Universidade do Minho
Escola de Engenharia

José Carlos Vaz da Silva

Segmentação de clientes B2B e previsão estratégica de oportunidades futuras com Inteligência Artificial

Julho de 2022

Segmentação de clientes B2B e previsão estratégica de oportunidades futuras com IA
José Carlos Vaz da Silva
UMinho 2022



Universidade do Minho
Escola de Engenharia

José Carlos Vaz da Silva
PG42970

Segmentação de clientes B2B e previsão estratégica de oportunidades futuras com Inteligência Artificial

Dissertação de Mestrado
Mestrado em Engenharia de Sistemas

Trabalho efetuado sob a orientação dos
Professor Doutor Cesar Analide de Freitas e Silva da Costa Rodrigues
Professor Doutor Bruno Filipe Martins Fernandes

Julho de 2022

DIREITOS DE AUTOR

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.



Atribuição CC BY

https://creativecommons.org/licenses/by/4.0/

AGRADECIMENTOS

Gostaria de começar por agradecer aos professores Bruno Fernandes e César Analide, pela oportunidade e por todo o apoio, com conhecimento, sugestões construtivas e de melhoria.

Também gostaria de agradecer aos professores e colegas do mestrado que me apoiaram em vários momentos mais difíceis e que foram determinantes na concretização deste objetivo pessoal e profissional.

Para os meus colegas profissionais, em particular para a administração da inCentea, uma referência especial, pelo incentivo, pelo apoio e pela disponibilidade em todo o processo.

Finalmente, quero agradecer à minha família especialmente à Maria João, a minha esposa e aos meus filhos Bruno, João e Maria, por todo o apoio e compreensão.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

Segmentação de clientes B2B e previsão estratégica de oportunidades futuras com Inteligência Artificial

Muitas empresas não conseguem maximizar o uso da informação que as fortaleceriam no mercado, como o uso do CRM para obter perceções sobre a procura e as necessidades dos clientes. A mineração de dados é o processo de extração de conhecimento a partir dos dados. O conhecimento da segmentação do comportamento dos clientes é útil para as empresas na definição do mercado alvo e no desenvolvimento de uma estratégia de marketing. Uma vez que o modelo RFM não leva em consideração a fidelização dos clientes, o modelo LRFM foi aplicado para fazer as alterações e para comparar com o novo modelo LRFMVP que tem em atenção a Variedade dos produtos e serviços e a Periodicidade das transações. No primeiro passo, os dados são extraídos para o modelo LRFM e agrupados em 4 *clusters* usando o algoritmo *K-Means*. No segundo passo, o potencial do cliente é considerado e os dados firmográficos são extraídos para o modelo LRFMVP, sendo de seguida agrupados usando o algoritmo *K-Means*. A seleção do número de agrupamentos é realizada aplicando o método do cotovelo e resulta em 3 *clusters*, sendo um deles composto por 219 clientes muito valiosos, com todas as dimensões do modelo LRFMVP acima da média. Adicionalmente, são procuradas as relações de interesse no grande conjunto de dados disponível, usando o algoritmo APRIORI. Observamos que o novo modelo LRFMVP é mais descritivo do que o modelo LRFM, obtém um conjunto de regras de associação superior ao modelo LRFM e melhora a percentagem média de confiança, de 93.67% para 94.38% mantendo a percentagem do *lift* em 148.67% para ambos os modelos. Os clientes *Platinium*, são clientes fidelizados (L=Muito Antigo), com transações muito frequentes (F=Muito Frequente) e valor Monetário mais alto (M=Mais alto), já a Periodicidade das transações é razoável (P=Razoável), face aos restantes. Terminado o processo de segmentação dos clientes e análise descritiva, o modelo foi treinado para ter capacidade de predição na prospeção de novos clientes, obtendo 91,5% de *accuracy*. Por fim, foram desenvolvidos os modelos de recomendação para ações de *cross selling* e para a melhor perceção do "cesto de compras". O modelo proposto define um conjunto de métricas direcionadas para a segmentação do cliente, o que melhora o conhecimento dos clientes. Os resultados permitirão definir estratégias de marketing mais assertivas para a fidelização dos clientes e para gerar oportunidades de negócio futuras.

Palavras-Chave: Agrupamentos, Gestão do relacionamento com o cliente, Marketing baseado na conta, Recomendação, Segmentação de clientes, Valor da vida útil do cliente.

ABSTRACT

B2B customer segmentation and strategic foresight of future opportunities using Artificial Intelligence

Many companies are unable to maximize the use of information that would strengthen their market, such as using CRM to gain insights into customer demand and needs. Data mining is the process of extracting knowledge from data. Knowledge from customer behaviour segmentation is useful for companies in setting the target market and developing a marketing strategy. Since the RFM model does not take the customers' loyalty into consideration, the LRFM model has instead been applied for making amendments and to compare with the new LRFMVP model, which considers the Variety of products and services and Periodicity of transactions. The first step, the data is extracted to LRFM model and then it is clustered using K-Means algorithm. In the second step, the customer's potential is considered, and the firmographics data is extracted to LRFMVP model and then it is clustered using K-Means algorithm. The selection of the number of clusters is performed by applying the elbow method and results in 3 clusters. The cluster_0 consists of 219 very valuable clients, with all dimensions of the LRFMVP model above average. Additionally, the relationships of interest in the large available data set are searched using the APRIORI algorithm. We observed that the new model, LRFMVP is more descriptive than the LRFM model, obtains a set of rules for association with the LRFM model and improves the average confidence percentage, from 93.67% to 94.38% for the same percentage of lift, which is 148.67% in both models. Platinium customers are loyalty (L=Very Old), with very frequent transactions (F=Very Frequent) and higher Monetary Value (M=Higher), while the Periodicity of transactions is reasonable (P=Reasonable), compared to the others. After the process of segmentation of customers and descriptive analysis, the model was trained to predict in the prospection of new customers, obtaining 91.5% of accuracy. Finally, recommendation models were developed for cross selling actions and the best perception of the "market basket". The proposed model defines a set of metrics aimed at customer segmentation, which improves the customers knowledge. The outcomes will allow to define more assertive marketing strategies for customer loyalty and to generate future business opportunities.

Key words: Clustering, Customer Relationship Management, Account Based Marketing, Recommendation, Customer segmentation, Customer Lifetime Value.

CONTEÚDO

RESUMO vii

ABSTRACT viii

LISTA DE ABREVIATURAS xi

LISTA DE FIGURAS xii

LISTA DE TABELAS xiv

LISTA DE EQUAÇÕES xv

1. INTRODUÇÃO 1

1.1. Contexto e Motivação 1

1.2. Objetivos 2

1.3. Metodologia 3

1.4. Estrutura do documento 5

2. ESTADO DA ARTE 6

2.1. Gestão do relacionamento com clientes 6

2.2. Segmentação de clientes 7

2.2.1. Pré-segmentação: O que é um mercado? 8

2.2.2. Variáveis para segmentação B2B 9

2.3. Segmentação B2B: obstáculos e desafios 10

2.4. Análise do valor do cliente 11

2.4.1. RFM 11

2.4.2. LRFM 12

2.4.3. LRFMP 13

2.4.4. RFMV 13

2.5. Customer Lifetime Value 14

2.6. Customer Portfolio Analysis 14

2.7. Data Mining 15

2.8. Machine learning 17

2.8.1. K-means 17

2.8.2. PCA 18

2.8.3. Apriori 19

2.8.4. Elbow Method 19

2.9. Sumário 20

3. COMPREENSÃO E PREPARAÇÃO DOS DADOS 23

3.1. Compreensão e exploração dos dados 23

3.2. Preparação e análise dos dados 29

3.2.1. Pré- processamento dos dados 29

3.2.2. Análise dos dados 35

3.2.3. Análise das caraterísticas e da correlação dos dados 39

4. DESENVOLVIMENTO E AVALIAÇÃO DO MODELO 42

4.1. *K-means*, segmentação em *clusters* 42

4.2. Customer lifetime value 43

4.3. Customer portfolio analysis 43

4.4. Novo modelo LRFMVP 44

4.5. Métodos preditivos e técnicas de tunning 46

4.6. Métodos prescritivos e de recomendação 47

5. DISCUSSÃO DOS RESULTADOS 52

5.1. Modelo LRFM 52

5.2. Novo modelo LRFMVP 58

6. CONCLUSÕES E TRABALHO FUTURO 65

REFERÊNCIAS BIBLIOGRÁFICAS 69

LISTA DE ABREVIATURAS

ABM	Account-based marketing
AI	Artificial Intelligence
ARR	Annual Recurring Revenue
B2B	Business-to-business
CLV	Customer Life time Value
CPA	Customer Portfolio Analysis
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relationship Management
DM	Data Mining
ERP	Enterprise resources planning
FMCG	Fast-Moving Consumer Goods
IA	Inteligência Artificial
KD	Knowledge discovery
LRFM	Length, Recency, Frequency, Monetary
LRFMP	Length, Recency, Frequency, Monetary, Periodicity
LRFMVP	Length, Recency, Frequency, Monetary, Variety, Periodicity
RFM	Recency, Frequency, Monetary
RFMC	Recency, Frequency, Monetary, Category
RFMV	Recency, Frequency, Monetary, Variety

LISTA DE FIGURAS

Figura 1 - Ciclo da metodologia CRISP-DM 5
Figura 2 - Classificação de técnicas de Data Mining em CRM, Adaptado de Ngai, 2009 16
Figura 3 - Estrutura do dataset 23
Figura 4 - Visualização geral do dataset 25
Figura 5 - Exploração geral do dataset 26
Figura 6 - Valor mediano, anual, de transações 26
Figura 7 - Número de clientes e documentos por ano 27
Figura 8 - Tag's das zonas 27
Figura 9 - Tag's das marcas 28
Figura 10 - Tag's dos Códigos de Atividade Económica 28
Figura 11 - Manipulação do CAE 30
Figura 12 - Estrutura base do dataset 30
Figura 13 - Tratamento de dados anómalos com algoritmo H2O 31
Figura 14 - Modelo supervisionado auto-binner 32
Figura 15 - Equação L RFM V P Score 33
Figura 16 - Resultados para as dimensões do modelo LRFM 33
Figura 17 - Resultados ordenados por LRFM score 34
Figura 18 - Escalões para classificação ABCD 34
Figura 19 - Resultados do modelo L RFM P 35
Figura 20 - Resultados ordenados por RFMP_2Y_Score 36
Figura 21 - Resultados ordenados por RFM_score 37
Figura 22 - ARR ao longo de todo o tempo de análise 37
Figura 23 – LinePlot do ARR ao longo de todo o tempo de análise 38
Figura 24 - ARR por ano e por mês 38
Figura 25 - Exemplo de aplicação do PCA em três dimensões 39
Figura 26 - Matriz de correlação 40
Figura 27 - Definição dos parâmetros para filtrar a correlação 41
Figura 28 - Matriz de correlação, após filtros 41
Figura 29 - Clusters obtidos, para o modelo LRFM 43

Figura 30 - Atribuição do portefólio a cada cluster. 44
Figura 31 - Visualização do resultado pelo método de cotovelo 45
Figura 32 - Custer's obtidos com a aplicação do método de cotovelo 45
Figura 33 - Tunning para avaliação dos melhores parâmetros 46
Figura 34 - Resultado da predição em novas contas 46
Figura 35 - Confusion Matrix do modelo 48
Figura 36 - Resultados de predição em clientes potenciais 48
Figura 37 - Confusion Matrix do modelo na previsão do ERP. 49
Figura 38 - Parâmetros de suporte e confiança 50
Figura 39 - Resultado da aplicação do algoritmo Apriori 51
Figura 40 - Regras de associação descobertas para fabricantes 51
Figura 41 . Scatter Plot do modelo LRFM 52
Figura 42 - Amostra de registos coloridos por cluster 54
Figura 43 - Amostra dos clientes mais valiosos 54
Figura 44 - Representação 3D da classificação dos clientes 55
Figura 45 - Representação 4D da classificação dos clientes 56
Figura 46 - Representação 3D dos clientes categorizados e classificados 57
Figura 47 - Representação gráfica da classificação dos clientes 57
Figura 48 - Representação 3D da classificação dos clientes, no modelo LRFMVP 58
Figura 49 - Descrição dos valores do rank obtido no modelo LRFMVP 59
Figura 50 - Regras obtidas pelo método Apriori para o Cluster_0 60
Figura 51 - Regras obtidas pelo método Apriori para o Cluster_1 61
Figura 52 - Regras obtidas pelo método Apriori para clientes Gold e Platinium 61
Figura 53 - Número de clientes por CAE e por Cluster 62
Figura 54 - Número de clientes por Zona e por Cluster 62
Figura 55 - Número de clientes por Categoria e por Cluster 63
Figura 56 - Distribuição dos clientes mais valiosos 63
Figura 57 - Box Plot com ARR por cluster 64
Figura 58 - Regras de associação descobertas no modelo LRFM 67
Figura 59 - Regras de associação descobertas no modelo LRFMVP 67

LISTA DE TABELAS

Tabela 1 - Portefólio de clientes 15
Tabela 2 - Detalhe do Dataset 25
Tabela 3 - Dimensões base para o modelo L RFM V P 31
Tabela 4 - Padrão do modelo LRFM e total de clientes por cluster 53
Tabela 5 - Padrão do modelo L RFM V P e total de clientes por cluster 59
Tabela 6 - Resultados experimentais nos modelos LRFM e LRFMVP 66

LISTA DE EQUAÇÕES

Equação 1 - Formula de cálculo do CLV 14
Equação 2 - Cálculo do erro absoluto médio (MAE), método do cotovelo 19
Equação 3 - Equações RFM Score e RFM Rank 32
Equação 6 - Cálculo da taxa de precisão do classificador (accuracy) 48

CAPÍTULO I

1. INTRODUÇÃO

1.1. Contexto e Motivação

Vivemos num período de forte aumento da concorrência entre as empresas e de disrupção dos modelos de negócio, potenciados em grande medida pela evolução tecnológica. Outra característica importante do mercado B2B (*business-to-business*) são os tipos de clientes, o seu processo de tomada de decisão e que predetermina as atividades para a construção de relações de longo prazo.

As empresas procuram novas estratégias que garantam manter e melhorar os seus resultados e estão a mudar a sua abordagem, mais centradas no cliente, nas suas necessidades e menos nos produtos (Wedel & Kamakura, 2000). A retenção de clientes é mais importante do que a aquisição de novos clientes. Segundo o princípio de Pareto (Riktesh Srivastava, 2016), 20% dos clientes contribuem mais para o resultado da empresa do que os demais. Os resultados das empresas podem ser melhorados por um modelo de segmentação de clientes. A segmentação de clientes pode ser criada através de uma variedade de características exclusivas do cliente para ajudar a personalizar planos de marketing, a identificar tendências, a planear o desenvolvimento de produtos e a fornecer os produtos relevantes. A segmentação de clientes ajuda a personalizar as mensagens para comunicar com os grupos realmente pretendidos.

Será particularmente importante usar a Inteligência Artificial, não só para cenários de marketing, mas também no contexto da transformação do negócio e nos modelos de trabalho para obter todo o potencial. Em Portugal, segundo Smith (2018), presidente e diretor jurídico da Microsoft, no relatório "Artificial Intelligence in Europe: Portugal", afirma que o país já abraçou com sucesso a inovação e as novas tecnologias inerentes à inteligência artificial (IA). No mesmo relatório, na última década, 2.9 milhões de euros foram investidos em startups portuguesas de inteligência artificial.

Apesar de Portugal estar ainda longe de atingir a maturidade plena de utilização de sistemas de IA nas organizações, as empresas reconhecem que este é um fator produtivo importante, sendo que dois terços das organizações inquiridas diz ser um tópico importante ao nível da administração e 86% considera que a inteligência artificial é tão importante ou mais do que as outras prioridades digitais (Smith, 2018).

Em resultado do relatório feito a estas organizações portuguesas, embora a IA não seja a maior prioridade digital para todos, esta tem ganho cada vez mais importância e as empresas relatam que estão a tomar as medidas necessárias para a implementação de práticas de IA nos seus negócios sendo que, no geral, todas se mostraram conscientes de que a IA irá mudar a forma como os ativos são gerados ao longo do seu ciclo de vida, bem como permitir uma utilização mais racional dos recursos e criar novos modelos de negócio.

Também e de acordo com a última análise de mercado da IDC, 40% das empresas inquiridas já implementaram projetos de IA, e mais de 80% planeiam fazê-lo num futuro próximo.

Perante o crescimento da inovação tecnológica e em particular da utilização da IA, torna-se relevante estudar a sua aplicabilidade, as vantagens e/ou desvantagens bem como que implicações terá nos negócios e nos relacionamentos das empresas com os seus clientes.

1.2. Objetivos

O desafio é: como aumentar a eficiência da força de vendas e aumentar a eficácia na concretização de novos negócios B2B?

Os objetivos deste projeto são:

- Como qualificar os dados dos clientes e dos potenciais clientes no CRM, com recurso a ferramentas analíticas?

- Como classificar a carteira de clientes? A classificação de clientes ABCD, será a adequada?

- Como segmentar a carteira de clientes, considerando a relação existente e o potencial de aquisição?

Como resultado esperados:

- Melhorar o conhecimento acerca dos clientes, tornando-se numa organização assertiva com os seus clientes.

- Segmentar os clientes, prever o potencial de negócio futuro na carteira e identificar potenciais clientes.

1.3. Metodologia

Para o desenvolvimento deste projeto em contexto empresarial, é necessário definir uma metodologia que vá de encontro com os seus objetivos. Deste modo, a metodologia usada para suportar este estudo e os seus respetivos objetivos é o CRISP-DM (CRoss-Industry Standard Process for Data Mining). A pesquisa em *Data Mining* abordou uma ampla gama de aplicações, entre as quais as vendas e a gestão do relacionamento com clientes (*CRM* - Customer Relationship Management).

De acordo com Martinez-Plumed et al. (2021), o CRISP-DM é uma metodologia que consiste em 6 fases que descrevem o ciclo de vida da ciência dos dados (Chapman, 2000). Ajudam a planear, organizar e depois implementar o projeto de ciência de dados.

As seis fases deste modelo são as seguintes:

1. Conhecimento do negócio (*Business Understanding*);

Esta fase inicial centra-se na compreensão dos objetivos e requisitos do projeto numa perspetiva de negócio, convertendo este conhecimento numa definição de problemas de mineração de dados e na conceção de um plano preliminar para atingir os objetivos.

2. Compreensão de dados (*Data Understanding*);

A fase de compreensão de dados começa com uma primeira recolha de dados e ajuda a familiarizar-se com os dados, a identificar problemas de qualidade dos dados, a obter conhecimento sobre os dados ou detetar subconjuntos interessantes em informação, até aqui, oculta.

3. Preparação de dados (*Data Preparation*);

A fase de preparação de dados abrange todas as atividades para a construção do conjunto de dados finais (dados que serão introduzidos na ferramenta de modelação) a partir dos dados brutos iniciais. As tarefas de preparação de dados são suscetíveis de serem executadas várias vezes e não em qualquer ordem prescrita. As tarefas incluem a seleção de tabelas, registos e atributos, bem como a transformação e limpeza de dados para ferramentas de modelação.

4. Modelação (*Modelling*);

Nesta fase, são selecionadas e aplicadas várias técnicas de modelação e os seus parâmetros são calibrados para valores ideais. Existem várias técnicas para o mesmo tipo de problema de mineração

de dados. Algumas técnicas têm requisitos específicos sobre a forma de dados, obrigando frequentemente a repetir a fase de preparação de dados.

5. Avaliação (*Evaluation*);

Antes de avançar para a implementação final do modelo, é importante avaliar mais aprofundadamente o modelo e rever os passos executados para construir o modelo para ter a certeza de que alcança adequadamente os objetivos do negócio. Um objectivo-chave é determinar se existe alguma questão importante do negócio que não tenha sido suficientemente ponderada. No final desta fase, deverá ser tomada uma decisão sobre a utilização dos resultados da extração de dados.

6. Implementação (*Deployment*).

A criação do modelo pode não significar o fim do projeto. Mesmo que o objetivo para o modelo seja cumprido e consigamos aumentar o conhecimento acerca dos dados, os conhecimentos obtidos terão de ser organizados e apresentados de forma que o cliente possa usá-lo. É comum aplicar os modelos aos processos de tomada de decisão de uma organização, por exemplo, na personalização em tempo real de páginas web, na atualização do CRM ou na extração contínua de dados para o marketing. Independente de quem implementa, é importante que compreenda antecipadamente que ações devem ser realizadas para realmente utilizar os modelos criados.

É uma metodologia que permite correções ao longo do desenvolvimento do sistema, durante todas as suas fases, o que se reflete numa grande vantagem. Ou seja, recuar para a etapa anterior no caso de serem detetadas possíveis hipóteses de melhoria e permitindo mudanças em tempo real, consoante o cenário vai evoluindo, ajudando também na obtenção de melhores tomadas de decisão baseadas em dados consistentes.

Para além disso, permite a resolução de problemas e melhorias no relacionamento com o cliente, pois vai de encontro às suas necessidades e práticas, sendo assim possível criar um serviço ou produto personalizado.

Na Figura 1 é representado o ciclo da metodologia CRISP-DM:

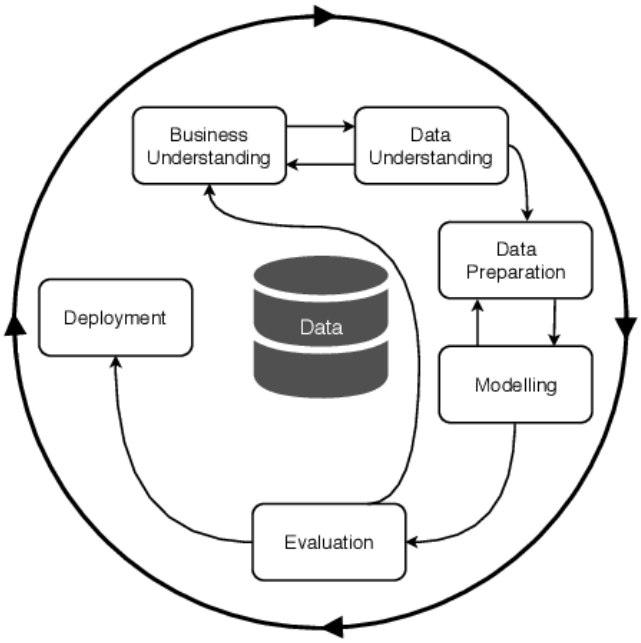


Figura 1 - Ciclo da metodologia CRISP-DM

1.4. Estrutura do documento

A dissertação está organizada em capítulos.

O estado da arte inicia o capítulo II, ao que se seguem a compreensão, a preparação e análise dos dados no capítulo III.

No capítulo IV, o desenvolvimento e experimentação do modelo, com aplicação de técnicas de tunning para avaliar os parâmetros que garantem os melhores resultados.

No capítulo V, serão discutidos os resultados, bem como os indicadores relevantes extraídos dos componentes de visualização, disponibilizados pela plataforma.

No capítulo VI, apresentamos as conclusões para os desafios deste estudo e que foram, como aumentar a eficiência da força de vendas e aumentar a eficácia na concretização de novos negócios empresariais. Inclui também, na conclusão, a comparação entre o modelo de análise do valor dos clientes, LRFM e o modelo proposto LRFMVP, bem como o trabalho futuro.

CAPÍTULO II

2. ESTADO DA ARTE

A relação entre empresas e clientes tornou-se um elemento central das empresas e, portanto, a existência de um mecanismo de gestão dessa relação é essencial. A identificação de cada cliente é extremamente importante e fulcral para a empresa selecionar as estratégias adequadas. A segmentação dos clientes é a parte mais significativa da identificação do cliente e inclui a divisão de todos os clientes em segmentos devidamente valorizados.

Esta seção começará por destacar a segmentação de clientes na base para pesquisas e desenvolvimentos substanciais em marketing, seguido por uma descrição de alguns métodos de gestão da relação com os clientes, que incluem a análise de valor do cliente e mecanismos de atendimento positivos.

Por fim, e como vivemos na era dos dados e da inteligência artificial, descreveremos alguns algoritmos e técnicas de extração de dados (*data mining*), aplicadas aos dados dos clientes para conseguirmos mais conhecimentos sobre o comportamento de compra do cliente, discutindo as suas características, bem como as vantagens.

2.1. Gestão do relacionamento com clientes

O conceito de Gestão do Relacionamento com Clientes (CRM) surgiu no início do século XX. O CRM ajuda a organização a adquirir conhecimento que lhe permita compreender melhor o comportamento e as relações dos clientes, e, por isso, melhorar a atração, a retenção e a rentabilidade dos clientes (Schaeffer & Rodriguez Sanchez, 2020).

A importância dos programas de CRM aumentou nos últimos anos devido a um ambiente competitivo, à especialização de vendedores e compradores e forçou as empresas a assegurarem uma dinâmica de fidelização dos clientes para obterem lucros mais elevados e ganharem uma maior quota do mercado do que os seus concorrentes (Yong Seog Kim et al., 2001).

Assim, um excelente programa de CRM pode ajudar as empresas a manterem os clientes existentes e atrairem novos. Além disso, as empresas também reforçam a eficácia do marketing e das vendas. Ravi Kalakota & Marcia Robinson, (1999) explicaram que o CRM deve integrar as várias áreas do relacionamento com o cliente na empresa, como o marketing, as vendas, os serviços e suporte técnico para as necessidades dos clientes, e geralmente utiliza as Tecnologias de Informação para ajudar a empresa a gerir o relacionamento com o cliente de forma sistemática, melhorando a fidelização do cliente e aumentando os lucros globais dos negócios.

2.2. Segmentação de clientes

O conceito de segmentação foi introduzido na comunidade científica por Wendell R. Smith em 1956 através da publicação do artigo: " Product Differentiation and Market Segmentation as Alternative Marketing Strategies.". Em 1956, definiu a segmentação do mercado como um processo que divide os clientes alvo de uma empresa em grupos, usando as suas semelhanças em características, comportamentos e necessidades de criar grupos homogéneos que são heterogéneos entre os outros grupos (Smith, 1956).

De acordo com Kotler et al. (2006), a segmentação dos clientes B2B baseiam-se nos mesmos princípios que a segmentação de marketing nos mercados de consumo, porque os compradores têm necessidades diferentes, é importante que as empresas dividam os seus mercados em segmentos mais pequenos para que o negócio possa direcionar-se exclusivamente a cada segmento.

Nas décadas de 1970 e 1980 foram desenvolvidos vários modelos de segmentação. Desde então, no geral, o número de contribuições tem sido relativamente estável. No entanto, o número de contribuições para a segmentação B2B em revistas conceituadas e revistas de primeira linha vem diminuindo nos últimos 35 anos (Mora Cortez et al., 2021). A pesquisa de segmentação B2B tem muitas questões essenciais ainda não tratadas. Num mundo em mudança, com o surgimento de novos mercados, os mercados tradicionais são afetados pela digitalização e os processos de compra estão evoluindo (Grewal et al., 2015), é problemático que haja pouca investigação e pouca literatura para lidar com estas mudanças de segmentação.

Também, faltam estudos empíricos que proporcionem uma maior compreensão de como as empresas aplicam, implementam e avaliam a segmentação e seu impato no desempenho das empresas. Além disso, há necessidade de diretrizes adicionais para segmentos robustos do mercado

B2B com base em novas condições que, por exemplo, as preocupações ambientais, a inteligência artificial, os valores sociais e as crises forneceram (Mora Cortez et al., 2021). A segmentação, embora ainda não perfeita, aporta valor às empresas e, como conceito-chave, tem sido útil, por exemplo, para desenvolver estratégias de marketing há mais de 40 anos e alocar recursos, além de compreender como os clientes devem ser direcionados e geridos (Morgan et al., 2019).

A evolução das tecnologias de informação e comunicação, especialmente nos sistemas de gestão de base de dados e extração de dados, tornaram mais importante a escolha das técnicas de segmentaçao. A vasta disponibilidade de dados e o desempenho ineficiente de técnicas estatísticas tradicionais em grandes volumes de dados, estimularam a pesquisa de ferramentas de segmentação mais eficazes na descoberta de informações úteis sobre os mercados e clientes. Assim, a descoberta de conhecimentos (KD) e a mineração de dados (DM) têm sido apontados como uma solução para o problema. O machine learning, a estatística, a inteligência artificial e as tecnologias de gestão de dados e conhecimento são incorporadas com KD e DM, fazendo uso de suas teorias e dos algoritmos.

Os especislistas de marketing estão interessados na aplicação destas tecnologias para solucionarem alguns dos problemas relacionados ao marketing, como a previsão e a segmentação, bem como outros sistemas de suporte a decisões de marketing, baseados em conhecimento, especialmente com técnicas de DM (Liao, 2003).

2.2.1. Pré-segmentação: O que é um mercado?

Na visão de uma empresa, o que é um mercado? Embora muitos artigos abordam a segmentação de mercados, poucos explicam o que é um mercado ou qual mercado que estão a segmentar. Os mercados podem ser vistos de forma diferente, seja do lado da oferta, do lado da procura ou da tecnologia (Goller et al., 2002). A visão mais comum e a que vai ser considerada neste estudo é o lado da procura, onde o cliente está no centro. Esta visão pode implicar clientes de diversas indústrias e diferenças entre clientes diretos e indiretos. Neste sentido, (Hlavacek, 1986) diferenciam-se entre indústria e mercado. Uma indústria refere-se a um amplo grupo de fabricantes na produção de uma gama de produtos relacionados.

Um segmento de mercado é um grupo mais distinto de clientes com requisitos semelhantes que podem ser satisfeitos por ofertas e tecnologias diferentes. O objetivo da segmentação de mercado é identificar segmentos de clientes relevantes e utilizar o *insight* para adequar as estratégias de

marketing às necessidades específicas de diferentes segmentos (Chéron & Kleinschmidt, 1985). A segmentação de mercado, seguida pela definição do *target* de mercado e pelo posicionamento no mercado, representa o primeiro passo no processo de marketing-alvo (Kotler & Keller, 2015).

2.2.2. Variáveis para segmentação B2B

Os consumidores têm uma variedade de diferenças de acordo com as suas características. Na literatura de marketing de consumo e industrial, podem ser encontradas várias variáveis de segmentação, tais como variáveis geográficas, demográficas, firmográficas, comportamentais, de tomada de decisões, comportamentos de compra, fatores de situação, personalidade, estilo de vida, psicográficos, e assim por diante (Kotler, 2006). A literatura sobre segmentação B2B propõe diversas variáveis para a criação de segmentos de mercado. Na década de 1980, uma categorização convencional das variáveis foi uma divisão em macro e microvariáveis. As macrovariáveis incluem tamanho, tipo de indústria, códigos de atividade, resultados e localização geográfica; variáveis que muitas vezes são facilmente observáveis e acessíveis a baixo custo ou publicadas por entidades especializadas e até governamentais (Chéron & Kleinschmidt, 1985). As variáveis mais utilizadas na segmentação estão no nivel macro (Abratt, 1993).

A segmentação baseada no setor de atividade é frequentemente usada como base para a segmentação baseada nas necessidades, uma vez que as necessidades dos clientes podem variar de acordo com sua indústria. Outra nuance no fluxo de macro-segmentação é baseada no tamanho do cliente e na importância estratégica, classificando os clientes em contas estratégicas, contas-chave ou contas nacionais (Rangan, Moriarty, & Swartz, 1992). Ao longo dos anos, houve uma mudança do foco das macrovariáveis para maior atenção às microvariáveis. A transição para uma abordagem de microsegmentação ganhou força desde o final da década de 1990, e foi dominada por um foco no comportamento de compra ou da transação de compra. As variáveis de segmentação resultam da transação de compra, do processo de compra e das características definidas, do real comportamento de compra por parte dos clientes e dos padrões de compra da estratégia de disponibilização (Frear, Alguire, & Metcalf, 1995).

Nos anos 2000, continuaram a prevalecer as microvariáveis. A falta de consideração das necessidades dos clientes como uma variável de segmentação foi determinada como primordial. Desde então, os investigadores proposeram variáveis mais centradas no cliente, incluindo na estratégia, na compreensão das necessidades do cliente, no uso do cliente, nos requisitos para

clientes e na importância da marca (Mora Cortez et al., 2021). As variáveis de segmentação do mercado foram classificadas em quatro grandes áreas, nomeadamente, variáveis geográficas, demográficas, psicográficas e comportamentais (Kotler, 2006).

Mais recentemente, os autores consideraram também o aumento da digitalização. Este fluxo considera web design, e-commerce e redes sociais. Neste sentido, Müller et al., (2018) investigam o marketing digital, social e móvel nas compras industriais com base em variáveis como frequência de compra, a função do comprador e o setor industrial.

2.3. Segmentação B2B: obstáculos e desafios

A segmentação de mercado B2B é o "Núcleo do bom marketing empresarial" (Bonoma & Shapiro, 1983). Apesar da utilidade identificada (Mora Cortez & Johnston, 2018);(Mora Cortez et al., 2021), ainda é necessário mais conhecimento sobre por que a segmentação do mercado B2B é "particularmente difícil de executar" e "com quem abordar, com que tipo de conteúdo" (Müller et al., 2018). Os estudos fornecem três razões para os obstáculos no domínio da segmentação de mercado B2B: (1) a grande maioria da literatura de segmentação de mercado lida com os mercados dos consumidores finais; (2) os profissionais de marketing industrial têm sido mais lentos para adotar a segmentação de mercado além dos meros segmentos tradicionais da indústria; e (3) o processo de compra em contextos B2B são mais formalizadas do que nos contextos de consumo (B2C), podem envolver múltiplos decisores (Müller et al., 2018).

Embora existam diferentes modelos de segmentação, os critérios utilizados para identificar segmentos de clientes são amplamente os mesmos. É importante mencionar as diferenças entre a segmentação dos mercados B2C e a segmentação do mercado B2B. Na segmentação do mercado de consumidor final os principais critérios para segmentação são geográficos, demográficos, psicográficos e comportamentais (Kotler, 2006). A segmentação do mercado industrial é mais complexa que a segmentação do mercado consumidor. Em geral, os produtos de consumo servem apenas a um propósito, enquanto os produtos industriais encontram múltiplas aplicações. Portanto, os clientes empresariais, interessados no mesmo produto, podem diferir substancialmente.

O principal desafio para uma segmentação eficaz continua a ser a resposta às cinco características centrais de um segmento: a medição, a sustentabilidade, a acessibilidade, a diferenciação e a execção (Kotler & Keller, 2015).

2.4. Análise do valor do cliente

A análise de valor do cliente é um método analítico para encontar as características dos clientes e faz uma análise mais aprofundada de clientes específicos para abstrair conhecimentos úteis a partir de grandes volumes de dados. As empresas aplicam o método de análise de valor aos clientes para identificar os clientes alvo e qual a sua contribuição atual e potencial.

O RFM foi desenvolvido pela primeira vez por Hughes, (1994), como um método para analisar o valor do cliente. A sua vantagem é extrair características dos clientes usando menos critérios (três dimensões) como atributos de segmentação para reduzir a complexidade do modelo de análise de valor do cliente. Além disso, do ponto de vista do comportamento de consumo, o modelo RFM é um método, de longa data, para medir a força do relacionamento com o cliente (Schijns & Schrode, 1996).

O custo de retenção é muito menos dispensiodo do que o custo de aquisição (Kotler, 2006) portanto, as empresas têm a intenção de usar a análise RFM para extrair informação das bases de dados e saber como os clientes gastam mais dinheiro e criam o maior valor para as empresas. Os diferentes grupos de clientes são segmentados com base no seu comportamento de consumo, através dos atributos RFM. Desta forma, garantimos que o valor do cliente em cada segmento é estabelecido objetivamente com base nos atributos RFM e não seja estabelecido subjetivamente. Ou seja, a segmentação do cliente é estabelecida apenas por três atributos (Recência - Frequência - valor Monetário da compra).

2.4.1. RFM

O modelo analítico RFM é proposto por Hughes, (1994) e é um modelo que diferencia os clientes importantes, em grandes quanidades de dados, por três variáveis (atributos), ou seja, intervalo de consumo do cliente, a frequência e o valor monetário de compra. O modelo RFM é descrito da seguinte forma:

(1) Recência da última compra (R), representa Recência, que se refere ao intervalo entre a ultima compra e o presente. Quanto menor for o intervalo, maior será o R.

(2) Frequência das compras (F), representa frequência, que se refere ao número de transações num determinado período, por exemplo, duas vezes no ano, duas vezes no trimestre ou duas vezes no mês. Muita frequência, o F é maior.

(3) Valor monetário das compras (M), representa o monetário, se refere-se ao consumo O valor do dinheiro num determinado período. Quanto maior o monetário é, meior é o M.

De acordo com a literatura, os clientes que gastam mais dinheiro ou compram com mais frequência, são também mais sensíveis às informações e mensagens que a empresa transmite. Da mesma forma, os clientes que compraram mais recentemente reagem melhor às iniciativas de marketing do que aqueles que não compraram recentemente. O pressuposto é que os clientes mais recentes, mais frequentes e com maiores gastos são os melhores clientes e assume que estes clientes vão atuar de forma semelhante no futuro. O método RFM tem atributos muito eficazes para a segmentação de clientes. A análise da RFM baseia-se na conhecida "Lei de Pareto" ou na 80/20. No caso da análise RFM, "80% das compras vêm de 20% dos clientes" ou seja, 20% dos clientes geram 80% das vendas". Estes resultados, embora pareçam irrealistas, são verificados na maioria das empresas (Hwang et al., 2004).

Embora o modelo RFM seja um bom método que diferencia, por três variaveis, os clientes importantes em grandes volumes de dados, existem estudos com algumas opiniões diferentes em relação às três variáveis do modelo RFM (Cheng & Chen, 2009). As três variáveis são iguais na importância, portanto, os pesos das três variáveis são idênticos (Hughes, 1994). Por outro lado, Stone (1995) indicou que as três variáveis são diferentes na importância devido à característica da indústria. Assim, os pesos das três variáveis não são iguais.

2.4.2. LRFM

Um método RFM desenvolvido por Arthur Hughes, (1994), é considerado a base para uma média de 80% de todas as técnicas de segmentação comportamental do cliente por investigadores (Kohavi & Parekh, 2004).

Uma análise RFM é realizada com a finalidade de agrupar clientes com padrões de gastos semelhantes. Também é importante notar que considerar a frequência média e os valores monetários por tempo unitário é uma abordagem mais adequada do que considerarmos valores totais.

Além disso, uma quarta variável foi adicionada à análise RFM por Buckinx & Van den Poel, (2005), ou seja, o atributo Antigudade (L), que é o tempo do relacionamento entre o cliente e a empresa.

De acordo com Mesforoush & Tarokh, (2013), o método LRFM é uma alternativa melhor do que o RFM a agrupar os dados.

2.4.3. LRFMP

Em alguns negócios, a periodicidade das transações dos clientes é um fator importante na análise do comportamento dos clientes. Recentemente foram publicados estudos, em setores diversos como o retalho, a banca mais tecnológica e a logística, aplicando um fator P adicional ao modelo LRFM.

Em 2017, ocorreu um contributo interessante, Peker et al. (2017) acrescentou a antiguidade e a periodicidade, ao modelo analítico RFM e propõe o modelo LRFMP. No modelo LRFMP, o fator P representa a periodicidade do retorno do cliente, que ajuda a identificar melhor os clientes críticos aos negócios e o modelo LRFMP aumenta também o desempenho do modelo de previsão de churn do cliente (S. Chen, 2018).

2.4.4. RFMV

O cliente é o ativo mais importante nos negócios. Por esta razão, as empresas estão a tentar tornar o seu negócio e os serviços centrados no cliente, especialmente no atual ambiente multicanal, onde as características comportamentais, sociais e demográficas dos clientes estão a mudar. Este ambiente dinâmico desperta maior interesse na gestão adaptativa do cliente para responder às necessidades dos clientes. Com a segmentação adequada, as empresas podem planear os produtos e serviços para direcionar os clientes interessados e construir uma relação mais próxima.

Os estudos mais recentes tentaram melhorar o conceito de RFM adicionando novas características ou aplicando técnicas de extração de dados. Embora o modelo RFM convencional fórmula bem a segmentação do cliente, geralmente ignora os itens adquiridos e sua categoria, o que pode fornecer informações significativas sobre o comportamento de compra do cliente.

O modelo RFMC, desenvolvido por Allegue et al., (2020) para segmentar os clientes num banco tunisino, e o modelo RFMV, desenvolvido em 2017 por Qadaki Moghaddam et al., (2017), adicionam um novo critério de segmentação B2B, a variedade (V) ou categoria (C) de produtos e serviços.

Os modelos RFMC e RFMV, embora apresentam diferentes técnicas, tratam a mesma especificidade no setor B2B, a necessidade de considerar a variedade e as categorias de produtos e serviços para melhorar a segmentação dos clientes.

2.5. Customer Lifetime Value

De acordo com Cheng, definir o valor da vida útil do cliente é um modelo baseado no cálculo da distância entre zeros e clusters de centro como um valor elevado e refere-se à maior parte da fidelização do cliente no interior (Cheng & Chen, 2009). O *Customer Lifetime Value* (CLV) é utilizado na avaliação do valor da rentabilidade do cliente. O CLV é calculado após segmentar clientes.
O valor CLV é calculado com base no *ranking* CLV especificado para cada segmento (Khajvand et al., 2011).

A forma de cálculo do CLV é $C^j = W_L\, C_L^j + W_R\, C_R^j + W_F\, C_F^j + W_M\, C_M^j$

Equação 1 - Formula de cálculo do CLV

Em que: C^j = Customer CLV rating

$C_L^j\ C_R^j\ C_F^j\ C_M^j$ = valores médios de L, R, F, M do cluster j

$W_L\ W_R\ W_F\ W_M$ = aos pesos atribuídos a L, R, F, M

2.6. Customer Portfolio Analysis

O portfólio de clientes é uma lista de grupos de clientes mutuamente exclusivos que consistem em toda a base de clientes do negócio (Buttle & Stan, 2015). O *Customer Portfolio Analysis* (CPA) é aplicado para analisar o segmento identificado utilizando o método de segmentação.

Na tabela 1, descreve os segmentos de clientes baseados no modelo analítico LRFM.

Segmento de cliente	Descrição do segmento
Platinium	Clientes com elevada fidelização. São os clientes mais valiosos, com o mais alto valor monetário médio, com elevada frequência e compras recentes.
Gold	Novos clientes, muito valiosos e com elevada frequência nas transações.

Silver	Clientes antigos com redução de atividade e valores médio LRFM abaixo da média.
Bronze	Clientes ocasionais, com valor médio LRFM, abaixo da média.

Tabela 1 - Portefólio de clientes

Ao analisar o CPA (*Customer Portfolio Analysis*), a empresa pode identificar os clientes que mais beneficiam a empresa.

Tal como para Thakur & Workman, (2016), os clientes em estudo são agrupados em 4 segmentos, nomeadamente ***Platinium, Gold, Silver*** **e** ***Bronze*** com base no valor de cada cliente, segundo o modelo de análise do valor dos clientes, LRFM.

2.7. Data Mining

O *Data Mining* (mineração de dados) é definido por Turban, (2007) como "o processo que utiliza técnicas estatísticas, matemáticas, de inteligência artificial e de *Machine Learning* para extrair e codificar informações úteis e, posteriormente, obter conhecimentos a partir de bases de dados de grande dimensão". Berry, (2004) também atribui uma definição semelhante no que diz respeito à extração de dados como sendo o processo de extração ou deteção de padrões ou informações ocultas a partir de grandes bases de dados.

Com dados abrangentes de clientes, a tecnologia de extração de dados pode fornecer inteligência empresarial para gerar novas oportunidades (Zhang, 1999). A aplicação de ferramentas de *Data Mining* em CRM é uma tendência emergente na economia global. Analisar e compreender comportamentos e características dos clientes é a base do desenvolvimento de uma estratégia competitiva de CRM, de modo a adquirir potenciais clientes e maximizar o valor do cliente.

De acordo com Swift, (2001), o CRM consiste em 4 dimensões:

(1) *Customer Identification*, é a dimensão que aprofundaremos neste estudo, e trata a classificação e a segmentação de clientes.

(2) *Customer Attraction*.

(3) *Customer Retention*.

(4) *Customer Development*.

Estas quatro dimensões podem ser vistas como um ciclo fechado de um sistema de gestão de clientes (Kracklauer, 2004). Partilham o objetivo comum de criar um conhecimento mais profunda dos clientes para maximizar o valor do cliente a longo prazo. As técnicas de *data mining* podem, portanto, ajudar a alcançar esse objetivo, extraindo ou detetando características e comportamentos desconhecidos do cliente, a partir de grandes bases de dados.

O *data mining* tem um aspeto gerador que consiste na construção de um modelo a partir de dados (Carrier, C. G., & Povel, O., 2003). Cada técnica de *data mining* pode executar um ou mais dos seguintes tipos de modelação de dados:

(1) Associação; (2) Classificação; (3) Segmentação; (4) Previsão; (5) Regressão; (6) Sequenciação; (7) Visualização.

Os sete modelos acima cobrem os modelos de *data mining* mencionados em vários artigos (Turban, 2007). Existem também várias técnicas de *Machine Learning* (aprendizagem automática) disponíveis para cada tipo de modelo de *data mining*. As escolhas das técnicas de extração de dados devem basear-se na característica dos dados e nos requisitos do projeto (Carrier, C. G., & Povel, O., 2003).

Existem vários algoritmos de *data mining* amplamente utilizados, neste projeto descreveremos apenas um modelo de segmentação dos dados.

Na figura 2, o esquema com a classificação gráfica das técnicas de extração de dados em CRM é adaptado do esquema proposto por Ngai et al., (2009) e que se baseia na literatura sobre técnicas de *data mining* em *CRM*.

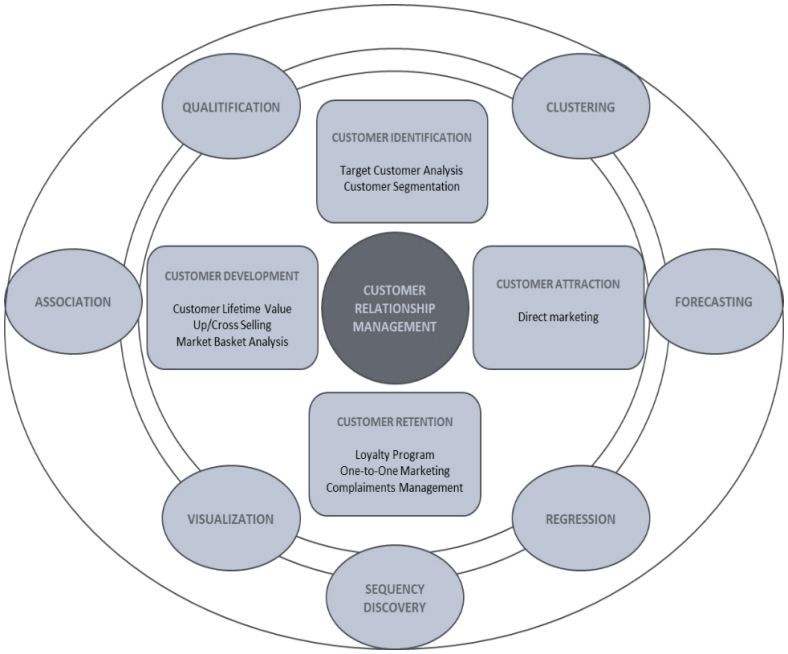


Figura 2 - Classificação de técnicas de Data Mining em CRM, Adaptado de Ngai, 2009

2.8. Machine learning

As técnicas estatísticas, tais como o *cluster* e a análise de componentes principais (PCA) combinadas com análise discriminatória (DA) ou regressão logística, têm sido tradicionalmente utilizadas para contruir os modelos de segmentação, mas a existência de grandes volumes de dados ligados a múltiplas características correlacionadas reduzem o real ajuste, robustez e interpretação destes modelos (Reinartz & Kumar, 2000).

Ao contrário dos modelos estatísticos, as técnicas de *machine learning* não confundem hipóteses restritivas a priori sobre dados e, combinadas com o armazenamento de dados de clientes existentes, poderão fornecer informações estratégicas para otimizar a eficiência da política promocional, utilizando tendências comportamentais anteriores para desenvolver modelos preditivos eficazes que antecipam futuras decisões de clientes (Norvig, 2003).

O *Machine learning* é um campo de informática que estuda a construção de algoritmos que aprendem com os dados, e podem fazer previsões sobre novos dados. Um processo de *machine learning* consiste em duas etapas: formação de um modelo de aprendizagem automática, seguido do processo de teste do modelo. O processo de teste é importante para ser capaz de reavaliar o desempenho do modelo e comparar diferentes modelos e algoritmos. O *machine learning* pode ser dividido em aprendizagem supervisionada, não supervisionada e de reforço. Diferem no objetivo geral do modelo, bem como nos requisitos dos dados utilizados para treinar e testar o modelo (Norvig, 2003).

Na aprendizagem não supervisionada, os dados não contêm quaisquer rótulos e são referidos como dados não rotulados, o que significa que a variável de resultados é desconhecida para todos os pontos de dados. O objetivo dessa aprendizagem é agrupar os dados em grupos de dados semelhantes sem nomear especificamente cada grupo. Isto pode dar novas informações sobre os dados e, portanto, ser útil em várias aplicações.

2.8.1. K-means

Embora a investigação de marketing inclua várias técnicas de segmentação de mercado, as técnicas de agrupamento são frequentemente mais utilizadas (Wang, 2005), especialmente o método de agrupamento *K-means*, e as suas variantes, porque podem acomodar grandes amostras associadas a estudos de segmentação de mercado. O *K-means* é um algoritmo de *machine learning* não

supervisionado, usado para segmentação dos dados num número (k) de *clusters* e foi introduzido por Mc Queen no ano de 1967.

A cada *cluster* é atribuído um ponto central, conhecido como *centroide*, que é calculado como a média de todos os pontos de dados que se referem ao *cluster*. O algoritmo *K-means* otimiza cada ponto de dados pertence ao *cluster* com o *centroid* mais próximo. Trata-se de um algoritmo iterativo e consiste em dois passos principais: atribuir um *cluster* a cada ponto de dados e atualizar o centroide de cada *cluster*. Estas são repetidas até que nenhum ponto de dados altera o *cluster* quando reatribuído (MacKay, 2002). A utilização do algoritmo *K-means* reque o conhecimento prévio do número de *clusters* k, que é o número de agrupamentos. Em muitas implementações de *machine learning* no mundo real, o k não é conhecido, uma vez que depende do conhecimento prévio dos dados. Existem várias abordagens para encontrar o k ideal para um determinado conjunto de dados, tais como testes gaussianos (Hamerly & Elkan, 2003). Neste projeto vamos usar o método do cotovelo (Elbow Method), para encontrar o número ideal de segmentos para os clientes mais valiosos, tendo em consideração os dados firmográficos e o modelo de análise do valor do cliente LRFMVP.

2.8.2. PCA

Na era da computação de alto desempenho, os investigadores são capazes de recolher uma grande quantidade de dados com muitas dimensões. A parte mais desafiadora é provavelmente analisar e fazer inferências com eficiência nesses conjuntos de dados com muitas dimensões. Para lidar com o número elevado de dimensões nos dados, os investigadores propuseram o uso de técnicas de redução de dimensões. Ao reduzir a dimensão dos dados, podemos concentrar as caraterísticas da informação em algumas dimensões. A técnica de redução de dimensões mais popular é a análise de componentes principais (PCA).

Esta técnica reduz a dimensionalidade dos dados mantendo a variação dos dados. Identifica os principais componentes que explicam a maior parte da variação dos dados para que cada ponto de dados possa ser representado com apenas um par de números, e consequentemente, desacoplá-lo para visualizar facilmente semelhanças e diferenças entre registos (Ringnér, 2008).

2.8.3. Apriori

Tendo em conta um grande conjunto de transações, o problema de identificação das regras de associação consiste em encontrar todas as regras de associação interessantes que tenham apoio e confiança superiores ao apoio mínimo especificado pelo utilizador e ao limiar mínimo de confiança, respetivamente. As regras de associação são usadas para detetar o uso comum de dados. Por exemplo, a compra de um produto quando outro produto é comprado representa uma regra de associação. Uma regra de associação é escrita como X =>Y, onde X e Y são o conjunto de itens.

Para gerar as regras de associação, o algoritmo Apriori é o mais utilizado (al Shorman & Hasan Jbara, 2017). A exploração das regras de associação é uma técnica importante na mineração de dados e é utilizada para encontrar associações em grandes quantidades de dados. Primeiro identifica os elementos individuais frequentes através das transações e depois estende-se aos conjuntos de elementos cada vez maiores até que os conjuntos de elementos resultantes atinjam um limiar de frequência especificado (suporte).

2.8.4. Elbow Method

O método do cotovelo consiste em executar o algoritmo *K-means* várias vezes com diferentes valores de k, ou seja, alterar o número de *clusters* para cada iteração.

Para cada iteração, calcula-se o erro absoluto médio (MAE - *Mean Absolute Error*), foi a medida de qualidade utilizada para medir a magnitude média dos erros num conjunto de previsões, sem considerar a sua direção. equação 2, da fórmula MAE:

$$\frac{1}{n}\sum_{j=1}^{n} \left|y_j - \hat{y}_j\right|$$

Equação 2 - Cálculo do erro absoluto médio (MAE), método do cotovelo

A ideia por trás do método do cotovelo é que o ganho marginal de adicionar um novo *cluster* irá para alguns k cair dramaticamente. Isto será claramente visível no gráfico como um ângulo (cotovelo), e o número de agrupamentos é encontrado nesse k (Bholowalia & Kumar, 2014).

2.9. Sumário

Nas últimas décadas, as mudanças sociais e económicas na interação entre organizações e clientes tornaram a gestão da relação com o cliente (CRM), um dos processos mais importantes na gestão dos negócios (seguido pela gestão de relações com parceiros de negócios). O CRM pode ser definido como a recolha, armazenamento e análise dos dados dos clientes de forma a aumentar a fidelização e o valor dos clientes, e aumentar os benefícios organizacionais (Hwang et al., 2004; Kandeil et al., 2014). De forma a aumentar a satisfação do cliente e impedir que os clientes saiam da organização, a organização deve focar-se na segmentação e na atenção às necessidades individuais dos clientes.

Em geral, a "segmentação do cliente" é o processo de divisão dos clientes em diferentes grupos baseados em informações geográficas, demográficas e firmográficas, com a finalidade de adotar estratégias adaptadas a cada grupo baseadas no consumo de bens e serviços e no histórico de compra dos clientes (Chang & Tsai, 2011; Kim et al., 2006).

O processo de segmentação é um fator essencial no CRM e no sucesso no negócio, ajuda a harmonizar os clientes, agrupando-os pelo seu valor e em grupos relacionados (Kandeil et al., 2014). Uma vez que não existe uma abordagem preferencial à segmentação de clientes, o melhor modelo é aquele que apresenta uma visão adequada dos clientes atuais e potenciais, e ajuda as organizações a alcançar eficazmente os mercados e o *feedback* adequado dos clientes (Lemon & Mark, 2006).

Diferentes técnicas e formas de análise de dados podem ser usadas para a segmentação do cliente; os mais comuns são o uso de técnicas de extração de dados e variáveis de comportamento do cliente. A mineração de dados é o processo de descoberta e extração de padrões ocultos a partir de grandes quantidades de dados (Parsa et al., n.d.).

O modelo RFM, que tenta identificar os clientes com base nas suas características comportamentais, é um dos mais utilizados (Y. L. Chen et al., 2009; Kandeil et al., 2014). Este modelo comportamental utiliza três critérios de transações de clientes: Recência, frequência e valor monetário. Estas categorias são interpretadas por ferramentas de mineração de dados e algoritmos. Assim, o modelo RFM prevê os próximos movimentos do cliente com base no seu comportamento (Cheng & Chen, 2009). Não é apenas considerado como um modelo de segmentação, mas também inclui o conceito de valor do cliente. É por isso que muitos estudos utilizam este modelo para a descoberta e análise do valor do cliente com base no comportamento de compra passado (Cheng & Chen, 2009).

Com base no seu comportamento de compra, o RFM revela que os clientes valiosos são aqueles com maior frequência e valor monetário e a menor Recência. Apesar da sua simplicidade de compreensão, interpretação e implementação (Hosseini et al., 2010; Wu et al., 2014), o modelo tem limitações, não

considera as características pessoais e variáveis demográficas dos clientes. Assim, foram realizados estudos com o objetivo de aumentar a precisão nos resultados do modelo RFM, adicionando variáveis que podem ser categorizadas em dois grupos de orientadas para o cliente e orientadas para o produto. No primeiro grupo, são tomados em consideração fatores como o valor vitalício do cliente (Allahyari et al. n.d.; Rajkumar Venkatesan & v. Kumar, 2004), são considerados e, no segundo grupo, os tipos de produtos são considerados para o desenvolvimento do modelo (Hosseini et al., 2010; Wu et al., 2014). Uma vez que a variedade de produtos é um dos principais parâmetros de avaliação do comportamento do cliente, estudar este fator no campo da comunicação negócio-a-negócio (B2B) representa uma nova abordagem vital. Assim, com base nas estratégias de venda de *up-selling* e *cross-selling* como as ferramentas importantes de aumentar o valor do cliente para as organizações, (Qadaki Moghaddam et al., n.d.) tenta melhorar o modelo RFM adicionando a variável de variedade de produto (V), o modelo RFMV visa adicionar a variável de variedade de produtos como uma nova variável comportamental no modelo RFM, e classificar os clientes da empresa na área da B2B.

Allegue et al. (2020) , propõe uma alternativa à segmentação clássica RFM no setor bancário, com uma nova segregação, denominada RFMC. Este modelo, segundo Allegue et al. (2020), representa melhor o comportamento dos clientes utilizando, não só os perfis dos clientes, mas também a categorização dos seus gastos e desta forma permite o uso de serviços que melhor satisfazem os desejos, expectativas e necessidades dos clientes, aproveitando ofertas que se adequam aos seus comportamentos de gastos. A maioria dos estudos abordam o negócio de consumidor final (B2C) e poucos estudos foram realizados na área do negócio empresarial (B2B). No domínio do B2B, a empresa terá maiores lucros em grande volume de compras. Portanto, compreender os clientes seria a chave de sucesso na sua área.

Alguns investigadores propuseram razões como a qualidade da comunicação, confiança, participação, satisfação, aumento da compra e alterações organizacionais como fatores eficazes de manutenção de clientes nas organizações B2B. Outros investigadores tentaram estabelecer novos modelos RFM, quer considerando variáveis adicionais, excluindo algumas das variáveis de acordo com a natureza do produto ou serviço. Por exemplo, Buckinx & van den Poel, (2005) introduziram a dimensão de Antiguidade da relação do cliente(L) no modelo RFM original. A antiguidade do cliente assume um papel fundamental na fidelização dos clientes e, posteriormente, o modelo LRFM tornou-se um dos modelos RFM mais utilizados, recebendo uma atenção considerável na literatura recente, por exemplo Kandeil et al., (2014) também utilizou técnicas de *clustering* LRFM para segmentar os clientes de um distribuidor B2B. Em alguns negócios, como o retalho, a periodicidade de compra é um fator importante na análise do comportamento dos clientes. Assim, o modelo LRFMP foi desenvolvido onde

o fator P implica a periodicidade de compra do cliente. Peker et al., (2017) aplicou o modelo, através da aplicação do algoritmo *k-means* para analisar os clientes de um retalhista na Turquia.

A revisão da literatura relevante mostra que existe uma escassez de investigação não só no contexto B2B em geral, mas também em setores que dependem fortemente da segmentação dos clientes e da gestão de relações. Na maioria dos estudos até à data, apenas um nível de agrupamento é usado, enquanto, neste estudo, duas camadas de agrupamento têm sido usadas para ter uma melhor compreensão dos clientes e escolher as estratégias mais adequadas. Por isso, numa breve referência, os clientes de uma *Fintech* no Irão são estudados através dos modelos LRFMP e são segmentados em *clusters* (Sheikh et al., 2019). É um contributo científico importante, em primeiro lugar, aumenta a literatura temática e é adequada para o modelo de negócio B2B, especialmente nas empresas prestadoras de serviços de pagamento. Em segundo lugar, propõe um método de agrupamento de vários níveis e que utilizado simultaneamente ajuda a melhor compreensão do comportamento dos clientes.

CAPÍTULO III

3. COMPREENSÃO E PREPARAÇÃO DOS DADOS

De acordo com Chapman, (2000) o conhecimento do negócio é o primeiro passo do processo CRISP-DM, que envolve uma compreensão clara do objetivo e do problema numa perspetiva de negócio.

O objetivo desta investigação é aplicar inteligência artificial para segmentar clientes empresariais e melhorar a previsão estratégica de oportunidades futuras.

3.1. Compreensão e exploração dos dados

Na segunda fase da metodologia debruçamos na recolha e análise dos dados, para os compreendermos. Os dados foram recolhidos a partir dos sistemas de gestão da empresa inCentea - tecnologias de gestao, SA.. A informação é relativa à operação comercial em Portugal. A figura 3, representa a estrutura dos registos obtidos a partir do CRM e do ERP da organização e compreendem a atividade entre o ano 2010 e Abril de 2022.

Columns: 48	Column Type	Column Index	Color Handler	Size Handler	Shape Han...	F
Data	Local Date	0				
Documento	String	1				
TipoDoc	String	2				
Série	String	3				
NumDoc	String	4				
NumLinha	String	5				
Venda	Number (double)	6				
Margem	Number (double)	7				
ModoPag	String	8				
Descrição	String	9				
Quantidade	Number (double)	10				
Artigo Código	String	11				
Artigo Descrição	String	12				
Tipo de Artigo Código	String	13				
Tipo de Artigo Descrição	String	14				
Família	String	15				
Subfamília	String	16				
Recorrência	Boolean value	17				

"NumDoc" (5/48)

Figura 3 - Estrutura do dataset

Os dados têm 317.883 registos e 48 dimensões para caterização da informação e que são descritas na tabela 2.

Dimensão	Tipo de dados	Breve descrição
Inc_nif	String	Número de identificação fiscal
Sector de Actividade	String	Setor de Atividade
País	String	País
Zona	String	Zona
Antiguidade	Number (integer)	Antiguidade
Data 1º Negócio	Local Date	Data 1º Negócio
Volume de Negócio	Number (integer)	Volume de Negócio da conta
Nº de Empregados	Number (integer)	Nº de Empregados da conta
Estado Conta	String	Estado da conta
Número	String	Número da conta
Dívida Vencida	Number (double)	Dívida Vencida
Condição de Pagamento nr dias	Number (integer)	Condição de Pagamento nr dias
Condição de Pagamento	String	Condição de Pagamento
Valida Inc.em Espera	Boolean value	Valida incidentes em Espera
Crédito Suspenso	Boolean value	Crédito Suspenso
Estado Info.Empresa	String	Estado da Informação empresa
Código Cliente	String	Código de Cliente
Abreviatura CC	String	Abreviatura da classificação comercial
Classificação Comercial	String	Classificação Comercial
Gestor	String	Gestor de conta
Cliente	String	Nome do cliente
Empresa	String	Nome da empresa do grupo inCentea
País Código	String	Código do País do cliente
Zona Descrição	String	Descrição da cona do cliente
Zona Código	String	Código da zona
Rubrica	String	Rúbrica do tipo de transação, ex: HardWare
Sigla Descrição Marca	String	Sigla descritiva da marca, ex: HPE
Sigla	String	Sigla interna da Marca, ex: HP
Marca Descrição	String	Descrição da Marca, ex: HP Enterprise
Marca Código	String	Código interno da marca

Recorrência	Boolean value	Transação recorrente ou não?
Subfamília	String	Subfamília do artigo
Família	String	Família do artigo
Tipo de Artigo Descrição	String	Descrição do tipo de artigo, ex: Serviço
Tipo de Artigo Código	String	Código do tipo de artigo
Artigo Descrição	String	Descrição do artigo
Artigo Código	String	Código do artigo
Quantidade	Number (double)	Quantidade do artigo na transação
Descrição	String	Descrição
ModoPag	String	Modo de pagamento
Margem	Number (double)	Margem da linha da transação
Venda	Number (double)	Valor da linha da transação
NumLinha	String	Número da linha
NumDoc	String	Número do documento
Série	String	Série do documento
TipoDoc	String	Tipo de documento
Documento	String	Descrição do documento
Data	Local Date	Data da transação

Tabela 2 - Detalhe do Dataset

Por motivos de confidencialidade e proteção de dados, as contas serão identificadas pelo número sequencial. Numa visualização geral dos dados, na figura 4, existem mais de 20.000 documentos diferentes e zero valores em falta, já os tipos de artigo são 24 e há valores não preenchidos em pelo menos 28.229 transações.

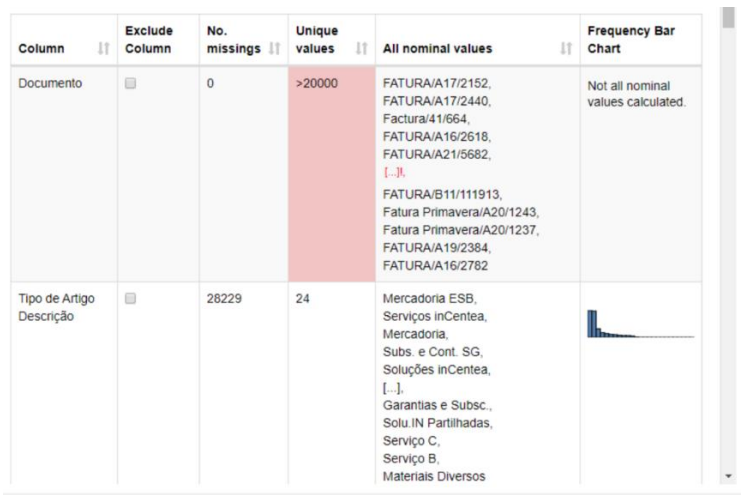


Figura 4 - Visualização geral do dataset

A exploração numérica dos dados permite-nos identificar a compreensão dos dados entre 2010 e 2022, entre outros dados relevantes, os 24 anos de antiguidade máxima de clientes e a média superior a 10 anos, são reflexo da longevidade da empresa em estudo e da fidelização dos seus clientes, conforme figura 5.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation
Estado Conta	☐	0	1	0.009	0.093
Recorrência	☐	0	1	0.352	0.478
Rubrica	☐	0	4	1.505	1.070
Year	☐	2010	2022	2016.737	3.200
Antiguidade	☐	0	24	10.799	7.611
Família	☐	0	236	46.950	24.269
Sigla	☐	0	769	73.611	110.598
Sector de Actividade	☐	0	666	173.561	141.346
Nº de Empregados	☐	0	5417	39.584	178.497
inc_AnoInformacaoEmpresa	☐	0	2022	1212.942	989.375

Figura 5 - Exploração geral do dataset

A análise ao número de transações e de clientes dá-nos uma ideia da dimensão da empresa e é muito significativo. A mediana do numero de transações anual é 23.537,5, com 2.370 clientes, figura 6.

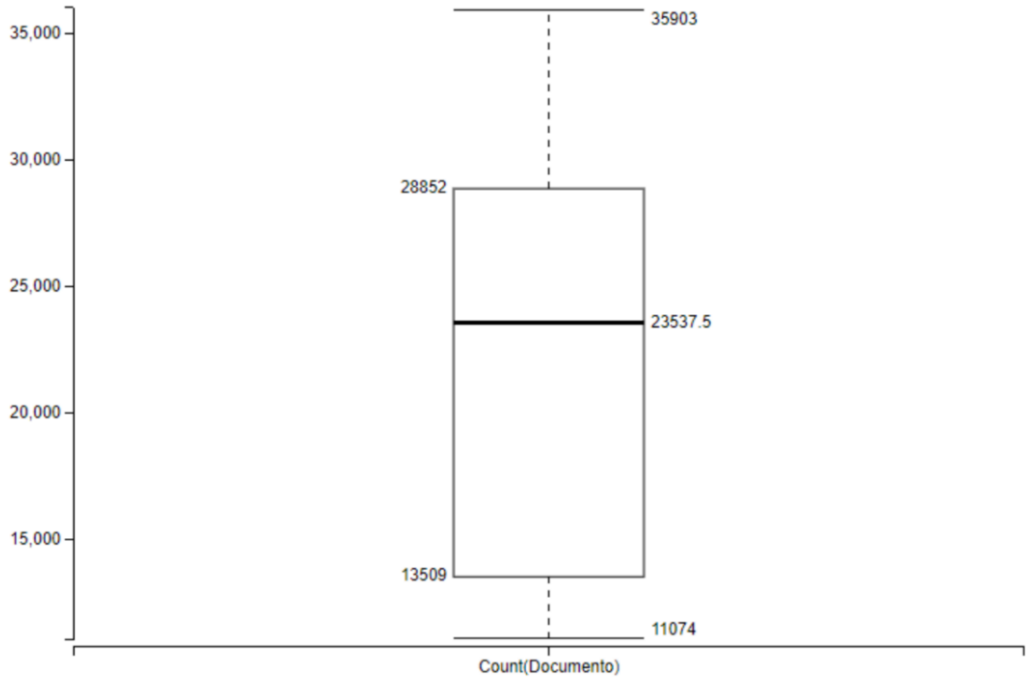


Figura 6 - Valor mediano, anual, de transações

Na figura 7, representa graficamente a verde o número de documentos de vendas emitidos, situado-se entre os 25.000 e os 35.000, nos ultimos 6 anos, mantendo estável o numero de clientes com os quais efetuou transações, situando-se nos 4.000 clientes.

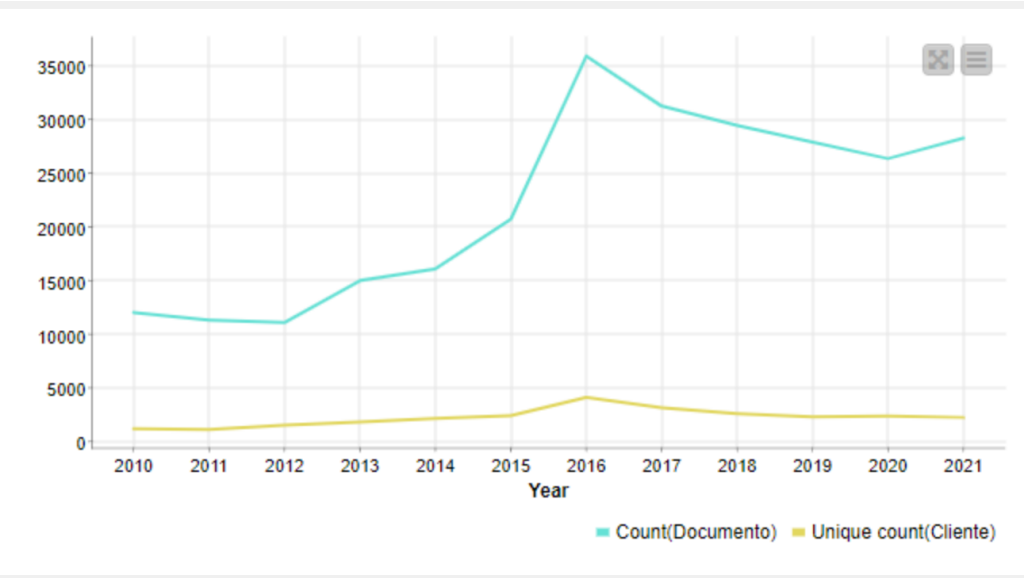


Figura 7 - Número de clientes e documentos por ano

A exploração de dados, na figura 8, mostra também a cobertura internacional da empresa, em países como Portugal, Angola, Moçambique, Cabo Verde e nos quais se destacam as cidades de Lisboa, Braga, Leiria, Luanda, Porto e Aveiro. Embora exista toda esta dispersão geográfica nos registos obtidos, neste estudo, consideramos apenas os dados de Portugal.



Figura 8 - Tag's das zonas

A figura 9 apresenta o vasto portefólio de fabricantes tecnológicos reresentados pela inCentea e que têm transações nos registos em análise. O maior destaque para algumas marcas como Microsoft, Sisqual, Primavera, Sage, Draytek, HP, MedicineOne ou SAP, significa que têm mais transações do que outras marcas. Quanto maior é o destaque, maior o número de transações, não valor, neste caso.



Figura 9 - Tag's das marcas

Para além da cobertura nacional e do portefóloio de fabricantes alargado, a inCentea é uma tecnológica multi setorial, como pode ser facilmente visualizado na figura 10, que lista os CAE's (código de atividade económica) associados aos clientes com transações.



Figura 10 - Tag's dos Códigos de Atividade Económica

3.2. Preparação e análise dos dados

A preparação de dados e as fases seguintes são essencialmente práticas e uso o KNIME que é uma plataforma modular e aberta para integração de dados, processamento, análise e exploração. O KNIME é cada vez mais utilizado por investigadores em várias áreas de *data mining* e *machine learning*.

Os dados fornecidos pela inCentea compreendem mais do que uma década entre 2010 e 2022. Neste período ocorreram várias alterações no catálogo de artigos e serviços, mas também alterações fiscais que poderão exigir um trabalho mais complexo para a normalização dos dados, bem como garantir que as contas de clientes são únicas e que estão devidamente preenchidas com os dados demográficos e firmográficas, nomeadamente o volume de negócios, o número de empregados e o setor ou código de atividade, que serão determinantes para a segmentação correta dos clientes.

3.2.1. Pré- processamento dos dados

Numa primeira fase da preparação dos dados, filtramos as colunas úteis e registos de Portugal, adicionalmente extraímos o ano na data de cada documento e extraímos o CAE da coluna Setor de atividade, figura 11.

Outra operação necessária foi o tratamento dos '*missing values*" nas fichas dos clientes, nomeadamente o Número de Empregados, Volume de Negócios, Setor de atividade e Antiguidade.

Esta operação incentivou a um trabalho massivo de atualização dos dados firmográficos, necessários para a identificação dos clientes *Corporate*, com volume de negócios superior a cinco milhões de euros. A identificação do documento resulta da concatenação do tipo de documento + série e número, dando origem a uma nova coluna, designada 'Documento'.

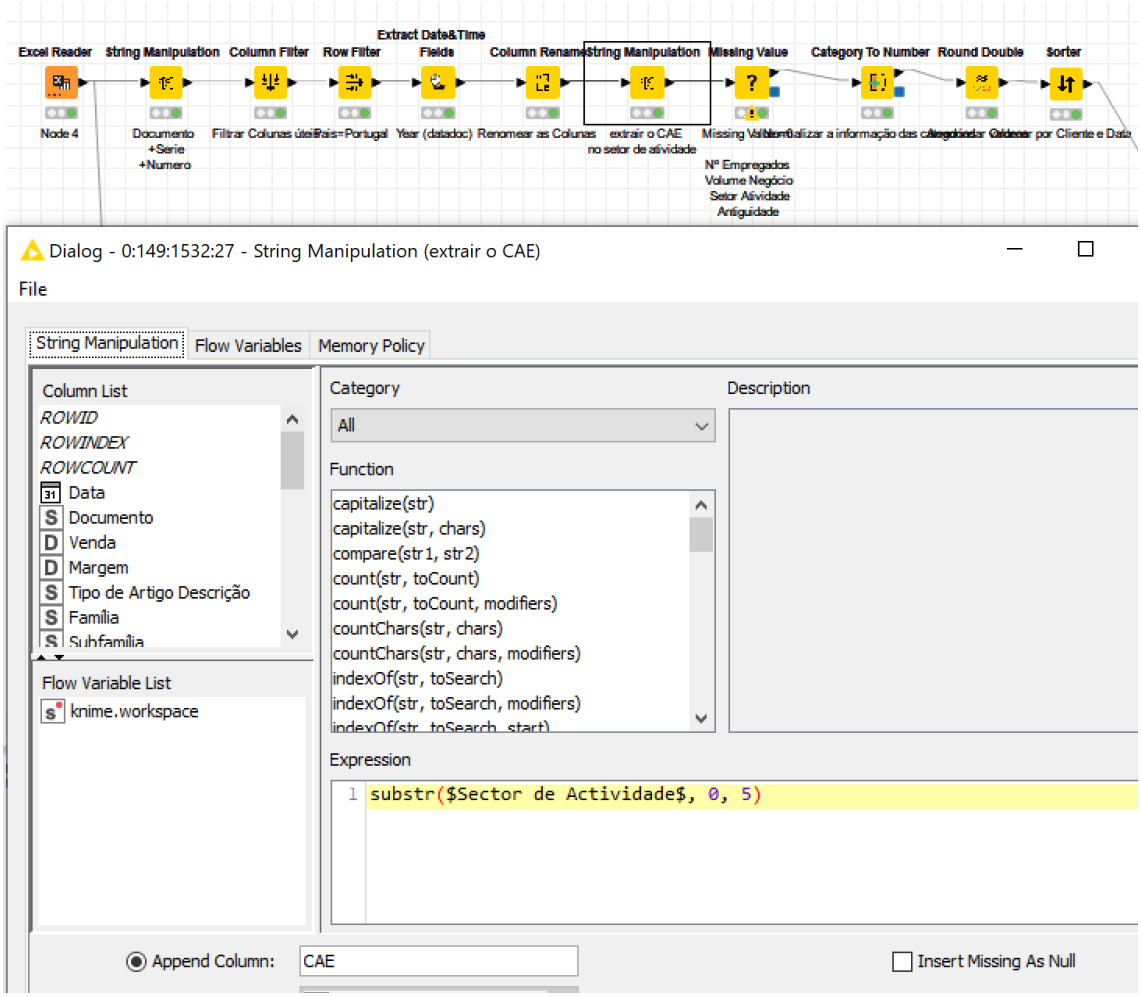


Figura 11 - Manipulação do CAE

Concluida a preparação dos dados fornecidos, ficamos com 274.006 registos e 29 colunas ou dimensões. Estes dados serão a base para as análises e processamentos seguintes, figura 12.

Row ID	Data	Documento	Venda	Margem	Tipo de...	Familia	Subfamilia	Recorr...	Marca Descrição	Sigla	Rubrica
Row302365	2017-01-25	Fatura Recibo/20171/112	12	2	?	90	ACE - Acessórios	false	?	408	0
Row130963	2019-05-16	FATURA/B19/2017	77	0	Mercadoria ...	23	11291 - Acess...	false	Outras	162	2
Row293466	2011-02-28	Factura/A11/110026	67,500	67,500	Serviço A	12	22001 - Imple...	false	Outras	54	1
Row48113	2012-03-05	FATURA/B12/120690	193	193	Serviços inC...	12	22001 - Imple...	false	inCentea	0	1
Row48114	2012-03-05	FATURA/B12/120690	211	211	Serviços inC...	12	22001 - Imple...	false	inCentea	0	1
Row35440	2012-04-30	FATURA/A12/120760	17	17	Serviços inC...	46	21004 - Sem C...	false	inCentea	0	1
Row20113	2015-09-30	NOTA DE CRÉDITO/A15...	-122	-122	Serviço A	42	26009 - Outros...	false	?	47	1

Figura 12 - Estrutura base do dataset

Uma das operações mais importante é identificar os *outliers*, que são os valores fora do padrão, neste caso concreto serão os volumes das transações ou margens das mesmas a analisar.

Usamos um modelo não supervisionado para detetar anomalias ou *outliers*, designado H2O e os registos dentro do quantil 0,995 foram marcados como 'anómalos', permitindo excluí-los das operações seguintes, figura 13.

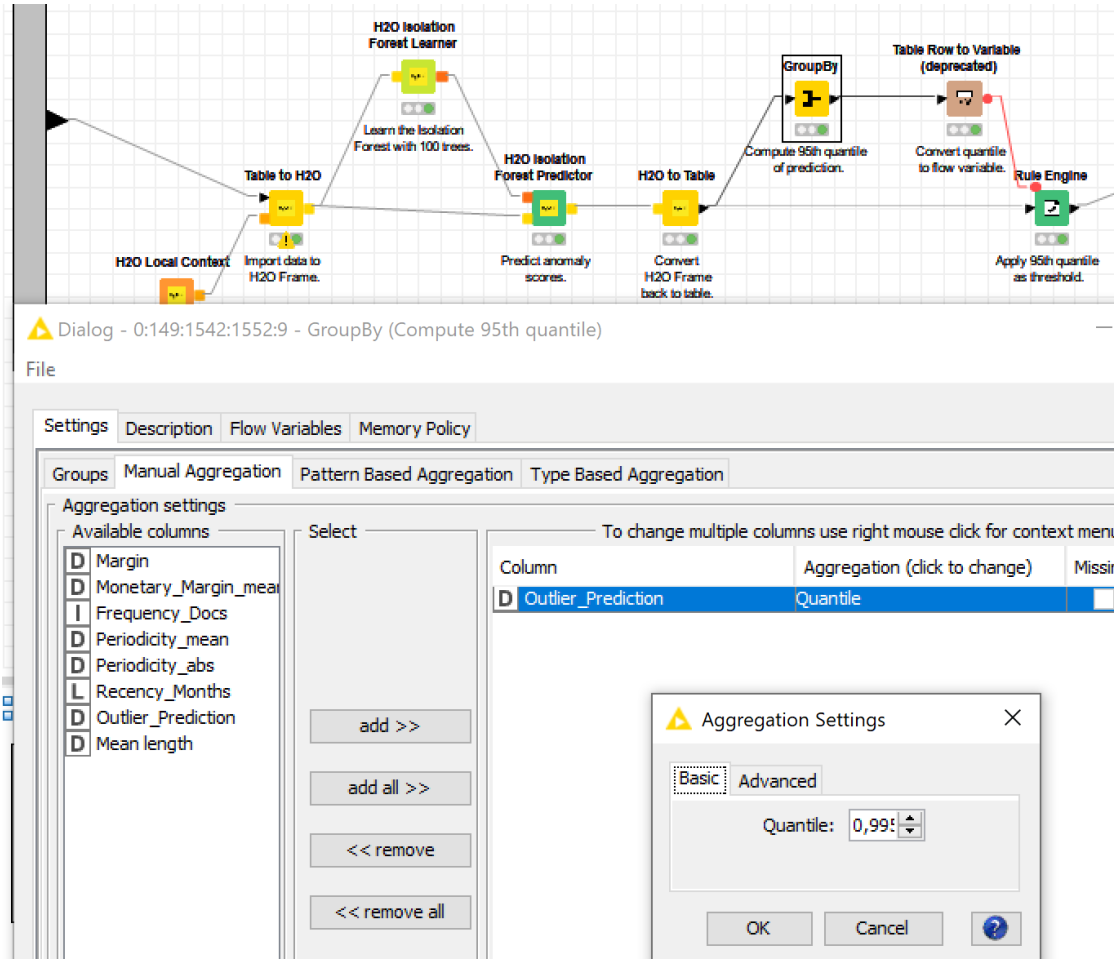


Figura 13 - Tratamento de dados anómalos com algoritmo H2O

Garantida a qualidade dos dados a processar, iniciamos uma das mais importantes fases de pre-processamento de dados que é calcular as novas dimensões que caraterizam o comportamento dos clientes e suportam a segmentação com base no modelo base RFM.

As dimensões que nos permitirão obter diferentes modelos, desde o base RFM, aos modelos que propomos, são de acordo com a tabela 3:

Dimensão	Designação	Observações
L	Lenght	Antiguidade do cliente
R	Recency	Recência, tempo que decorre desde a última transação
F	Frequency	Frequência de documentos no periodo
M	Monetary	Margem das transações no periodo
V	Variety	Gama de produtos comprados no periodo
P	Periodicity	Periodocidade das transações no periodo

Tabela 3 - Dimensões base para o modelo L RFM V P

Cada dimensão calculada é posteriormente dividida em 5 categorias, sendo atribuido o valor de 1 a 5, de acordo com o valor. Todos, à exceção do Recency, quanto maior o valor da dimensão, maior valor para a categoria. Significa que o Quantil com maior margem, terá categoria 5, já o 'Recency' maior terá a categoria 1, pois pretende-se que os clientes comprem de forma continua e atual.

Usamos o modelo supervisionado Auto-Binner para criar as dimensões L,R,F,M,V e P, figura 14.

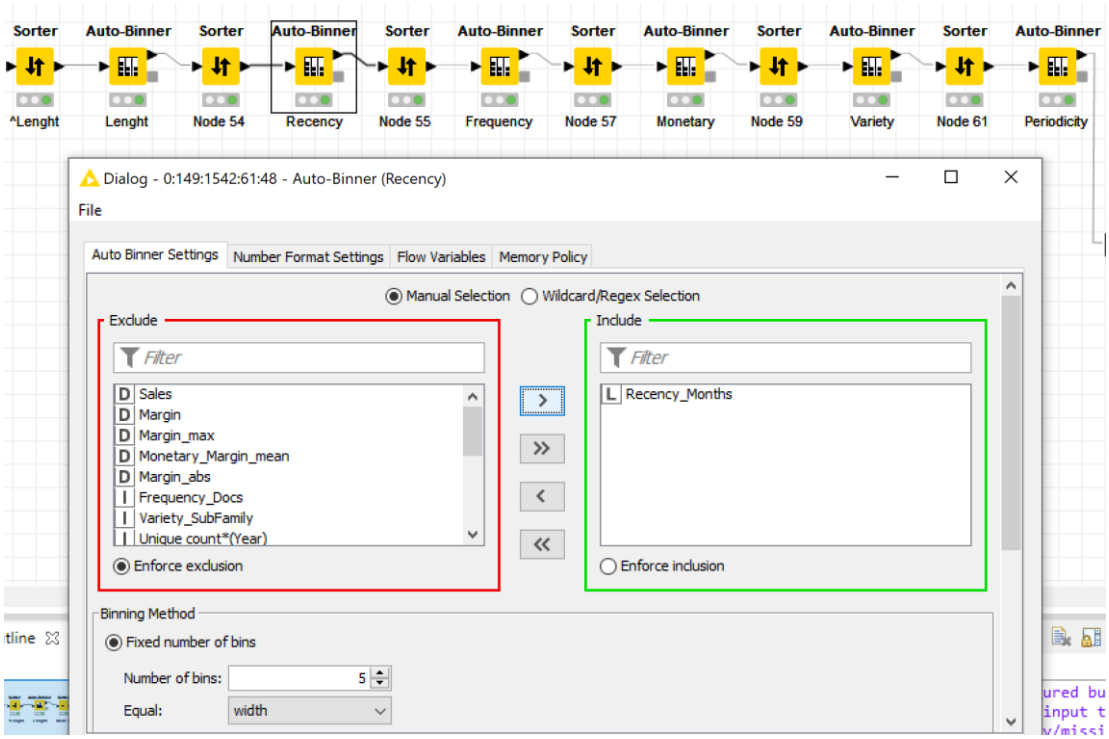


Figura 14 - Modelo supervisionado auto-binner

Seguindo a literatura, após obter estas dimensões e para maior apoio nas análises futuras, calculámos duas dimensõess adicionais, o *rank* e o *score*, de acordo com a fórmula da equação 3.

RFM Score = R + F + M

RFM Rank = 'R' + 'F' + 'M'

Equação 3 - Equações RFM Score e RFM Rank

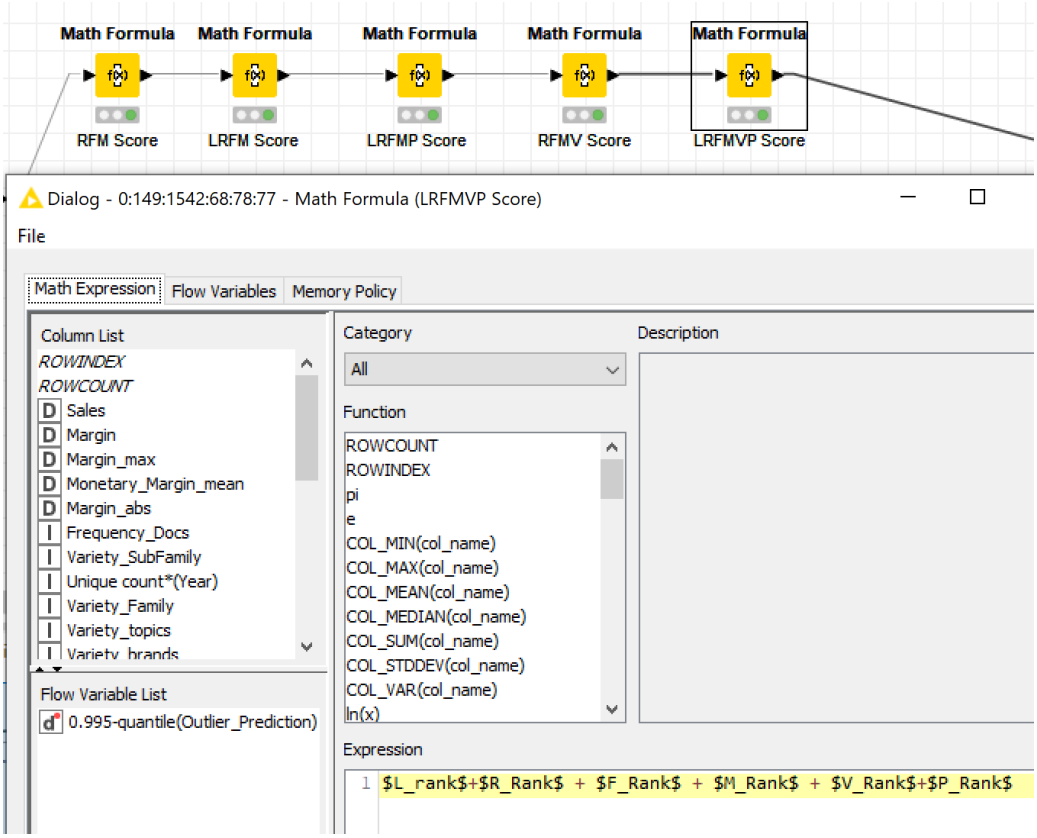


Figura 15 - Equação L RFM V P Score

Na figura 15, a equação de cálculo do LRFMVP *Score* e na figura 16 um exemplo de registos após o processamento das dimenões em análise.

Row ID	L_rank	R_Rank	F_Rank	M_Rank	LRFM_Rank	LRFM Score
Row5065	4	4	5	4	4454	17
Row5070	3	5	5	5	3555	18
Row5099	5	4	5	4	5454	18
Row5103	5	4	5	4	5454	18
Row5108	3	5	5	4	3554	17
Row5122	5	4	5	4	5454	18
Row5124	4	3	5	5	4355	17
Row5134	5	3	5	4	5354	17
Row5136	5	3	5	4	5354	17
Row5141	3	5	5	4	3554	17
Row5142	3	5	5	5	3555	18
Row5146	5	4	4	4	5444	17
Row5159	2	5	5	5	2555	17
Row5170	4	4	5	4	4454	17
Row5172	5	3	5	4	5354	17
Row5173	5	4	5	4	5454	18
Row5189	1	5	5	4	1554	15
Row5190	4	3	4	4	4344	15
Row5204	5	4	5	4	5454	18

Figura 16 - Resultados para as dimensões do modelo LRFM

Na figura 17 apresenta uma breve análise em que os clientes com maior margem média, são os clientes mais antigos e estes 324 clientes têm o LRFM *score* máximo, que é 20.

Row ID	LRFM Score	Customers	Lenght_Years_mean	Monetary_Margin_mean
Row14	20	324	16.559	72,171.127
Row13	19	330	11.703	35,870.376
Row12	18	403	10.05	29,382.588
Row11	17	389	9.496	20,830.278
Row10	16	537	8.229	15,431.056
Row9	15	539	8.019	5,521.174
Row8	14	604	8.088	3,476.553
Row7	13	614	7.875	1,903.161
Row6	12	698	7.734	1,326.467
Row5	11	739	7.524	684.191
Row4	10	716	7.528	365.648
Row3	9	798	6.787	168.286
Row2	8	725	6.546	76.854
Row1	7	555	5.225	31.526
Row0	6	299	4.545	20.201

Figura 17 - Resultados ordenados por LRFM score

Para além das dimensoes que avaliam o comportamento de compra dos clientes, no periodo global, propomos mais duas dimensões que completam os critéros de segmentação pretendidos pela empresa do estudo e que analisam o valor anual de compra recorrente do cliente, designado 'ARR' e o comportamento de compra do cliente nos 2 ultimos anos, sendo neste caso classificados de acordo com os escalões fornecidos pelo cliente.

Neste caso usamos um metodo supervisionado, '*Numeric Binner*' para efetuar a classificação comercial tradicional ABCD, de acordo com os escalões fornecidos pela empresa, sendo o A o escalão dos clientes com maior volume de compras, o D é o menor e o S é utilizado para clientes sem compras no periodo, conforme a figura 18.

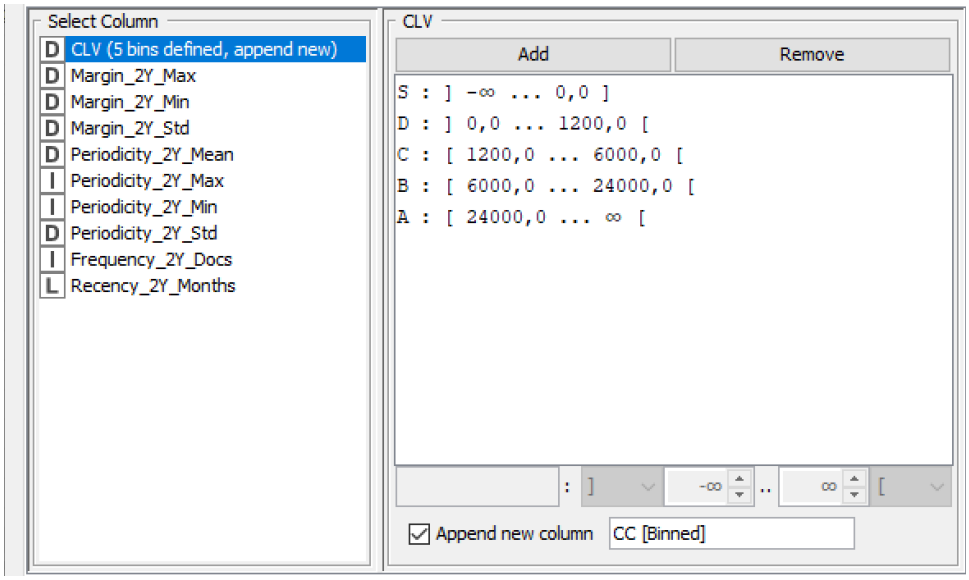


Figura 18 - Escalões para classificação ABCD

Na figura 19, um exemplo de registos após o processamento das dimensões para análise, em que substituímos a dimensão da (L) da antiguidade, que neste caso é conhecida, são 2 anos, pela

dimensão (P) para conhecermos a Periocidade de compra e com isto queremos perceber a influência dos novos modelos de negócio, baseados em mensalidades ou anuidades e que provocam mais transações e com menor periodicidade entre elas. O modelo a aplicar é o RFMP ao que acrescentamos '_2y' para distinguir de que se trata de uma análise aos 2 últimos anos e designa-se por RFMP_2y.

Quando comparadas com as dimensões históricas identificamos que os clientes que têm o RFMP_2y *score* mais alto, têm também um alto LRFM *score*, o que pode justificar a importância dos clientes fidelizados no negócio da inCentea, sendo que entre os mais antigos, surgem também clientes com 3 anos e mais, o que pode justificar a importância das estratégias de conquista de novos clientes.

Os clientes com RFMP_2y *score* máximo, têm transações de baixo valor e periocidade máxima e justificam a atenção da empresa para estratégias de venda recorrente, dada a importância que têm na avaliação das empresas tecnológicas e também a oportunidade da transformação do modelo de negócio tradicional que aposta na propriedade para o modelo com forte expansão baseado no 'pagamento por uso'.

Row ID	R_2Y_Rank	F_2Y_Rank	M_2Y_Rank	P_2Y_Rank	RFMP_2Y_Score	RFMP_2Y_Rank	LRFM Score	Lenght_Years	ARR
Row2297	5	3	1	5	14	5315	16	3	30,447
Row2281	5	2	1	5	13	5215	16	3	16,330
Row2280	5	2	1	5	13	5215	20	14	18,399
Row2287	5	2	1	5	13	5215	18	6	14,513
Row2277	5	2	1	5	13	5215	19	8	14,922
Row2294	5	2	1	5	13	5215	18	6	20,882
Row2288	5	2	1	5	13	5215	20	14	11,966
Row2284	5	2	1	5	13	5215	20	24	16,812
Row2290	5	2	1	5	13	5215	20	24	589
Row2285	5	2	1	5	13	5215	20	24	5,423
Row2286	5	2	1	5	13	5215	20	11	0
Row2291	5	2	1	5	13	5215	20	11	0
Row2276	5	2	1	5	13	5215	20	16	8,565
Row2282	5	2	1	5	13	5215	20	11	11,906
Row2292	5	2	1	5	13	5215	20	13	8,577
Row2278	5	2	1	5	13	5215	19	7	11,234
Row2283	5	2	1	5	13	5215	20	11	0
Row2151	5	1	1	5	12	5115	19	7	0
Row2190	5	1	1	5	12	5115	16	3	0
Row1587	5	1	1	5	12	5115	17	5	0
Row2240	5	1	1	5	12	5115	20	11	19,030
Row2245	5	1	1	5	12	5115	16	0	19,624

Figura 19 - Resultados do modelo L RFM P

3.2.2. Análise dos dados

Numa breve análise, na figura 20, 95% dos clientes estão entre os *score* 10 e 12, têm uma antiguidade média entre os 7 e 8 anos e compram entre 5 e 7 gamas de produtos diferentes. A margem média destes clientes vária entre os 1.894€ e os 2.040€. Nesta análise destacam-se 16 clientes, com antiguidade média de 13 anos e que compram 51 gamas de produtos e a margem média dispara para os 58.746€.

Outro aspeto que é alvo deste estudo é a importância e a influência do ARR no desempenho da organização e identificamos que estes 16 clientes têm um valor de faturação recorrente (ARR) muito superior e para isso contribui a variedade máxima na gama de produtos. Numa das seções deste

trabalho, vamos identificar a relação de compra destes clientes e ajudar a organização a recomendar, aos clientes semelhantes, os produtos que os outros compram e com isto aumentar a variedade em todos os clientes, aumentar a periodicidade e com isto aumentar o valor monetário, com especial atenção ao ARR, garantindo menor sasonalidade nos negócios.

Row ID	RFMP_2Y_Score	Customers	Lenght_Years_mean	Variety_SubFamily_Mean	Monetary_Margin_2y_mean	ARR_mean
Row9	14	1	3	33	80200	30447
Row8	13	16	13	51	58746	10007
Row7	12	7287	8	5	1894	373
Row6	11	371	7	7	2040	198
Row5	10	218	7	7	1919	107
Row4	9	198	6	6	872	5
Row3	8	154	6	5	827	0
Row2	7	21	11	4	86	4
Row1	6	3	10	5	52	0
Row0	5	1	10	3	80	0

Figura 20 - Resultados ordenados por RFMP_2Y_Score

Na figura 21 analisa os comportamento de compra dos cliente, segundo o módelo RFM e compara com as abordagens mais detalhadas que analisam a antiguidade, a variedade de produtos e tambem a variedade da gama de produtos. Os clientes mais valiosos e mais leais, com RFM *score* máximo, têm um valor médio de ARR muito superior aos restantes clientes, têm um LRFM *score* proximo do máximo, com 19 valores.

Simultaneamente são valiosos quando avaliados na Periodicidade e obtem um RFMP_2y *score* de 12, o mesmo acontece se juntarmos a avaliação da antiguidade (L), obtem um LRFMP *score* de 22, próximo do máximo que é 25 e curiosamente, se avaliarmos tambem pela Variedade da gama de produtos que compram, o LRFMVP é máximo e isto leva-nos a pensar que a variedade da gama de produtos tem especial influência nos resultados obtidos. Esta análise será particularmente útil para as equipas de marketing e de gestão de produtos, na elaboração de estratégias de compra conjunta, vulgo cross-selling e que tentaremos com este trabalho, através de técnicas machine learning, prescrever novos produtos a clientes semelhantes.

A importância do ARR, em todas as análises, será também objeto da aplicação de técnicas machine learning para prever qual é o ARR potencial para os clientes da carteira, mas também, e especialmente, para clientes novos e em prospeção, ajudando assim as equipas de marketing e comerciais a desenvolverem estratégias de conquista e de up-selling assertivas, numa bordagem 'Acount Based Marketing', que apresenta um conjunto de vantagens competitivas e do ponto de vista do custo vs beneficio, focada em contas com o CLV máximo.

Row ID	RFM Score	Customers	ARR_mean	LRFM Score mean	RFMP_2Y_Score mean	LRFMP Score mean	LRFMVP Score mean	Lenght_Years_mean
Row12	15	721	3628	19	12	22	25	10
Row11	14	539	334	18	11	21	23	9
Row10	13	573	164	16	12	19	21	8
Row9	12	519	125	15	11	18	20	8
Row8	11	607	37	14	12	17	19	8
Row7	10	608	31	13	12	16	18	7
Row6	9	710	6	12	12	15	17	8
Row5	8	766	4	11	12	14	16	7
Row4	7	766	1	10	12	14	15	8
Row3	6	713	0	9	12	13	14	8
Row2	5	877	0	8	12	12	14	8
Row1	4	637	0	8	12	12	13	8
Row0	3	234	0	8	12	13	14	11

Figura 21 - Resultados ordenados por RFM_score

Para melhor compreensão, aplicamos técnicas de análise do ARR e também, na figura 22, é visivel a tendência crescente da faturação recorrente e podemos ainda inferir que este modelo de negócio é mais sensivel aos momentos de crise. Em 2011 Portugal mergulhava numa profunda crise económica e sendo a faturação recorrente resultado de modelos de subscrição, são mais voláteis e flexiveis porque são pagos de acordo com o uso. Neste anos, as empresas puderam suspender o pagamento dos serviços e aplicações em subscrição e voltaram a ativar esses serviços quando necessários.

Estes modelos, exigem às empresas prestadoras mais atenção para evitar que os clientes desliguem os serviços e também por isso conhecer bem os clientes e atuar proativamente é uma das principais preocupações das empresas e para tal necessitam de ferramentas de lhe permitam agir antecipadamente, sempre possivel com ferramemtas preditivas.

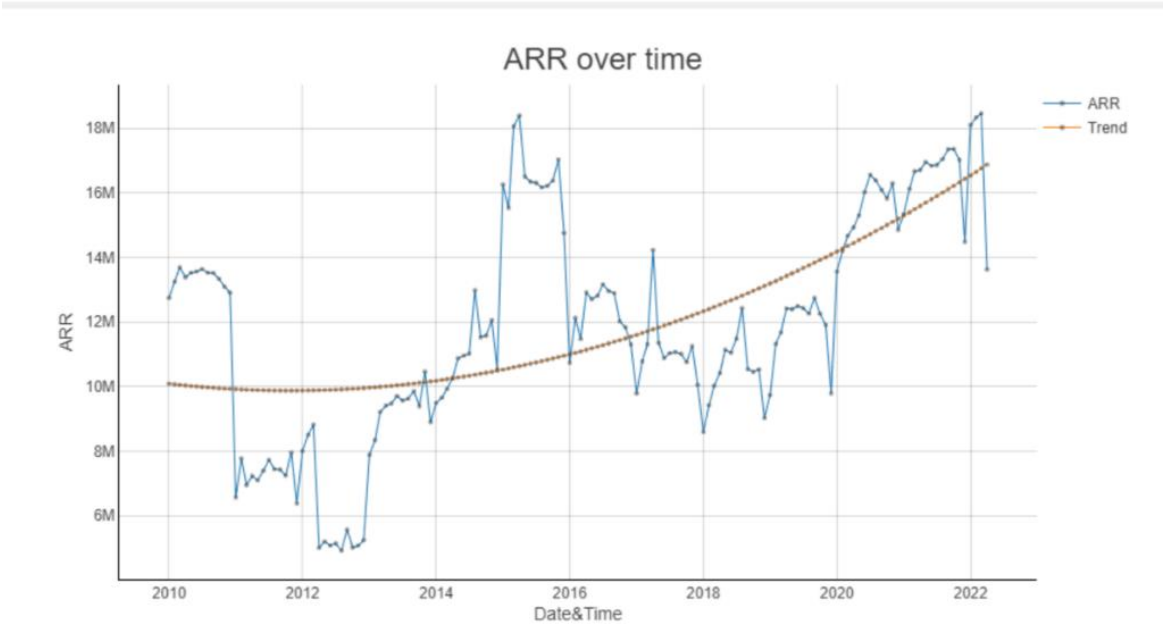


Figura 22 - ARR ao longo de todo o tempo de análise, gráfico Cohort

Para testar a fiabilidade da informação, adicionalmente exploramos o ARR com um artefacto base da ferramenta Knime, o *LinePlot* e verificamos o comportamento semelhante, figura 23.

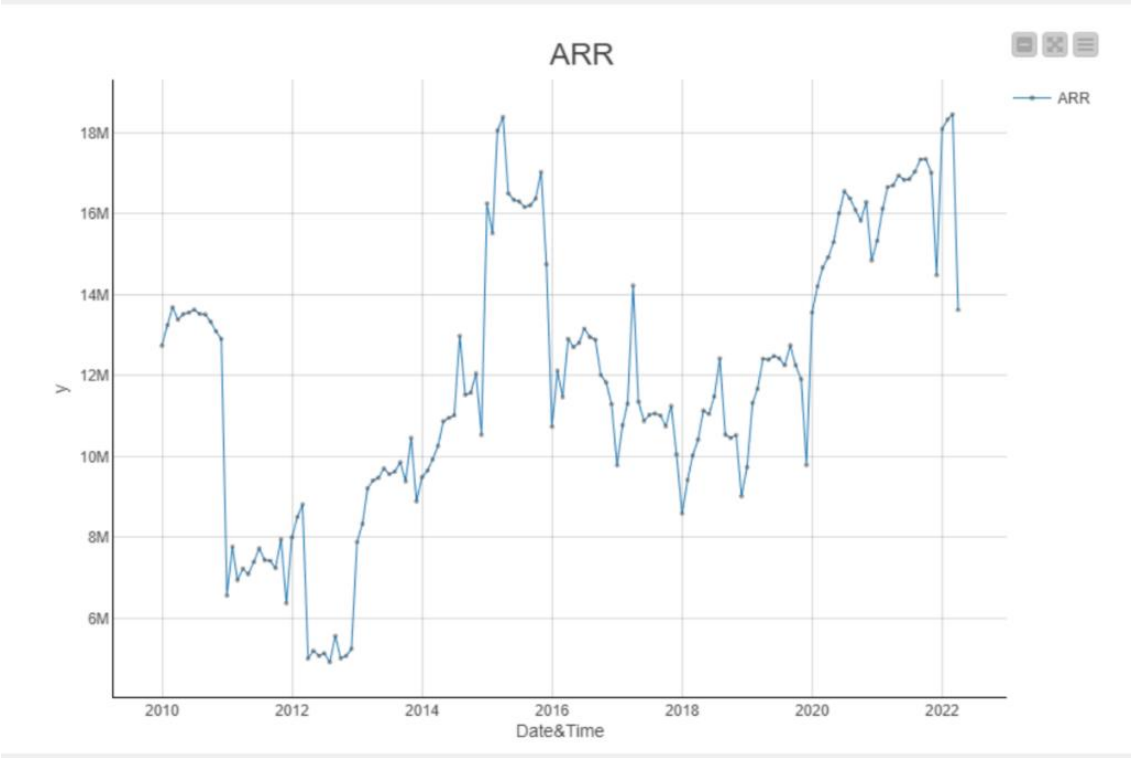


Figura 23 –ARR ao longo de todo o tempo de análise, LinePlot

A análise seguinte, figura 24, apresenta a faturação recorrente uma distribuição mensal e permite-nos identificar o comportamento ao longo dos anos.

O ano 2021, é o melhor ano e apresenta um valor de faturação recorrente mensal, vulgo MRR, estável ao longo do ano. Dezembro de 2021 foi compensado pelo mês de Janeiro de 2022 e os 3 meses iniciais indicam novos máximos, acompanhando assim a tendencia de crescimento do ARR nas empresas, sobretudo técnológicas.

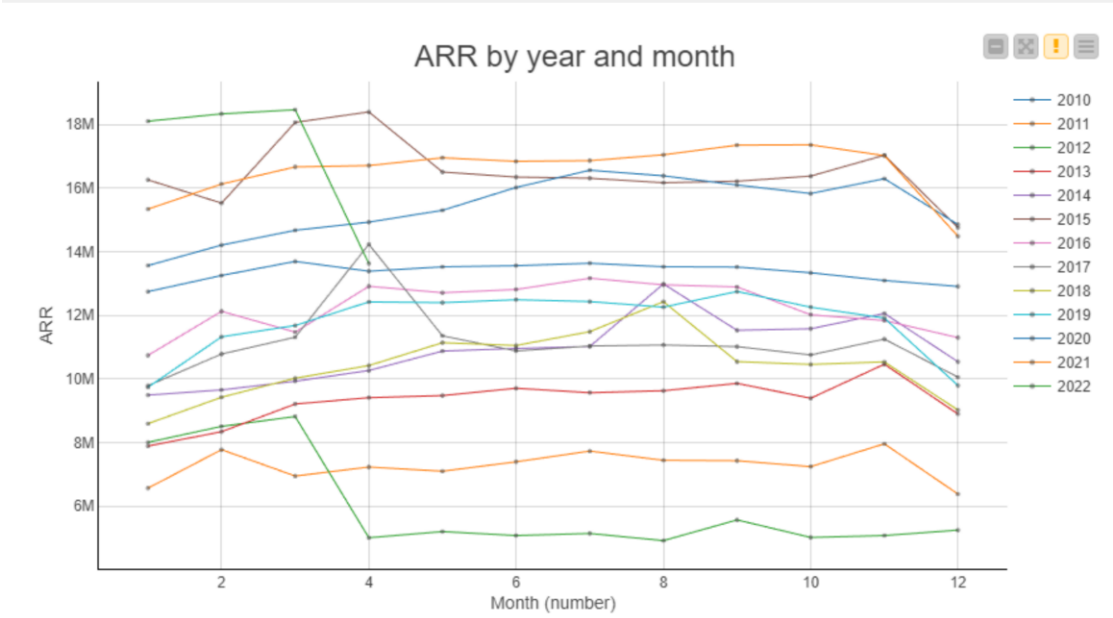


Figura 24 - ARR por ano e por mês

3.2.3. Análise das caraterísticas e da correlação dos dados

A seleção dos atributos é fundamental para a melhoria dos modelos de machine larning e uma das formas de reduzir a dimensão dos dados. Para interpretar grandes conjuntos de dados, são necessários métodos para reduzir drasticamente a sua dimensionalidade de forma interpretável, de modo que a maior parte da informação nos dados seja preservada.

Muitas técnicas foram desenvolvidas para este fim, mas a análise principal de componentes (PCA) é uma das mais antigas e mais utilizadas, figura 25. A sua ideia é simples — reduzir a dimensionalidade de um conjunto de dados, preservando ao mesmo tempo o máximo de "variabilidade" (ou seja, informação estatística) possível.

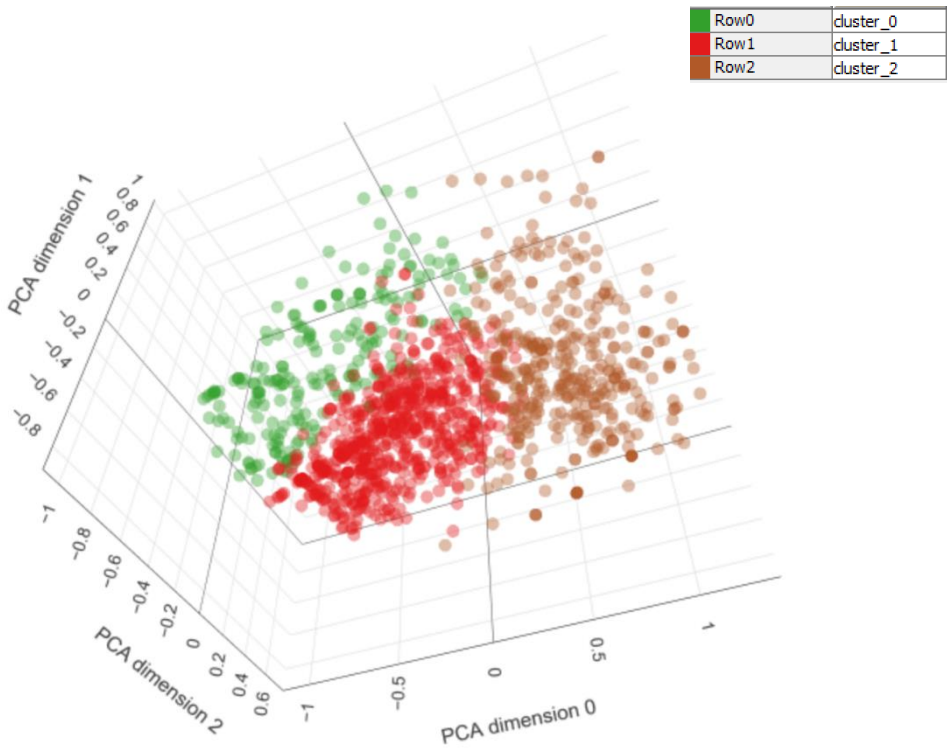


Figura 25 - Exemplo de aplicação do PCA em três dimensões

Outro método e uma das formas mais fáceis de visualizar a correlação de dados é através dos *heatmaps* (mapas de calor), figura 26. Estes tipos de gráficos permitem ver como as características estão relacionadas entre si, através de uma matriz onde linhas e colunas são definidas por cada característica e cada célula tem a quantidade de correlação (valores entre -1 e 1) entre as características.

As características mais relevantes e a identificação do valor de saída, são uma boa prática filtrar apenas as caraterísticas que se correlacionam com um valor absoluto superior ao valor de saída, uma vez que ajuda o modelo a dar respostas mais precisas e não incluir "lixo" no modelo.

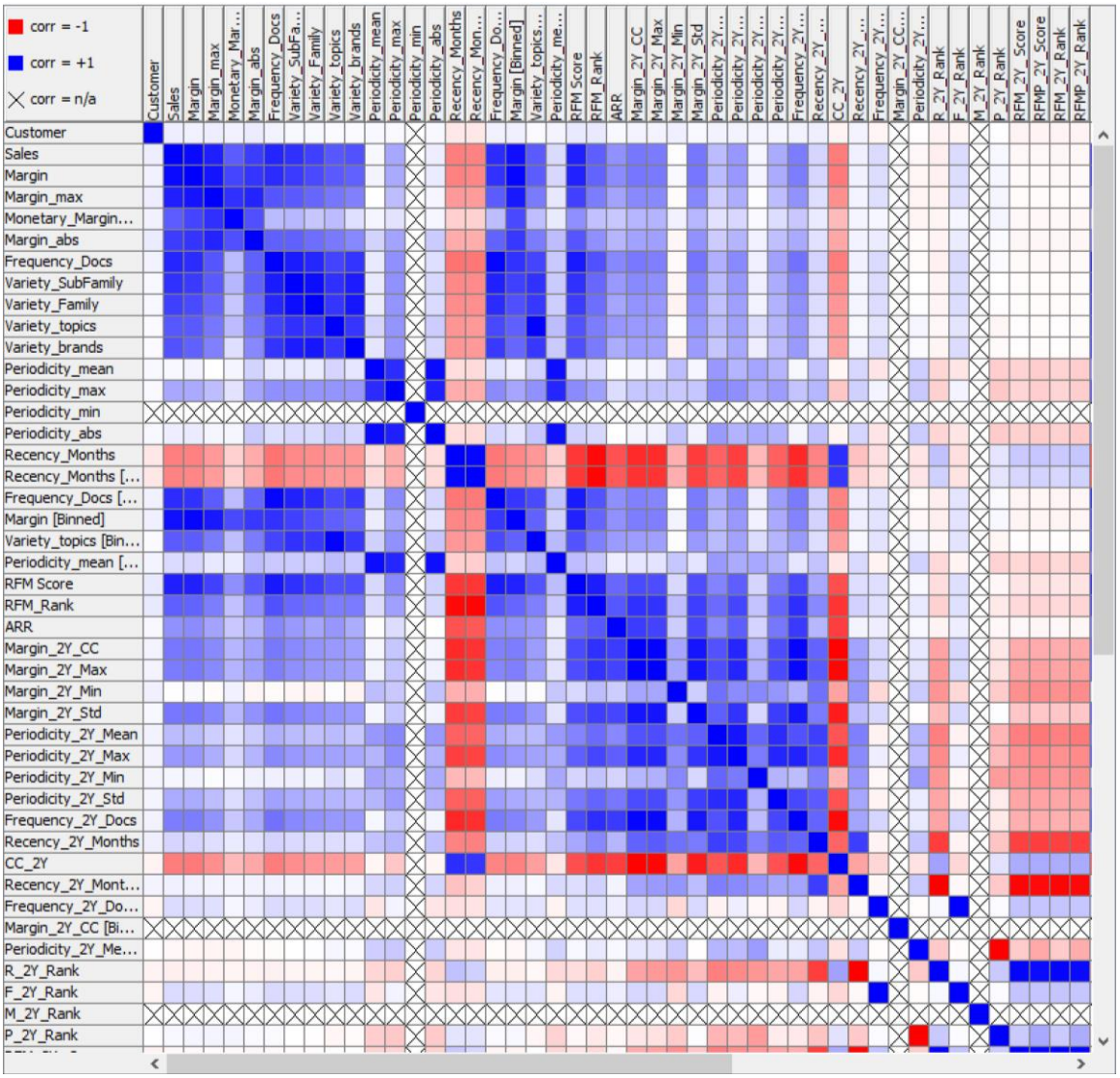


Figura 26 - Matriz de correlação

De acordo com os estudos citados, vamos filtrar as carateristicas com correlação superior a 0.6 e após calcular, excluiremos 51 colunas que não trariam valor para o modelo em estudo, figura 27.

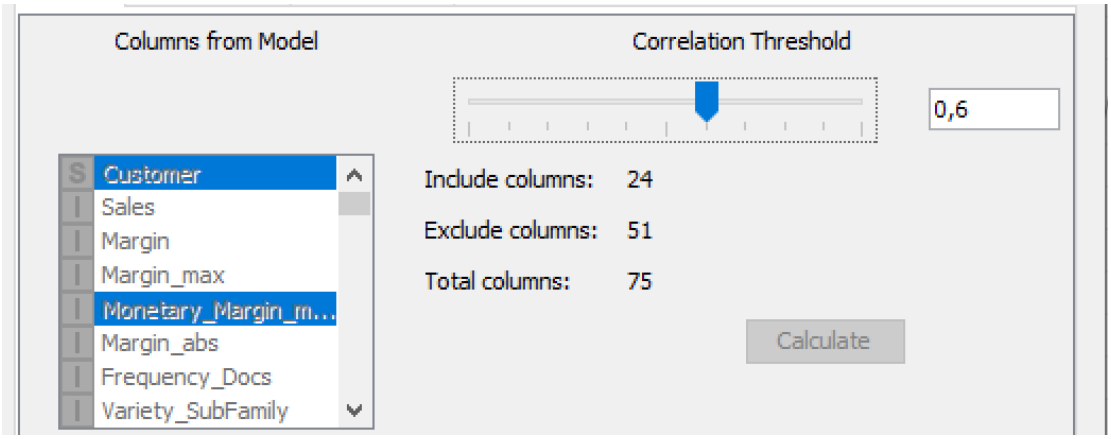


Figura 27 - Definição dos parâmetros para filtrar a correlação

Após esta operação, o mapa de calor apresenta menos carateristicas. Neste caso as características que tinham uma correlação absoluta superior a 0,6, conforme a figura 28.

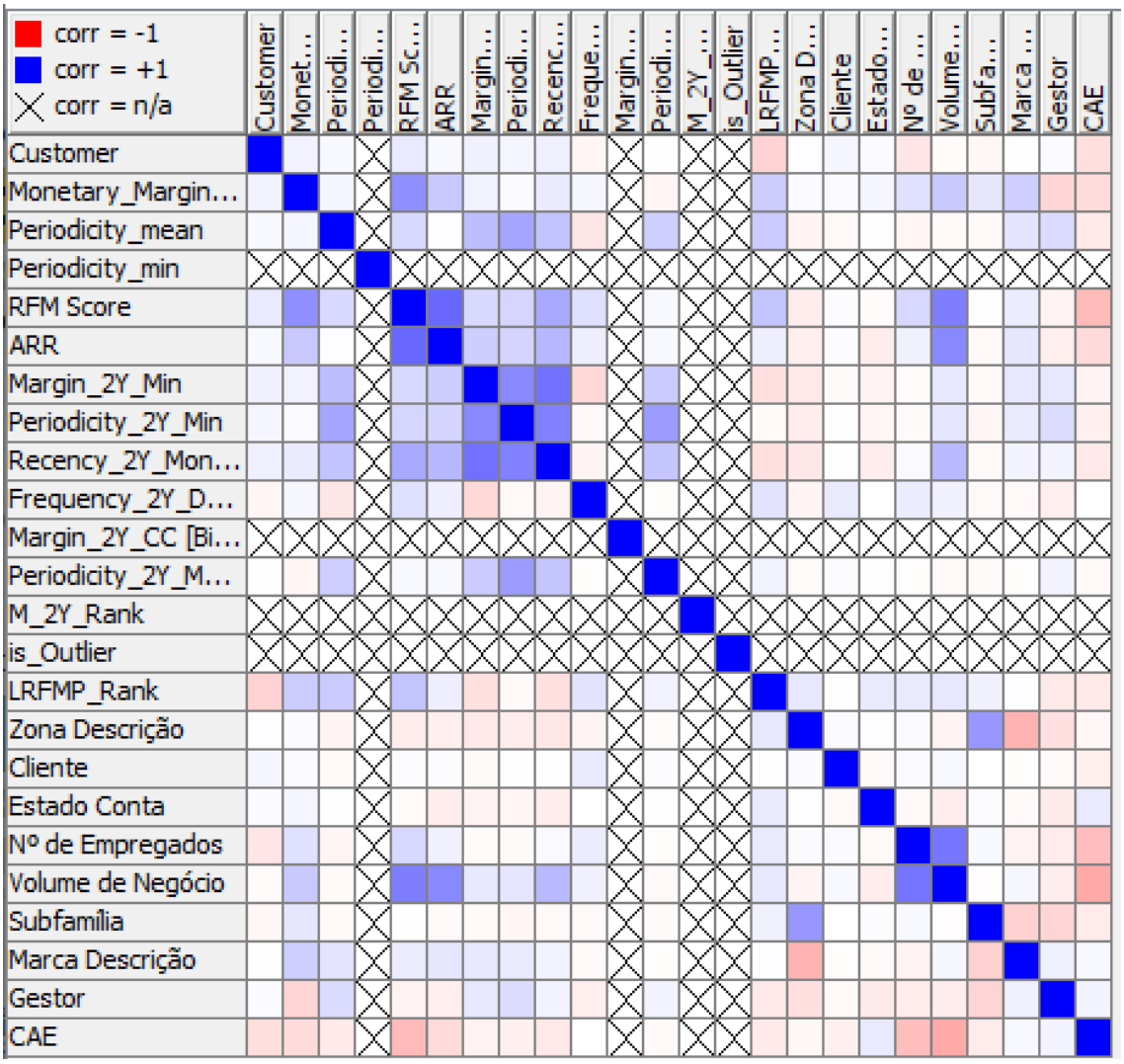


Figura 28 - Matriz de correlação, após filtros

CAPÍTULO IV

4. DESENVOLVIMENTO E AVALIAÇÃO DO MODELO

O estudo propõe a categorização dos clientes aplicando duas fases de segmentação.

Numa primeira fase, a segmentação com base no valor LRFM, agrupando os clientes similares. O algoritmo de *clustering k-means* será aplicado para obtermos esta segmentação.

Na segunda fase deste estudo concentra nos clientes, considerados pela empresa, com mais potencial e que são os clientes classificados como Corporate e Business e também os clientes de categoria 'A' que resultam da classificação tradicional ABCd, e passa a considerar os dados firmográficos destes clientes, nomeadamente o setor de atividade, o volume de negócios e o numero de colaboradores.

Para compreendermos melhor o comportamente de compra destes clientes, proponho um novo modelo, que considera a Variedade de produtos e serviços e também a Periodicidade das transações e designa-se por modelo LRFMVP.

4.1. *K-means*, segmentação em *clusters*

Os clientes são agrupados em 4 segmentos que identificam o seu comportamento de compra, com base nos respetivos valores LRFM.

O parâmetro K foi definido para 4, uma vez que quatro (2x2) possíveis combinações de entradas (LRFM) podem ser obtidas atribuindo ↑ ou ↓, de acordo com os valores médios de L, R, F, M de um *cluster* sendo inferior ou superior à média global. Se o valor médio R (F, M) de um *cluster* exceder a média global R (F, M), então foi incluída uma seta ascendente ↑, caso contrário e seta descendente ↓. Por exemplo, L↑R↑F↓M↓ representa que o valor médio de Recência de um segmento de cliente é superior à média global, enquanto os valores da frequência e da média monetária são inferiores às médias globais.

Na figura 29 apresenta o resultado, enumerando quatro *clusters*, cada um com o número correspondente de clientes, os seus valores médios de L, R, F e M.

Row ID	S Cluster	I Customers	D LRFM Mean	D LRFM Std Desviation	D LRFM Minimum	D LRFM Maximum	D L Mean	D R Mean	D F Mean	D M Mean
Row0	cluster_0	2205	17.34	1.8	14	20	4.43	3.88	4.55	4.49
Row1	cluster_1	1526	13.49	1.68	10	17	1.75	4.22	3.82	3.69
Row2	cluster_2	2153	8.3	1.52	6	13	1.98	2.99	1.76	1.57
Row3	cluster_3	2386	10.78	1.93	7	14	4.55	1.43	2.33	2.47

Figura 29 - Clusters obtidos, para o modelo LRFM

Os grupos de clientes gerados neste processo permitem identificar os clientes mais valiosos e avaliar o valor futuro dos clientes através do modelo CLV (*customer lifetime value*).

4.2. Customer lifetime value

O modelo CLV é processado através da multiplicação dos pesos por cada valor do LRFM, usando um processamento de análise hierárquica.

Neste estudo, de acordo com a investigação de Reinartz & Kumar, (2000), para os parâmetros WL, WR, WF, WM, vamos usar respetivamente os valores 0.238, 0.088, 0.326 e 0.348, para simplificação do processo e porque a tipologia de clientes é a mesma destes estudos.

A forma de cálculo do CLV é $C^j = W_L\, C_L^j + W_R\, C_R^j + W_F\, C_F^j + W_M\, C_M^j$

Em que: C^j = Customer CLV rating

$C_L^j\ C_R^j\ C_F^j\ C_M^j$ = valores médios de L, R, F, M do *cluster* j

$W_L\ W_R\ W_F\ W_M$ = aos pesos atribuídos a L, R, F, M

Após processamento do CLV podemos classificar os clientes num portefólio de categorias (CPA – *customer portefólio analysis*) que inclui os clientes *Platinium*, que são os mais valiosos e compradores, os clientes *Gold*, clientes de primeira viagem e gastadores, os *Silver* são clientes de agitação e menos frequentes e os clientes incertos e menos valiosos, são classificados como *Bronze*.

4.3. Customer portfolio analysis

Figura 30, apresenta o dados do CLV processados para cada *cluster*. De acordo com Rajkumar Venkatesan & v. Kumar, (2004), permite identificar os clientes mais valiosos e mais leais, associados ao portefólio de gestão dos clientes.

Row ID	S Cluster	I Custom...	D L Mean	D R Mean	D F Mean	D M Mean	D CLV	S CPM
Row0	cluster_0	2205	4.425	3.875	4.555	4.488	4.441	Platinium
Row1	cluster_1	1526	1.75	4.225	3.82	3.691	3.318	Gold
Row3	cluster_3	2386	4.549	1.428	2.329	2.473	2.828	Silver
Row2	cluster_2	2153	1.982	2.99	1.758	1.571	1.855	Bronze

Figura 30 - Atribuição do portefólio a cada cluster.

4.4. Novo modelo LRFMVP

Na segunda fase deste estudo detalha apenas os clientes, considerados pela empresa, com mais potencial e que são os Corporate e Business e também os clientes 'A' que resultam da classificação tradicional ABCD, e compara-os com os dados firmográficos.

Para compreendermos melhor o comportamente de compra destes clientes, proponho um novo modelo, que considera a Variedade de produtos e serviços e também a Periodicidade das transações.

Estas dimensões já foram estudadas de forma separada, como resposta à escassêz de estudos empiricos no sempremento B2B, e mais especificamente em empresas tecnológicas que vêm os seus modelos de negócio tradicionais a serem substituidos por modelos de pagamento por uso, através de subscrições mensais e anuais.

Este modelo assenta na recorrência, é mais dinámixo e flexivel, o que obriga as empresas ao acompanhamento permanente dos clientes, na tentativa de aumentar as subscrições e evitar que os clientes mudam para outros fornecedores.

O modelo proposto é LRFMVP e que pode ser analisado pelo seguinte exemplo: L↑R↑F↑M↑V↓P↓ representa que é um cliente antigo (L), o valor médio de Recência de um segmento de cliente é superior à média global, os valores da Frequência e da média Monetária são superiores às médias globais, enquanto os valores médios da Variedade e da Periodicidade são inferiores á média.

Outro exemplo, L↓R↑F↑M↓V↓P↑ representa um cliente novo, com alta Periodicidade, em que o valor médio de Recência e frequência de um segmento de cliente é superior à média global, enquanto os valores da média monetária e da variedade são inferiores às médias globais.

O algoritmo de *clustering* k-means será aplicado para obtermos esta segmentação.

Os clientes são agrupados em k segmentos que identificam o seu comportamento de compra, com base nos respetivos valores LRFMVP e nos dados firmográficos do cliente, nomeadamente o setor de atividade, o volume de negócios, o numero de funcionários e a região onde se localiza a empresa.

Para identificação do parâmetro k aplicamos o método do Cotovelo (*Elbow Method*) e a medida de qualidade MAE.

O erro absoluto médio (MAE - *Mean Absolute Error*), foi a medida de qualidade utilizada para medir a magnitude média dos erros num conjunto de previsões, sem considerar a sua direção.

Fórmula MAE:

$$\frac{1}{n}\sum_{j=1}^{n} \lfloor y_j - \hat{y}_j \rfloor$$

Na figura 31 visualizamos o resultado obtido e a visualização, pelo método cotovelo, da maior descida até k=2.

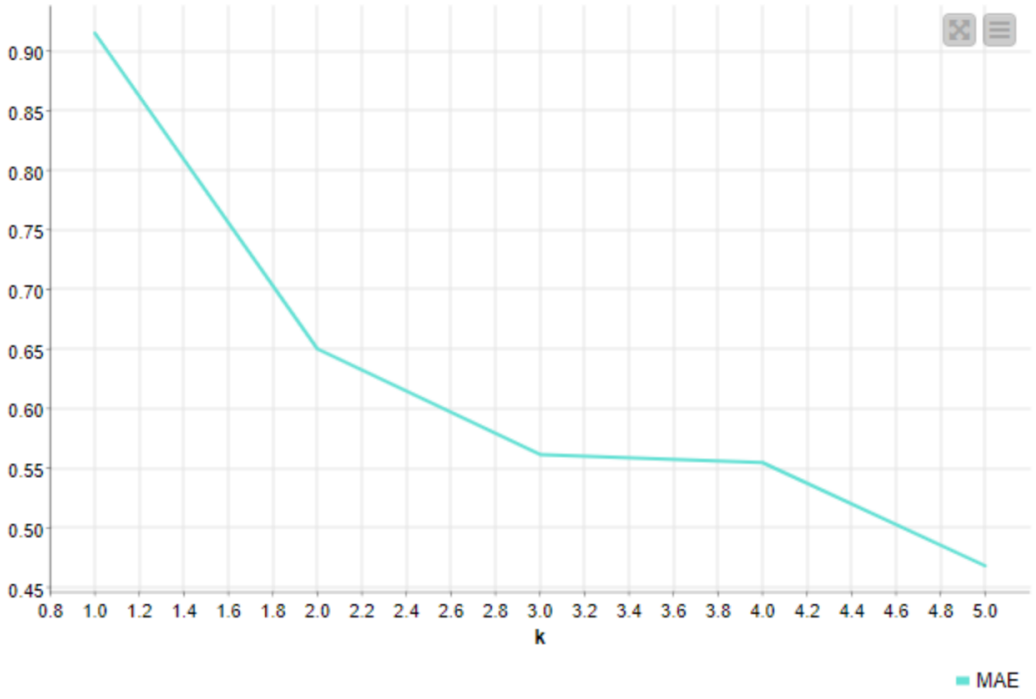


Figura 31 - Visualização do resultado pelo método de cotovelo

A figura 32 apresenta o resultado, enumerando os 3 *clusters*, cada um com o número correspondente de clientes, os seus valores médios de L, R, F, M, V, P e também a mediana do Volume de Negócios, sendo um dado firmográfico base para seleção dos clientes.

Row ID	Cluster	Count	LRFMVP Mean	L Mean	R Mean	F Mean	M Mean	V Mean	P Mean	Sales	VN Median
Row0	cluster_0	219	26	5	5	5	5	3	3		7,387,000
Row1	cluster_1	708	23	3	5	5	5	3	3		5,337,395
Row2	cluster_2	376	16	2	4	3	3	1	3		11,913,500

Figura 32 - Custer's obtidos com a aplicação do método de cotovelo

4.5. Métodos preditivos e técnicas de *tunning*

Terminado o processo de segmentação dos clientes, passamos a um importante passo que é a segmentação e catalogação de clientes em prospeção, com base na informação da carteira. Neste processo é evidenciada toda a capacidade e utilidade da inteligência artificial aplicada aos negócios e ao marketing, numa tarefa complexa e que só pode ser realizada de forma massiva e em tempo real, com recurso a técnicas e ferramentas de inteligentes. Sendo um dos objetivos primordiais deste estudo, a previsão de oportunidades futuras, vamos aplicar modelos preditivos a uma base de dados de prospeção todo o conhecimento adquirido na segmentação e classificação dos clientes atuais. Para garantir maior nível de fiabilidade na predição do potencial destas contas de prospeção, foram aplicadas técnicas de *tunning* para avaliar os parâmetros que garantem os melhores resultados e para isso é importante conseguir altos valores *accuracy*, se possível acima dos 90%, conforme figura 33.

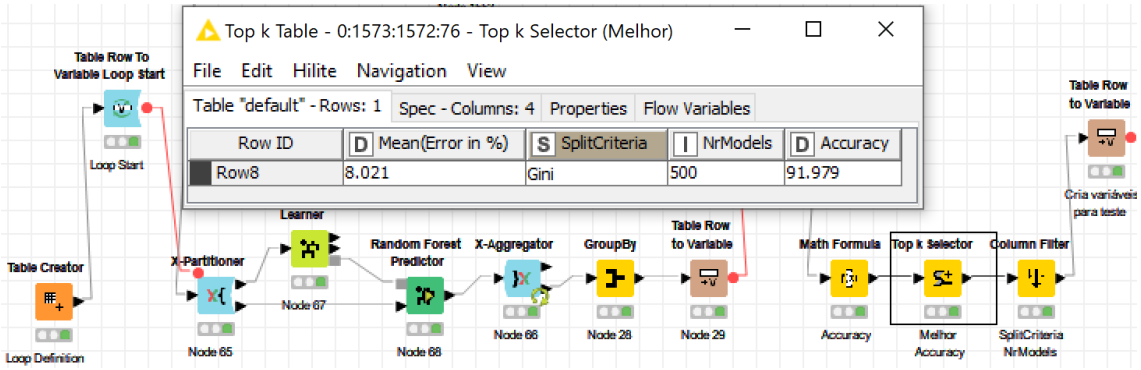


Figura 33 - Tunning para avaliação dos melhores parâmetros

Para além da caraterística preditiva do modelo é igualmente importante que identifique 'Clientes' na base de dados de prospeção e que os catalogue de acordo com a informação do sistema.

Na figura 34 mostra o resultado da aplicação do modelo preditivo à base de dados de prospeção. Os valores apresentados, estão normalizados, por uma questão de confidencialidade e menor aproximação á realidade.

Row ID	Customer	Category	Vendas e Serviços Prestados (€)	Empregados	Idade	CAE ...	Prediction (CPM)	Prediction (CPM) ...	Prediction (ARR)	Prediction (ARR) ...
Row0	Cliente	Corporate	1,886,063,068.36	515	28	46762	Silver	1	0	0
Row1	Prospect	Corporate	1,453,503,910	236	102	46711	Platinium	0.004	0.002	0.001
Row2	Prospect	Corporate	1,084,009,099.06	105	93	46711	Platinium	0.003	0.001	0
Row3	Prospect	Corporate	803,103,870.3	66	44	46214	Platinium	0.002	0.053	0.093
Row4	Cliente	Corporate	778,365,841.97	2051	91	46390	Gold	1	0	0
Row5	Prospect	Corporate	686,339,410	114	58	46711	Platinium	0.003	0	0
Row6	Cliente	Corporate	613,752,734.64	322	23	46350	Platinium	1	0	0
Row7	Prospect	Corporate	601,290,327	407	30	46460	Platinium	0.002	0.042	0.053
Row8	Cliente	Corporate	592,874,244.92	687	15	46711	Bronze	1	0	0
Row9	Prospect	Corporate	543,186,498.48	291	62	46460	Platinium	0.002	0.049	0.058
Row10	Prospect	Corporate	435,952,677.25	146	30	46510	Platinium	0.004	0.005	0.001
Row11	Prospect	Corporate	372,215,438.47	476	14	46350	Platinium	0.005	0.002	0
Row12	Prospect	Corporate	367,351,529.17	79	39	46900	Platinium	0.001	0.026	0.021
Row13	Prospect	Corporate	351,000,000	1008	33	46390	Platinium	0.004	0.004	0.004
Row14	Prospect	Corporate	319,415,485.72	197	31	46510	Platinium	0.004	0.008	0.003
Row15	Prospect	Business	317,849,111.54	44	46	46460	Platinium	0.002	0.146	0.24
Row16	Cliente	Corporate	276,695,508	1974	117	46690	Platinium	1	0	0
Row17	Prospect	Corporate	256,379,744.11	320	24	46332	Platinium	0.002	0.049	0.102

Figura 34 - Resultado da predição em novas contas

4.6. Métodos prescritivos e de recomendação

Neste momento já responderíamos aos objetivos principais deste estudo, contudo vamos elevar a complexidade e apresentar um modelo prescritivo que apoie os responsáveis pelo marketing e comercial na identificação de oportunidades e recomendação iniciativas de cross-selling.

Para a empresa em estudo, identificou 2 aspetos que mais importantes.

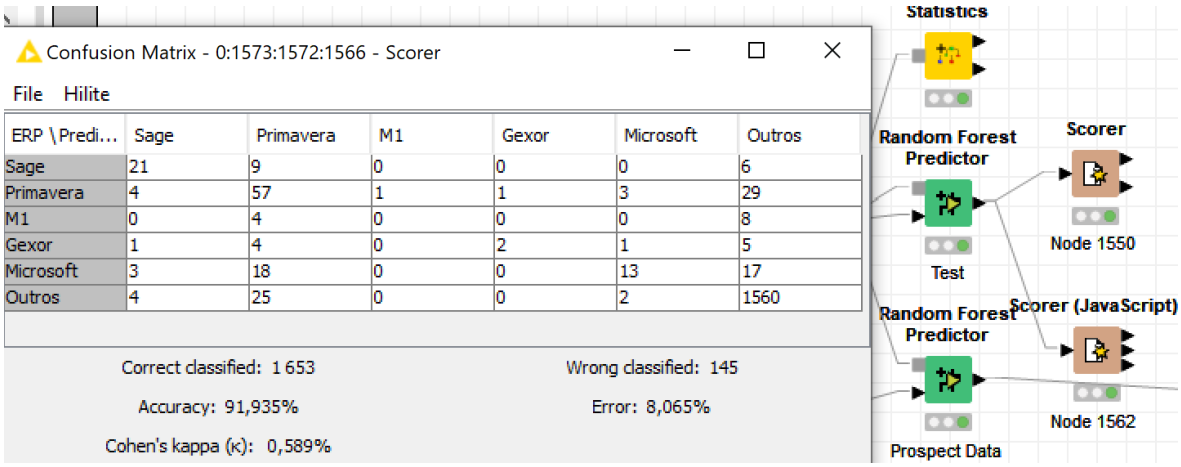
Primeiro, recomendar o sistema de gestão (ERP) mais adequado a estes potenciais clientes, com base no comportamento de compra dos atuais clientes e a experiência da empresa ao longo dos 35 anos, nos vários setores de atividade e na dimensão das empresas. A informação acerca do ERP mais adequada ao potencial, condiciona toda a oferta sequente e também as equipas que devem endereçar a futura oportunidade de negócio.

Segundo, propor iniciativas de *cross selling*, através das relações de compra dos clientes atuais.

Para responder usamos métodos diferentes. No primeiro aspeto utilizaremos um método supervisionado com grande capacidade de processamento, o *random forest predictor,* que nos vai prescrever o ERP que devemos propor prioritariamente.

Para melhorar o resultado e garantir melhor *accuracy*, criamos uma abordagem de treino do algoritmo *random forest* e foi feita uma *cross-validation* de 10 validações, onde, no fim, é calculada a média dos erros de todas as validações. Foi também realizado um ciclo para otimizar o *random forest* que consiste em treinar o modelo com todas as combinações dos seguintes parâmetros: *Split Criterion* a variar entre *Information Gain*, *Information Gain Ratio* e *Gini Index*, e *Number of models* com valores de 100, 200 e 500. O modelo foi treinado com os dados dos clientes previamente segmentados e classificados, similares aos potenciais clientes da base de dados de prospeção. De seguida

O modelo apresenta 91,94% de *accuracy*, para um erro inferior a 8,07% e aplicamos este modelo aos dados de prospeção para prever qual o sistema de gestão mais indicado para cada potencial cliente, conforme figura 35.



ERP \ Predi...	Sage	Primavera	M1	Gexor	Microsoft	Outros
Sage	21	9	0	0	0	6
Primavera	4	57	1	1	3	29
M1	0	4	0	0	0	8
Gexor	1	4	0	2	1	5
Microsoft	3	18	0	0	13	17
Outros	4	25	0	0	2	1560

Figura 35 - Confusion Matrix do modelo

Na figura 36, mostra um conjunto de registos da base de potenciais clientes em prospeção, com a sugestão do ERP.

Row ID	Customer	Category	Vendas e Serviços Prestados (€)	Empregados	Idade	CAE Pri...	Prediction (ERP)
Row398	Prospect	Corporate	17,455,766.13	106	30	46620	Primavera
Row476	Prospect	Corporate	14,907,784.75	90	30	46390	Microsoft
Row484	Prospect	Business	14,538,743.79	29	34	46460	Gexor
Row502	Prospect	Business	14,091,379.37	44	10	46390	Primavera
Row535	Prospect	Business	13,193,394.2	28	17	46630	Outros
Row546	Prospect	Business	12,930,740.54	33	23	46731	Outros
Row588	Prospect	Corporate	12,072,910.05	149	71	46491	Sage
Row616	Prospect	Corporate	11,277,615.47	70	35	46690	Primavera
Row628	Prospect	Business	11,028,243.57	42	56	46740	Primavera
Row629	Prospect	Corporate	11,008,922.43	85	54	46690	Primavera
Row703	Prospect	Business	10,068,878.62	37	22	46311	Primavera
Row23	Prospect	Corporate	196,542,629.84	158	24	46460	Outros
Row58	Prospect	Business	88,150,444.45	26	16	46771	Outros
Row85	Prospect	Corporate	71,855,012	104	57	46331	Outros
Row101	Prospect	Business	63,434,475.67	15	47	46731	Outros
Row174	Prospect	Corporate	39,722,966.43	138	33	46740	Outros
Row179	Prospect	Corporate	38,377,256.91	75	52	46460	Primavera
Row245	Prospect	Corporate	26,838,028.04	74	53	46430	Outros
Row286	Prospect	Corporate	23,972,735.54	71	31	46311	Outros
Row317	Prospect	Business	21,415,780.28	12	13	46732	Outros
Row322	Prospect	Business	21,101,683.86	38	20	46382	Outros
Row409	Prospect	Business	17,028,032.44	20	7	46732	Primavera
Row446	Prospect	Business	15,597,564.69	35	52	46630	Primavera
Row456	Prospect	Business	15,380,592.52	44	23	46750	Outros
Row506	Prospect	Business	13,938,560	41	14	46510	Primavera
Row524	Prospect	Business	13,513,277.58	31	30	46494	Primavera

Figura 36 - Resultados de predição em clientes potenciais

Até agora assumimos que as métricas simples, como a taxa de precisão (*accuracy*), são usadas para medir o desempenho do modelo. O *accuracy* é uma métrica popular porque é muito fácil de medir, mas pode ser demasiado simplista para a aplicação de técnicas de *data mining* em determinados problemas reais de negócio.

$$accuracy = \frac{\text{número de classificações corretas}}{\text{número total de classificações}}$$

Equação 4 - Cálculo da taxa de precisão do classificador (accuracy)

O *accuracy* é igual a (1- a taxa de erro). É uma métrica de avaliação comum muito usada em estudos de *data mining* porque reduz o desempenho do classificador para um único número, e é muito fácil de medir. Infelizmente, é simplista e tem alguns problemas bem conhecidos (Reitor, 1998). Para

compreender estes problemas, precisamos de uma forma de os decompor e contarmos os diferentes tipos de decisão correta e incorreta tomadas pelo classificador. Para isso usamos a *confusion matrix*.

Na *confusion matrix*, figura 37, a diagonal contém as contagens de decisões corretas. Os erros do classificador são os falsos positivos (casos negativos classificados como positivos) e falsos negativos (positivos classificados como negativos).

Confusion Matrix

	Gexor (Pre...	M1 (Predict...	Microsoft (...	Outros (Pre...	Primavera (...	Sage (Predi...	
Gexor (Act...	2	0	1	5	4	1	15.38%
M1 (Actual)	0	0	0	8	4	0	0.00%
Microsoft (...	0	0	13	17	18	3	25.49%
Outros (Act...	0	0	2	1560	25	4	98.05%
Primavera (...	1	1	3	29	57	4	60.00%
Sage (Actual)	0	0	0	6	9	21	58.33%
	66.67%	0.00%	68.42%	96.00%	48.72%	63.64%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
91.94%	8.06%	0.589	1653	145

Figura 37 - Confusion Matrix do modelo na previsão do ERP.

A matriz de confusão separa as decisões tomadas pelo classificador e discrimina como um *ERP* está a ser confundido com outro. Desta forma, os diferentes tipos de erros podem ser tratados separadamente. Podemos diferenciar entre os valores verdadeiros para cada *ERP* e o previsto pelo modelo.

Como exemplo para o *ERP* Primavera, o modelo classificou corretamente 57 vezes e 38 vezes confundiu. Classificou 60% das vezes corretamente. Já para o *ERP* 'Outros', classificou 98% da vezes corretamente. Não sendo conclusivo o motivo da performance do modelo, seria recomendável testar com outros algoritmos de classificação. O facto de ter maior *cross-validation*, uma *accuracy* de 91,94%, mas um *score* menor, significa que as alterações que foram feitas só fizeram *overfitting* ao modelo de teste. Ou seja, o modelo adaptou-se bem ao modelo de testes e apresentava melhorias significativas, mas a capacidade de generalizar novos dados é baixa e não reflete as melhorias. A *confusion matrix* é uma alternativa para avaliação da performance de decisão do modelo, que permite obter maior detalhe e informação útil para clarificar os decisores sobre as decisões de recomendação do modelo.

No segundo problema usamos um método que utilizada técnicas de *data mining* para descobrir regras de associação e conseguir assim promover produtos diferentes em clientes, através de estratégias de *cross selling*. Neste estudo vamos utilizar o agrupador de produtos, que é a subfamília e que se enquadra melhor no objetivo do estudo que é identificar as regras de associação nas 'gamas de produtos', figura 38.

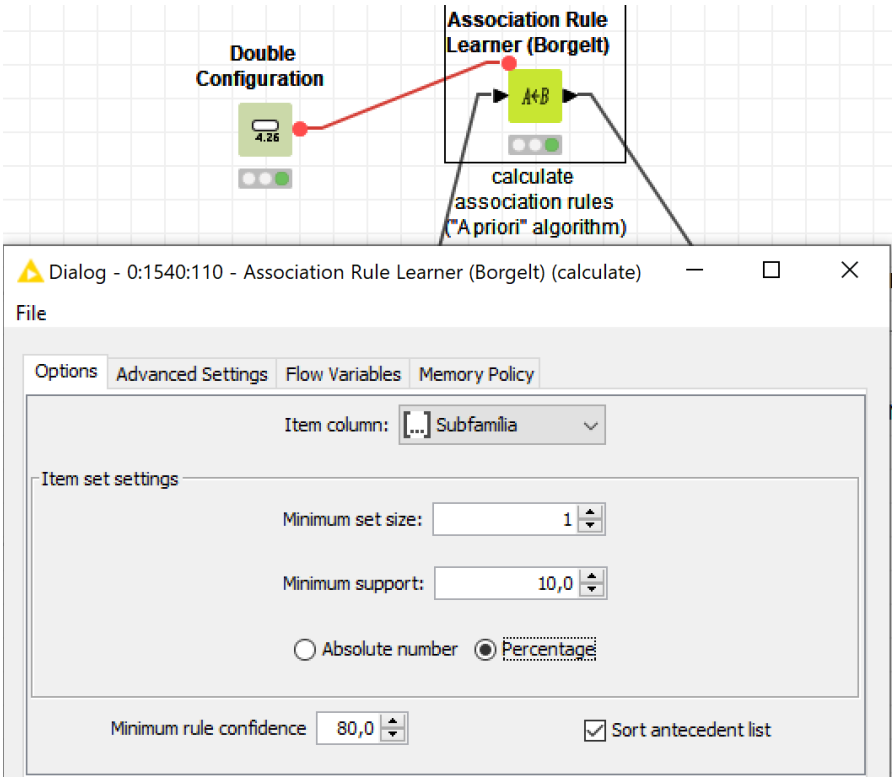


Figura 38 - Parâmetros de suporte e confiança

O resultado deste método depende das medidas de interesse selecionados. O suporte é de 10% e a confiança de 80%. Significa que em 10% das transações existe uma forte relação entre as regras de associação encontradas e que 80% dos clientes que compraram o produto da regra '*antecedent*', também compraram o produto da regra '*consequent*'.

Na figura 39, apresentamos uma lista dos clientes que compraram [Microsoft 365, Antivírus e CSA], e provavelmente também compraram [210002 - CSM]. Com 90.9% de confiança, o que significa que em 90.0% das vezes, esta condição é verificada. A relação entre antecedente e consequente permite considerar a oferta de outros produtos, que podem ser novos, substitutos ou menos mais inovadores, para aumentar o número de soluções e serviços que vendemos aos clientes.

Row ID	Conseq...	Antecedent	ItemS...	RuleLift%	RuleLift	RuleConfiden...
Row11517	21008 - CSM,...	[11621 - Elementos de Cobre,12103 - Microsoft 365,13191 - Renovação CSA - faturação directa,...]	29	282.42	2.824	96.7
Row10712	21002 - CSM	[11481 - Proteção de Perímetro,12103 - Microsoft 365,13191 - Renovação CSA - faturação directa,...]	30	206.8	2.068	90.9
Row11445	21002 - CSM	[12103 - Microsoft 365,12901 - Anti-vírus Perpétuo (1 ou 3 anos),13191 - Renovação CSA - faturação directa,...]	30	206.8	2.068	90.9
Row10702	21008 - CSM,...	[11481 - Proteção de Perímetro,11621 - Elementos de Cobre,12103 - Microsoft 365,...]	29	264.77	2.648	90.6
Row11514	21002 - CSM	[11621 - Elementos de Cobre,12103 - Microsoft 365,13191 - Renovação CSA - faturação directa,...]	29	206.15	2.062	90.6
Row10623	21008 - CSM,...	[11293 - Acessórios - discos,11481 - Proteção de Perímetro,12103 - Microsoft 365,...]	28	263.88	2.639	90.3
Row11054	21002 - CSM	[12103 - Microsoft 365,12903 - Segurança Perpétuo (1 a 3 anos),13191 - Renovação CSA - faturação directa,...]	28	205.47	2.055	90.3
Row11440	21002 - CSM	[12103 - Microsoft 365,12901 - Anti-vírus Perpétuo (1 ou 3 anos),13191 - Renovação CSA - faturação directa,...]	28	205.47	2.055	90.3

Figura 39 - Resultado da aplicação do algoritmo Apriori

Neste estudo consideramos também a descoberta de regras de associação entre fabricantes e marcas. Esta análise é muito importante, uma vez que permite descobrir relações e comportamentos de compra dos clientes por cada fabricante ou marca e dota as equipas de marketing com conhecimento para gerarem futuras propostas de negócio personalizadas com ofertas cruzadas e complementares.

Na figura 40, identificamos que se o cliente comprou HP, é provável que compre também Microsoft, Primavera e *StorageCraft*. Com 93.9% de confiança, o que significa que em 93.9% das vezes, esta condição é verificada.

Row ID	Consequent	Antecedent	ItemSetSupport	RuleLift%	RuleConfidence%	RelativeBodySetSupport%
Row1115	HP	[Microsoft,Primavera,StorageCraft,...]	31	237.24	93.9	11.1
Row1117	HP	[Microsoft,Primavera,StorageCraft,...]	31	237.24	93.9	11.1
Row1123	HP	[Primavera,StorageCraft,Watchguard,...]	31	237.24	93.9	11.1
Row1124	HP	[Primavera,StorageCraft,Watchguard]	31	237.24	93.9	11.1
Row1178	HP	[Microsoft,Primavera,Watchguard,...]	40	210.45	83.3	16.1
Row1180	HP	[Microsoft,Primavera,Watchguard]	40	210.45	83.3	16.1

Figura 40 - Regras de associação descobertas para fabricantes

CAPÍTULO V

5. DISCUSSÃO DOS RESULTADOS

Os resultados apresentados neste capítulo utilizam os melhores parâmetros possíveis nos respetivos modelos, bem como a fiabilidade de cada modelo, que foram mencionados anteriormente.

Os resultados em análise foram implementados na mesma plataforma utilizada para o desenvolvimento do modelo, o Knime.

5.1. Modelo LRFM

Para criar uma relação entre os clientes, foi utilizado o algoritmo não supervisionado, *k-means*, para analisar ligações entre clientes e inferir os seus padrões e tendências. Vale referir que este algoritmo é o mais usado em modelos para segmentação de clientes. O K significa que coloca um número arbitrário de centroides e que neste caso é escolhido (quatro). Estes foram os pontos de dados distribuídos utilizando quatro *clusters* em duas dimensões, na figura 41.

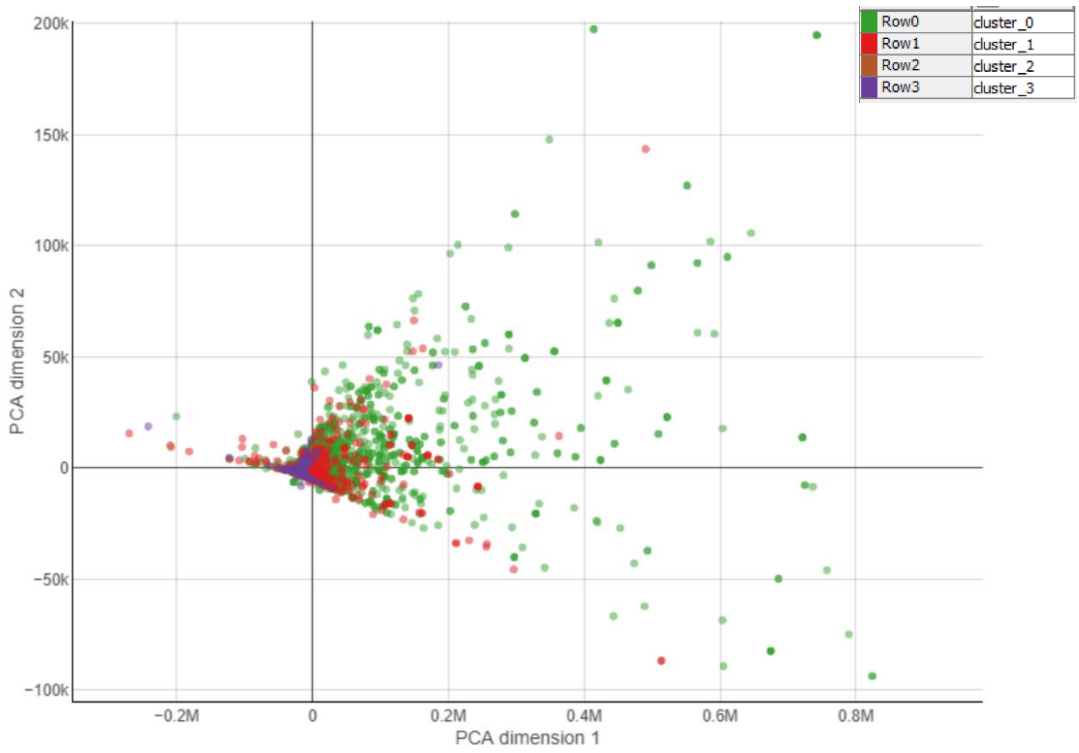


Figura 41 . Scatter Plot do modelo LRFM

Para entender melhor o agrupamento, é essencial olhar para eles e recolher a informação que é exclusiva de cada *cluster*, conforme detalhado nos pontos anteriores.

Vamos analisar os clientes de acordo com o comportamento de compra traduzido pelo modelo LRFM e quanto à sua lealdade e ao valor.

Cluster	Número de Clientes	Padrão LRFM	Tipo de Cliente
Cluster_0	2205	L↑ R↑ F↑ M↑	Platinium
Cluster_1	1526	L↓ R↑ F↑ M↑	Gold
Cluster_2	2153	L↓ R↓ F↓ M↓	Silver
Cluster_3	2386	L↑ R↓ F↓ M↓	Bronze

Tabela 4 - Padrão do modelo LRFM e total de clientes por cluster

As duas últimas colunas da tabela 4 mostram o padrão LRFM para cada *cluster* e o respetivo tipo de cliente. Enquanto o cluster_3 contém o número máximo de clientes (2386 clientes), o cluster_1 inclui o mínimo, 1526 clientes.

O segmento de cliente cluster_0 contém os clientes mais valiosos, porque é composto por clientes antigos, que recentemente fizeram compras regulares, e também têm uma média mais alta de frequência de compra e valor de compra.

É seguido pelo cluster_1, que pode incluir clientes novos, que compraram recentemente, sai frequentes e têm uma média do valor de compra alta.

O Cluster_2 (L↓M↓) pode incluir clientes pela primeira vez, que visitaram recentemente a empresa, com maior Recência e menor frequência de compra e menor valor monetário. São os clientes menos valiosos, eventualmente perdidos (churn). Algo pode ter corrido mal com estes clientes e, portanto, parece ser um indicador de probabilidade de agitação

Os clientes do cluster_3, são em maior número, fizeram um significa número de compras com elevados valores monetários, são clientes antigos, e deixaram de comprar há bastante tempo. É necessário contactar com estes clientes, ou seja, enviar um e-mail e planear um programa de reativação do cliente, ou seja, sugestão de promoção.

Para maior facilidade na leitura, na figura 42, os registos foram 'coloridos' de acordo com a classificação do cliente, que corresponde aos *clusters* obtidos.

Row ID	LRFM_Rank	LRFM Score	L_rank	R_Rank	F_Rank	M_Rank	Cluster	CPM
Row3371	4244	14	4	2	4	4	cluster_0	Platinium
Row3372	5144	14	5	1	4	4	cluster_3	Silver
Row3373	5144	14	5	1	4	4	cluster_3	Silver
Row3374	4244	14	4	2	4	4	cluster_0	Platinium
Row3375	4343	14	4	3	4	3	cluster_0	Platinium
Row3377	5144	14	5	1	4	4	cluster_3	Silver
Row3379	5144	14	5	1	4	4	cluster_3	Silver
Row3380	3443	14	3	4	4	3	cluster_1	Gold
Row3381	4343	14	4	3	4	3	cluster_0	Platinium
Row3382	5144	14	5	1	4	4	cluster_3	Silver
Row3383	2444	14	2	4	4	4	cluster_1	Gold
Row3385	4343	14	4	3	4	3	cluster_0	Platinium
Row3386	3443	14	3	4	4	3	cluster_1	Gold
Row3387	5144	14	5	1	4	4	cluster_3	Silver
Row3388	3443	14	3	4	4	3	cluster_1	Gold
Row3389	5144	14	5	1	4	4	cluster_3	Silver
Row3392	4244	14	4	2	4	4	cluster_0	Platinium
Row3397	3344	14	3	3	4	4	cluster_1	Gold
Row3400	5144	14	5	1	4	4	cluster_3	Silver
Row3401	5144	14	5	1	4	4	cluster_3	Silver
Row3402	5144	14	5	1	4	4	cluster_3	Silver

Figura 42 - Amostra de registos coloridos por cluster

Terminada a primeira fase da segmentação com recurso a métodos machine learning e técnicas de *data mining* é possivel compararmos os resultados obtidos, com os modelos de classificação tradicionais, e neste caso o método ABCD, que é utilizado para a classificação dos clientes pelo volume de margem libertada. Para esta análise concentramos nos clientes mais valiosos, os mais antigos e leais e os mais recentes, com elevado LRFM *score*.

Numa primeira análise, e visualizando, na figura 43, na coluna da direita é bem distinta a média de antiguidade dos clientes do cluster_0, com média de antiguidade superior a 12 anos e do cluster_1, com média de antiguidade inferiror a 4 anos.

Row ID	Cluster	CPM	CC_2Y	Category	Custom...	ARR M...	LRFM M...	RFM_2...	LRFMV...	L Mean	R Mean	F Mean	M Mean	Lenght_Years Mean
Row0	cluster_0	Platinium	A	Business	5	13,528.4	19.2	7	27.4	4.2	5	5	5	9.8
Row1	cluster_0	Platinium	A	Corporate	56	10,804.73	19.62	7.16	27.36	4.68	4.95	5	5	14.98
Row2	cluster_0	Platinium	A	Not Classified	196	10,322.46	19.27	7.14	26.67	4.31	4.98	4.98	5	11.57
Row3	cluster_0	Platinium	B	Business	15	2,009.6	18.8	7	26.07	3.93	5	4.87	5	11.87
Row4	cluster_0	Platinium	B	Corporate	30	2,804.33	19.6	7	26.67	4.63	4.97	5	5	13.7
Row5	cluster_0	Platinium	B	Not Classified	286	2,629.65	19.16	6.92	26.26	4.27	4.94	4.96	4.99	11.57
Row6	cluster_0	Platinium	C	Business	15	259.6	18.47	6.27	24.33	4.33	4.47	4.87	4.8	11.33
Row7	cluster_0	Platinium	C	Corporate	43	142.72	18.7	6.35	24.49	4.33	4.7	4.81	4.86	10.09
Row8	cluster_0	Platinium	C	Not Classified	461	431.61	18.64	6.54	24.59	4.22	4.75	4.84	4.83	10.9
Row9	cluster_0	Platinium	D	Business	9	111.67	18.56	5.11	25.11	4.78	4.22	4.89	4.67	11.22
Row10	cluster_0	Platinium	D	Corporate	47	44.83	17.83	6.04	23.26	4.43	4.49	4.64	4.28	12.51
Row11	cluster_0	Platinium	D	Not Classified	642	93.02	17.65	5.96	22.46	4.45	4.51	4.48	4.21	11.68
Row12	cluster_0	Platinium	P	Business	11	0	17.18	7	23.73	4.64	3.09	4.73	4.73	14.36
Row13	cluster_0	Platinium	P	Corporate	95	0	16.24	7	22.13	4.24	2.87	4.62	4.51	10.98
Row14	cluster_0	Platinium	P	Not Classified	1,222	2.32	16.15	7	21.67	4.41	2.95	4.45	4.35	10.64
Row15	cluster_0	Platinium	S	Not Classified	12	0	17.75	4.17	23.5	4.25	4	4.75	4.75	10.5
Row16	cluster_1	Gold	A	Business	2	11,276	16.5	7	23	1.5	5	5	5	3
Row17	cluster_1	Gold	A	Corporate	5	8,170	15.6	7	21.6	1	5	4.6	5	2
Row18	cluster_1	Gold	A	Not Classified	52	10,351.06	15.71	6.9	22.19	1.13	4.9	4.67	5	2.44
Row19	cluster_1	Gold	B	Business	6	672.33	15.17	6	22	1.33	4.5	4.33	5	3
Row20	cluster_1	Gold	B	Corporate	15	2,131.27	15.27	6.67	21.47	1.2	4.87	4.2	5	2.67
Row21	cluster_1	Gold	B	Not Classified	117	2,489.48	15.31	6.61	21.45	1.22	4.82	4.26	5	2.77
Row22	cluster_1	Gold	C	Business	1	0	14	7	17	1	5	3	5	3
Row23	cluster_1	Gold	C	Corporate	13	79.15	14.15	6.38	19	1.15	4.77	3.85	4.38	2.69
Row24	cluster_1	Gold	C	Not Classified	315	452.76	14.48	6.55	19.86	1.37	4.77	3.99	4.36	3.03
Row25	cluster_1	Gold	D	Business	2	133.5	12.5	7	16	1	5	3.5	3	3.5
Row26	cluster_1	Gold	D	Corporate	20	49.6	13.4	5.8	17.45	1.4	4.45	3.75	3.8	3.7
Row27	cluster_1	Gold	D	Not Classified	697	73.74	13.46	6.19	18.06	1.8	4.62	3.79	3.25	4.24
Row28	cluster_1	Gold	P	Business	11	0	13.45	7	19.18	2.64	3.09	4.27	3.45	5.64
Row29	cluster_1	Gold	P	Corporate	40	0	12.88	7	17.82	2.38	3.28	3.85	3.38	5.3
Row30	cluster_1	Gold	P	Not Classified	675	0	12.88	7	17.81	2.33	3.39	3.77	3.38	5.27
Row31	cluster_1	Gold	S	Not Classified	9	0	13	4.89	18.56	1.33	4.11	3.56	4	4.11

Figura 43 - Amostra dos clientes mais valiosos

Os clientes *Platinium*, são os mais valiosos de acordo com o resultado da segmentação em quatro *clusters*, segundo o modelo LRFM. Se concentrarmos a análise destes clientes e compararmos com a classificação tradicional ABCD verificamos que 642 clientes atingem o escalão D, que é o mais baixo, sendo o A o mais alto e o número mais elevado de clientes, 1222, não compraram nos últimos

2 anos e por esse motivo são classificados como P (potenciais) ou S (sem compras). Estes 1222 clientes têm o LRFM *Score* baixo para o portefólio *Platinium*, 16,15%, significa que estes clientes tiveram uma relação comercial interessante com a empresa, e também tiveram uma relação de fidelização média superior a 10 anos, mas não compraram recentemente como demonstra o parametro R (Recency) com o valor de 2,95, que é um valor negativo no *rank* LRFM, sendo positivo a partir de 3. Estes clientes têm em comum, com os 461 clientes classificados C, com os 286 clientes classificados B e com os 196 clientes classificados A, o facto de não estarem classificados quanto aos dados firmográficos e por esse motivo são categorizados como '*Not Classified*'. Tratando-se dos clientes mais valiosos, que são os clientes *Platinium*, este modelo de análise do valor do cliente ganha maior relevância em conjunto com a categorização do potencial do cliente e da classificação tradicional ABCD, são complementares e em conjunto contribuem para uma maior assertividade das iniciativas de marketing segmentadas às contas que podem ser focadas na retenção, no desenvolvimento ou na reativaçao. Os cerca de 2000 clientes com classificação D, P e S, que são clientes com menor atividade nos últimos dois anos são claramente candidatos a iniciativas de reativação.

Numa visão 3D, na figura 44, é também perceptivel que os clientes *Platinium* têm maior dispersão, o que pode ser motivo de maior dinamismo e exigência na gestao. Já os clientes *Silver* e *Bronze*, são clientes com pouca atividade, normalmente clientes tipo D, escalão minimo, no método tradicional de classificaçao ABCD, já o P e S, significa que não compram, pelo menos há dois anos.

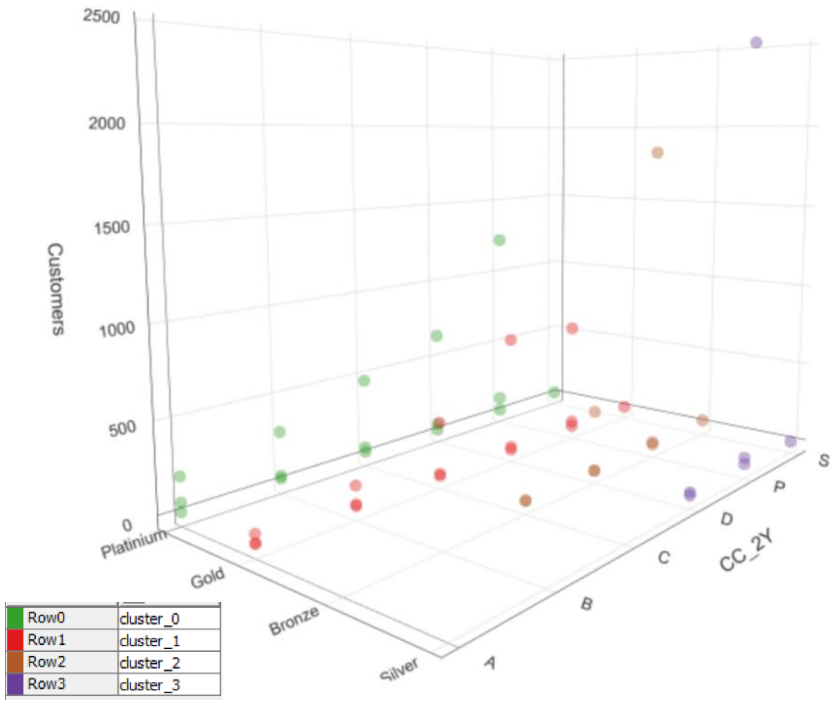


Figura 44 - Representação 3D da classificação dos clientes

Outra análise importante acerca dos *clusters* e da consequente classificação do portefólio de clientes (*Platinium, Gold, Silver* e *Bronze*) é o comportamento de compra recorrente (ARR) e destacam-se bem os clientes *Platinium* e *Gold* nesta análise, na figura 45. É tambem destacado os LRFM *score* mais altos para estes clientes.

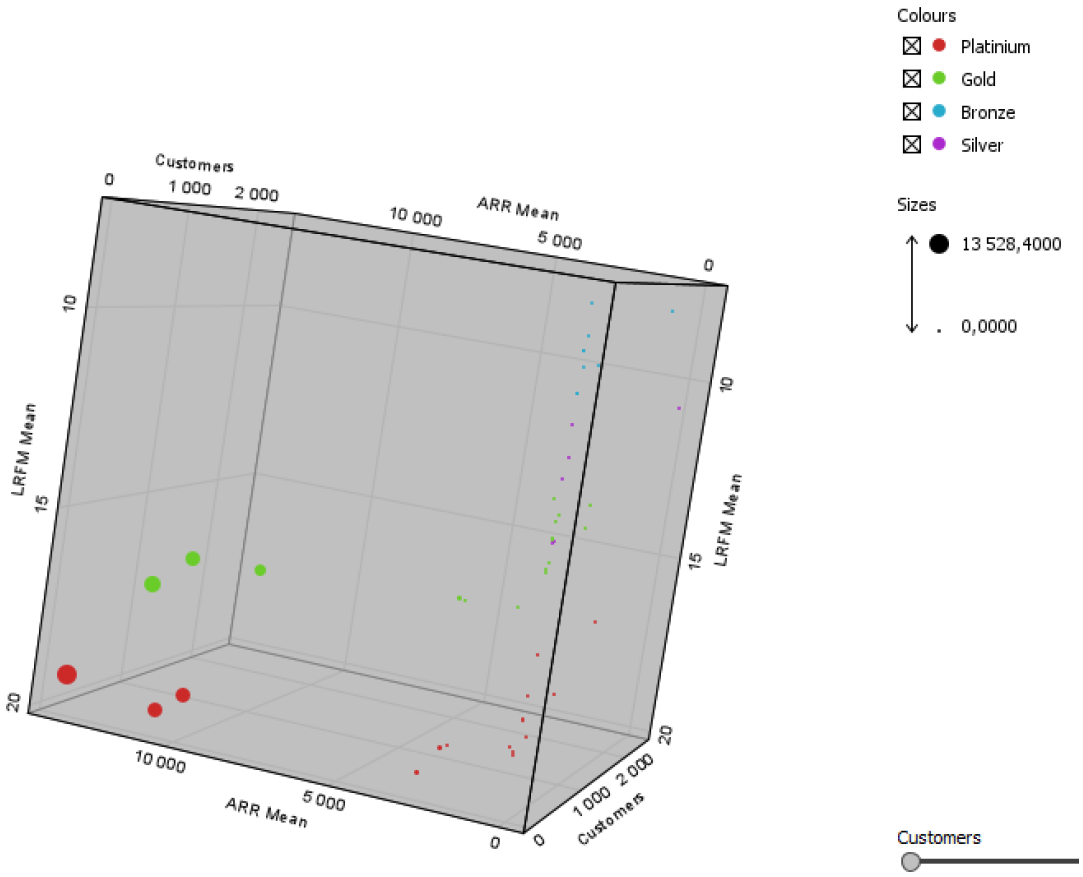


Figura 45 - Representação 4D da classificação dos clientes

A segunda fase da segmentação, considera também o potencial dos clientes, baseados nos dados firmográficos, para além do comportamento de compra obtido pelo método LRFM, e as variantes analisadas neste estudo. Vamos analisar a relação entre o comportamento de compra, o potencial do cliente e a classificação comercial do CRM, baseada no método tradicional ABCD.

Os clientes *Platinium* a côr verde, e os clientes *Gold* a côr vormelha, são maioritariamento clientes classificados A e B, mas estão distribuidos por todas as classificações, como identificamos anteriormente neste estudo, bem como sugerimos a possibilidade de iniciativas de recuperação. Mas o que realça mais, na figura 46, é a existência de um número elevado de clientes não classificados, significa que não está atualizada a informaçã firmográfica em CRM.

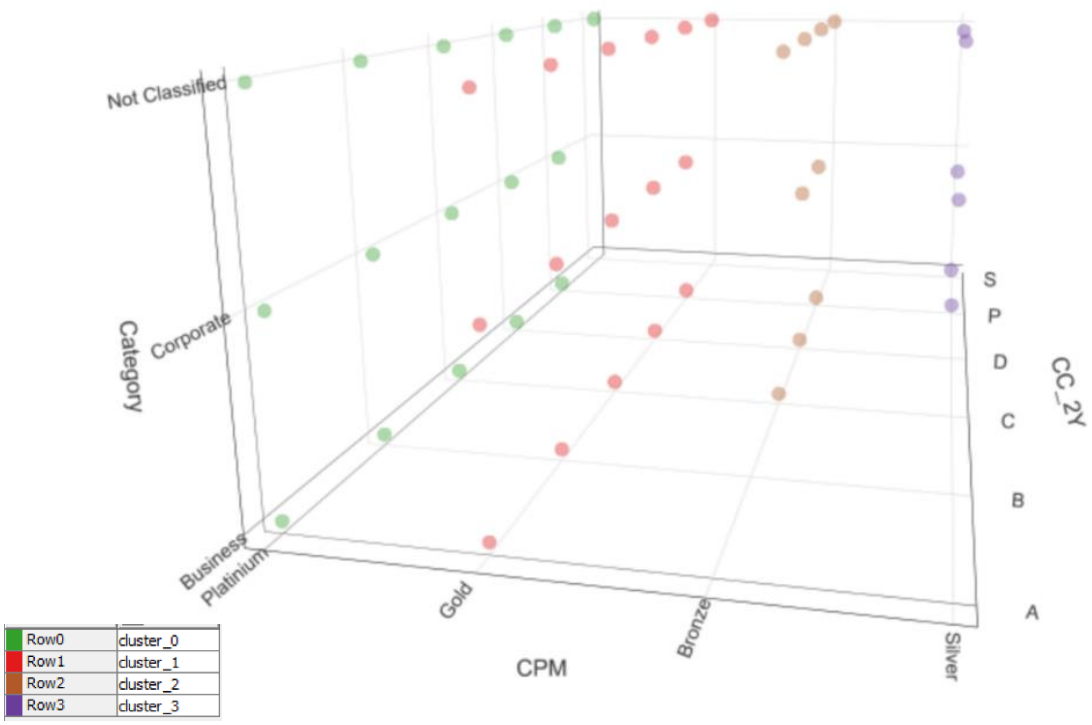


Figura 46 - Representação 3D dos clientes categorizados e classificados

Sabendo da importância da informação firmográfica para agrupar clientes similares, analisamos, no grafico representado na figura 47, a parcela de clientes que requerem a atualização e verificamos que é muito significativa, pelo que sugerimos a atualização.

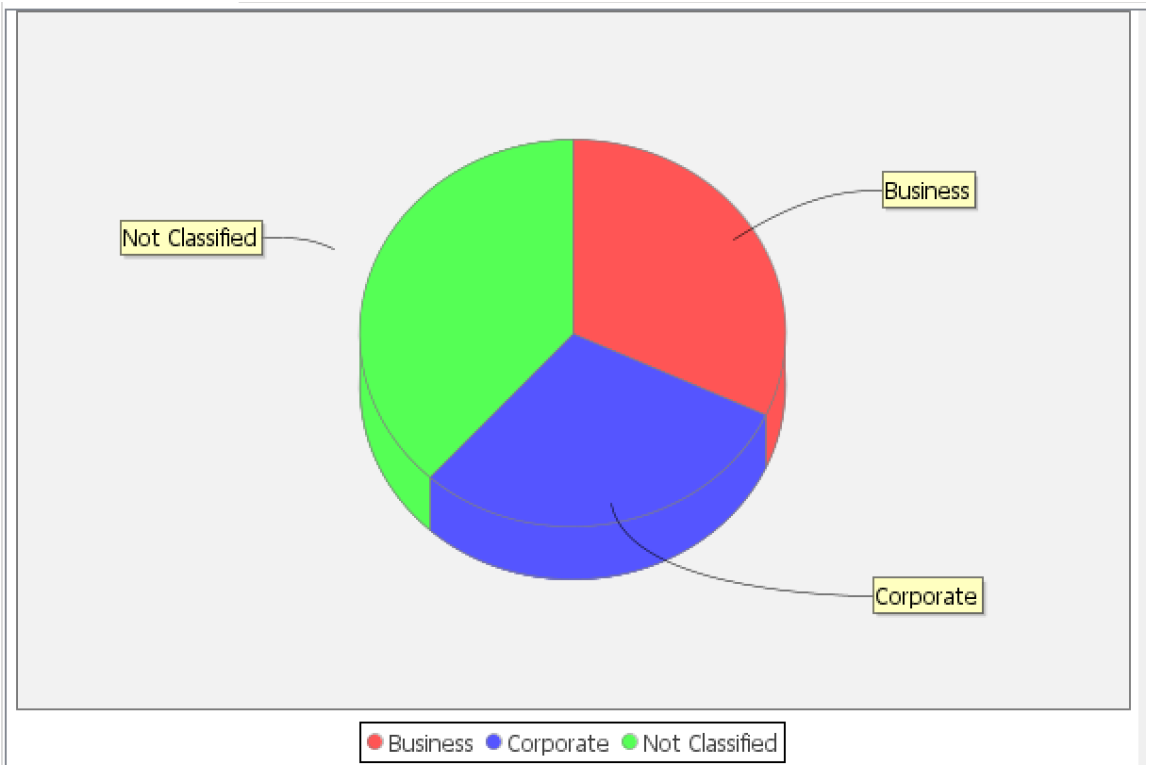


Figura 47 - Representação gráfica da classificação dos clientes

5.2. Novo modelo LRFMVP

O modelo LRFMVP inclui 5 critérios de análise e, portanto, permite avaliar a lealdade dos clientes, mas também permite avaliar a variedade dos produtos e serviços que compra e com que periodicidade ocorrem as transações, para além do modelo tradicional RFM. Para segmentação dos clientes utilizamos o mesmo algoritmo que foi utilizado no modelo LRFM, o *K-means*, que é o algoritmo mais usado na segmentação de clientes pela sua similaridade, para analisar ligações entre os clientes e inferir os seus padrões e tendências.

O método do cotovelo foi aplicado para identificar o k ideal. O resultado do cálculo do melhor erro médio absoluto foram três *clusters*, identificados graficamente em três dimensões na figura 48, com os pontos de dados distribuídos nos três *clusters*.

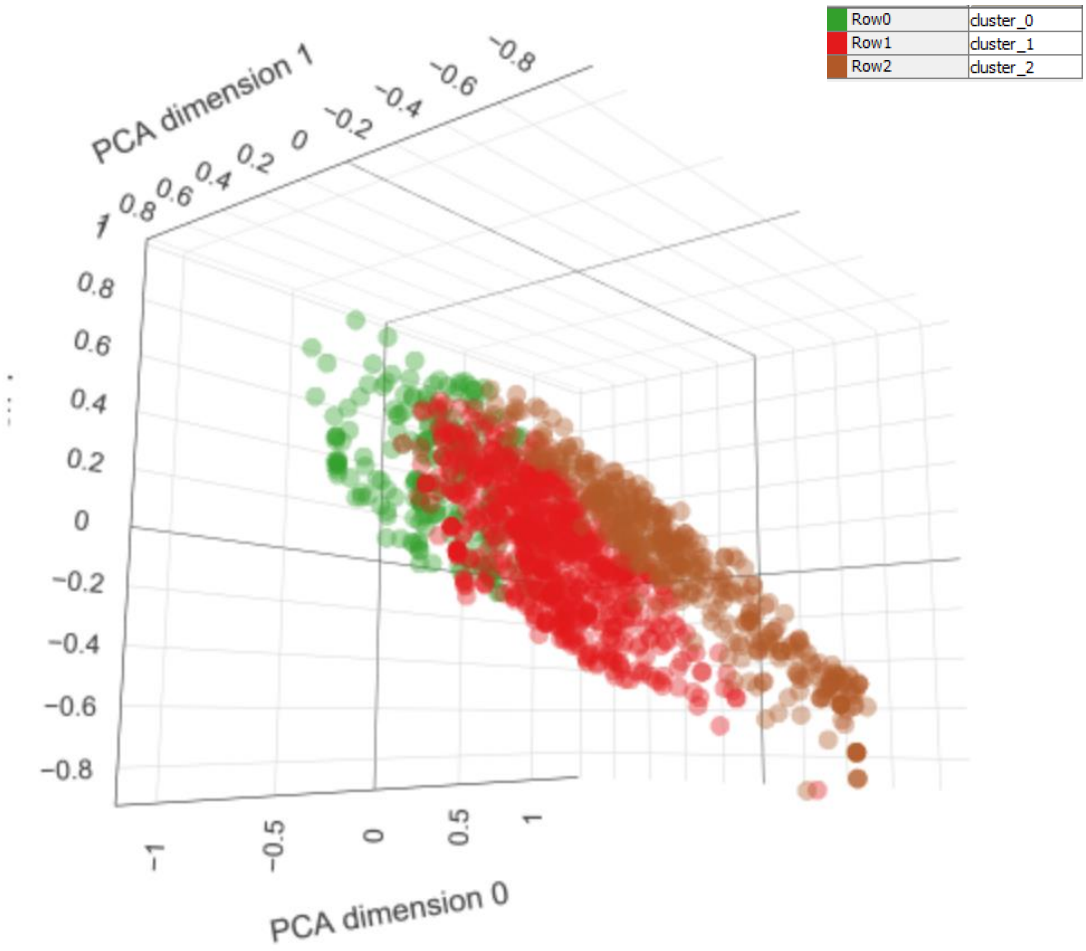


Figura 48 - Representação 3D da classificação dos clientes, no modelo LRFMVP

O modelo LRFMVP oferece mais informação e permite analisar, com maior detalhe, o comportamento de compra dos clientes mais valiosos. Este detalhe, na tabela 5, permite perceber melhor a lealdade e o valor do cliente, mas também a forma como acontece. O conhecimento pormenorizado de cada

cliente e simultaneamente a dos clientes similares é crucial para melhorar a gestão dos clientes e para a aplicação de métodos de marketing inovadores e direcionados à conta.

Cluster	**Número de Clientes**	**Padrão LRFMVP**
Cluster_0	219	L↑ R↑ F↑ M↑ V↑ P↑
Cluster_1	708	L↓ R↑ F↑ M↑ V↑ P↑
Cluster_2	376	L↓ R↑ F↑ M↑ V↓ P↑

Tabela 5 - Padrão do modelo L RFM V P e total de clientes por cluster

O padrão LRFMVP para cada *cluster* é bastante elucidativo e conclusivo. Logo à partida tratando-se dos clientes mais valiosos, os parâmetros RFMP estão acima da média. A variedade de produtos assume especial relevância na composição do "cabaz de compra". Conclui-se também que a antiguidade é um fator importante para o valor do cliente, mas acima de tudo é o barómetro da lealdade e da capacidade de fidelização dos clientes. Embora mais de metade dos clientes são mais novos do que a média, também nos clientes mais valiosos se verifica a regra de pareto e cerca de 70% do negócio é feito em 30% dos clientes.

Recorrendo a técnicas de *data mining* para descobrir as regras de associação, neste caso utilizando o algoritmo *Apriori*, conseguimos estudar o comportamento dos clientes, em cada segmento. Estas técnicas são poderosas ferramentas, que fornecem um nível de conhecimento só alcançável com recurso a inteligência artificial e que garantem níveis de conhecimento mais aprofundados. O comportamento em cada parâmetro obedece ao um *rank* de 1 a 5, em que 1 é o pior e 5 o melhor. Adicionalmente e para cada parâmetro descrevemos a correspondência de cada valor do *rank*, conforme a figura 49.

Rank	S Lenght	S Recency	S Frequency	S Monetary	S Variety	S Periodicity
5	L=Muito Antigo	R=Muito Recente	F=Muito Frequente	M=Mais alto	V=Total Variedade	P=Muito Elevada
4	L=Antigo	R=Recente	F=Bastante Frequente	M=Elevado	V=Elevada Variedade	P=Bastante Elevada
3	L=Médio	R=Algum tempo	F=Frequente	M=Considerável	V=Diversa	P=Razoável
2	L=Novo	R=Bastante Tempo	F=Ocasional	M=Baixo	V=Concentrada	P=Baixa
1	L=Muito novo	R=Muito Tempo	F=Raramente	M=Insignificante	V=Exclusiva	P=Rara

Figura 49 - Descrição dos valores do rank obtido no modelo LRFMVP

Para melhor compreensão dos clientes mais valiosos, os 3 segmentos de clientes são analisados individualmente.

O segmento do Cluster_0, na figura 50,

Row ID	Consequent	Antecedent	ItemSetSupport	RuleLift%	RuleConfidence%
Row9	Platinium	[F=Muito Frequente,L=Muito Antigo,P=Razoável]	45	100.92	100
Row11	Platinium	[F=Muito Frequente,P=Razoável]	45	100.92	100
Row17	Platinium	[L=Muito Antigo,M=Mais alto,P=Razoável]	44	100.92	100
Row19	Platinium	[M=Mais alto,P=Razoável]	44	100.92	100
Row29	Platinium	[F=Muito Frequente,L=Muito Antigo,M=Mais alto,...]	45	100.92	100
Row34	Platinium	[F=Muito Frequente,M=Mais alto,P=Bastante Elevada,...]	46	100.92	100
Row49	Platinium	[F=Muito Frequente,L=Muito Antigo,P=Bastante Elevada,...]	45	100.92	100
Row53	Platinium	[F=Muito Frequente,P=Bastante Elevada,R=Muito Recente...	46	100.92	100
Row65	Platinium	[L=Muito Antigo,M=Mais alto,P=Bastante Elevada,...]	45	100.92	100
Row69	Platinium	[M=Mais alto,P=Bastante Elevada,R=Muito Recente,...]	46	100.92	100
Row81	Platinium	[L=Muito Antigo,P=Bastante Elevada,R=Muito Recente,...]	45	100.92	100
Row84	Platinium	[P=Bastante Elevada,R=Muito Recente,V=Total Variedade]	46	100.92	100
Row93	Platinium	[F=Muito Frequente,L=Muito Antigo,M=Mais alto,...]	50	100.92	100
Row97	Platinium	[F=Muito Frequente,M=Mais alto,P=Bastante Elevada,...]	51	100.92	100
Row109	Platinium	[F=Muito Frequente,L=Muito Antigo,P=Bastante Elevada,...]	50	100.92	100
Row112	Platinium	[F=Muito Frequente,P=Bastante Elevada,V=Total Variedade]	51	100.92	100
Row121	Platinium	[L=Muito Antigo,M=Mais alto,P=Bastante Elevada,...]	50	100.92	100
Row124	Platinium	[M=Mais alto,P=Bastante Elevada,V=Total Variedade]	51	100.92	100
Row133	Platinium	[L=Muito Antigo,P=Bastante Elevada,V=Total Variedade]	50	100.92	100
Row135	Platinium	[P=Bastante Elevada,V=Total Variedade]	51	100.92	100
Row157	Platinium	[Corporate,F=Muito Frequente,L=Muito Antigo,...]	44	100.92	100
Row160	Platinium	[Corporate,F=Muito Frequente,M=Mais alto,...]	44	100.92	100
Row169	Platinium	[Corporate,F=Muito Frequente,L=Muito Antigo,...]	44	100.92	100
Row171	Platinium	[Corporate,F=Muito Frequente,V=Total Variedade]	44	100.92	100
Row177	Platinium	[Corporate,L=Muito Antigo,M=Mais alto,...]	44	100.92	100
Row179	Platinium	[Corporate,M=Mais alto,V=Total Variedade]	44	100.92	100
Row185	Platinium	[Corporate,L=Muito Antigo,V=Total Variedade]	45	100.92	100

Figura 50 - Regras obtidas pelo método Apriori para o Cluster_0

O Cluster_0, constituído por 219 clientes muito valiosos, com todas as variáveis do modelo LRFMVP acima da média e conseguimos identificar que os clientes *Platinium*, têm transações muito frequentes (F=Muito Frequente), são cliente muito antigos (L=Muito Antigo) e a Periodicidade das transações é razoável (P=Razoável), face aos restantes. Estes clientes têm o valor Monetário mais alto (M=Mais alto).

O segmento do Cluster_1, na figura 51,

Row ID	Consequent	Antecedent	ItemSetSupport	RuleLift%	RuleConfidence%
Row7	Gold	[L=Muito novo,R=Muito Recente]	147	282.47	100
Row8	Gold	[L=Muito novo]	170	282.47	100
Row0	Platinium	[L=Muito Antigo]	149	154.8	100
Row3	Platinium	[L=Antigo]	152	154.8	100
Row11	Platinium	[F=Muito Frequente,L=Médio]	151	147.01	95
Row12	Platinium	[L=Médio]	157	135.02	87.2
Row59	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,Platinium,...]	145	127.06	100
Row62	F=Muito Frequente	[P=Bastante Elevada,Platinium,V=Diversa]	157	127.06	100
Row71	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,Platinium,...]	158	127.06	100
Row74	F=Muito Frequente	[P=Bastante Elevada,Platinium,R=Muito Recente]	169	127.06	100
Row40	Platinium	[Corporate,F=Muito Frequente,V=Diversa]	149	126.04	81.4
Row78	F=Muito Frequente	[P=Bastante Elevada,Platinium]	243	126.02	99.2
Row75	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,Platinium]	217	125.9	99.1
Row30	F=Muito Frequente	[Corporate,P=Bastante Elevada,Platinium]	152	125.41	98.7
Row6	R=Muito Recente	[Gold,L=Muito novo]	147	124.86	86.5
Row9	R=Muito Recente	[L=Muito novo]	147	124.86	86.5
Row63	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,R=Muito Recente,...]	160	124.72	98.2
Row14	Platinium	[F=Muito Frequente,P=Razoável]	128	124.62	80.5
Row67	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,V=Diversa]	204	124.62	98.1
Row70	F=Muito Frequente	[P=Bastante Elevada,V=Diversa]	219	124.22	97.8
Row66	F=Muito Frequente	[P=Bastante Elevada,R=Muito Recente,V=Diversa]	171	124.16	97.7
Row37	Platinium	[Corporate,F=Muito Frequente,M=Mais alto,...]	129	124.04	80.1
Row34	F=Muito Frequente	[Corporate,P=Bastante Elevada]	209	122.94	96.8
Row79	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,R=Muito Recente]	247	122.59	96.5
Row31	F=Muito Frequente	[Corporate,M=Mais alto,P=Bastante Elevada]	180	122.3	96.3
Row10	F=Muito Frequente	[L=Médio,Platinium]	151	122.21	96.2
Row82	F=Muito Frequente	[P=Bastante Elevada,R=Muito Recente]	274	121.73	95.8

Figura 51 - Regras obtidas pelo método Apriori para o Cluster_1

No Segmento Custer_1 algo de surpreendente, mas marcante, os clientes *Gold* são L= muito novos e os clientes *Platinium*, L= muito antigos. Este conhecimento, são alcançáveis com recurso a ferramentas sofisticadas ou grandes investimentos nos desenvolvimentos de soluções de análise e estatística, permite à empresa do estudo, desenvolver estratégias de marketing diferenciadas para cada *cluster*.

Row ID	S Consequent	[...] Antecedent	I ItemSetSupport	D RuleLift%	D RuleConfidence%
Row2	L=Muito novo	[Gold,R=Muito Recente]	68	180.37	84
Row3	Gold	[L=Muito novo,R=Muito Recente]	68	177.04	81.9
Row1	V=Exclusiva	[P=Muito Elevada]	83	148.21	88.3
Row0	V=Exclusiva	[Bronze]	69	144.78	86.2

Figura 52 - Regras obtidas pelo método Apriori para clientes Gold e Platinium

No segmento do Cluster_0, apresentado na figura 52, foram geradas apenas 4 regras para as medidas de interesse, comuns aos 3 *clusters*, 20% de suporte mínimo e confiança de 80%, sendo que os clientes L=Muito novo (considerando que já fazem parte do grupo selecionado dos clientes mais valiosos), são *Gold* e o R= Muito recente, o que significa que são clientes novos, com transações muito recentes.

Outro aspeto, menos surpreendente, mas igualmente importante, os clientes com V=exclusiva, ou seja, que trabalham uma gama muito exclusiva de produtos, em alguns casos um único produto, têm uma periodicidade muito elevada e são bronzes, o que nos pode sugerir serem clientes de soluções *Saas* ou subscrições de baixo valor. Pode ser interessante explorar este segmento para novas soluções *cloud* de maior valor, aumentando a variedade e o valor das transações.

A seguir estudamos a distribuição dos clientes mais valiosos pelos principais dados firmográficos, para percebermos onde a empresa tem mais potencial e consiga desenvolver ações que potenciem mais negócio, nas áreas que já domina.

Na figura 53, apresenta a distribuição das contas no Top 20 dos setores de atividade com mais clientes. O setor 41200 tem 31 clientes, distribuídos pelos no cluster_1 e cluster_2. O setor 25734 tem 9, dos mais valiosos clientes.

Frequency	cluster_0	cluster_1	cluster_2	Total
11021		10		10
14131		10		10
22292		16		16
25110		7		7
25120		9		9
25734	9			9
29320		7		7
41200		17	14	31
42990			8	8
43992		9		9
45110			11	11
46382		7		7
46690		14	8	22
46900		12		12
49410		10		10
70220		12		12
82990		9		9
86100		9		9
Total	9	158	41	208

☑ Frequency
☐ Expected
☐ Deviation
☐ Percent
☐ Row Percent
☐ Column Percent
☐ Cell Chi-Square

Max rows: 20

Max columns: 10

Figura 53 - Número de clientes por CAE e por Cluster

Importante também identificar os padrões e o potencial dos clientes por região, num país relativamente pequeno, mas muito diversificado, sendo que os maiores clientes se concentram em Leiria, Lisboa e Porto, conforme a figura 54.

Frequency	cluster_0	cluster_1	cluster_2	Total
?		98	37	135
Aveiro	7	60	41	108
Braga		30	31	61
Coimbra	11	29	14	54
Faro		7		7
Leiria	123	71	46	240
Lisboa	33	157	77	267
Madeira		11		11
Porto	12	108	50	170
Santarém	17	14	17	48
Setubal		21	17	38
Viseu		12	8	20
Zona Norte		56	12	68
Total	203	674	350	1 227

☑ Frequency
☐ Expected
☐ Deviation
☐ Percent
☐ Row Percent
☐ Column Percent
☐ Cell Chi-Square

Max rows: 15

Max columns: 10

Figura 54 - Número de clientes por Zona e por Cluster

Dos 1.303 clientes, os mais valiosos, apenas 137 contas da categoria corporate estão no cluster_0, podendo indicar o enorme potencial da carteira com 762 clientes corporate e a pertinência para aplicação de metodologias de marketing baseadas nas contas devidamente segmentadas com o modelo LRFMVP, conforme a figura 55.

Frequency	cluster_0	cluster_1	cluster_2	Total
Business	64	308	118	490
Corporate	137	370	255	762
Not Classified	10	19	2	31
iGo	8	11	1	20
Total	219	708	376	1 303

☑ Frequency
☐ Expected
☐ Deviation
☐ Percent
☐ Row Percent
☐ Column Percent
☐ Cell Chi-Square

Figura 55 - Número de clientes por Categoria e por Cluster

A figura 56 apresenta a distribuição dos clientes mais valiosos, pelas 2 dimensões de análise firmográficas, Zona, pela Categoria e o pelo respetivo *cluster* da segmentação.

O cluster_1, tem o maior número de clientes classificados como corporate e business.

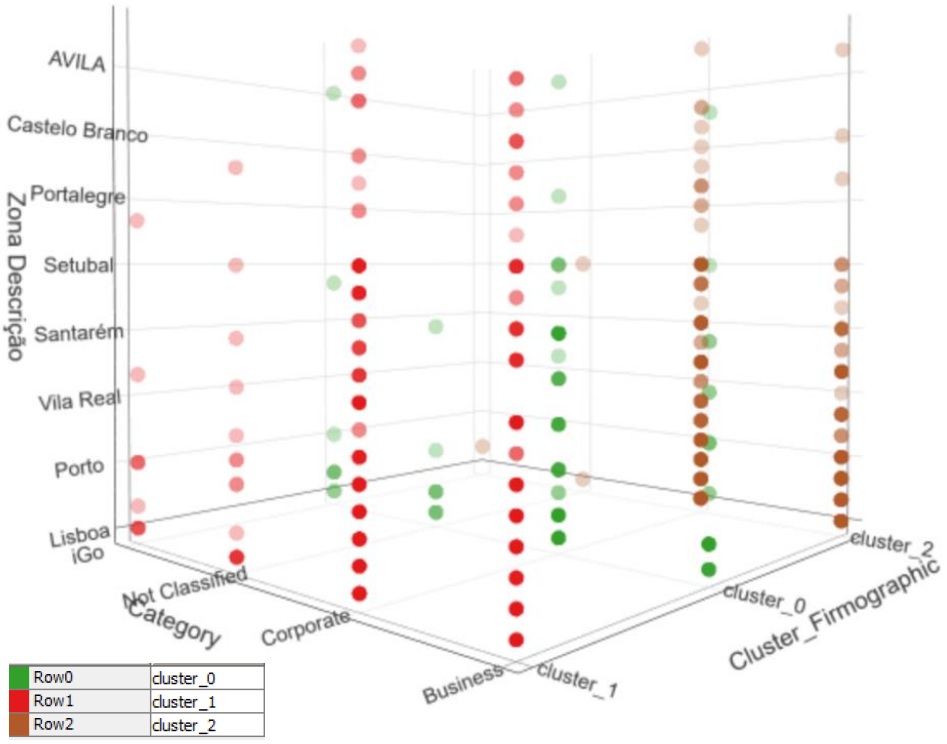


Figura 56 - Distribuição dos clientes mais valiosos

A análise na figura 57, demonstra em certa medida uma das dificuldades que os vários investigadores referem em todos os estudos, por um lado a escassez de estudos empíricos e por outro lado a

natureza do negócio B2B é muito diferente do mercado de consumo e retalho, é muito diversificado e os padrões de compra dos clientes são muito difíceis de replicar.

Conseguimos identificar que, mesmo que os clientes do cluster_0 sejam similares e agrupados num segmento, o seu volume de compra difere imenso e por esse motivo são identificados vários '*outliers*'.

Mesmo assim é possível identificar que o ARR dos clientes do cluster_0 muito superior ao valor dos clientes do cluster_1 e incomparavelmente superior aos clientes do cluster_2. Ao longo do estudo vai ficando demonstrada a importância da faturação recorrente na empresa em estudo e é um dos fatores de maior valorização das empresas tecnológicas.

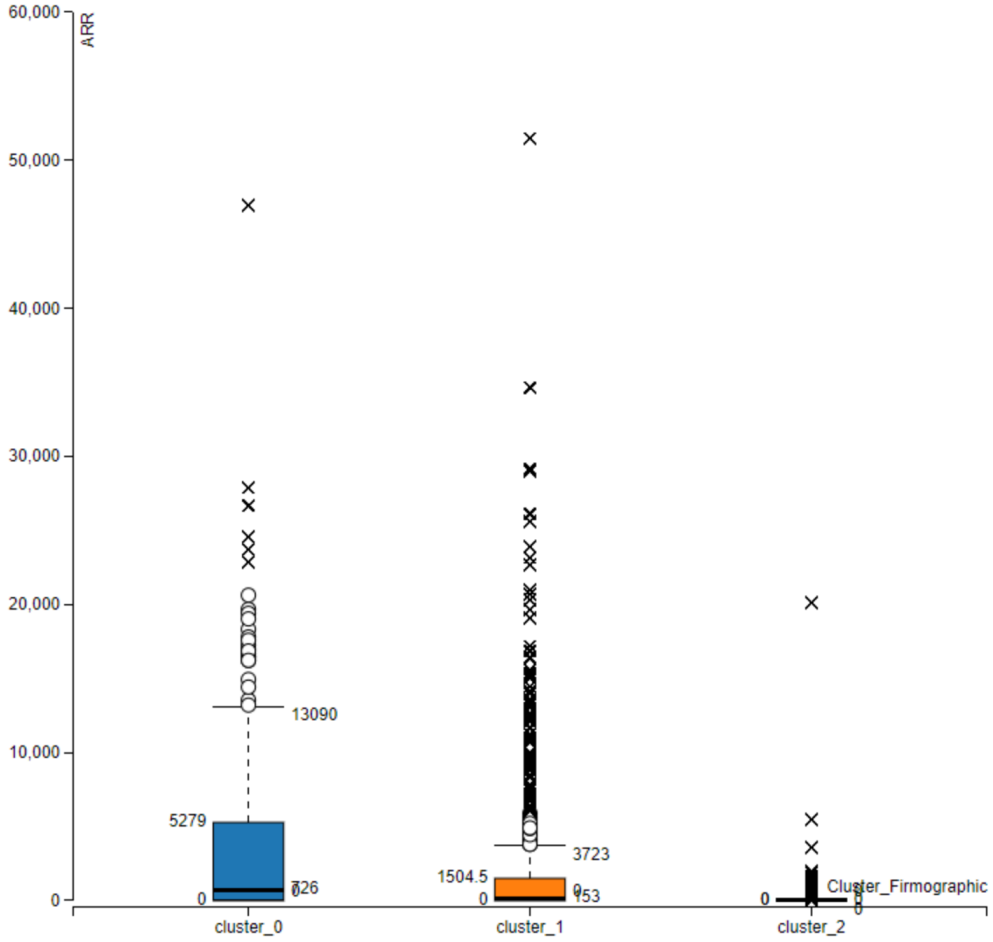


Figura 57 - Box Plot com ARR por cluster

CAPÍTULO VI

6. CONCLUSÕES E TRABALHO FUTURO

Como aumentar a eficiência da força de vendas e aumentar a eficácia na concretização de novos negócios empresariais, foram os desafios para este trabalho de investigação.

À partida foram também definidos resultados esperados, que visavam em primeiro lugar melhorar o conhecimento acerca dos clientes e com isso aumentar a assertividade na relação com os clientes, (Gentner et al., 2018), e em segundo lugar usar metodologias e técnicas inovadoras, com critérios de mercado atuais, para segmentar os clientes, prever o potencial de negócio futuro na carteira e identificar potenciais clientes (Heras Barberá et al., 2006).

Após a realização de investigação, e do estado da arte, percebemos que teríamos muitas dificuldades, pois a literatura e os modelos estudados estão muito concentradas em negócios de consumidor final, com foco no retalho, também designado por FMCG (*Fast-Moving Consumer Goods*) e ambicionamos com este trabalho de investigação acrescentar algo de novo, daí surge o novo modelo LRFMVP.

Terminadas as tarefas de processamento de dados ficou evidenciada a importância dos clientes mais valiosos, mas também ficou clara a assertividade na estratégia da empresa, em querer conhecer e segmentar os clientes pelo seu potencial. Por um lado, os *scores* obtidos no processamento dos modelos LRFM e LRFMVP permitem obter o valor da vida útil do cliente, por outro lado os dados firmográficos, que incluem o volume de negócios, o número de empregados e o setor de atividade, são da maior importância, pois permitem identificar, entre outras informações, quais são os clientes Corporate neste estudo, são os que têm volume de negócios superior a 5 milhões de euros.

Numa perspetiva de gestão dos clientes e de '*CRM insights'*, ficou claro que o modelo analítico tradicional de classificação ABCD não é descritivo, já que considera apenas um critério, neste caso a margem das vendas. Contudo e para que a empresa consiga, em tempo real, conhecer o comportamento de compra dos clientes e possa desenvolver ações proativas de retenção, desenvolvimento ou de reativação de clientes que deixaram de comprar, este modelo pode ser complementar ao modelo de segmentação dos clientes proposto, com recurso à utilização de inteligência artificial, é um insight importante e que deve estar disponível e atualizado no CRM.

Além do CRM, e do ponto de vista das implicações na gestão, foi possível gerar conhecimento capaz de responder a algumas das ambições dos responsáveis pelo marketing, como a criação de iniciativas de marketing personalizadas e para isso muito contribuíram a aplicação do modelo de *clustering, K-means*, que segmentou os clientes pela sua similaridade em cada cenário e o modelo de análise do valor dos clientes, adaptado do modelo RFM. As várias análises com modelos adaptados do RFM permitem identificar que a Periodicidade das transações é uma variável da análise do comportamento de compra do cliente com elevada influência nos resultados alcançados e que os clientes mais fidelizados com maior antiguidade, medidos pela variável L do modelo, tem *scores* mais altos no modelo e por isso são os clientes mais valiosos. Estas conclusões resultaram do estudo da análise de associação, que permitiu encontrar relações de interesse no grande conjunto de dados disponível, bem como a capacidade para a interpretação descritiva dos *clusters* gerados nos pontos anteriores. Observamos que o conjunto de regras de associação obtidas no modelo LRFMVP são superiores, em comparação com as regras de associação obtidas no modelo LRFM, mas melhora a percentagem média de confiança, de 93.67% para 94.38%, para a mesma percentagem de interesse da regra que é de 148.67%.

Modelo	**Número de regras**	**% de Confiança**	**% de Interesse (lift)**	**% de suporte relativo**
LRFM	13	93.67%	148.67%	30.75%
LRFMVP	30	94.38%	148.67%	30.83%

Tabela 6 - Resultados experimentais nos modelos LRFM e LRFMVP

Na tabela 6, comparando os dois modelos, usando as regras de associação, observamos que o novo modelo, LRFMVP é mais descritivo do que o modelo LRFM para ação de retenção e desenvolvimento. Destaco a interpretação das regras de associação descobertas no novo modelo de análise do valor dos clientes, designado LRFMVP, e nas suas vantagens, nas conclusões acerca da segmentação mais detalhada ou não, por exemplo se analisarmos o valor do cliente com o método LRFM. As regras, na figura 58, estão ordenadas com base no *Lift*, que mede o interesse da regra.

Row ID	Consequent	Antecedent	ItemSe...	RuleLift	RuleLift%	RuleConfidence%
Row1	Platinium	[F=Muito Frequente,L=Muito Antigo,M=Mais alto]	417	1.637	163.66	100
Row3	Platinium	[L=Muito Antigo,M=Mais alto]	467	1.633	163.31	99.8
Row2	Platinium	[F=Muito Frequente,L=Muito Antigo]	453	1.633	163.3	99.8
Row4	Platinium	[L=Muito Antigo]	569	1.502	150.2	91.8
Row10	F=Muito Frequente	[M=Mais alto,Platinium,R=Muito Recente]	555	1.49	148.96	94.4
Row9	M=Mais alto	[F=Muito Frequente,Platinium,R=Muito Recente]	555	1.466	146.56	94.4
Row11	F=Muito Frequente	[M=Mais alto,Platinium]	744	1.444	144.42	91.5
Row0	M=Mais alto	[F=Muito Frequente,L=Muito Antigo,Platinium]	417	1.429	142.93	92.1
Row5	M=Mais alto	[F=Muito Frequente,L=Muito Antigo]	417	1.426	142.62	91.9
Row7	F=Muito Frequente	[Corporate,M=Mais alto,Platinium]	441	1.426	142.62	90.4
Row8	F=Muito Frequente	[Corporate,M=Mais alto,R=Muito Recente]	369	1.424	142.38	90.2
Row12	M=Mais alto	[F=Muito Frequente,R=Muito Recente]	692	1.417	141.75	91.3
Row6	M=Mais alto	[Corporate,F=Muito Frequente,Platinium]	441	1.4	140.03	90.2

Figura 58 - Regras de associação descobertas no modelo LRFM

Para parâmetros iguais, o *data set* igual, com o modelo LRFMVP conseguimos obter mais detalhe, com maior número de regras de associação, como confirma a figura 59, em comparação com a figura 57.

Row ID	Consequent	Antecedent	ItemSetSupport	Rul...	RuleLift%	RuleConfidence%
Row1	Platinium	[F=Muito Frequente,L=Muito Antigo,M=Mais alto]	417	1.637	163.66	100
Row3	Platinium	[L=Muito Antigo,M=Mais alto]	467	1.633	163.31	99.8
Row2	Platinium	[F=Muito Frequente,L=Muito Antigo]	453	1.633	163.3	99.8
Row8	F=Muito Frequente	[P=Bastante Elevada,Platinium,R=Muito Recente]	419	1.574	157.44	99.8
Row11	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,Platinium]	494	1.565	156.55	99.2
Row12	F=Muito Frequente	[P=Bastante Elevada,Platinium]	523	1.554	155.44	98.5
Row15	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada,R=Muito Recente]	527	1.549	154.88	98.1
Row19	F=Muito Frequente	[M=Mais alto,P=Bastante Elevada]	632	1.527	152.74	96.8
Row16	F=Muito Frequente	[P=Bastante Elevada,R=Muito Recente]	562	1.514	151.35	95.9
Row7	M=Mais alto	[F=Muito Frequente,P=Bastante Elevada,Platinium,...]	407	1.508	150.82	97.1
Row9	M=Mais alto	[P=Bastante Elevada,Platinium,R=Muito Recente]	407	1.505	150.46	96.9
Row4	Platinium	[L=Muito Antigo]	569	1.502	150.2	91.8
Row27	F=Muito Frequente	[M=Mais alto,Platinium,R=Muito Recente]	555	1.49	148.96	94.4
Row10	M=Mais alto	[F=Muito Frequente,P=Bastante Elevada,Platinium]	494	1.467	146.66	94.5
Row26	M=Mais alto	[F=Muito Frequente,Platinium,R=Muito Recente]	555	1.466	146.56	94.4
Row13	M=Mais alto	[P=Bastante Elevada,Platinium]	498	1.456	145.62	93.8
Row14	M=Mais alto	[F=Muito Frequente,P=Bastante Elevada,R=Muito Recente]	527	1.456	145.6	93.8
Row20	F=Muito Frequente	[P=Bastante Elevada]	693	1.454	145.44	92.2
Row28	F=Muito Frequente	[M=Mais alto,Platinium]	744	1.444	144.42	91.5
Row21	M=Mais alto	[F=Muito Frequente,R=Muito Recente,V=Diversa]	421	1.443	144.3	92.9
Row6	F=Muito Frequente	[Corporate,P=Bastante Elevada]	395	1.436	143.64	91
Row0	M=Mais alto	[F=Muito Frequente,L=Muito Antigo,Platinium]	417	1.429	142.93	92.1
Row5	M=Mais alto	[F=Muito Frequente,L=Muito Antigo]	417	1.426	142.62	91.9
Row24	F=Muito Frequente	[Corporate,M=Mais alto,Platinium]	441	1.426	142.62	90.4
Row22	F=Muito Frequente	[M=Mais alto,R=Muito Recente,V=Diversa]	421	1.426	142.58	90.3
Row25	F=Muito Frequente	[Corporate,M=Mais alto,R=Muito Recente]	369	1.424	142.38	90.2
Row17	M=Mais alto	[P=Bastante Elevada,R=Muito Recente]	537	1.423	142.29	91.6
Row29	M=Mais alto	[F=Muito Frequente,R=Muito Recente]	692	1.417	141.75	91.3
Row18	M=Mais alto	[F=Muito Frequente,P=Bastante Elevada]	632	1.416	141.6	91.2
Row23	M=Mais alto	[Corporate,F=Muito Frequente,Platinium]	441	1.4	140.03	90.2

Figura 59 - Regras de associação descobertas no modelo LRFMVP

O modelo LRFMVP permite observar que a Periodicidade é importante e está associada a várias regras de associação, já a Variedade tem menos influência no modelo, e isso pode justificar-se pelo facto de estudarmos um portefólio de soluções relativamente estável ao longo dos anos.

Ao detalharmos as regras mais importantes, percebemos o maior detalhe e que nos permite avaliar melhor a carteira de clientes e a importância do fator Periodicidade na definição dos vários indicadores, pelo que sugerimos que seja o modelo a adotar em empresas tecnológicas e que têm muitos anos de existência, bem como um portefólio muito alargado de produtos e serviços.

Terminado o processo de segmentação dos clientes, o modelo foi treinado para ter capacidade de predição sobre as novas contas de prospeção e os resultados foram muito positivos. O modelo apresenta 91,5% de *accuracy*, para um erro inferior a 8,5%. A aplicação deste modelo a novas contas de prospeção, identificando as contas que já são clientes e para as restantes sugere qual o sistema de gestão mais indicado, bem como o potencial de negócio e o potencial de faturação recorrente, foram identificados como enormes ganhos para as futuras ações de marketing, são capacidades só alcançáveis com a aplicação de técnicas e métodos de Inteligência Artificial.

Por fim, os modelos de recomendação desenvolvidos, permitiram observar a enorme aplicabilidade em ações de cross selling e a melhor perceção do "cesto de compras", tão relevante para o trabalho das equipas de marketing e na gestão de clientes.

Se por um lado deste trabalho de investigação resultaram um conjunto de conclusões bem recebidas e com aplicabilidade para os objetivos traçados, há tarefas e áreas, que no futuro, podem ser incluídas. O mercado empresarial, B2B, está muito interessado em melhorar o conhecimento acerca dos interesses e da personalidade dos compradores com base nas suas preferências de compra passada, nos hábitos de pesquisa e nos seus dados tecnológicos e psicográficos, para além dos dados demográficos, firmográficos, que foram considerados neste trabalho de investigação.

Além disso, pode ser feito um trabalho mais extenso e profundo na modelação, uma vez que foram implementados os algoritmos e artefactos mais comuns, mas outras coisas poderiam ser feitas para criar uma previsão mais precisa do modelo, bem como comparar com abordagens e algoritmos diferentes. Uma opção seria treiná-los com dados de empresas especializadas ou usar dados treinados e aplicá-los ao modelo.

BIBLIOGRAFIA

REFERÊNCIAS BIBLIOGRÁFICAS

Abratt, R. (1993). Market segmentation practices of industrial marketers. *Industrial Marketing Management*, *22*(2), 79–84. https://doi.org/10.1016/0019-8501(93)90033-4

al Shorman, D. M., & Hasan Jbara, D. (2017). An Improved Association Rule Mining Algorithm Based on Apriori and Ant Colony approaches. In *IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org ISSN* (Vol. 07). www.iosrjen.org

Allahyari, R., 1+, S., & Fathalizade, E. (n.d.). *Customer Segmentation based on Modified RFM Model in the Insurance Industry*.

Allegue, S., Abdellatif, T., & Bannour, K. (2020). RFMC: A spending-category segmentation. *Proceedings of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE*, *2020-September*, 165–170. https://doi.org/10.1109/WETICE49692.2020.00040

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. In *International Journal of Computer Applications* (Vol. 105, Issue 9).

Buckinx, W., & van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*(1), 252–268. https://doi.org/10.1016/j.ejor.2003.12.010

Chang, H. C., & Tsai, H. P. (2011). Group RFM analysis as a novel framework to discover better customer consumption behavior. *Expert Systems with Applications*, *38*(12), 14499–14513. https://doi.org/10.1016/j.eswa.2011.05.034

Chen, S. (2018). Estimating Customer Lifetime Value Using Machine Learning Techniques. In *Data Mining*. InTech. https://doi.org/10.5772/intechopen.76990

Chen, Y. L., Kuo, M. H., Wu, S. Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications*, *8*(5), 241–251. https://doi.org/10.1016/j.elerap.2009.03.002

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3 PART 1), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

Chéron, E. J., & Kleinschmidt, E. J. (1985). A review of industrial market segmentation research and a proposal for an integrated segmentation framework. *International Journal of Research in Marketing*, *2*(2), 101–115. https://doi.org/10.1016/0167-8116(85)90027-8

Gentner, D., Stelzer, B., Ramosaj, B., & Brecht, L. (2018). Strategic Foresight of Future B2B Customer Opportunities through Machine Learning. *Technology Innovation Management Review*, *8*(10), 5–17. https://doi.org/10.22215/timreview/1189

Goller, S., Hogg, A., & Kalafatis, S. P. (2002). A new research agenda for business segmentation. *European Journal of Marketing*, *36*(1/2), 252–271. https://doi.org/10.1108/03090560210412782

Grewal, R., Lilien, G. L., Bharadwaj, S., Jindal, P., Kayande, U., Lusch, R. F., Mantrala, M., Palmatier, R. W., Rindfleisch, A., Scheer, L. K., Spekman, R., & Sridhar, S. (2015). Business-to-Business

Buying: Challenges and Opportunities. *Customer Needs and Solutions*, *2*(3), 193–208. https://doi.org/10.1007/s40547-015-0040-5

Hamerly, G., & Elkan, C. (2003). *Learning the k in k-means*.

Heras Barberá, S., Juan´, J., García-Pardo, J., Ramos-Garijo, R., Palomares, A., Julián, V., Rebollo, M., & Botti, V. (2006). *LNCS 4224 - CBR Model for the Intelligent Management of Customer Support Centers*.

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications*, *26*(2), 181–188. https://doi.org/10.1016/S0957-4174(03)00133-7

Kandeil, D. A., Saad, A. A., & Youssef, S. M. (2014). A two-phase clustering analysis for B2B customer segmentation. *Proceedings - 2014 International Conference on Intelligent Networking and Collaborative Systems, IEEE INCoS 2014*, 221–228. https://doi.org/10.1109/INCoS.2014.49

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57–63. https://doi.org/10.1016/j.procs.2010.12.011

Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, *31*(1), 101–107. https://doi.org/10.1016/j.eswa.2005.09.004

Kohavi, R., & Parekh, R. (2004). Visualizing RFM segmentation. *SIAM Proceedings Series*, 391–399. https://doi.org/10.1137/1.9781611972740.36

Kotler, P. (2006). *From Mass Marketing to Mass Customization*.

Lemon, K. N., & Mark, T. (2006). Basis of Customer Segmentation. *Journal of Relationship Marketing*, *5*(3), 55–69. https://doi.org/10.1300/J366v05n02_04

Liao, S. H. (2003). Knowledge management technologies and applications - Literature review from 1995 to 2002. In *Expert Systems with Applications* (Vol. 25, Issue 2, pp. 155–164). https://doi.org/10.1016/S0957-4174(03)00043-5

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Mesforoush, A., & Tarokh, M. J. (2013). Customer Profitability Segmentation for SMEs Case Study: Network Equipment Company. In *Int. J. Research in Industrial Engineering* (Vol. 2, Issue 1). www.nvlscience.com/index.php/ijrie

Mora Cortez, R., Højbjerg Clarke, A., & Freytag, P. V. (2021). B2B market segmentation: A systematic review and research agenda. In *Journal of Business Research* (Vol. 126, pp. 415–428). Elsevier Inc. https://doi.org/10.1016/j.jbusres.2020.12.070

Morgan, N. A., Whitler, K. A., Feng, H., & Chari, S. (2019). Research in marketing strategy. In *Journal of the Academy of Marketing Science* (Vol. 47, Issue 1, pp. 4–29). Springer New York LLC. https://doi.org/10.1007/s11747-018-0598-1

Müller, J. M., Pommeranz, B., Weisser, J., & Voigt, K. I. (2018). Digital, Social Media, and Mobile Marketing in industrial buying: Still in need of customer segmentation? Empirical evidence from Poland and Germany. *Industrial Marketing Management*, *73*, 70–83. https://doi.org/10.1016/j.indmarman.2018.01.033

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. In *Expert Systems with*

Applications (Vol. 36, Issue 2 PART 2, pp. 2592–2602). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2008.02.021

Parsa, T., Asiabi, K., & Tavoli, R. (n.d.). *A Review of Different Data Mining ....*

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence and Planning*, *35*(4), 544–559. https://doi.org/10.1108/MIP-11-2016-0210

Qadaki Moghaddam, S., Abdolvand, N., & Rajaee Harandi, S. (n.d.). *A RFMV Model and Customer Segmentation Based on Variety of Products*.

Rajkumar Venkatesan, & v. Kumar. (2004). *A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy Maximizing CLV*.

Ravi Kalakota, & Marcia Robinson. (1999). *e-Business Related papers Springer Text s in Business and Economics Elect ronic Commerce A Managerial and Social Net ... maidul islam Springer Text s in Business and Economics Elect ronic Commerce 2018 A Managerial and Social Net wo....*

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, *64*(4), 17–35. https://doi.org/10.1509/jmkg.64.4.17.18077

Riktesh Srivastava. (2016). *Identification of customer clusters using RFM Model: A case of diverse purchaser classification Currently working on text mining View project Riktesh Srivastava Identification of customer clusters using RFM Model: A case of diverse purchaser classification*. https://www.researchgate.net/publication/287815433

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Schaeffer, S. E., & Rodriguez Sanchez, S. V. (2020). Forecasting client retention — A machine-learning approach. *Journal of Retailing and Consumer Services*, *52*. https://doi.org/10.1016/j.jretconser.2019.101918

Sheikh, A., Ghanbarpour, T., & Gholamiangonabadi, D. (2019). A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. *Journal of Business-to-Business Marketing*, *26*(2), 197–207. https://doi.org/10.1080/1051712X.2019.1603420

Smith, B. (2018). *Artificial Intelligence in Europe: Portugal*.

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. In *Source: Journal of Marketing* (Vol. 21, Issue 1).

Thakur, R., & Workman, L. (2016). Customer portfolio management (CPM) for improved customer relationship management (CRM): Are your customers platinum, gold, silver, or bronze? *Journal of Business Research*, *69*(10), 4095–4102. https://doi.org/10.1016/J.JBUSRES.2016.03.042

Wedel, Michel., & Kamakura, W. A. (Wagner A. (2000). *Market segmentation: conceptual and methodological foundations*. Kluwer Academic.

Wu, H. H., Lin, S. Y., & Liu, C. W. (2014). Analysing patients' values by applying cluster analysis and LRFM model in a pediatric dental clinic in Taiwan. *Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/685495

Yong Seog Kim, Street, W. N., & Menczer, F. (2001). An evolutionary multi-objective local selection algorithm for customer targeting. *Proceedings of the IEEE Conference on Evolutionary Computation, ICEC*, *2*, 759–766. https://doi.org/10.1109/cec.2001.934266