

the products.

After some exploration, the following insights were noteworthy:

- All suspected fraud orders used the"TRANSFER" payment method.
- Customers with fraudulent orders had much less orders than those with no such orders, having 2.84 and 8.63 orders per Customer ID respectively.

An ideal filter for fraud would focus on orders that use the "TRANSFER" payment method on accounts with a low number of orders.

Focusing on the individual customers with identical full names. The top 10 names with the most occurrences of suspected fraud have been plotted in Figure 14.

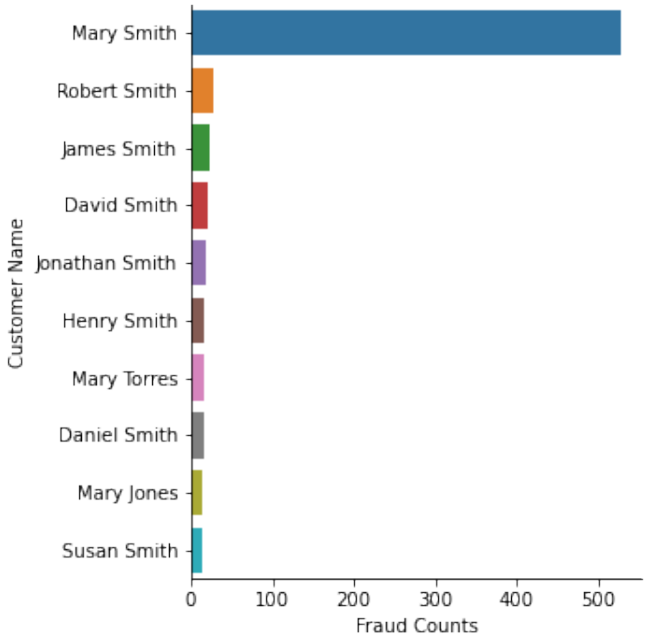

Figure 18

528 of all counts of suspected fraud were detected from customers named "Mary Smith". This proportion which is much higher than anyone else in the list. Additionally 8 out of the top 10 customers have the surname Smith. Given that "Smith" is the most common surname in the US (according to NameCensus), this fact can partially account for the possibility that many distinct customers with the same name could have been flagged for fraud. This is supported by the fact that there were 662 instances of customers named "Mary Smith." If anonymised names are correctly mapped to any repeated names in the creation of this dataset, this suggests that the system for creating user

accounts could be easily abused in the event of fraud. However, this scenario depends on how the dataset was created – it cannot be determined if names were randomly anonymised and Mary Smith happened to come out top by chance. Unfortunately, there was no further information provided in the schema.

All given data sets were joined to create a general data set. This data was then cleaned and processed to create a heat map of correlations:

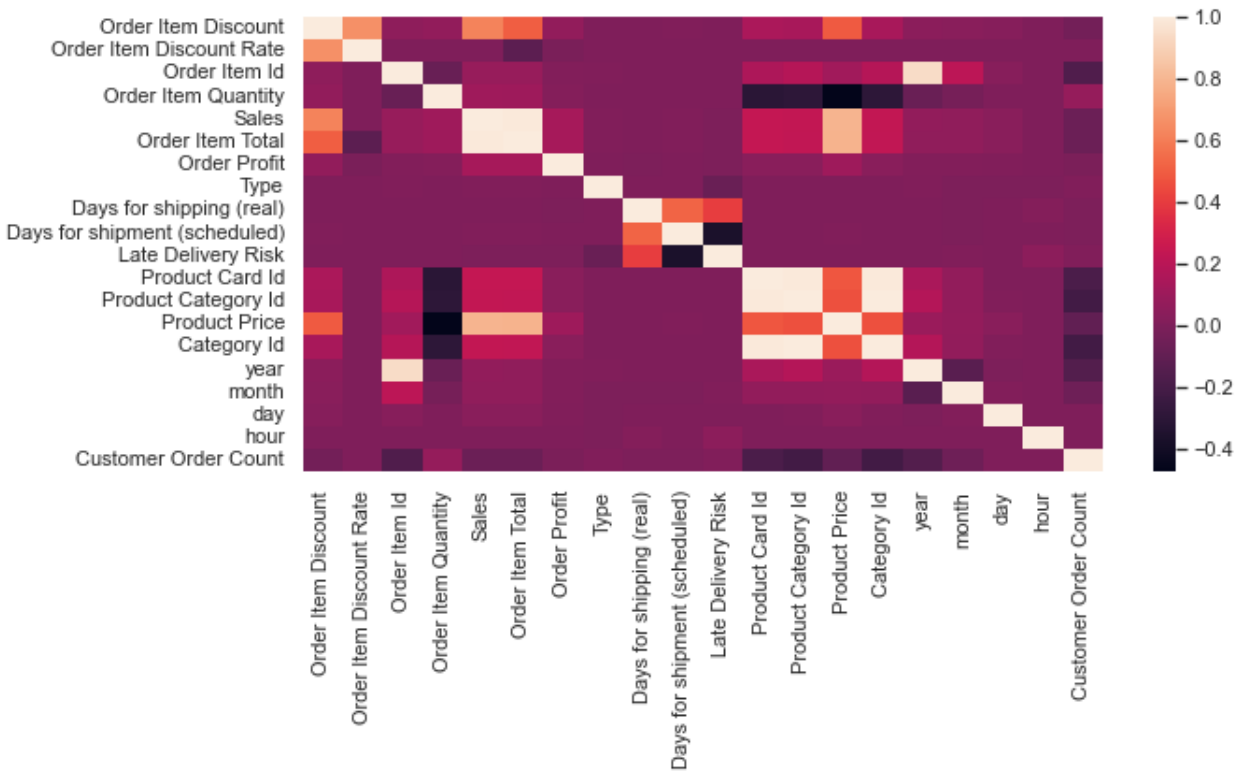

Figure 19

From this heat map, no new insights are discovered, only basic correlations are observed, such as sales and product price being correlated. Subsequently, with the use of random forests, we can determine the importance of factors that put orders at risk of suspicion for fraud. The first most obvious feature is the product status (availability of stock), followed by the country of the customer and the quantity ordered. Despite it being possible to train a model to detect future fraud cases, we have decided not to pursue it and leave it for an idea as future development – mostly due to suspected fraud cases only making up 2.15% of lost profits. Having to train and update the model as well as dealing with false positives would be an unnecessary compromise for a small improvement in sales, and fraudsters can easily change their tactics.

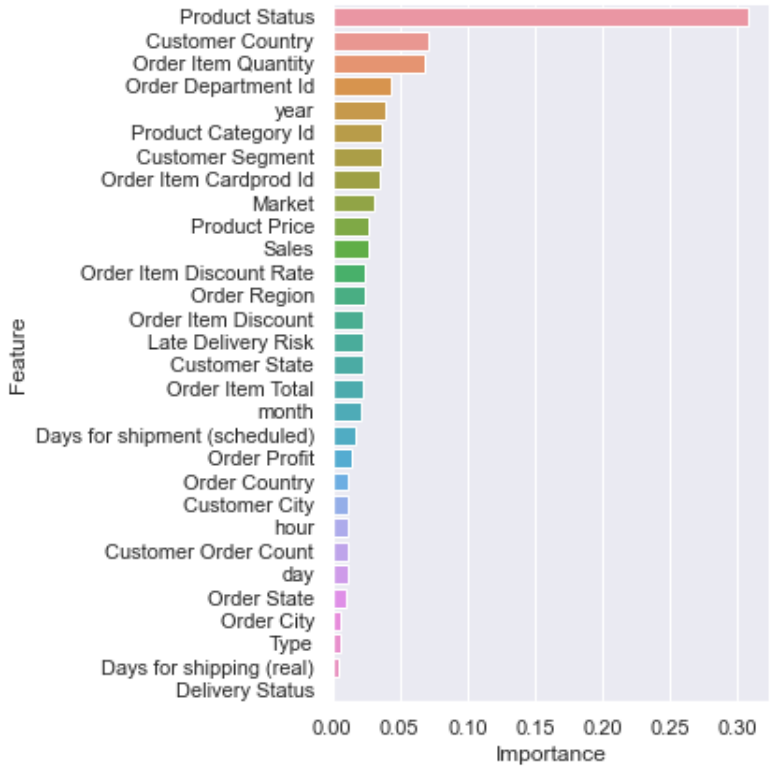

Figure 20: Random Forest Feature Importances