Explaining and Predicting Customer Churn

Master in Business Analytics and Big Data - Capstone Project

Group K members:

Quirijn Bolhuis, Julian Krauth,
Diego Mata, Julius Prestin,
Diego Scheibling, Thomas Truyts,
Thomas Van den Borg

7th July 2022

Executive Summary

Business and problem description

The construction solutions and equipment provider ClientCo face a multitude of challenges. Weak sales online, a performance decline in Western Europe and Southeast Asia as well as liquidity problems are the main issues the company is confronted with in the client faced business.

Customer analysis

Based on a part of the company's sales spanning from 09/2017 to 09/2019 a detailed customer, product and transaction analysis was performed, in order to gain insights on client groups, retention, lifetime value and churn probability with the goal to make meaningful business recommendations to ClientCo.

Also, a RFM analysis was performed where the customers were divided into three different groups based on three KPIs: recency, frequency and monetary. This was done by using K-means clustering. Subsequently, the customers were segmented in five different customer groups based on the RFM scores, deriving from the features.

Usually about 40% of the clients become retained customers. However, looking only at clients that buy through the channel online, the number increases to about 60% showing the potential of this channel despite its relatively low share in revenues. Using the retention rates to explore the average lifetime for each client group (by main used channel) the life-time value for each customer was analysed showing a small group of clients making up a large amount of the company's revenue which can be considered as the most important client group.

Using the information generated from the other parts, the probability of churning was determined. This churn was determined by different thresholds based on the product group a client bought. Later a classification model was developed, taking multiple metrics into account, to predict if a client will churn or not with an accuracy of 95%. Having this information, specific clients can be targeted and possibly retained as active customers.

In a final step, two specific client groups (1. Conservative Core and 2. Digital Future) were segmented into subgroups, using unsupervised machine learning to make meaningful marketing recommendations and ultimately increase the company's revenue.

Recommendations

The ClientCo is facing several key challenges including insufficient availability of liquidity, the creation of a strong online presence, the prevention of regional business decline and the adaption of changing regulations that require more frequent assortment reviews.

Consequently, a two-fold strategy is recommended that combines a restructuring and a strategy refreshment part. The restructuring phase has the intention to strengthen the core body of the business. As part of this, a product cluster of 167,102 products was identified for removal that contributes USD 124 m in sales and 38 m units sold. In addition, a branch cluster of 58 branches was identified for closure that contributes USD 38 m in sales and 26 k units sold, therefore, USD 16 m marked as churned sales.

The strategy refreshment has the intention to identify the most important customer segments that need to be prioritized. For considering the characteristics and preferences of the actual customer base, a conservative core cluster combines the most relevant clients, making up to 5% of totals clients (8,669) and up to 50% of the total sales (USD 5.3 bn). For contemplating the emerging trend of digitisation, a digital cluster combines customers that are already

predominantly using online channels. This cluster combines 4% (3,786) of total customers and 8% (USD 717 m) of totals sales.

In general, all clusters in the conservative core show a low online exposure, which needs to be increased gradually during the next years. This can be done during customer contact by introducing the existing online platform and giving the right incentives.

The Digital Future segment shows a high adoption of the online channel. Therefore, it is important to extract their knowledge and understand the need of improvement. Ultimately, online purchase data should be used to enhance the future-oriented product portfolio, branch network and supply chain.

To implement the above-mentioned recommendations, an implementation roadmap was designed based on three main phases: Project kick-off, restructuring phase, strategy refreshment phase.

Following the link to the notebooks and datasets:

https://drive.google.com/drive/folders/1sbh8qnM4bZ94kfF6tllyU3qySetuVSiy?usp=sharing

Table of content

List of Figures

List of Tables

Glossary

%	Percent
$	Dollar
AT	Average timespan
Bn	Billion
C	Client
CAC	Customer acquisition cost
CLV	Customer lifetime value
CSV	Comma-seperated values
D	Discount rate
datetime64	data type date time 64-bit
DSO	Days sales outstanding
e.g.	For example
F	Frequency
float64	data type float 64-bit
int64	data type integer 64-bit
K	Thousand
LP	Last purchase
M	Monetary
M	Million
p1	Product category 1
p2	Product category 2
p3	Product category 3
PT	PowerTransformer
R	Recency
USD	US Dollar
YtY	Year-to-Year

1 Background

1.1 Introduction

ClientCo is a global construction and renovation products company with global presence across international markets such as Europe, North America, Latin America, and Asia-Pacific. Their main business is to acquire construction products from multiple suppliers and resell them, through their own network of different distribution points, to other firms that do construction work. Their business is considerably large consisting of 15.5 billion dollars revenue in 2018 and over 31,000 employees worldwide.

The product range of ClientCo spans from small equipment pieces such as ear plugs, gloves, masks, hammers, etc. to heavy equipment such as climate control systems, power generators and vapor barriers. Also, ClientCo suppliers consists of multiple manufacturing groups, both local and international that have a large customer pool from small businesses to global construction conglomerates.

Currently, ClientCo has a strong presence in multiple markets and is the dominating player when it comes to construction equipment and construction resources. ClientCo's network consists of more than one thousand branches in more than fifteen countries in Latin America, Asia, North America, and Europe. The global network is ClientCo's strength since it consists of more than 450k active customers ranging from contractors, end-user, industrial and tertiary businesses, allowing them to have a very diverse end-market. Due to its multinational presence, ClientCo has a very robust line of products, both, under their own brand and third-party brands.

Sales channels are very important to ClientCo due to the nature of its business; therefore, they have sales executives, multiple on-premises branches, e-commerce website, and a wide range of web applications for their products. One of the strongest markets, deriving from their multinational portfolio, is Central America as they are one of the leading corporations in the region.

1.2 Problem Statement

1.2.1 Internal

ClientCo has a strong performance in most of its markets. However, it is currently underperforming in their Western Europe and Southeast Asia markets. One of the challenges ClientCo is currently facing in these regions is their financial liquidity shortage, hindering them to fund larger projects. Furthermore, it makes them have considerable opportunity costs for ventures that require robust funding. ClientCo´s increasing liabilities affects the company´s performance.

Another hurdle ClientCo is confronted with, is its failure to exploit their e-commerce business since they have a very limited online presence. Consequently, their main competitors are gaining market shares as the industry grow in that space. In this context, as ClientCo goes through the digitalization of its operations to comply with the demands of its consumers, it is exposed to potential threat of cyberattacks. In the digitalisation process, ClientCo also needs to consider data protection requirements, which demands a sophisticated IT-architecture and -infrastructure, hence, large investments.

1.2.2 External

1.2.2.1 Business perspective

One of the many challenges of the construction industry is the increasing regulation due to the sizeable environmental impact this industry has. Not only the industry is under high scrutiny from environmental groups but also the regulations are constantly changing, which ceases industry players to adapt and plan production. ClientCo´s large product portfolio is affected by these regulations, since it consists of items which do not meet the sustainable requirements. This increments constant inventory audits to ensure that ClientCo always complies with country specific regulations. Another challenge are currency exchange rates which affects the operations of ClientCo, given the multinational nature of their business. International currency instability increases the accruals and in the daily business the constant exchange rate fees.

1.2.2.2 Geopolitical perspective

Besides the problems deriving from the business side, there are also negative externalities which could have an impact on ClientCo's business. The lately placed tariffs in the "trade war" between the United States and China raise uncertainties in the construction industry whether contractors are substituting products or redesign processes, making many materials redundant. This handicaps ClientCo in their product portfolio planning and can farther increase procurement costs. Moreover, the United Kingdom leaving the European Union on October 31st, 2019, poses a risk for ClientCo of restricted sales and material procurement to and from the United Kingdom. Furthermore, the refuse of President Nicolás Maduro to step down and let inaugurate Juan Guiadó as the new president of Venezuela, could lead to tensions or even civil war. This poses the risk of restrained oil exports, which could increase oil prices and hence, the operating costs of ClientCo.

1.3 Hypotheses

As cited in the problem statement, ClientCo is facing several challenges related to finance, legislation, and digitalisation. These challenges unite in the problem of customers churning from ClientCo which further accelerates the downward spiral. ClientCo contacted Group K to approach this issue by explaining first the existing problems, in a second instance to predict which customers are likely to churn, and finally develop actionable recommendations.

Based on the preliminary conducted research, consisting of a background analysis and the problem statement, the question arises "Why are ClientCo's customers churning and how can this be prevented?". Group K states the following hypothesis why customer churn:

- Product portfolio is too large and does not meet the customers' needs.
- Channels are wrongly handled as the point of sales for the client and do not correspond with customer demands.
- Branches are operating inefficient by offering wrong product catalogues and vending via unsuitable sale channels

For this project and the upcoming analysis, all the features contained in the dataset are going to be relevant. The explorations are going to find relationships between features, identifying patterns about why customers churn, and label them accordingly. The group of "churning clients" are going to be studied deeper. Specifically, products and branches are going to be identified, which must be withdrawn from the product portfolio, respectively branches which must be closed.

2 Analysis

2.1 Exploratory Data Analysis

2.1.1 Description of the dataset

2.1.1.1 General Description

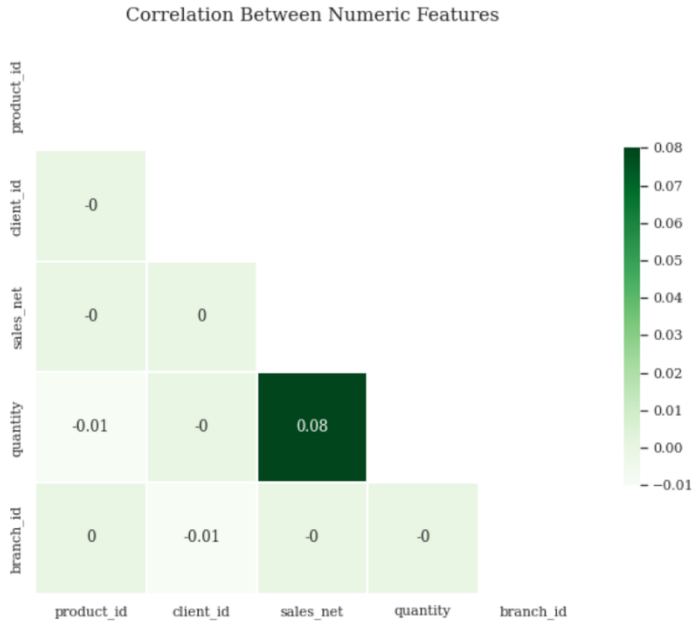
For this capstone project a subset of the client's data was delivered to explain and predict customer churn. The dataset included datapoints from September 22nd, 2017, to September 22nd, 2019, and was delivered in transactional form as CSV-file. The total size of the CSV-file was 4,33 GB, containing 64m rows. The dataset contained in its raw form the following fields and data types illustrated in table 1.

Table 1: Feature overview

Field	Data type	Description
date_order	datetime64	Date on which the order was placed
date_invoice	datetime64	Date of the invoice
product_id	int64	Unique product ID
client_id	int64	Unique client ID
sales_net	float64	Net revenue in currency
quantity	int64	Number of items sold in that transaction
order_channel	object	Channel in which the order was placed
branch_id	int64	Unique store ID

The exploration for missing values resulted in one missing data point in the column "date_invoice". In a further step the correlation between the features was examined which can be seen in figure 1.

Figure 1: Feature Correlation



The correlation matrix displays that none of the variables included in the dataset have a significant correlation. Moreover, analysing the dataset further, it includes 582 distinct branches, selling over 607k products to 170k different clients.

2.1.1.2 Data handling

The data was initially explored and analysed using Power BI and Python. The size of the dataset (4.33 GB and 64m rows) was challenging to deal with locally. In order to ease the loading and saving of the dataset, the dataset is transformed from CSV into a Feather file, which is also able to store the datatypes. When performing some transformations (e.g., joining two tables or applying functions to each row) the processing-time was very high and with the possibility of the kernel crashing due to RAM limitations. To enable to run these transformations different libraries and data structures have been tried (PySpark on local cluster, Sparse DataFrames and Dask) but these were insufficient. Ultimately, Google Colab Pro + offers the requirements

needed for this project. This enabled access to >50GB of RAM and parallel working several notebooks. Separate Python notebooks per analysis topic are created, to structure the work.

Despite the capabilities of Google Colab Pro+, some of the analyses are still not feasible to be ran, most notably the product association. The number of unique products (>600k) and baskets (9m) are not possible to represent in standard data formats and building the rules require a lot of computational power due to the high number of combinations.

2.1.2 Clients

2.1.2.1 General insights

As mentioned in the preceding sections, the dataset contains information of about 170.5k different clients. The customer with the highest turnover has a total revenue share of 0.28%. The largest ordering client accounts for 0.82% of total quantity. Due to this large customer base, there is no cluster risk. However, the top ten percent of clients, in terms of net sales, realise 75.0% of total revenues and 77.1% of total quantity. This implies that there are many clients with smaller turnovers and quantity order.

2.1.2.2 Channel overview by customer

In this section the channel overview will be further discussed.

Table 2: Overview channel usage by customer

Channel	Customer usage in %	Net sales	Share net sales in %	Quantity	Share net quantity
Phone	79.5%	6'417'631'674.00	64.5%	3'466'099'328.00	62.2%
Store	86.0%	2'613'007'608.00	26.3%	1'425'659'232.00	26.2%
Online	12.5%	897'202'346.00	9.0%	650'595'181.00	11.5%
Sales rep.	0.7%	7'826'328.00	0.0%	3'016'358.00	0.0%
Other	5.1%	9'186'545.00	0.0%	174'494.00	0.0%

As shown in the table 2, customers of ClientCo have clear preference to buy products through the phone and store channel, for which the phone channel accounts roughly two thirds of total revenue streams. The same applies for quantities. Exploring the channels more deeply, the sales to quantity ratio seems to be the highest in phone, indicating that higher priced products are sold more through phone compared to store. In contrast, low valued products are mainly sold online.

The top 10% revenue generating customers purchase products mainly via the phone, store, and online channel. Sales representatives seem not to target the top revenue generating clients, since only eight clients utilise this channel, generating $67.5k net sales. Despite roughly half of the top clients buying items through the online channel, it generates 81% more revenue compared to the stores.

2.1.3 Products

2.1.3.1 General insights

For the product analysis, only product_ID has been considered with an average price larger than $0.00. In general, the company has a product portfolio containing 607.4k different products of which 605.5k have average prices higher than $0.00. In the subset provided, the prices range between $0.00, rounded to two digits, and $435'438.00. The top 10% sales revenue products, generate net sales of $8.9 bn accounting for 90% of total sales. Hence, with 90% of the product catalogue, the company generates roughly 10% of total revenue. Examining the top 10% of products by quantity sold, $5.6 bn items have been dispatched, representing more than 99% of total quantity sold. The average price of these items is $20.37, whereas for the remaining 90%

of the product portfolio the average price is $210.19, indicating that the majority of the turnover is generated with products being classified in product category 1 according to the churn propensity model (see chapter 2.4).

2.1.3.2 Channel overview by product

Phone With 572k different items vended, this channel sells almost all products of the product catalogue with an average price of $193.6. 64.5% of total revenue is generated and 61.3% of total quantity sold through this channel. Examining the top 10% net sales products, they account for 87.1% of the channels revenue and 86% quantities sold.

Store This channel sells 25.7% of the product catalogue with average price of $105.9. The store channel contributes 26.3% to the total revenue generated and 27.1% to the total quantity sold. The top 10% net sales generating products are responsible for 90.6% of total sales and 89.8% of total quantities sold. This implies that 90% of the products sold through this channel are dispatched in low quantities while using storage space.

Online Similar to the store channel, about 24% different products are sold via this channel. Furthermore, the online channel contributes 9% to the total revenue stream and 11.5% of quantities sold. With an average product price of $83.9 and a higher quantity to revenue ration than other channels, it indicates that most of products sold through this

channel are deriving from product categories 1and 2. The distribution of total sales and quantity share among the top 10% selling products is similar to the phone and store channel, 91.3% revenue and 87.4% quantities sold being accountable to them.

Sales representative Through the sales staff, only 1.56% of the product portfolio are sold, accounting for 0.08% of net sales, 0.05% of total quantity vended, and average price of $97.69. This channel seems to have the most distributed sales across the products_id's sold through this channel. Specifically, 62.5% revenue and 60% quantities dispatched are accountable to the top 10% net sales generating products, making it the most balanced channel.

Other Only 83 items are sold through this channel with the lowest average price across all channels ($76.37). 174.5k quantity is dispatched through "other", generating total net sales of $9.18m. This channel can be marked as the least important.

2.1.3.3 Developments over time

For comparing the developments in average prices, revenue, and quantity the timeframe between October 2017 and August 2019 was selected, since as off and until this month's data for the whole month is available.

In general, the months May and September were in this two-year period notably weak sales months over all channels, whereas in March and August the strongest revenue streams could be observed. Analysing the net sales more deeply, Q4 in 2018 marked a strong closing YtY. Contrary to this, the ClientCo was struggling in Q1 2019 with negative rates in January and February but was able to increase its net sales YtY over the rest of the year with highlights in April and July, showing double digit YtY net sales growth. Overall, the firm was able to increase its average prices by 2,99% (net 0.84% after inflation), resulting in an YtY upward trend of 4,94% in net sales, which aligns with the competitor Hilti Group, showing a growth rate of 4,3% in net sales.

Analysing the corresponding for quantity, the organisation experienced likewise in quantities a strong YtY closing in Q4 2018. In contrast, Q1 to Q3 2019 was challenging for ClientCo with an YtY average growth rate of 1,74%, due to a strong April with double digit growth. In general, the company was able to increase its total YtY quantity dispatched by 2.63% over the two-year period.

2.1.4 Branches

2.1.4.1 General insights

The dataset contains information of 582 different branches. The branch with the highest turnover has a total revenue share of 1.30% and 1.95% quantity stake. The top 10% revenue generating branches, account for 36.3% of total revenue streams and 41.65% quantity sold. Analysing the 50% least performing branches in terms of revenue, they are accountable for 17% of total revenue and 14.24% of total quantity sold. 50% of the revenue is generated with the 103 most performing branches. These figures indicate that most branches are underperforming.

However, insightful is that these 103 most performing branches sell lower valued products with average prices of $35.36 compared to the remaining branches with average prices of $88.71. The branch with most customers has 3'768 clients and is accountable for $40.6m revenue and 22.3m items sold. Further, 58% of the top 10% most valuable clients are within the top 10% branches with most clients. Moreover, the branches with top 10% most clients contribute 29.6% to the total revenue stream and account for 31.75% of total quantity dispatched. Another insight of this analysis is that 74 branches have less than 10 clients, generating $48.5m revenue, taking a 0.49% stake of the total revenue stream.

2.2 RFM

RFM-analysis is commonly used to segment customers into homogenous groups. The clustering of these customers requires the analysis of three quantitative variables: recency-, frequency-, and monetary value. Frequency and monetary value provide insights into the value of customers. Lower frequency scores indicate higher demand and loyalty, and low monetary scores mean purchases of higher amounts. Recency provides information about customer satisfaction and engagement. A higher score indicates a higher possibility of being churned.

2.2.1 RFM features

The values to segment the clients in the dataset used are calculated as follows (per client):

$$Recency = date_{last} - date_{last\_purchase}$$

$$Frequency = \sum purchases$$

$$Monetary = \sum_{i}^{n} sales\_net_i$$

With $i = purchase$.

Figure 2 shows the correlation matrix between the three new values. Monetary and frequency are highly correlated because when a client buys more often, it is probable that this client also spends more. (~80%).

Figure 2: Correlation matrix between the RFM features

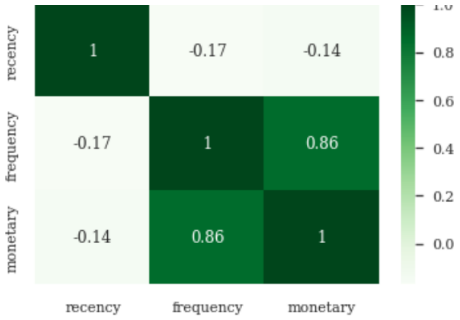
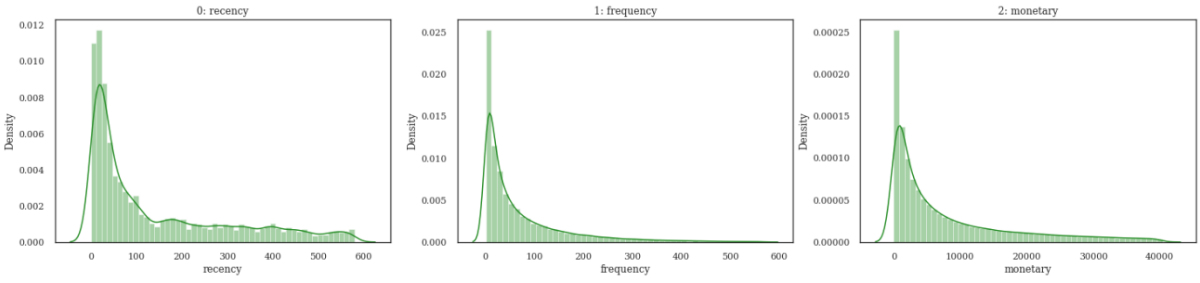


Figure 3 shows the distribution of the three variables. The distributions for all features are highly skewed. This means that in a subsequent step these variables need to be normalized, to utilize them for clustering the clients. However, this transformation is not necessary for the segmentation.

Figure 3: The distribution of the three features



2.2.2 Clustering

2.2.2.1 Clustering method one

For segmenting customers regarding their RFM values, the variables are distributed with a scoring system of 1 to 4 based on the quantiles (25%, 50%, 75%, 100%). Consequently, two new columns are made based on the following formulas:

$$RFM_{score} = R_{score} + F_{score} + M_{score}$$

with the score ranging from 3 to 12 where 3 is the best- and 12 the worst customer; and

$$RFM_{group} = string(R_{score}) + string(F_{score}) + string(M_{score})$$

with '111' being the best and '555' being the worst score/string.

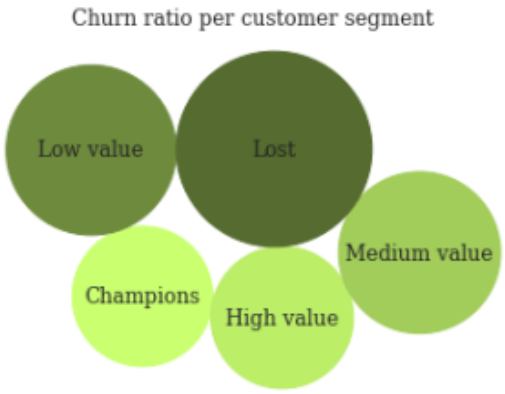
The client based dataframe is initially clustered into five customer groups based on their *RFM_score*.

Table 3: The five customer groups based on RFM_score

RFM_score	**Segment**
>10	Lost customers
>8	Low value customers
>5	Medium value customers
>3	High value customers
>0	Champions

Figure 4 represents the segments with their churn rate. The size of the bubble is proportional to the number of clients in that particular segment, whilst the darker the color the higher the churn rate.

Figure 4: Segments with their churn rate and number of clients within segments



Here it is clear that the customer segment *lost* has a higher churn rate and consequently a darker color. Contrary, the segment *champions* are the best customers and therefore, their churn rate has the lowest value.

2.2.2.2 Clustering method two

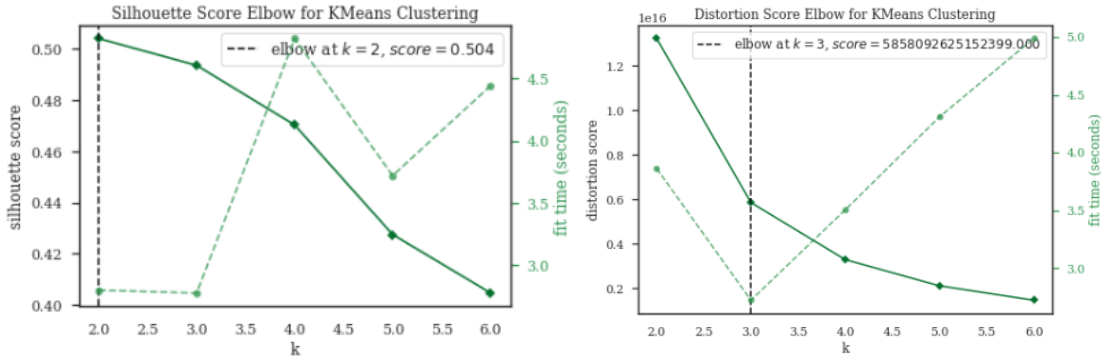
Next, a second clustering is performed based on the *RFM features (recency, frequency, monetary)* that can be used for customer segmentation based on consumer behavior. The chosen algorithm for this cluster method is K-means. For fully exploit this method, the outliers are for all included features are removed by using the following formulas:

$$outliers > Q_3 + 1{,}5 \cdot IQR$$
$$outliers < Q_1 - 1{,}5 \cdot IQR$$

Thereafter, the features are scaled using a standard scaler from the Scikit learn library for them to be proper as an input to the K-means. This algorithm divides the data into clusters where each client belongs to the mean of the nearest cluster. The number of clusters is decided based

on the elbow technique, which runs the K-means clustering for multiple k values (ranging from 2 to 6). The right side of figure 5 below shows the k-means distortion score for the given k-values. As the number of clusters increases, the sums of square distances are becoming less. The optimal k value is decided where the elbow curve is bending. In this case the sum of square distances decreases dramatically at k = 3. The left side of the figure below indicates the silhouette score projected to the different k values. The silhouette score is a metric, ranging between -1 and 1, used to calculate the quality of the clusters. It considers the intra-cluster distance and the nearest-cluster distance; in this case 0,494. Based on both graphs the optimal number of clusters is reasoned to be at k = 3.

Figure 5: Silhouette and distortion score according to the different k-values



The distribution of the clients across the clusters is as follows: cluster 0 contains roughly 56%, cluster 1 19%, and cluster 2 25% of clients. The following figure 6 displays the silhouette score for each sample on a per-cluster basis, visualizing the density of and the separation between clusters. Clusters with higher values have wider silhouettes.

Figure 6: Silhouette plot of K-means clustering

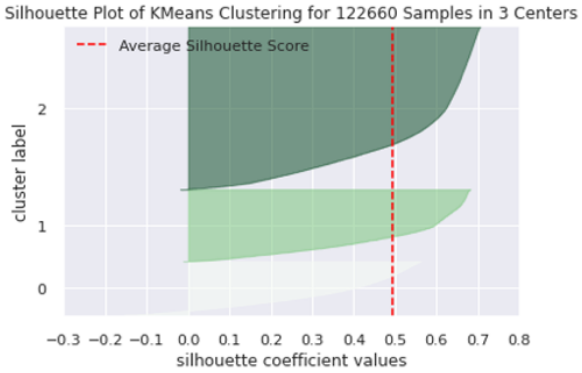
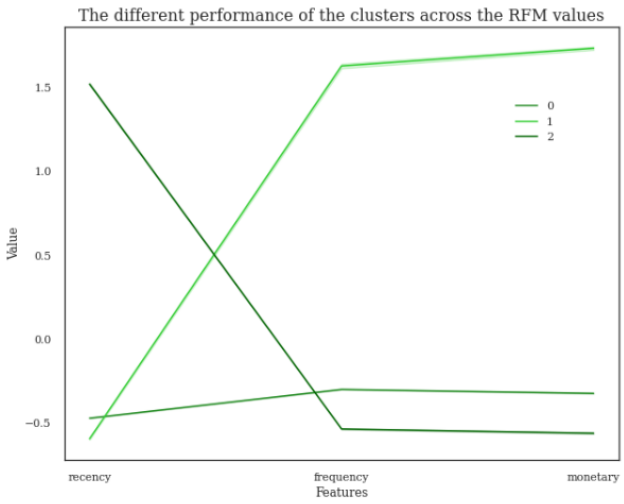


Figure 7: Different performance of the clusters across the RFM values



Further insights into the behaviour of the features inside the different clusters provides figure 7. Cluster 1 and 2 behave in the opposite way, while clients in cluster 0 score low on all features and can therefore, be considered as the best clients.

2.3 Retention Curve

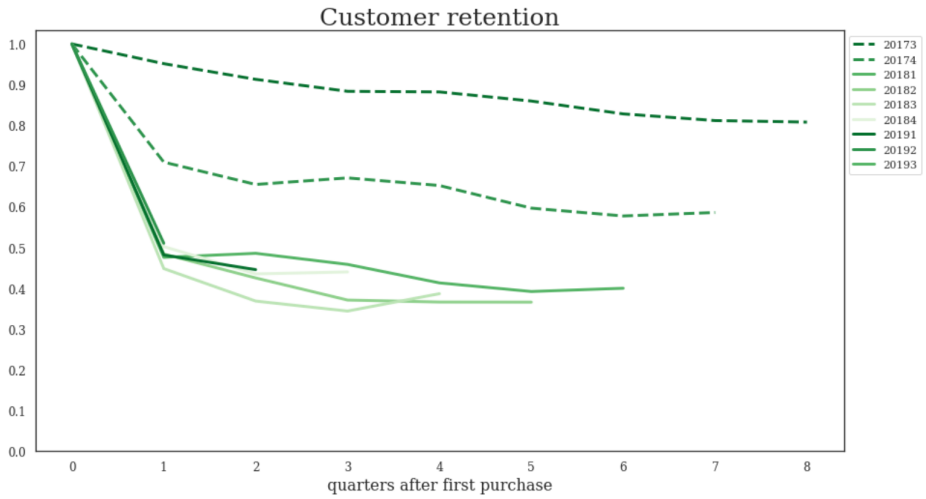
Analysing the buying behaviour of certain customer groups and channels over time, can help to understand the client base better and gain detailed insights as well as generate targeted marketing recommendations.

2.3.1 Technical steps

For creating retention curves, the clients are grouped into cohorts. The individual client was assigned a cohort based on the quarter and year the customer first purchased one of the products, resulting total of 9 cohorts. The cohorts range from 3rd quarter 2017 to 3rd quarter 2019. For the further exploration, information about the channel of first contact is included to enable a more detailed analysis in the subsequent steps.

The retention dataframe was built by grouping the cohorts and looking at the count of customers from this cohort that continued buying in the quarters following their first purchase. The individual purchase amount and frequency of these transactions were not considered important for this analysis since we consider a client buying in the quarter as a continuation of his relationship with us.

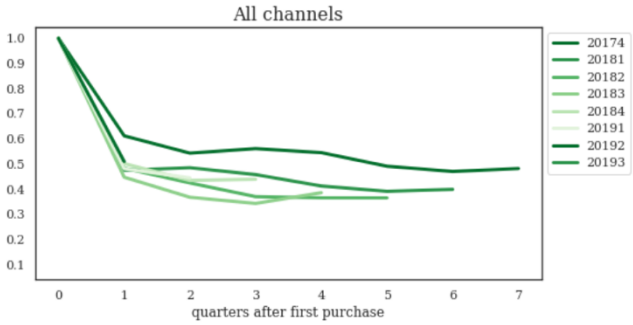
Figure 8: ClientCo customer retention curves (grouped by cohort)



As indicated in the first part of this report, the data is based on a fixed timeframe of an active business. Classifying loyal clients as new clients with the first purchase in the dataset/timeframe may lead to misinterpretations. As indicated in Figure 8 – the first two cohorts are clients where their first purchase was in the months of September to December in 2017. With regards to the understanding of an active customer, which will be analysed in more detailed in the following chapters, a client is expected to return in a short period of time. Without prior knowledge, one cannot be certain that clients who bought in September or October (~ first 5 weeks in our dataset) of 2017 to be classified as first-time customers. In the more comprehensive retention curves shown later, this group will be excluded from the charts and the analysis, since the uncertainty prevents any meaningful insights.

2.3.2 Interpretation of retention

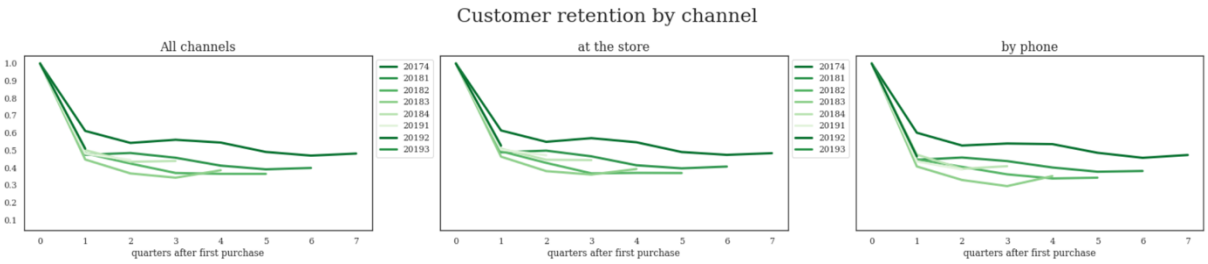
Figure 9: Customer retention over all channels



Looking at the overall retention (without clients first recorded in Sep or Oct of 2017) a clear trend can be identified. Retention declines mostly in the first two to three quarters after the initial sale, with a long-term retention of about 40-50%, without any large difference between the cohorts. Cohort 20174 shows the highest relative retained number compared to the other cohorts. This can still be connected to some past loyal customers first being recorded in the November or December of 2017.

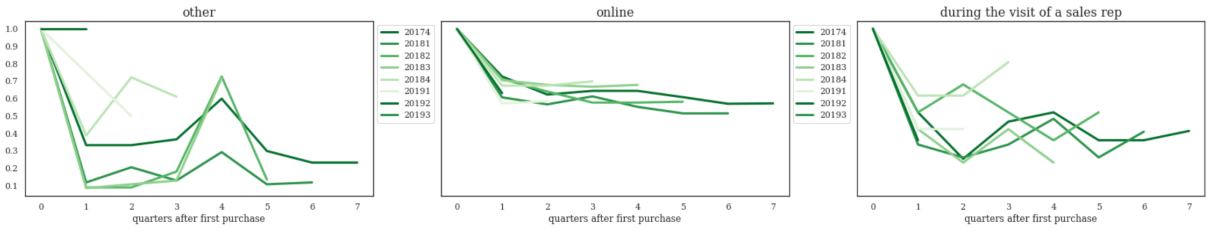
Looking at retention curves based on the channel of the first interaction with the client can provide some interesting insights.

Figure 10: Retention curves based on channel (total, store, and phone) of first contact



While store and phone display a similar development as the overall client base, the three other channels show differing developments.

Figure 1111: Retention curves based on channel (other, online and sales rep) of first contact



Clients that initially bought online mark the highest overall retention rate with a long-term rate of about 60%, which could be understood as a higher loyalty level, compared to the other channels. Sales rep shows firm fluctuations, which could be based on irregular visits or, as indicated in previous chapters, on the nature of products being sold by a sales rep, being mostly high-volume sales that do not necessarily result in the sale of other smaller items. Although the remarkable spike in quarter four in the other channel, the low amount of revenue in this channel (0.1% of total sales) does not incentivise further analysis.

2.4 CLV

2.4.1 Technical steps

The customer lifetime value (CLV) is a metric for the total revenue or income a business can expect from a customer over the course of the relationship. The average revenue or profit is measured over fixed timespan. The CLV is compared to the acquisition costs to be able to estimate the profitability of the customers and the business's potential for long-term growth. The following formula is used to calculate the CLV.

$$CLV = \frac{profit\ margin * average\ lifetime}{(1+d)^{average\ lifetime}} - CAC$$

where:

*profit margin = average yearly sales * profit margin (%)*

average lifetime = (1 / (1 – retention rate))

CAC = customer acquisition cost

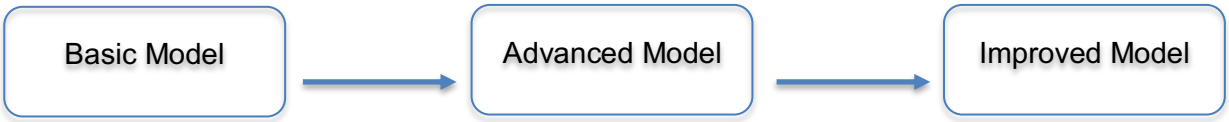
d = discount factor

The revenues documented in the dataset were adjusted to represent average annual sales per client. This conversion allowed to find the customer lifetime in years as well. The average yearly sales are reduced to the average yearly profit to get an idea on how much the company is averagely making for every customer. The basic profit margins are based on the three different price categories. Assumed is that the company makes higher margins on more expensive products. The average lifetime of a customer is calculated using the customer retention rates for all five channels derived from the customer retention analysis (see table below). The discount factor is assumed to be 5%.

Table 4: Average retention rate after first year per channel

Channel	Customer retention rate
Online	0,641
Phone	0,410
Store	0,443
Sales representative	0,490
Other	0,284

2.4.2 CLV Models



To calculate the CLV we developed three different models, moving from simple to more advanced. The basic model assumes five different profit margins for the five different channels and assumes that the acquisition cost is equal to 2 euros. The second model advances the profit margin calculation by additionally considering client size and sales channel visible in the table on the next page. The client size is based on the total sales per client which is then grouped by quantile into small, medium, and big clients. For the client sizes is presumed that the company makes higher margins on smaller clients, due to for instance discounts granted to big/established clients. The profit margin is further differentiated by sales channel assuming that profit margins through online and phone sales are more favorable, and margins through representatives and fairs (other) less favorable, all with respect to store sales. The final model builds upon the second model but further complicates the acquisition cost approximation by assigning different CAC values for the different sales channels (see the table on acquisition cost on the next page) based on the channel through which the company had the first contact with the client. The acquisition cost is calculated as a percentage of the sales margin per client, following the logic that bigger clients are more costly to acquire. Assumed is that customer acquisition through representatives and fairs is rather costly, while in-store and online recruitment are cheaper alternatives, all with respect to acquiring clients by phone.

Table 5: Assumptions for CLV calculation

Profit margin bases	**Profit margins (%, multipliers)**
Product category	*% (model 1/2/3)*
1	10%
2	16%
3	22%
Client size	*Multiplier (model 2/3)*
Small	1,2
Medium	1,0
Big	0,8
Sales channels	*Multiplier (model 2/3)*
Online	1,3
Phone	1,2
Store	1,0
Representative	0,9
Other	0,8

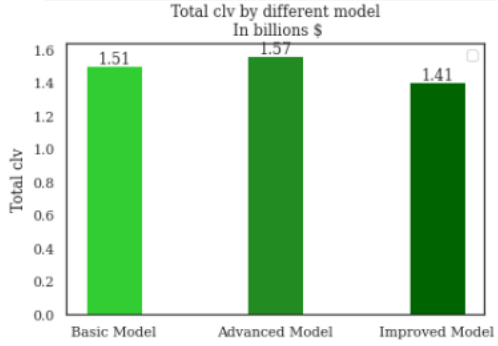
*6e.g.: big customer buys product from category 1 online --> 10% * 0,8 * 1,3 = 10,4% profit margin*

Acquisition cost bases	**Acquisition cost as % of annual sales**
Sales channels	*Multiplier (model 3)*
Online	0,020
Phone	0,030
Store	0,025
Representative	0,035
Other	0,040

7e.g.: customer with average annual spendings of $1.000 that first bought at the company through the online channel is acquired for $20

2.4.3 Interpretation of CLV

Figure 12: Total CLV of the different models



The graph on the left visualizes the different total CLV for the different models. The improved model appears to be the most conservate approach in approximating the CLV. The advanced model estimates the aggregated CLV of all customers at 1.57 billion.

The graphs below provide insights into the clients with the highest and lowest CLVs. The client with the highest CLV is estimated to be worth more than 8 million over the whole period of the customer relationship. On the other hand, the right graph indicates negative values for the least engaging customers. These negative values are relatively small and can be explained by product returns.

Figure 13: Top and bottom 5 clients regarding CLV

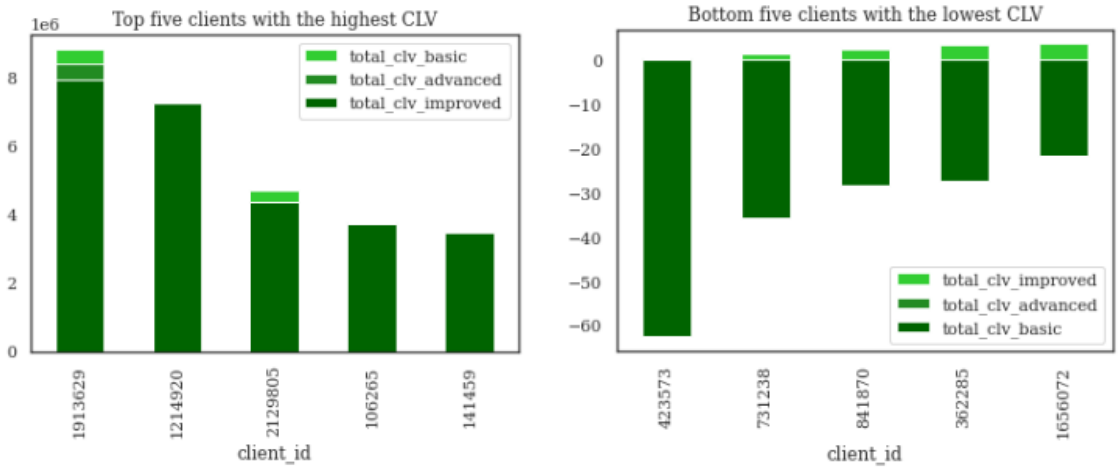
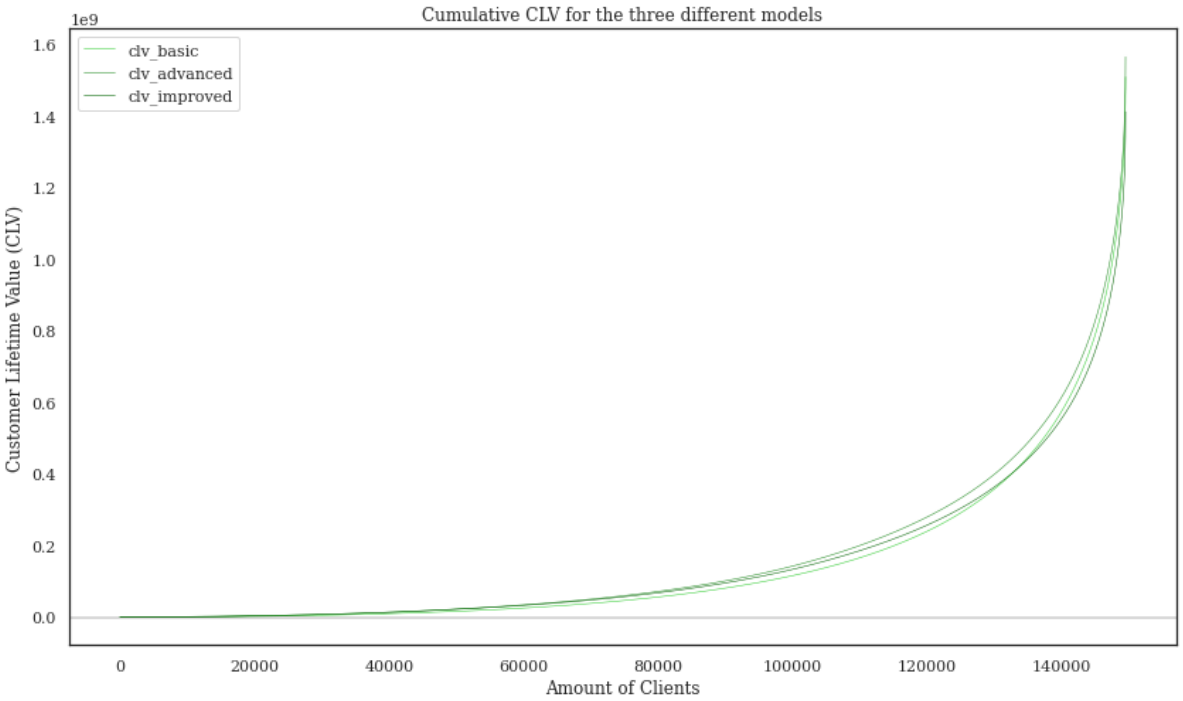


Figure 14: Cumulative Customer-lifetime value



The profitability chart above indicates that a minority of the customers is responsible for most of the cumulated customer lifetime value for the company. Due to the scale the negative CLVs are not visible in this graph.

2.5 Churn propensity Model

A propensity model helps to identify the likelihood of a customer's churn by analysing their past behaviours. This can help us detect clients with a risk of churning and apply the appropriate strategies to handle them.

2.5.1 Labelling

Firstly, a propensity model has to be built with a labelled dataset. Consequently, the notebook *Propensity_labeling.ipynb* is used to label each client as churn (1) or non-churn (0).

2.5.1.1 Technical steps

To gather all client and transaction-based information into one dataframe, the *transaction.feather* and the *rfm.feather* are joined. To label a client as churn or non-churn, the difference in time between his purchases is calculated to start creating the average time between purchases. The chosen strategy to label a client as churn is to set up time period thresholds based on product categories. As indicated in the Exploratory Data Analysis section, the product categories are divided as follows:

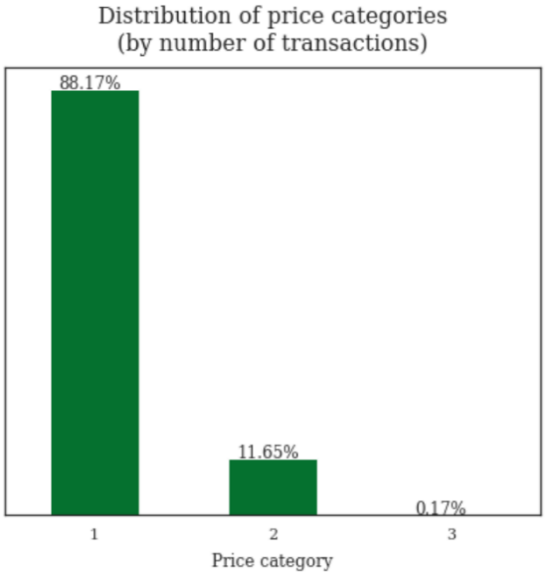
Table 6 8: Different product categories

Product categories	Price ranges
Product category 1	0 < x <= 50
Product category 2	50 < x <= 1000
Product category 3	1000 < x <= 500000

The thresholds are decided based on domain knowledge and analyses of product catalogues of ClientCo's main competitors (Hilti and Makita).

On figure 12, there are more products that belong the price category one in ClientCo's catalogue than price category three.

Figure 1515: Distribution of price categories



This is done on purpose because with pandas.qcut it would have returned equally distributed counts of products per product category. Using price ranges based on domain knowledge for categorising products enables to have for instance nails and hammers together in one product category and power generators in another. This increases the meaningfulness of the product categories since the association between products in the categories are significantly higher than with pandas.qcut. Also, to consider in this context that negative prices are not taking into account while creating the different price categories. The assumption is made that negative prices explain product returns, consequently these transactions are dropped out of the dataset since they only represent 3,77% of the transactions and returns (repeated customers) are not considered when labelling churn, analysis done in EDA *(Kushwaha, Kumar, & Meleet, 2021)*. After, to calculate the time period thresholds the values are calculated per product category over all clients.

$$AT_{ci} = \frac{LP_{ci} - FP_{ci}}{purchases_{ci} - 1}$$

With AT = average timespan, LP = last purchase data, FP = first purchase data and *purchases* = number of purchases. $c \in$ {client_id}, $i \in$ {product_category}. This results in an average timespan each client buys each product category. The thresholds are calculated as follows:

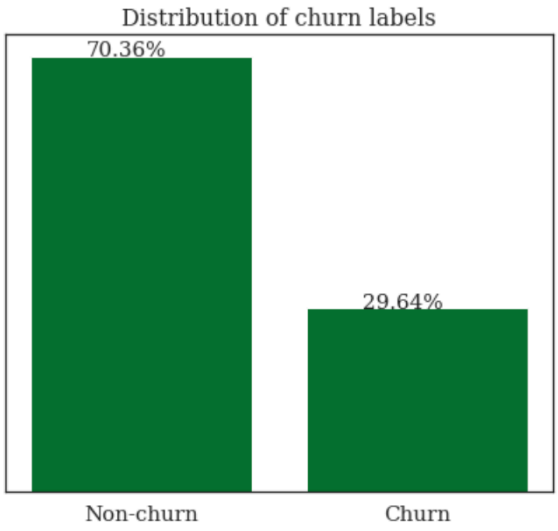
$$Threshold - product_1 = P(97.5,\ AT_{c1}) = 99\ days$$

$$Threshold - product_2 = P(97.5,\ AT_{c2}) = 270\ days$$

$$Threshold - product_3 = P(97.5,\ AT_{c3}) = 450\ days$$

The 97,5th percentile is a commonly used point in a business research distribution in statistics. Churn will be labelled only considering the last purchase of each client. That is why, for each client the number of days is calculated between the last date of the data set *(22/09/2019)* and his last purchase. With this method, the machine learning churn prediction model will not include future purchase dates and no information is leaked into the algorithm. Figure 13 shows the resulting churning distribution.

Figure 1616: Distribution of churn labels



2.5.2 Churn prediction model

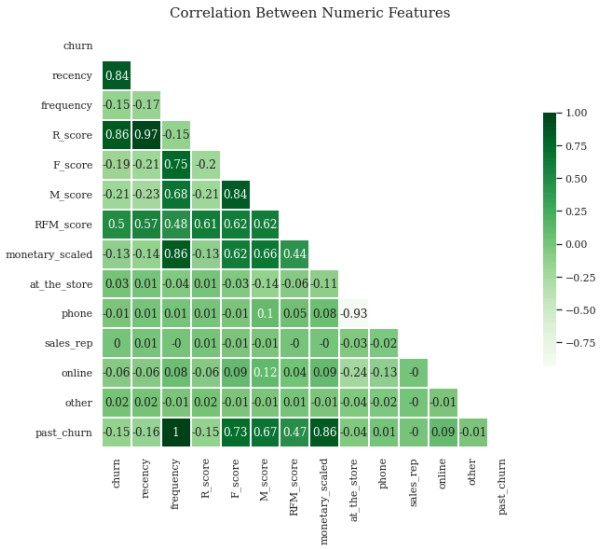
After having labelled the dataset, a machine learning model can be made to get deeper insights in the features and predict whether new clients are likely to churn.

2.5.2.1 Pre-processing

The dataframes that are used for the propensity model are a merge of the *client_churn.feather* and the *rfm.feather*. In this dataset there are no missing values and consists of 166.521 different clients with 16 features. The following matrix represents the correlation between the numerical features. It is clear that the added variables in the RFM model (R_score, F_score, M_score, and RFM_score) are highly correlated with each other. Further, order channel *'phone'* and *'at_the_store'* have a high correlation of –0,93.

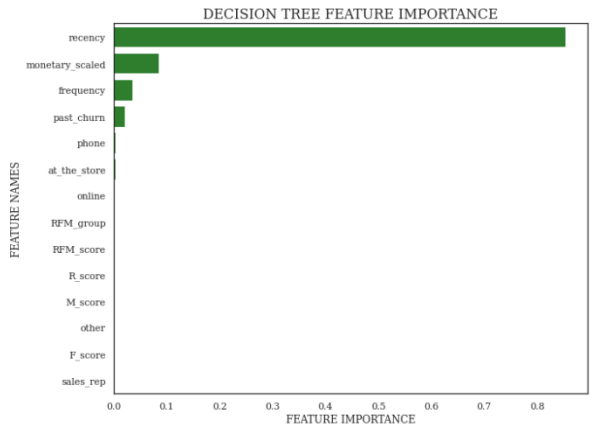
In this correlation matrix, another variable is added in the model *'past_churn'* which is calculated in the *propensity_labeling.ipynb.* This variable explains how many times a client churned before the last purchase. As shown in the correlation matrix, the variable has a 100% correlation with the frequency value. However, it has been decided to not delete this variable because as shown in the feature importance figure (figure 15), the variable explains the target variable.

Figure 1717: Correlation matrix of the numerical features



To look at the importance of each feature, a baseline decision tree model is executed with the following results:

Figure 1818: Feature importance based on decision tree



	feature	importance
0	recency	0.852826
1	monetary_scaled	0.084353
2	frequency	0.035054
3	past_churn	0.020775
4	phone	0.003091
5	at_the_store	0.002586
6	online	0.000659
7	RFM_group	0.000269
8	RFM_score	0.000150
9	R_score	0.000139
10	M_score	0.000066
11	other	0.000024
12	F_score	0.000008
13	sales_rep	0.000000

The dataframe shows that the variable *'recency'* has an importance of 89,91% which makes sense because the labelling is based on the time period after the last purchase. Following up is the variable *'frequency'* where the thresholds of the churn are based on. So, it is clear that the model extracted the labelling strategy that is chosen before.

2.5.2.2 *Modelling and evaluation*

Then following pipelines with their respective scores are executed (On top of these models, PCA was executed but this did not give valuable results):

Table 7: Model results churn prediction

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Decision tree	94%	93%	93%	93%
Decision tree scaled + PT	94%	93%	93%	93%
Random Forest	95%	94%	94%	94%
Extra Trees	95%	94%	94%	94%
Gradient Boosting	95,43%	94,20%	94,96%	94,57%
Gradient Boosting scaled	70%	35%	50%	41%
LGBM	95,52%	94,53%	94,78%	94,65%
LGBM with SMOTE	95,53%	93,65%	95,42%	94,46%

2.5.2.3 *Model insight*

In the following figure the means of the used features for the final LGBM model are listed. To show the lack of the prediction model (95% accuracy), the columns are split into means for the wrongly predicted clients and the correctly predicted clients. This shows that the model is less accurate for clients with a low value for *past_churn* and frequency.

Figure 1919: lack of prediction of final model

	mean_wrong	mean_correct
recency	184.153644	144.849443
frequency	60.898884	406.762318
R_score	2.030204	1.595085
F_score	1.011819	1.157479
M_score	1.020355	1.183374
RFM_score	4.062377	3.935939
monetary_scaled	0.000377	0.002195
at_the_store	0.632961	0.629381
phone	0.344714	0.336511
sales_rep	0.001970	0.000566
online	0.019698	0.032534
other	0.000657	0.001007
past_churn	11.872620	121.793531

3 Recommendations

3.1 Results

As outlined in the hypotheses (Chapter 1), the ClientCo is facing several key challenges that need to be solved in order to stay competitive in the future construction solutions and equipment market. Based on these trends, products are becoming more environmental-friendly, supply chains are becoming more digital, and products are more likely to be rented out as the rental market is picking up steam.

Taking into account the external trends as well as the performed data analysis, the main challenges for ClientCo include the sufficient availability of liquidity, the creation of a strong online presence, the prevention of regional business decline (e.g. Europe, South East Asia) and the adaption of changing regulations (environmental concern) that require more frequent assortment reviews.

Consequently, a two-fold strategy is recommended that is coming with a restructuring phase (underperforming portfolio/branch cleaning) and a strategy refreshment phase (reallocation of current customer focus).

The restructuring phase has the intention to strengthen the core body of the business and make it more resilient and agile for the upcoming changes inside and outside the organization.

During this phase, the product portfolio and branches will be analysed, and not-performing units downsized in order to encounter the liquidity problem and the upcoming assortment reviews. Products with low contribution to the company's performance have been identified by using the following filters:

- Total sales below USD 500 with quantity below 10 units
- Churned sales rate of more than 33%
- Days Sales Outstanding (DSO) of more than 90 days

As an outcome, a product cluster of 167,102 products was identified for removal that contributes USD 124 m in sales and 38 m units sold, thereof USD 38 m marked as churned sales, which are sales that belong to a churned customer.

To streamline the logistics expenditures and optimise the supply chain, the branch network was analysed based on its ability to serve main customers and future channels. Branches with low contribution to the company's performance have been identified by using the following filters:

- Total sales below USD 5 m, quantity below 100,000 units, and online channel coverage below 50%
- Total sales below USD 10 m, churned sales rate of more than 20% and online channel coverage of less than 50%

As an outcome, a branch cluster of 58 branches was identified for closure that contributes USD 38 m in sales and 26 k units sold, thereof USD 16 m marked as churned sales.

After restructuring the existing portfolio, the strategy phase has the intention to identify the most important customer segments that need to be prioritized when reallocating the strategic focus of ClientCo.

For this phase, a two-step process is followed: In a first step, two main customer segments, the conservative core and the digital future, were identified based on their importance for the

current business and for the desired future orientation. This was achieved, by filtering the total customers based on carefully defined quantitative thresholds.

In a second step, for each of those segments several sub-clusters were identified as relevant customer niches that need to be addressed with individual concepts.

Segment I: Conservative Core

To consider the characteristics and preferences of the actual customer base, a conservative core cluster combines the most relevant clients making up to 5% of totals clients (8,669) and up to 50% of the total sales (USD 5.3 bn). This cluster was defined by applying the following filters:

- Cumulated sales of more than USD 50 k
- RFM segment "Champion"
- CLV of USD 15 k or higher
- Online channel usage of less than 50% to create no overlap with the second cluster
- Days Sales Outstanding (DSO) of less than 30 days
- No churned customers

Within the Conservative Core cluster, three main niches were identified: The prime, the whale and the dolphin.

Prime

The prime niche includes 2,694 customers with a combined sales volume of USD 1.2 bn. This cluster has the highest DSO with around 6 days, which is 50-60% higher than the two other clusters. With an average CLV of USD 41 k per customer the prime segment is balanced between the two others. In terms of channel distribution, the prime channel is phone with 74%,

followed by store (21%) and online (5%). The product mix includes all three product categories with majority share in product category 2 (p1: 31%, p2: 53%, p3: 16%). With 16% in the highest product category 3, the average price customers are paying in this cluster is 200% compared to the others.

Whales

The whales niche includes 4,321 customers with cumulated sales of USD 3.5 bn, which represents 35% of the total sales of ClientCo in the considered period. The average sales per customer are around USD 818 k, distributed on the three main channels phone: 64%, store 30% and online 6%. The product mix includes all three products (p1: 55%, p2: 40% and p3: 5%), giving this segment an average CLV of USD 59 k.

Dolphins

The dolphins niche shows similar characteristics compared to larger client. The cluster includes 1,654 customers with cumulated sales of USD 613 m. However, averages sales and quantity sold per customer are 40% down. This is also a consequence of the product mix that only includes product category 1 and 2 (p1: 58%, p2: 42% and p3: 0%) and leads to a lower average CLV of USD 30 k.

Recommendations: In general, all clusters in the conservative core show a low online exposure, which needs to be increased gradually during the next years. As the phone remains the main order channel until now, resources (from existing call centers/phone branches) should be allocated to build a task force that assists customers in their online journey. This can be done during customer contact via phone or in the store by introducing the existing online platform

and giving the right incentives. Customers can be incentivized by special online offers, detailed product description (product category 3) that are exclusively available online or enhanced digital marketing. Furthermore, an equipment rental model should be introduced with focus on product categories 2 and 3. This model can be explicitly offered to prime customers, as they show the highest interest in high-value products.

To reach this group and offer a more extended offering for product category 3, resources (from existing call centers /phone branches) should be used to build a task force that assist customers in using the online channel for future orders. Detailed product information should be also made available more and more exclusively online to persuade customers to use the existing platform:

Segment II: Digital Future

To consider the emerging trend of digitization, a digital cluster combines customers that are already predominantly using online channels. This cluster combines 4% (3,786) of total customers and 8% (USD 717 m) of totals sales. The cluster was defined by applying the following filters:

- Use of online channel is larger than 50%
- No churned customer

Within the Digital Future cluster, three main niches were identified: Small product digital natives, remote whales, opportunists

Small product digital natives

The niche only includes 261 customers with cumulated sales of USD 500 k, which is only 0.1% of the Digital Future cluster and results in a low CLV of USD 2.6 k. However, with an average

online channel usage of above 90% it gives valuable insights in the online business requirements and shows an enormous upside potential for the CLV. Compared to the following niches, the small product digital natives buy only products from product category 1, resulting in a 40% lower average price per customer, while also having the lowest DSO with 2.65 days.

Remote whales

Like the whales in the conservative core, the remote whales display a customer niche with large customers. The cluster includes 1,018 customers and cumulated sales of USD 579 m, with average sales of USD 569 k per customer. The online channel reaches 70%, while still 24% of orders are made via phone. Regarding the product mix, all product categories are bought, with an average share of 10% for product category 3, which lifts the CLV up to USD 77k (p1: 46%, p2: 44% and P3: 10%). Moreover, this cluster has the highest DSO of 3.44 days, which exceeds the small product digital natives by 30%.

Opportunists

The opportunists niche shows similar characteristics compared to the remote whales. The cluster includes 2,507 customers with cumulated sales of USD 137 m. However, average sales and quantity sold per customer are only 90% down. This is also a consequence of the product mix that only includes product category 1 and 2 (p1: 54%, p2: 46% and p3: 0%) and leads to a lower average CLV of USD 11 k. Moreover, online channel sales are slightly higher with 75%, while phone is down to 15% (10% store) compared to the remote whales.

Recommendations: In general, the Digital Future segment shows a high adoption of the online channel with at least 70% coverage per cluster. These customers know the benefits of the

existing digital platform and can work as ambassador. Therefore, it is important to extract their knowledge and understand what services and offerings are running well and which ones need further improvement. For this purpose, an additional task force can be used to set up a comprehensive feedback program. In a next step, purchase data of existing online customers should be analysed to understand which products are bought online and in which combinations. This information should be paired together with the existing branch network to optimize the existing supply chain and infrastructure. This way inefficient branches for in-store purchases or purchases via phone could be extended or merged to dark store facilities that are handling solely online orders.

3.2 Implementation

To implement the above-mentioned recommendations, an implementation roadmap was designed based on three main phases: Project kick-off, restructuring phase, strategy refreshment phase.

Project kick-off:

- Setup of project team to align tasks and distribute responsibilities
- Onboard relevant stakeholders to align understanding and rationale of the project
- Internal communication

Phase 1: restructuring:

- Product and branch network analysis to understand which products and branches should be ultimately eliminated
- Product phase-out by product relevance, stock availability and contractual obligations
- Store closure by size, network relevance and contractual

Phase 2: strategy refreshment:

- Product portfolio optimization to understand recent development in the market and competitor environment
- Setup/enhancement of digital platform to offer a competitive solution
- Roll-out of marketing campaign to generate awareness