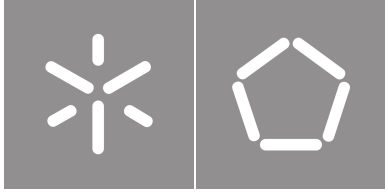
Universidade do Minho
Escola de Engenharia

Joel Costa Carvalho

MLOps aplicado à análise comportamental dos clientes no ambiente de um ERP

Julho de 2022



Universidade do Minho
Escola de Engenharia

Joel Costa Carvalho

MLOps aplicado à análise comportamental dos clientes no ambiente de um ERP

Dissertação de Mestrado
⟨Mestrado em Engenharia Informática⟩
⟨Inteligência Articial⟩

Trabalho efetuado sob a orientação de:
Rui Manuel Ribeiro Castro Mendes

Julho de 2022

Agradecimentos

Durante este percurso, recebi bastante suporte e ânimo para concretizar o desafio o melhor possível. Primeiramente, gostaria de agradecer à PRIMAVERA por proporcionar-me as condições ideais para o desenvolvimento da minha dissertação, especialmente aos meus supervisores, André Martins e Miguel Domingues. Singularmente ao André, pelos desafios constantes, pelas críticas construtivas, pela partilha de conhecimento e dedicação. Também agradecer ao meu Professor-Orientador da Universidade do Minho, Rui Mendes, pelo suporte e orientação constante.

Agradecer aos meus pais, Eduardo e Glória, pela paciência e conforto, à minha irmã Leonor e à minha avó Fernanda pelo amor. À Mariana pelas frases motivacionais e por acreditar sempre em mim.

Aos meus amigos de longa data, Bruno, Francisco, Henrique e Marco, pelo incentivo extra e amizade incondicional. Por último, um agradecimento particular ao meu amigo e colega de trabalho Fábio, por toda a motivação e companhia diária.

Joel Costa Carvalho

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.

Vila Verde, 3 de julho de 2022
(Local) (Data)

Joel Costa Carvalho
(Joel Costa Carvalho)

"Let's go invent tomorrow instead of worrying about what happened yesterday." (Steve Jobs)

Resumo

MLOps aplicado à análise comportamental dos clientes no ambiente de um ERP

A PRIMAVERA é uma empresa portuguesa pioneira no desenvolvimento de soluções de gestão para Windows, nomeadamente os ERP's *(Enterprise Resource Planning)*. Um ERP, sendo um software de gestão empresarial, envolve um grande volume de informação. Por este motivo, a extração de dados relevantes acerca dos clientes pode-se tornar complexa, agravando-se com o crescimento exponencial do volume de negócio.

O presente documento detalha todo o processo de criação de modelos baseados em Inteligência Artificial que diligenciam interpretar e prever a periodicidade e o comportamento financeiro dos clientes do ERP, com o propósito de tornar o negócio inteligente e permitir obter resultados complexos de modo perspicaz.

À vista disto, foi implementada uma solução baseada na classe de modelos *Buy 'Til You Die (BTYD)*, monitorizada recorrendo a *Machine Learning Operations*, capaz de analisar o desempenho dos clientes e produzirem previsões probabilísticas. Transversalmente, dispondo da aplicação do caso de estudo, *Customer Lifetime Value*, obtém-se a capacidade de evidenciar os melhores clientes, futurar valores de transações e receitas e identificar clientes em risco de abandono transacional *(churn)*.

Para concluir, este projeto permitiu ainda segmentar os clientes, potenciando a ligação com os mais leais e limitar custos associados a marketing mal distribuído, com a finalidade de auxiliar a empresa em estudos estatísticos e financeiros.

Palavras-chave: *Enterprise Resource Planning*, Inteligência Artificial, *Machine Learning Operations*, *Customer Lifetime Value*

Abstract

MLOps applied to customer behavior analysis in an ERP environment

PRIMAVERA is a pioneering Portuguese company in the development of management solutions for Windows, namely ERPs (Enterprise Resource Plannings). ERP, as a business management software, can involve a large volume of information. For that reason, the extraction of relevant data about the customers might be very complex, and it becomes worse with turnover growth.

This dissertation details every stage of the model creation process based on Artificial Intelligence to understand and predict the periodicity and financial behavior of ERP customers, to become the business smart, and allow complex results to be obtained insightfully.

Besides that, a solution based on the Buy 'Til You Die (BTYD) model class was implemented, monitored using Machine Learning Operations, capable of analyzing customer performance and producing probabilistic forecasts. Using the application of the case study, Customer Lifetime Value, the ability to highlight the best customers, future transaction values, and revenues and identify customers at risk of transactional churn abandonment is obtained.

In conclusion, this project also made it possible to segment customers, enhancing the connection with the most loyal and limiting costs associated with poorly distributed marketing, in order to assist the company in statistical and financial studies.

Keywords: Enterprise Resource Planning, Artificial Intelligence, Machine Learning, Customer Lifetime Value

Índice

Índice de Figuras

Índice de Tabelas

Glossário

add-ons Consiste em complementos para que os utilizadores possam estender e modificar um software [79]. Similar às extensões do Google Chrome. 3, 12

Big Data Conjunto de dados extremamente amplos e complexos, que tendem a crescer rapidamente. Dificilmente manipulados por métodos tradicionais. 48

CRM *Customer Relationship Management (CRM)* consiste numa abordagem empresarial para interpretar e influenciar o comportamento do cliente por meio de uma comunicação significativa para melhorar a aquisição, retenção, fidelidade e lucratividade do cliente. O objetivo é ir ao ínfimo detalhe nas relações com os clientes e maximizar o seu valor dentro da organização [29]. 15, 18

Data Lake Repositório de dados que contém dados de múltiplos formatos. Estes formatos podem manter o seu formato original (formatos em bruto) [36]. 48

Databricks A Databricks é uma empresa norte-americana de software empresarial, fundada pelos criadores do Apache Spark. A empresa também desenvolveu o Delta Lake e o Koalas, que consiste em projetos de *open-source* que abrangem *Data Engineering, Data Science* e *Machine Learning*. O software está direcionado para múltiplas indústrias, desde automóvel, educação, tecnológica, marketing, entre outros [20]. 29

Data Engineering Focado na colheita e análise dos dados, assiste no fluxo e no acesso da informação, concebe interfaces e mecanismos apropriados [40]. 23

Delta Lake Ferramenta *open-source* para a manipulação e gestão de *Data Lakes* [22]. 48

DevOps Ciclo de trabalho que visa melhorar a comunicação entre todas as fases da construção de software (desde o planeamento, construção, testes, lançamento e monitorização), auxiliando as equipas na gestão de versões e padronizam regras mantendo um alinhamento entre os intervenientes. 23, 24

Distribuição de Poisson Consiste numa variável aleatória discreta que determina probabilidade de uma série de eventos ocorrer num certo período de tempo, ou seja, se estes eventos ocorrem independentemente de quando ocorreu o último evento [14]. 19

Feature Store Repositório de *features* que permite reutilizar *features* já trabalhadas. Armazenadas como *features tables* [24]. 38

Hyperopt *Package* Python para otimização de hiperparâmetros *(tuning)*. Este procedimento consiste em atingir a melhor combinação de valores entre os hiperparâmetros, de forma a obter o máximo desempenho num período de tempo razoável. Considerado uma das parte mais complexas no desenvolvimento de modelos, tem papel de elevada importância na *accuracy* da previsão de um algoritmo. O *Databricks Runtime* é composta por uma versão otimizada e aprimorada do mesmo [5, 6]. 37, 41, 43, 72

Kaggle Plataforma online para toda a comunidade envolvida no ciclo de vida de *Machine Learning*. 10

Lifetimes *Package* Python para análise comportamental de clientes. Permite aplicar os modelos mais capacitados para a realização de previsões de compras e do calculo da métrica CLV [26] 19, 21, 22, 37, 38

On-premises software Software instalado e executado nos computadores dos clientes, toda a informação fica guardada localmente. 72

Siglas

AI Artificial Intelligence 5, 6, 8, 11, 13, 26, 34
API Application Programming Interface 48, 73
AutoML Automated Machine Learning 9, 10, 11, 34

BG Beta-Geometric 19, 20
BG/BB Beta-Geometric/Beta-Binominal 19, 20
BTYD Buy 'Til You Die 18, 31, 72

CI/CD Continuous Integration/Continuous Delivery 27
CLI Command-Line Interface 28
CLV Customer Lifetime Value xviii, 15, 16, 17, 18, 21, 22, 31, 32, 34, 36, 37, 42, 43, 51

Dev Development 23

ERP Enterprise Resource Planning 2, 3, 11, 34

GG Gamma-Gamma 20, 21, 43

KPI Key Performance Indicator 27

MAE Mean Absolute Error 23
MBG Modified Beta-Geometric 19, 20
ML Machine Learning 6, 7, 9, 10, 11, 13, 23, 24, 26, 28, 29, 32, 46
MLE Maximum Likelihood Estimation 70
MLOps Machine Learning Operations 1, 23, 24, 25, 26, 28, 31, 34, 46

MSE Mean Square Error 23, 41, 43, 47, 60, 62, 73

NBD Negative Binomial Distribution 19, 20

Ops Operations 23

PALOP Países Africanos de Língua Oficial Portuguesa 2

RFM Recency, Frequency and Monetary 18, 21, 31, 32, 37
RH Recursos Humanos 2
RMSE Root Mean Square Error 23
ROI Return on investment 18

TI Tecnologia da Informação 11

UCs Use Cases 13
UI User Interface 29

WRFM Weighted RFM 18

1 Introdução

O estudo e a implementação da dissertação foi realizada em contexto empresarial e tem como finalidade a aplicação de MLOps no estudo comportamental dos clientes em larga escala, de forma a, reter dinheiro de forma astuciosa e identificar os principais clientes.

Este capítulo introdutório contextualiza ao leitor o propósito, motivações, objetivos, planeamento e estrutura do documento. A secção 1.1 identifica o problema na PRIMAVERA e detalha o intuito da empresa no mercado. A secção 1.2 delineia os esforços principais que impulsionam este trabalho e apresenta o problema abordado por esta tese. A secção 1.3 pormenoriza os objetivos para a concretização do projeto. Por fim, a secção 1.4 explica a estrutura do restante documento.

1.1 Contexto

No mundo empresarial, tal como na PRIMAVERA, problemas como a identificação dos melhores clientes, a perda de dinheiro em marketing mal atribuído, o abandono dos clientes ou preços dos produtos demasiado custosos são constantes e conduzem a contrariedades não desejadas. Então para que haja suporte neste problema e estas contrariedades sejam minimizadas, hoje em dia, as empresas recorrem ao poderio das novas tecnologias, nomeadamente da *Artificial Intelligence*.

Na PRIMAVERA o objetivo passa por colmatar essas adversidades garantindo o sucesso empresarial. Assim sendo, o processo caracteriza-se por estudar e comparar a finalidade dos diversos casos de uso aplicáveis à instituição, desde, segmentação de clientes ou produtos, sistemas de recomendação, *churn* e prevenção de fraude(s), e por fim garantir o desenvolvimento de um modelo capaz, com base em *Machine Learning Operations (MLOps)*.

1.1.1 PRIMAVERA BSS

A PRIMAVERA é uma empresa tecnológica portuguesa que disponibiliza soluções empresariais de gestão através das suas propostas inovadoras e que se enquadra às exigências fiscais e legais de cada país e capacidade de adaptação da oferta ao contexto particular de cada cliente, suportada em plataformas

tecnológicas extensíveis amplamente configuráveis. Tem como missão simplificar a vida nas organizações, aumentando a criação de valor. De modo a, explorar possibilidades, transpor barreiras e inovar a gestão empresarial.

Fundada no ano de 1993, a PRIMAVERA foi a primeira empresa a desenvolver soluções de gestão para Windows, com o objetivo de inovar, simplificar e acelerar processos de negócio. Presentes na Península Ibérica, com sede em Braga e escritórios em Lisboa, Leiria e Madrid. Relativamente ao continente africano estão presentes em 3 dos 5 Países Africanos de Língua Oficial Portuguesa (PALOP), em Angola com sede em Luanda, em Cabo Verde com sede na Cidade da Praia e em Moçambique com sede em Maputo [9].

A Figura 1 apresenta de forma diagramática o peso das vendas da PRIMAVERA e alguns dos seus clientes por regiões.

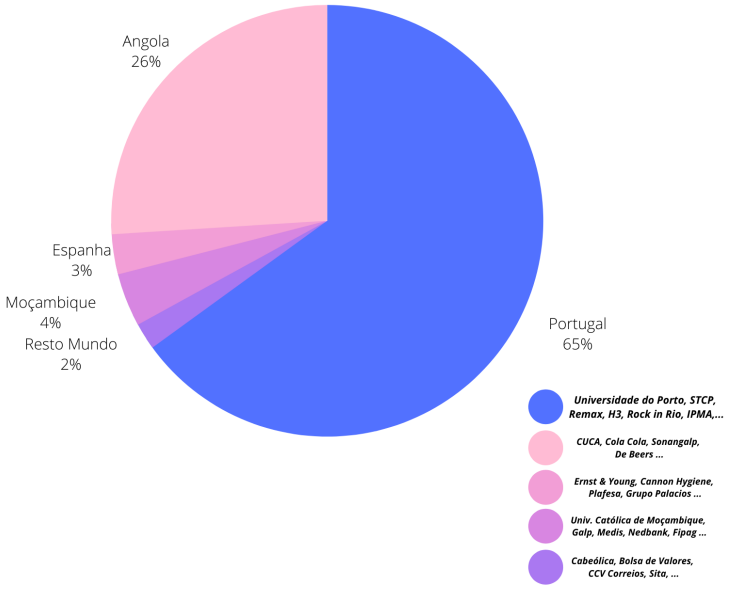


Figura 1: Peso das vendas e clientes [8, 7]

1.1.2 ERP

Enterprise Resource Planning (ERP) é um software de gestão de processos de negócio que gere e integra as atividades de finanças, cadeia de fornecimento, operações, comércio, relatórios, fabrico e Recursos Humanos (RH) de uma empresa, desta forma, toda a informação de vários sistemas estão agregados num sistema unificado [57]. Além disso, sem a utilização de um sistema ERP, a empresa terá que lidar com muitos fornecedores de software, aumentando custos com licenças, suporte técnico, servidores, entre outros [3].

1.1.2.1 Vantagens

A Tabela 1 inúmera os benefícios principais da utilização de um ERP:

Tabela 1: Vantagens da utilização de um ERP

Vantagens
[3] Apoiar na comunicação entre os intervenientes
[3] Evitar anomalias, como: processamento de salários
[3] Apoiar na tomada de decisões, como no auxílio da elaboração de estratégias
[3] Agilizar a obtenção de dados referentes a determinados cenários, interna e externamente
[3] Diminuir o tempo de entrega do(s) produto(s) ou serviço ao cliente
[3] Ajudar a lidar com grandes volumes de informação
[3] Melhor gestão de tarefas e fluxo de trabalho, evitando redundância de atividades
[3] Manter ou atualizar a legislação em vigor mediante um estado, ou país
[67] Agilizar a execução de processos, minimizando atrasos e a quantidade dos mesmos
[67] Aumentar a escalabilidade: sistema estruturado e modular com a possibilidade de adicionar *add-ons*

1.1.2.2 Desvantagens

A Tabela 2 inúmera os inconvenientes principais da utilização de um ERP:

Tabela 2: Desvantagens da utilização de um ERP

Desvantagens
[3] Risco de prejuízo ou queda de desempenho com erros inesperados do sistema
[3] Dependência, possíveis problemas com suporte e manutenção (ex: encerramento das atividades do "dono"do ERP)
[3] Adaptação e treino por parte dos colaboradores poderá levar a perdas de tempo superiores às esperadas
[3] O sistema pode exigir mudanças em determinados aspetos
[3] Poderá ser necessário realizar um estudo para perceber se a solução oferece uma relação custo-benefício positivo
[3] Ao longo do tempo, atualizações e adição de módulos podem levar a alguma lentidão
[67] Custos elevados com personalização e implementação, principalmente para pequenas empresas
[67] Complexidade acrescida em alguns módulos

Apesar destas complicações, comum a qualquer sistema ou software, o uso do sistema de gestão ERP garante maior e melhor estabilidade de negócio.

1.2 Motivação

Um dos pontos fortes dos produtos da PRIMAVERA está no apoio à gestão que coloca à disposição dos seus clientes e permitem automatizar e configurar, conforme o perfil de cada utilizador, um conjunto de relatórios e *dashboards* sobre os dados existentes. Contudo, num mundo cada vez mais complexo, não

basta analisar os dados do passado, é necessário projetar a evolução da atividade da empresa, detetar erros, desvios ou oportunidades bem como automatizar processos.

A PRIMAVERA tem desenvolvido vários projetos ao abrigo da iniciativa "ERP Inteligente" que visa a construção de soluções progressivamente mais autónomas, que permitam aos seus clientes focarem-se essencialmente na estratégia das suas empresas, dedicando menos energia à gestão operacional do negócio. A evolução tecnológica da *Artificial Intelligence* pode ser explorada no sentido de tornar as soluções do ERP mais inteligentes.

Importa que com o desenvolvimento industrial e tecnológico mais recente, sobretudo das plataformas de *Artificial Intelligence* e dos algoritmos de *Machine Learning*, surja um novo impulso para a modernização das áreas mais tradicionais da gestão das organizações.

1.3 Objetivos

De forma a garantir o sucesso deste projeto, é necessário detetar os tópicos mais relevantes, sendo assim, é expectável que a listagem seguinte seja cumprida:

- Compreensão da importância da aplicabilidade de *Artificial Intelligence* num ERP;
- Interpretação os conceitos de MLOps e AutoML;
- Estudo de casos de uso e identificação caso de uso com mais utilidade para a empresa;
- Interpretação caso de uso selecionado *(Customer Lifetime Value)*;
- Estudo do MLOps e compreender o *workflow* de MLflow;
- Análise da Fábrica de Dados da PRIMAVERA;
- Desenvolvimento prático, exploração de MLOps e produção dos modelos;
- Revisão, melhoria e descrição escrita do trabalho desenvolvido.

1.4 Estrutura

Este documento de dissertação encontra-se dividido em 7 capítulos, incluindo o capítulo atual (1). No capítulo 2 é apresentado o estado de arte e uma revisão da literatura, de forma a contextualizar o conteúdo mais relevante para a dissertação. O capítulo 3 descreve os problemas associados ao projeto e descrimina o planeamento das tarefas realizadas e respetivo diagrama Gantt. O capítulo 4 contém toda a informação pormenorizada do processo de desenvolvido e implementação dos modelos. O capítulo 5 descrever os resultados e abordagens tentadas relacionadas com a implementação dos modelos. Por último, o capítulo 6 concluí o projeto desenvolvido e aponta dificuldades e objetivos futuros.

2

Estado de Arte

O pospositivo capítulo 2, irá descrever fundamentos teóricos e conceitos científicos relacionados com o núcleo do desenvolvimento do projeto. Posto isto, a secção 2.1 descreve o conceito de *Artificial Intelligence* e a sua estrutura. A secção 2.2 detalha o propósito de *Machine Learning*, explicação do processo de aprendizagem, tipos de modelos existentes e introdução a conceitos básicos. A secção 2.3 introduz conceitos relacionados com o AutoML, fundamenta a sua utilização e componentes principais. A secção 2.4 baseia-se no estudo do impacto de *Artificial Intelligence* aplicação a um ERP. A secção 2.5 específica o caso de uso implementado sobre a organização. A secção 2.6 engloba todo o processo envolvente na utilização de MLOps, desde a sua estrutura às ferramentas inerentes. Finalmente a secção 2.7 sumariza este capítulo.

2.1 *Artificial Intelligence*

Artificial Intelligence (AI) resume-se a um ramo das ciências da computação, que foca essencialmente na aplicação de inteligência a sistemas computacionais. Por outras palavras, AI permite construir máquinas inteligentes através da utilização de algoritmos, de modo a atingir um propósito ou uma solução [61].

AI é caracterizada pela adaptabilidade a múltiplas indústrias, portanto é aplicada por diversas empresas. Saúde, Finanças, Educação, Transporte, Desporto e Segurança são exemplos de indústrias ativas no mercado [76].

2.1.1 Estrutura

Segundo Tom Taulli [76], em representação de uma estrutura não pormenorizada de AI, observe-se a Figura 2 que representa as duas grandes categorias internas da *Artificial Intelligence: Machine Learning* e *Deep Learning*.

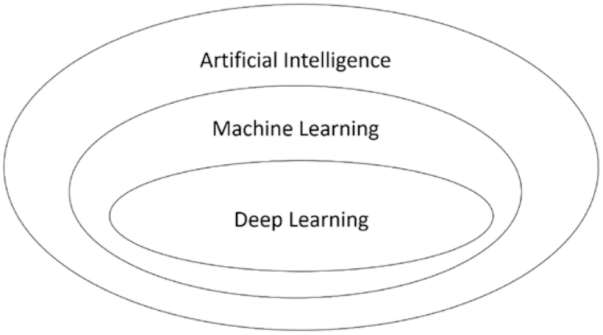


Figura 2: Estrutura *high-level* de IA [76]

É inegável que ao detalhar infimamente, diversos autores fundamentam uma estrutura para AI bem mais composta. Podendo ainda ser representada por mais camadas, como: *Robotics, Planning, Vision, Speech, Natural Language Processing, Experts Systems*, entre outros [19, 71].

2.2 *Machine Learning*

Retrocedendo séculos, *Machine Learning (ML)* surgiu com base na teoria que os computadores convencionais podem aprender determinadas tarefas sem recurso à programação. O processo iterativo de ML é crucial para que as máquinas se adaptem a novos dados, permitindo a otimização de cálculos e fornecer resultados adequados. Tipicamente, o software desenvolvido baseia-se num conjunto de regras e condições pré-definidas para obter resultados fornecidos pelas máquinas. Então se as máquinas reconhecerem padrões e disponibilizarem e explicarem resultados? Consideramos isto ML, ou seja, ML é um método que através de dados, mecanismos e algoritmos produzem resultados com o mínimo de intervenção humana. Contudo, ML depende dos dados de entrada e disponibiliza um resultado baseado numa determinada tarefa, este formato de entrada necessitam de ser laborados com base no objetivo final. Estes dados de entrada são denominados *features* [19].

2.2.1 Processo de Aprendizagem

Veja-se a Figura 3, que apresenta de forma diagramática um paralelismo bastante comum e que todos cruzam na vida. Quando estudamos na escola aprendemos com base nos professores, isto chamamos treino *(training)*, no final da aprendizagem necessitamos de testar esse treino *(testing)*, para validarem os nossos conhecimentos. O resultado final *(score)* é baseado numa avaliação *(evaluation)*, normalmente a avaliação é realizada considerando um limite *(threshold)* para ser aprovado *(baseline)*, mediante este *score* determinamos se é necessário reformular o trabalho ou avançar para o próximo passo *(deployment)*. Essencialmente esta é considerada a base da aprendizagem de uma máquina [19].

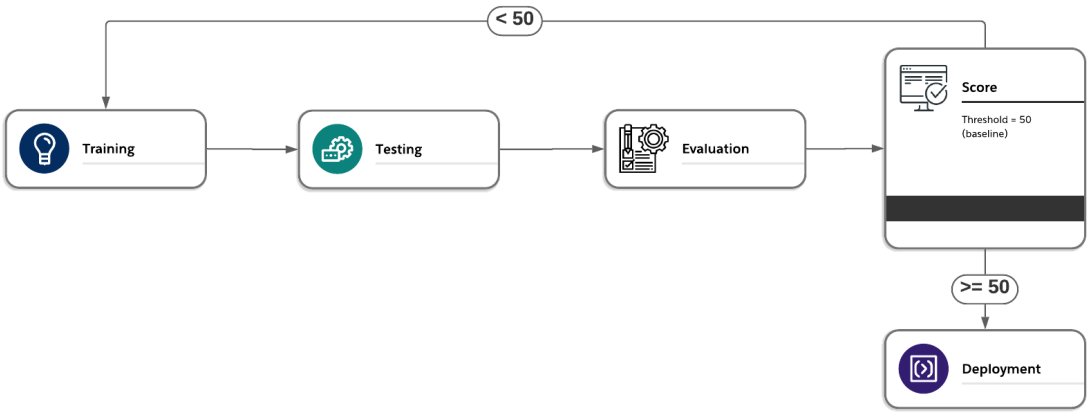


Figura 3: Arquitetura do Processo de Aprendizagem

2.2.2 Tipos de Modelos

De acordo com Emmanuel Raj, atualmente, encontram-se cerca de 15 tipos de técnicas para ML, categorizadas em 4 tipos, intituladas de *Learning Models, Hybrid Models, Statistical Models* e *Human-In-The-Loop Models*, a seguinte matriz reflete o enquadramento de cada uma delas, como visível na Figura 4 [65].

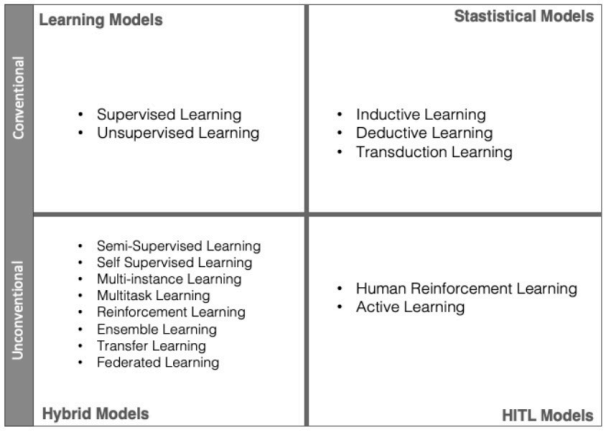


Figura 4: Matriz dos Tipos de Modelo [65]

2.2.2.1 *Learning Models*

Em virtude do contexto do desenvolvimento deste projeto, a subseção seguinte detalha o que consiste os *Learning Models* que identificamos na matriz anterior.

Supervised Learning Traduzindo e como o nome sugere, é uma aprendizagem com supervisão baseada num resultado específico ou devidamente identificado. Essencialmente, os dados estão inicialmente rotulados com o objetivo final. O objetivo final pode ser um atributo numérico, binário (sim/não) ou

multi-classe até 2 resultados (email com/sem spam) [19]. Este paradigma de aprendizagem está, normalmente, dividido em duas categorias [48]:

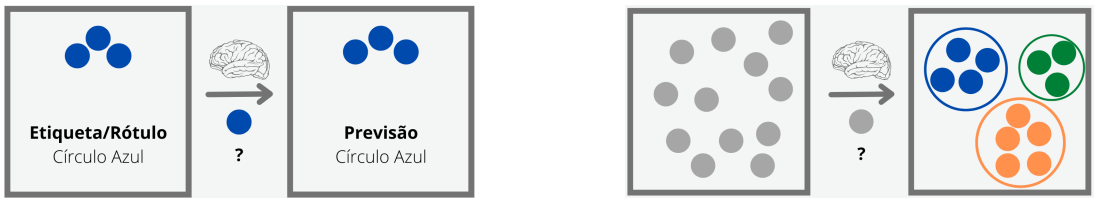
- ***Classification***: Processo de categorizar um determinado conjunto de dados em classes, basicamente reconhece padrões num *dataset* e tira conclusões de como deve ser etiquetadas (Ex: Deteção de fraude).
- ***Regression***: Usado quando o resultado consiste num valor real ou contínuo (Ex: Previsão de casos de COVID-19).

Unsupervised Learning Neste contexto os dados não se encontram etiquetados/padronizados, assim sendo o objetivo da traduz-se na identificação de padrões, deduzindo estruturas e relação entre as *features* presentes no *dataset*. Inclusivamente, pode ser utilizado para expor regras que definem grupos, como segmentação de clientes. Este paradigma de aprendizagem está, normalmente, dividido em duas categorias [19]:

- ***Clustering***: Consiste em organizar os dados em grupos concisos, como agrupar clientes que compram o mesmo produto.
- ***Association***: Consiste em associar comportamentos, como agrupar clientes que compram o mesmo produto e também não gostam de um determinado produto.

O processo de aprendizagem não é tão linear, comparativamente com o *supervised learning* e torna o processo de AI mais complexo de solucionar. Contudo, se a solução for bem laborada e concluída, obtém-se um produto bem mais poderoso [63].

A Figura 5 exibe, de forma diagramática, duas figuras representadas (a) e (b), em que a primeira descreve o *supervised learning*, em que os dados iniciais contém um rótulo predefinido (Círculo Azul) e o resultado previsto após a aplicação do modelo corresponde a esse mesmo rótulo e a última figura descreve o *unsupervised learning*, em que os elementos não etiquetados (bolas cinzentas) e após a aplicação desta técnica de aprendizagem, os elementos encontram-se todos agrupados por cores, salientado o poderio e a vantagem da mesma.



(a) Representação de *Supervised Learning*

(b) Representação de *Unsupervised Learning*

Figura 5: Comparação entre os *Learning Models*

2.2.3 Conceitos

A listagem seguinte, clarifica ao leitor alguns conceitos inerentes ao estudo de *Machine Learning* aplicados ao tema desenvolvido:

- ***Feature***: Propriedade mensurável de um objeto em análise. Num conjunto de dados, como um *dataset*, estas propriedades são representadas como colunas [25]. Caso o *dataset* seja composto por professores, o género, número mecanográfico ou a área de ensino são exemplos de *features*.
- ***Prediction***: Método, tal como o nome indica, que realiza uma previsão com base em eventos passados.
- ***Accuracy***: Métrica/nome que identifica a precisão de um modelo;
- ***Missing Values***: Valores indisponíveis em alguns atributos/variáveis, isto acontece por diversas razões tais como erros na recolha ou elementos opcionais no preenchimento dos dados [19]. Para o seu tratamento variadas técnicas reparam estes valores, como: eliminação, utilização de um valor *default*, substituição por um das medidas de tendências centrais (média, mediana ou moda), *Random Forest*, *Listwise deletion*, *Pairwise deletion*, *Maximum likelihood*, entre muitos mais [50].
- **Normalização**: Técnica que assegura que as *features* numéricas utilizadas pelo modelo têm o mesmo peso, de forma a manter uma representação neutra [19]. Isto significa que, os dados são reescalados para um intervalo como de $[0, 1]$ ou $[-1, 1]$.
- ***Outlier***: Corresponde a valores atípicos (muito pequenos ou muito grandes) num determinado conjunto de dados, ou seja, é uma observação que apresenta um grande afastamento das demais da série, ou que é inconsistente [2].

2.3 AutoML

Automated Machine Learning (AutoML) disponibiliza métodos e processos para desenvolver ML, de forma a melhorar a eficiência e acelerar o processo, independentemente do nível de conhecimento. Os algoritmos trabalham sobre os dados e obtêm certos padrões, no final, o *output* disponibiliza o resultado de todas as abordagens realizadas [19]. A sua utilização torna ML mais acessível, aplica diversos modelos/métodos sequencialmente e acelera a pesquisa e o desenvolvimento.

Considerado bastante diversificado, este contém pacotes e métodos direcionados a diferentes elementos da organização, segue algumas etapas automatizadas com a sua utilização:

- Processamentos dos dados (interpretação, preparação e visualização);
- Seleção do modelo;

- Otimização de hiperparâmetros;
- Avaliação das métricas;
- Monitorização e análise de resultados.

A Figura 6 representa, genericamente, as etapas internas de *Machine Learning* automatizadas, desde a preparação dos dados até à monitorização dos resultados.

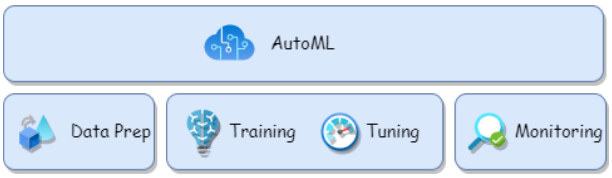


Figura 6: Fluxo genérico do processo do AutoML

2.3.1 Utilidade

Tendo em consideração pesquisas do Kaggle, informação confusa e desorganizada está no topo dos problemas identificados pela população. Quando um *dataset* se encontra nessas condições e necessita de algum tratamento especial através de modelos ML, o tempo gasto para a limpeza, manipulação e tratamento das *features* é consideravelmente elevado, neste sentido o *data scientist* irá consumir demasiado tempo laboral Como a otimização de hiperparâmetros, treino, validação e testes são considerados etapas cruciais para o desempenho dos modelos, o AutoML oferece resultados robustos, de forma automática, para a construção de modelos ML [19].

2.3.2 Componentes Principais

Essencialmente, AutoML contém os seguintes componentes como núcleo funcional [19].

1. ***Pré-processamento automatizado dos recursos:*** No tratamento de problemas ligados a ML, geralmente existem vários tipos de dados em que cada um deles deve ser tratado antes de ser treinado. Como:

    a) Tratamento de dados numéricos, aplicação de um intervalo numa determinada escala.

    b) Tratamento de dados categóricos, aplicando métodos como *label encoding*, *binary encoding* ou *one-hot encoding*.

    c) Tratamento de imagens, múltiplas transformações como segmentação.

    Quando um determinado *dataset* é composto por centenas ou milhares de *features* o processo manual torna-se demasiado lento. AutoML automatiza este processo.

2. ***Seleção do algoritmo automatizado:*** Uma vez realizado processamento das *features*, é necessário identificar os algoritmos adequados para o treino e avaliação do modelo. Como:

    a) Com algoritmos de *clustering*, como k-means, k-medoids, DBSCAN, entre outros. O utilizador poderá não estar familiarizado com todos eles e não identificar o algoritmo que lhe oferece melhor performance. Com a utilização do AutoML, este processo torna-se simples e sugere qual(ais) algoritmo(s) utilizar.

3. ***Otimização de hiperparâmetros:*** Qualquer que seja o algoritmo de ML tem um ou mais hiperparâmetros, estes hiperparâmetros necessitam de ser afinados de modo a melhorar a precisão do modelo. Como:

    a) Definir uma escala dos hiperparâmetros desejados, sugerindo a melhora da performance do modelo.

2.4 *Artificial Intelligence* Aplicada a um ERP

Este capítulo descreve como a AI está em constante transformação nos sistemas ERP, a evolução ao longo das décadas, a aplicabilidade da implementação em alguns módulos e que problemas soluciona.

O sistema de gestão ERP é um modelo de gestão empresarial moderno baseado em Tecnologia da Informação (TI) [28], a sua utilização tornou-se uma parte fundamental para o negócio de uma organização, abrilhantando a performance das empresas. Com o impacto dos ERPs, independentemente da dimensão da indústria, e o desenvolvimento da AI o melhor dos dois mundos irão afetar positivamente as operações diárias, melhorando substancialmente as tarefas humanas. À vista disso, o desenvolvimento de novas tecnologias também impulsiona a necessidade de reduzir o custo das operações do negócio, potenciando a eficiência das operações e a competitividade entre as organizações [39].

Em 1993, Gadallah e Elmaraghy citaram *"Artificial intelligence has just begun to appear in ERP applications and the addition of the new technology is still a relatively new phenomenon, but the possibilities artificial intelligence adds to the ERP systems must be said to be unlimited"* [39, 35], demonstrando o potencial da AI integrada a um ERP.

A cada dia passado, a AI transforma as tarefas diárias mais básicas, de maneira a suprimir os erros humanos [39].

O conceito de *Enterprise Resource Planning (ERP)* remonta aos anos 60, porém durante as últimas décadas, estas soluções têm evoluído mediante o poder dos dados das organizações. Com o objetivo de tornar esses dados em informação relevante, potenciando a relação com o cliente, este processo torna-se exaustivo e neste ponto é que AI emerge, assistindo em [39]:

- Automatizar e melhorar tarefas complexas;
- Avaliar dados em tempo-real;

- Ajustar comportamentos;
- Otimizar a gestão de ativos;
- Melhorar a performance operacional;
- Diminuir tempos de resposta;
- Melhorar eficiência e precisão.

Observando na Figura 7, nos anos 60, as organizações desenvolveram e implementaram sistemas de computação centralizados, principalmente automatizando o sistema de controlo de inventário utilizando o *Inventory Control Package*. Na década de 70, surgiu o *Material Requirements Planning*, que envolveu o planeamento do produto e os requisitos fundamentais baseado no cronograma de produção [67]. Na década posterior, foram introduzidos processos de fabrico. No início da década de 90, surgiram os primeiros ERPS, estes seguiam a base dos anteriores e introduziram no seguimento do negócio, os processos de distribuição, finanças, recursos humanos, transporte, manutenção e entre outros. Por fim, no decorrer dos anos, novos módulos e funções (*add-ons*) foram adicionadas, dando origem ao *Extended ERP* [67].

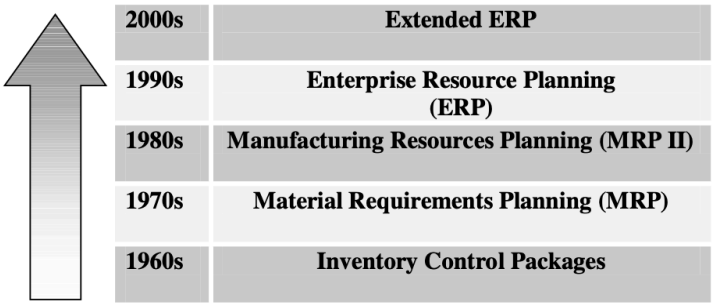


Figura 7: Evolução do sistema ERP [67]

Uma organização mantendo este nível de sistema, garante um fluxo de trabalho eficiente tirando o máximo valor do trabalho das máquinas e dos humanos, essencial para melhorar a eficiência e ter uma vantagem competitiva no mercado atual.

Em representação de alguns módulos, a Figura 8 representa de forma esquemática 3 camadas no contexto da gestão de negócio. A camada inferior apresenta os processos físicos, desde compra, produção e vendas, a camada interna corresponde aos módulos de RH e Finanças, como gestão de encomendas, gestão de qualidade e gestão de vendas, finalmente a camada superior representa a informação em formato de *dashboards* [41].

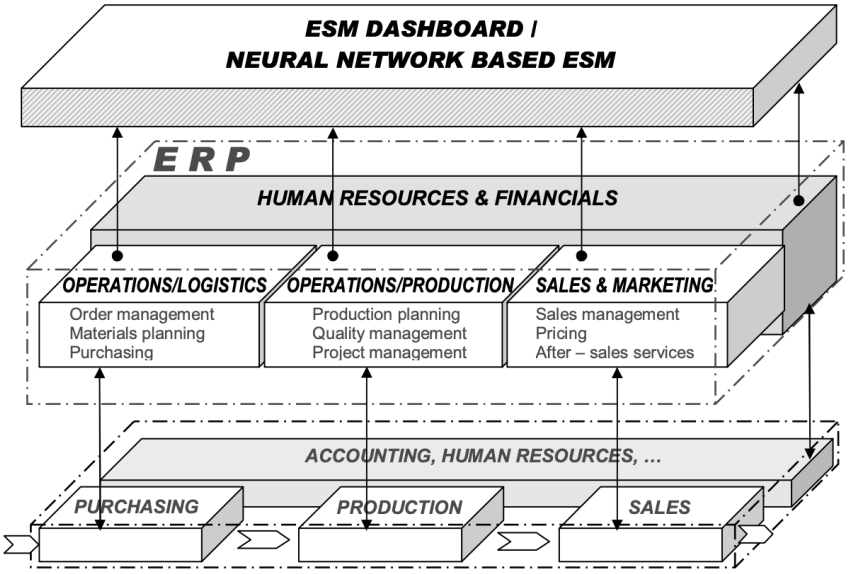


Figura 8: Processo de negócio ERP com os módulos de RH e Finanças [41]

Em suma, a utilização de um sistema de gestão ERP moderno espelha-se como a espinha-dorsal *(backbone)* para uma infinidade de empresas. Contudo, a sua introdução consiste numa grande mudança na gestão empresarial [28], nomeadamente para pequenas e médias empresas, que por vezes a sua implementação poderá causar esforço e um reajuste na organização [41].

2.4.1 *Use Cases* aplicados a um ERP

Com base nos *Use Cases (UCs)* disponibilizados [21, 66], a subsecção subsequente descreve e exemplifica uma grande diversidade de UCs que podem ser aplicados a um ERP, aperfeiçoando o desempenho e o "raciocínio"de um sistema de gestão.

Churn Ex: Quantos clientes a empresa perdeu ou irá perder no próximo ano ou trimestre?

É uma taxa/métrica de rotatividade, que permite identificar o número de indivíduos que saem de uma determinada organização. Por exemplo, quantos deles a empresa perdeu ou irá perder por mês, trimestre ou ano? Consiste em analisar informação do cliente, tal como: histórico de compras, produtos utilizados, entre outros fatores de risco, de forma a entender os clientes e criar experiências personalizadas [66].

Cognitive Robotic Process Automation Ex: Chatbots.

Tornar o workflow inteligente, isto é, com a introdução e poder de AI e ML tomar decisões de forma automática. Decrementando do tempo consumido ou eliminando repetição de eventos, aumentando significativamente a produtividade. Consequentemente, torna o sistema mais inteligente, adaptado, confiável e na vanguarda da tecnologia [60].

Customer Lifetime Value Ex: Emails de marketing personalizados e ofertas especiais.

Medida que estima a receita média gerada por cada cliente num determinado espaço temporal. Realiza uma distinção entre clientes, prevendo os clientes que irão gerar mais lucro. Maximiza o negócio e otimiza a relação com os clientes [64].

Customer Segmentation Ex: Antecipar vendas de uma cidade.

Criação grupos de clientes de acordo com a sua área geográfica, encontrar padrões significativos, entender quais são os pontos chaves dos clientes e que produtos os cativam mais, com esta segmentação fina rapidamente melhoramos a experiência dos utilizadores [62].

Fraud Detection and Prevention Ex: Monitorizar constantemente o volume de dados.

Identifica padrões não usuais e atua de forma preventiva sobre os impactos. Ações fraudulentas podem destruir receitas e dificulta entender os clientes legítimos. Com estas táticas, as empresas podem agir proativamente minimizando o impacto financeiro negativo [72].

Next Best Action Ex: Antecipar necessidades.

Aplicar uma certa ação, num exato momento e direcionada a um cliente perfeito. Usar padrões de eventos reais, comportamento de compra, interações de mídia social e outros insights para decidir quais ações devem ser tomadas para cada cliente, a fim de aumentar a fidelidade, intensificar as interações com sua organização e gerar receitas [55].

Pricing Optimization Ex: Modelos de preços diários ou personalizados por cliente.

Preços baixos aumentam o risco das margens de negócio e preços elevados podem afugentar clientes. Então é importante otimizar preços olhando para a área geográfica, as variações do mercado ou preferências [15].

Text Mining Ex: Extrair conteúdo de formulários, pedidos e chats.

Processo de transformação de texto não estruturado, isto é, compreender e adquirir conhecimento de conteúdo textual [52].

Cross-Selling and Up-Selling Ex: Oferecer produtos personalizados de modo a fortalecer a relação com o cliente.

Dar relevância aos clientes, recolher informação sobre histórico e comportamento de compras, independentemente da fonte ou do formato dos dados, após isso oferecer produtos personalizados de modo a fortalecer a relação com o cliente [44].

Recommendation Engines Ex: SBC: Sugestão de um livro semelhante a um já lido por mim e FC: Sugestão de um livro lido por um utilizador com um perfil similar ao meu.

Permitem prever ou mostrar itens que o utilizadores poderá estar interessado. Este UC subdivide-se em dois componentes: Sistemas Baseados em Conteúdo (SBC) e Filtragem Colaborativa (FC). Contudo, há plataformas capazes de sugerir um modelo híbrido (Netflix) [43].

2.5 *Customer Lifetime Value*

Baseado na pesquisa anterior e na opinião dos interessados, o *use case* nomeado para perquirição foi o *Customer Lifetime Value (CLV)*.

Com base em diversos estudos, está provado ser mais dispendioso adquirir do que reter clientes [49, 53, 74]. Consequentemente, avaliar os clientes e manter o valor dos clientes torna-se um fator crítico que decide o sucesso ou insucesso do negócio. A investigação debruça-se sobre a análise de segmentação de clientes e melhorar o marketing de forma apropriada de forma a potenciar o valor dos clientes [12]. Atualmente observamos maior complexidade e competitividade de negócio, levando as empresas a inovarem as tecnologias retendo os melhores clientes e consequentemente satisfazê-los [12].

Perante isto, surge o CLV surge do problema de CRM [29] e com o objetivo estimar a faturação esperada por cada cliente num determinado período [78]. Além disso, esta métrica permite [64]:

- Utilizar informação vinda das decisões de marketing, de modo a aplicar recursos apropriadamente;
- Segmentar de clientes, com a finalidade de identificar os clientes mais e menos rentáveis;
- Compreender o balanço atual do negócio;
- Estimar a quantidade de compras por cliente num determinado período;
- Estimar a probabilidade de um cliente se tornar permanentemente inativo.

Uma diversidade enorme de empresas recorre ao CLV regularmente para controlar e supervisionar as estratégias e avaliar o sucesso do seu mercado [11]. Por exemplo, em outubro de 2005, a empresa eBAY contabilizou 168 milhões de clientes registados, mas apenas 68 milhões (aprox. 40%) é que continuavam ativos dentro da aplicação [51]. Da mesma forma, a Netflix percebeu que alguns clientes são impacientes, e que aproximadamente após 2 anos de subscrição cancelam o plano, por estatísticas sobre a análise comportamental dos clientes, a Netflix calcula que reduziu a taxa de cancelamento em 4% [38].

Ao longo dos anos, o CLV tem recebido um aumento de atenção. Recentemente, a avaliação ou previsão desta métrica tornou-se uma questão fundamental para a pesquisa de marketing e vários estudos sobre a previsão do CLV foram propostos [78].

Importa salientar que uma má abordagem conduz o modelo a uma previsão enfraquecida e poderá levar, consequentemente, a uma má segmentação de clientes, ao alocar clientes fora do seu mercado resulta num aumento significativo dos custos de marketing [78]. Desta forma, a Microsoft [45], recomenda que para o cálculo do CLV dos próximos 12 meses seja necessário entre 18 a 24 meses de histórico.

Veja-se a Figura 9 como representação do objetivo do CLV, a dimensão dos círculos representa o valor despendido, de outro modo, passa por estimar quando e que valor cada cliente irá gastar após a data limite pré-definida (*thresold date*).

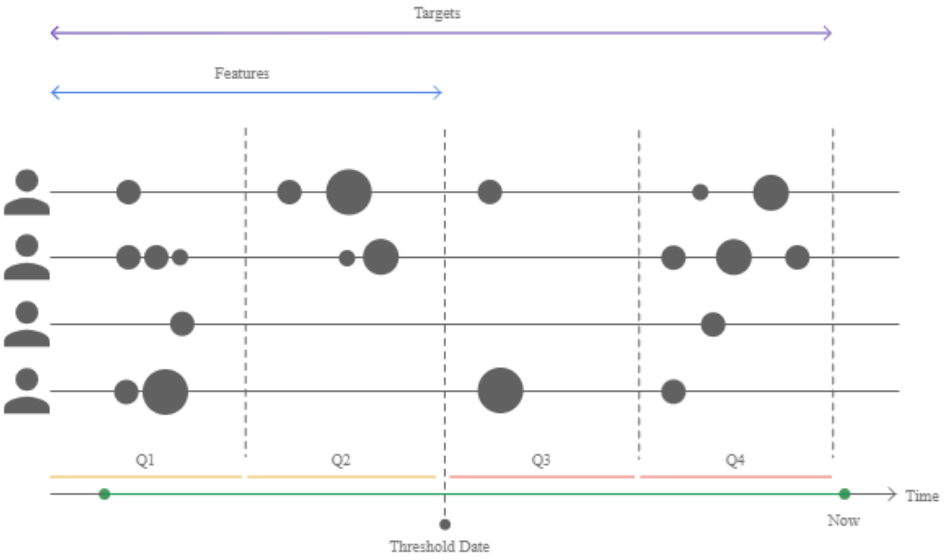


Figura 9: Objetivo do CLV [64]

Em suma, a capacidade de prever o momento de transações futuras e a receita gerada, pode ser crucial para acelerar e otimizar o negócio.

2.5.1 Cálculo do CLV

Tipicamente, permite diferenciar clientes através dos gastos monetários e identificar quais são os mais valiosos. A dificuldade passa por prever as receitas futuras quando o tempo ou o lucro das futuras operações são desconhecidas [11]. Apesar de existir determinadas variações, Gupta e outros autores [16, 68], propuseram o cálculo detalhado do CLV representado por [11, 42]:

$$CLV = \sum_{t=0}^{T} \frac{(P_t - C_t)r_t}{(1+i)^t} - AC$$

Onde:

- P_t - preço pago pelo cliente num determinado tempo t
- C_t - custo direto de atendimento ao cliente
- r_t - taxa de desconto ou custo de capital para a empresa
- i - taxa de desconto ou custo de capital para a empresa
- AC - custo de aquisição
- T - tempo estimado de CLV

De forma menos detalhada, é possível representar por:

$$CLV = \sum_{t=0}^{T} \frac{(ValorMedio_t - TaxaRetencao_t)}{(1 + TaxaDesconto)^t}$$

2.5.2 Métodos para Cálculo

Os modelos seguintes detalham estratégias para o cálculo do CLV com base no histórico de vendas. Contudo, por serem bastante elementares não têm capacidade de realizar previsões futuras.

2.5.2.1 Modelo de Agregação

Considerado o mais simplista e mais antigo método para o cálculo do CLV. Este método assume um valor médio gasto constante e uma taxa de *churn* equivalente para todos os clientes, desta forma, o resultado poderá ser altamente variável e não corresponde à realidade para alguns clientes, conduzindo a estimativas pouco realistas [73].

$$CLV = (VM * F/c) * ML$$

2.5.2.2 Modelo Cohort

Bastante rudimentar, este modelo é baseado no histórico e consiste no agrupamento dos clientes com características semelhantes, tipicamente a primeira data de compra por mês. De outra forma, o resultado traduz-se na receita média por grupo. Contudo, este modelo supera o Modelo de Agregação, visto que o anterior considera que todos os clientes pertencem ao mesmo grupo [73].

$$CLV_t = (VM_t * F_t/c_t) * ML_t$$

Onde:

- VM - valor médio por compra
- F - número de compras realizadas (frequência)
- c - percentagem de *churn*
- ML - margem de lucro

2.5.3 Métodos para Previsão

De forma a calcular o CLV, pesquisas e documentação sugerem inúmeros métodos. Contudo, a relação entre o comportamento das compras e o *lifetime value* do cliente não é específico. Diferentes

modelos para medir o CLV causam diferenciação nas estimativas das expetativas de comportamento de compra de cada cliente no futuro [11].

2.5.3.1 Modelo RFM

Bult e Wansbeek [10], desenvolveram o modelo *Recency, Frequency and Monetary (RFM)*, considerado o mais capaz e intuitivo dentro do CRM [29]. Este modelo baseia-se no comportamento das compras definidas pelo histórico de cada cliente, constituído por 3 elementos base, *Recency*, *Frequency* e *Monetary*, estes definem-se por:

- ***Recency:*** período desde a última compra, sendo que o valor mais baixo corresponde à maior probabilidade de um cliente repetir a compra;
- ***Frequency:*** número de compras realizadas dentro de um intervalo temporal, uma frequência mais elevada corresponde a um maior grau de lealdade;
- ***Monetary Value:*** total gasto durante um certo período ou valor médio gasto por transação, quanto maior for esse indicador mais ênfase deverá ser dada ao cliente.

A sua utilização minimiza o custo de marketing, potencia o aumento do *Return on investment (ROI)*, diminui a taxa de *churn* e obtém-se melhor compreensão do mercado.

A Tabela 3 apresenta o modelo RFM aplicado em dias sobre 2 clientes, o cliente 1 efetuou a 11.ª compra há 5 dias e gasta em média 50 € por transação, enquanto o cliente 2, provavelmente está ausente e não efetua nenhuma compra há 365 dias. Desta forma, concluímos que o cliente 1 é mais valioso que o cliente 2.

Tabela 3: Exemplo da aplicação do modelo RFM

ID	**R**ecency	**F**requency	**M**onetaty Value
1	5	11	50
2	365	2	10

Mais recentemente, alguns autores propuseram o modelo *Weighted RFM (WRFM)*, que consiste na atribuição de pesos a cada variável mediante as características da indústria envolvida [29].

2.5.3.2 Modelos probabilísticos

Baseado sobretudo no histórico de compras realizadas, o modelo *Buy 'Til You Die (BTYD)* disponibiliza modelos estatísticos que capturam características comportamentais de clientes não contratuais (compras podem ocorrer em qualquer momento, como compras numa loja online) ou quando a instituição não tem a capacidade de observar diretamente a taxa de *churn*, de modo a, obter previsões futuras de clientes

repetidos (com pelos menos 2 compras), forçando os clientes a comprar até "morrerem"(tornarem-se inativos).

O modelo é descrito por:

- Processo de transação (*Buy*):
  - Enquanto ativo, o cliente realiza compras aleatórias em torno da taxa de transação;
  - A taxa de transação é variada para cada cliente.
- Processo de inatividade (*Till You Die*):
  - Cada cliente tem uma probabilidade de inatividade;
  - A probabilidade de inatividade é variada para cada cliente.

A biblioteca Lifetimes encapsula a complexidade matemática por detrás dos seguintes modelos: *Pareto/Negative Binomial Distribution (NBD)*, *Beta-Geometric (BG)/NBD*, Modified Beta-Geometric (MBG)/NBD e *Beta-Geometric/Beta-Binominal (BG/BB)* [64, 56].

Pareto/NBD Modelo desenvolvido, em 1987, por Schmittlein, Morrison, e Colombo, considerado um modelo poderoso para o cálculo do fluxo de transações [33], este foi o primeiro a incluir o processo de abandono dos clientes, este compõe-se por uma técnica que deteta o comportamento de compra do cliente no passado e usa-o para prever a probabilidade do cliente permanecer ativo no negócio. Por outras palavras, estima a probabilidade de atividade e o número de transações dos clientes. Para que seja realizada a previsão, uma premissa essencial refere-se à independência entre o lucro para cada transação e o número de transações de um determinado cliente [11, 64].

O modelo Pareto/NBD é baseado em 5 pressupostos [30]:

- Enquanto ativo, o número de transações efetuadas por cliente, num determinado período t é realizado através da Distribuição de Poisson com média λt;
- Heterogeneidade na taxa de transação λ segue uma distribuição Gamma com parâmetro de forma r e parâmetro de escala α.
- Cada cliente tem "tempo de vida"não observado de tamanho r, neste instante o cliente torna-se inativo através da distribuição exponencial com uma taxa de *churn* μ.
- Heterogeneidade na taxa de *churn* segue uma distribuição Gamma com parâmetro de forma s e parâmetro de escala β.
- As taxas de transação e de *churn*, λ e μ, variam de forma independente entre os clientes.

BG/NBD Desenvolvido em 2005, este modelo produz resultados muito similares ao modelo antecedente, visto que consiste numa extensão do mesmo. A principal diferença entre os dois modelos é que o modelo BG/NBD mapeia a curva de sobrevivência (*churn*) para uma distribuição beta-geométrica em vez de uma distribuição pareto. Isto significa que, os autores propuseram a substituição do processo de *churn* num espaço contínuo por espaço discreto, ou seja, após cada compra, a probabilidade de *churn* é individualizada por cliente [51]. Internamente, o BG/NBD simplifica a complexidade matemática envolvente. Este modelo é considerado dos mais influentes no estudo comportamental dos clientes, graças à sua fácil interpretação e precisão [4].

O modelo BG/NBD é baseado em 5 pressupostos, sendo os 2 primeiros equivalentes ao modelo Pareto/NBD [30]:

- Enquanto ativo, o número de transações efetuadas por cliente, num determinado período t é realizado através da Distribuição de Poisson com média λt;
- Heterogeneidade na taxa de transação λ segue uma distribuição Gamma com parâmetro de forma r e parâmetro de escala α.
- Após cada transação, os clientes tornam-se inativos com base na probabilidade p. Portanto, a cada momento que o cliente deixe de comprar, essa probabilidade é distribuída pelas transações através de uma distribuição geométrica (deslocada) através da função massa de probabilidade.
- Heterogeneidade p segue uma distribuição beta com a função densidade.
- A taxa de transação λ e a probabilidade de *churn* p, variam de forma independente entre os clientes.

MBG/NBD Tal como descreve o nome, este modelo consiste numa mudança do modelo BG/NBD. Concretizado no ano de 2007, o presente modelo elimina alguma inconsistência do modelo BG/NBD entre clientes com um número elevado de compras e permite a integração com clientes sem transações repetidas. Apesar deste modelo ser bastante menos recorrente para o cálculo do CLV, poderá ser útil para alguns tipos de *datasets*.

BG/BB Assim como o modelo BG/NBD, em múltiplos aspetos é similar ao Pareto/NBD, contudo a grande diferença é que as transações futuras são tratadas como transações discretas, em vez de transações esporádicas [56]. Desenvolvido em 2010 e considerada ótimo para organizações sem fins lucrativos com padrão de doação pré-definidos.

Em 2005, Fader [34], apresentou o modelo *Gamma-Gamma (GG)* que permite prever o valor médio gasto por transação efetuada por clientes com os gastos superiores a 0. O modelo GG é baseado nos seguintes pressupostos:

- O valor monetário das transações varia aleatoriamente em torno do seu valor médio de transação, independentemente do tipo de moeda.
- Os valores médios de transação diferem entre clientes, mas não diferem temporalmente por indivíduo.
- Os valores médios de transação são independentes do processo de transação.

Assim concluí-se que, com a integração da biblioteca Lifetimes é possível calcular o CLV através da associação dos modelos abordados anteriormente:

$$(Pareto|BG) + GG = CLV$$

A sua utilização conduz à resposta de 2 questões inquietantes para os empresários, quanto e quando [51]. Diversos investigadores elogiaram o modelo Pareto/NBD devido à sua utilidade e ao seu excelente comportamento [30]. Importa salientar que para a projeção de resultados minimamente satisfatórios, é necessário um amplo histórico de uma comunidade de clientes, caso contrário o resultado poderá ser critico e desajustado. Para a sua implementação é importante ter em consideração os seguintes pontos [69]:

- Aplicar o modelo de forma individualizada a grupos de clientes, como, por exemplo, a aplicação baseada em RFM;
- Assumir que a sua utilização está voltada para um grupo de clientes idêntico ao histórico observado;
- Para além dos modelos de previsão de compras futuros deve ser implementado um modelo (GG) para previsão de valor médio de compra para o cálculo do CLV.

Contudo, esta biblioteca processa informação apoiada em clientes repetidos, de outro modo, que tenham realizado pelo menos 2 compras, caso contrário, estes são descartados. O seu desenvolvimento segue uma nomenclatura específica inspirada nas métricas do modelo RFM, definida por:

- ***Recency:*** esta métrica difere daquela detalha na subsecção 2.5.3.1 normalmente utilizada pela comunidade de marketing [31]. Então, esta representa a diferença temporal entre a primeira e a última compra, sendo que o valor mais elevado corresponde à maior probabilidade de um cliente repetir uma compra;
- ***Frequency:*** número de compras repetidas realizadas dentro de um intervalo temporal, a frequência mais elevada corresponde a um maior grau de lealdade (primeira compra é excluída);
- ***Monetary Value:*** valor médio gasto por transação, quando maior for esse indicador mais ênfase deverá dar ao cliente (primeira compra é excluída);

- ***T:*** diferença temporal entre a data da primeira compra e a última data registada no *dataset*, ou seja, tempo desde a primeira compra.

Assim, observe-se a Figura 10 que apresenta de forma diagramática a nomenclatura definida pela biblioteca Lifetimes:

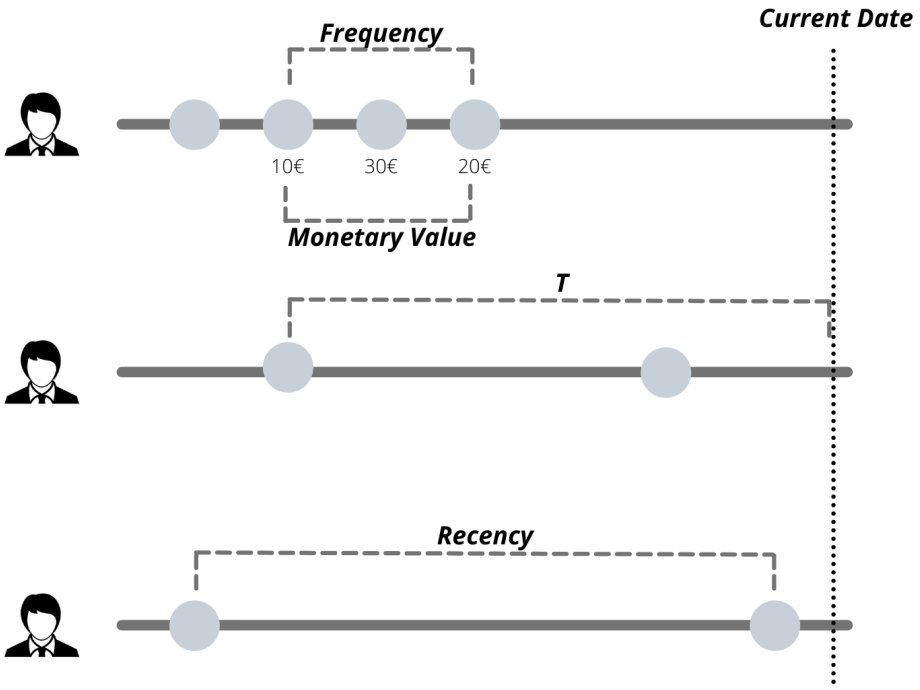


Figura 10: Nomenclatura Lifetimes

A principal limitação destes modelos está relacionada com as transações sazonais (como a época natalícia), uma das soluções para minimizar os danos, consiste na aplicação de *cohorts*[1] ou incorporar efeitos sazonais de séries temporais. Um estudo sobre o comportamento dos modelos [46], concluiu que apesar de não existir um modelo ideal, a utilização dos modelos Pareto/NBD e BG/NBD, alcançaram resultados estáveis, com desvios padrão relativamente baixo. Além disso, outra investigação [33], com a utilização do Pareto/NBD, analisou um *dataset* de 78 semanas, utilizando 39 semanas para calibração (treino) e 39 semanas para validação (teste) para realizar o *tracking* de transações acumuladas, concluíram que o modelo produziu previsões bastante confiáveis.

2.5.3.3 Modelos de *Computer Science*

Para além de modelos destinados ao CLV, encontra-se uma panóplia de abordagens de *data mining*, *machine learning (clustering/regression)* e estatística não paramétrica que permitem atingir o mesmo objetivo. Isto inclui, modelos de *neural network*, modelos aditivos generalizados, modelos *deep neural*

[1] Conjunto de indivíduos que partilham características comuns no período de tempo

network e *support vector machine* [11]. Contudo, calcular o CLV é bastante complexo e o uso de modelos comuns, como modelos de regressão, que tenta prever o comportamento futuro com base apenas em medidas observáveis, é problemático e inadequado. Estes são projetados para prever o "comportamento"no próximo período. Contudo, ao calcular o CLV, o interesse vai além do próximo período, então para fazer a previsão para o terceiro mês são usados dados do segundo mês. À medida que a previsão alarga as métricas de RFM não podem ser especificadas e os resultados tornam-se não confiáveis. A melhor abordagem consiste em realizar os cálculos com a utilização de modelos probabilísticos sobre o comportamento dos clientes, em que o comportamento observado resulta de um processo aleatório gerido por características latentes. Estas são apenas indicadores do perfil comportamental dos clientes. Diferentes características produzirão valores distintos das métricas de RFM, e portanto, previsões diferentes. Por vezes, ligeiras alterações no comportamento passado pode levar a enormes diferenças na expectativa futura [32].

Para finalizar, como medidas de precisão do modelo, Glady [1] usou o *Mean Absolute Error (MAE)* e o *Root Mean Square Error (RMSE)* (raiz quadrada de *Mean Square Error (MSE)*) entre o valor real e a previsão do valor na vida dos clientes, claro está que quanto menor forem os resultados destas operações, melhor em ambos os casos [11].

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} |y_i - x_i| . \quad MSE = \frac{1}{n} \Sigma_{i=1}^{n} \left(y_i - x_i\right)^2 . \quad RMSE = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} \left(y_i - x_i\right)^2} .$$

Onde:

- n - número total de observações
- x_i - valor atual
- y_i - valor previsto

2.6 MLOps

Criar soluções de ML para diversos problemas pode tornar-se uma tarefa árdua, então o *Machine Learning Operations (MLOps)* consiste na padronização e simplificação da gestão do ciclo de vida de ML [77] composta por várias pessoas/grupos, cada um com a sua tarefa. Implementa um conjunto de práticas de engenharia de ML que visa unificar o desenvolvimento de sistemas de ML *(Development (Dev))* e a operação de sistemas de ML *(Operations (Ops))*. Na prática, MLOps defende a automação e monitoriza todos os passos da construção do sistema de ML, incluindo integração, teste, lançamento, implementação e gestão de infraestrutura (similar ao *benchmarking*) [13].

Portanto, MLOps resulta da combinação de *Machine Learning*, DevOps, e *Data Engineering*, observe-se a Figura 11, tornando o desenvolvimento do sistema ML confiável e eficiente [40].

$$ML + DevOps + DataEngineering = MLOps$$

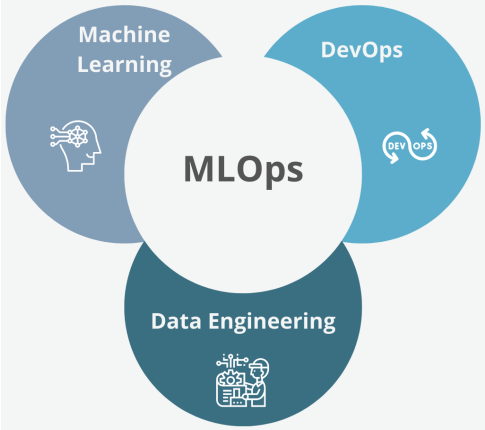


Figura 11: Estrutura MLOps

De modo empírico, consiste na aplicação dos benefícios de DevOps a sistemas ML, para:

- Encurtar os ciclos de desenvolvimento;
- Aumentar a velocidade de implementação;
- Ajudar as organizações a produzir modelos ML com qualidade que melhoram a qualidade de negócio;
- Ajustar diferentes versões a diferentes modelos de negócio.

Além disto, os dados estão em constante mudança o que significa que os modelos de ML aprendem continuamente a adaptar-se, perante a complexidade deste ambiente torna o uso de MLOps crucial [77].

2.6.1 Estrutura

O objetivo principal do MLOps é garantir às organizações uma boa construção de dados e um ótimo modelo para solucionar problemas de negócio, extraindo a melhor performance e transparência possível, evitando custos e tempo exageradamente consumido. Visível na Figura 12, este sistema pode ser segmentado em: *Small Data Ops, Big Data Ops, Large-scale Ops* e *Hybrid Ops*. Estas categorias resultam do tamanho da equipa, do modelo do negócio, ferramentas e infraestrutura das operações [65].

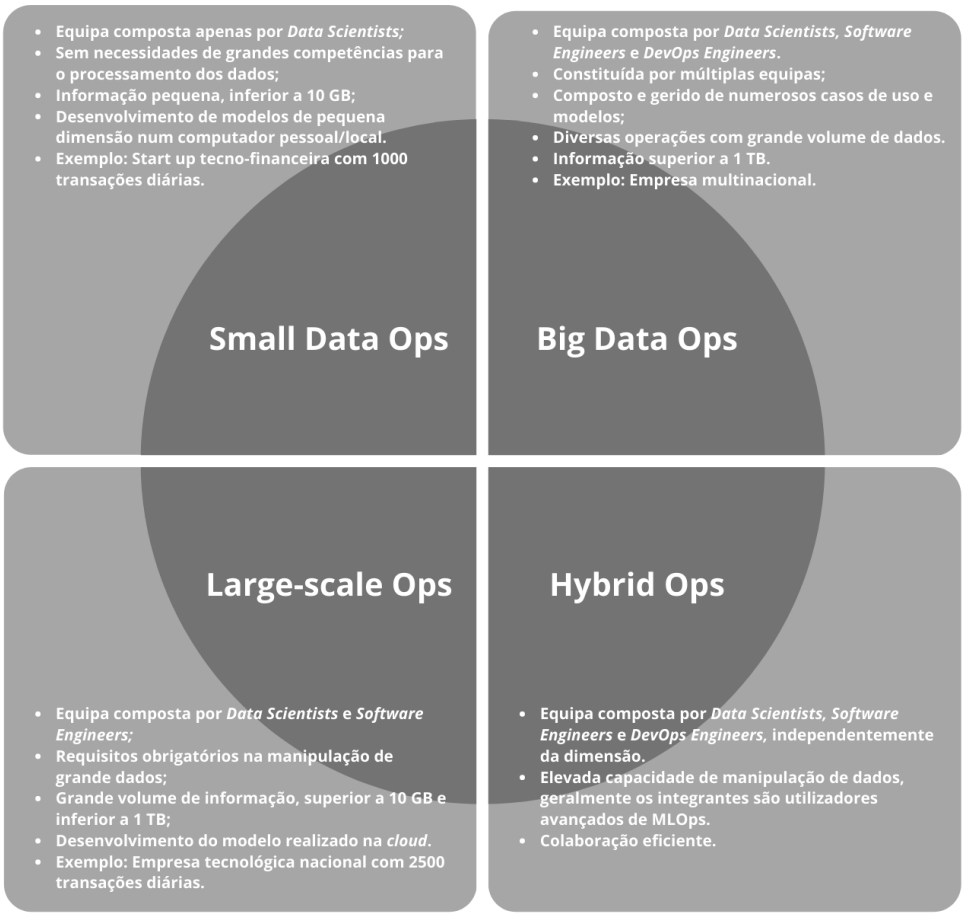


Figura 12: Categorias MLOps

2.6.2 Vantagens

A Tabela 4 inúmera os benefícios principais da utilização de MLOps:

Tabela 4: Vantagens da utilização de MLOps

Vantagens
Controlo de versões dos modelos como: *rollbacks* e auditorias, em caso de falhas
Tracking dos hiperparâmetros, percebendo quando, onde e porquê cada hiperparâmetro foi alterado
Encurtar o ciclo de produção do modelo, consequentemente diminuindo custos
Monitorização da performance e dos resultados, percebendo qual o melhor modelo temporalmente
Padronizar fluxos de criação de novos modelos, facilitando recentes implementações
Desenvolvimento simplificado com alta performance em qualquer área de negócio

2.6.3 Mitigação do Risco

Segundo Mark Treveil [77], MLOps é a solução para a mitigação do risco, dado que qualquer modelo necessita de monitorização contínua e ajustes. Contudo, estas práticas contêm um custo associado,

consequentemente é necessário realizar uma avaliação de custo-benefício para cada *use case*.

2.6.3.1 Estimativa do Risco

Em ML, o risco é altamente volátil. Sendo necessário analisar meticulosamente de forma a precaver riscos como: o modelo ficar interrompido por tempo indeterminado, o modelo prever um resultado fora do contexto devido a uma ação não planeada e a *accuracy* diminuir ao longo do tempo [77].

Visualizando a matriz da Figura 13, a avaliação do risco é baseada em 2 métricas agregadas, a probabilidade e o impacto. Esta avaliação deve ser realizada no início de cada projeto e reavaliada ao longo do mesmo devido a mudanças não previamente estudadas [77].

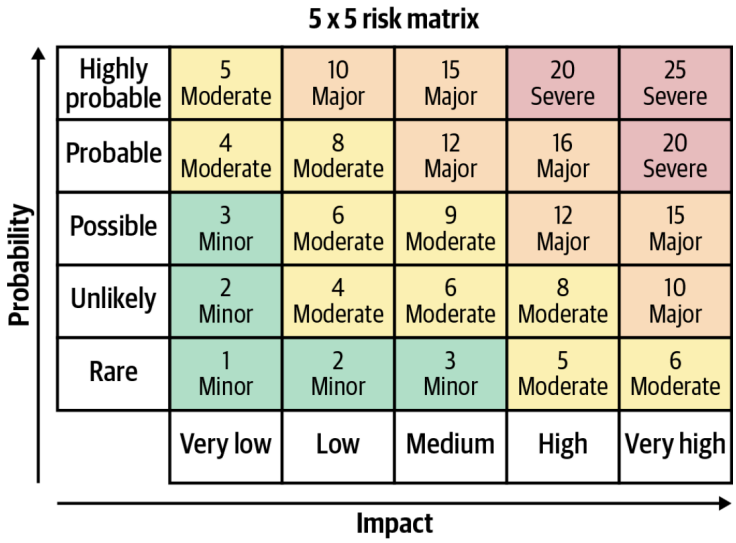


Figura 13: Tabela representativa da avaliação do risco [77]

MLOps não é só essencial para mitigar o risco dos modelos, mas também é importante para o *deploy*. Alterar um dos vários modelos de produção requer disciplina, é extremamente importante manter um *tracking* das versões (principalmente na fase de estruturação), compreender se os modelos treinados são melhores que as versões anteriores (mantendo sempre os melhores em produção) e assegurar que o modelo não se degrada em produção [77].

2.6.4 *Personas*

Primeiramente os modelos de ML são desenvolvidos por *data scientists*, mas observando a dimensão de todo o processo de MLOps, estes não são os únicos envolvidos. MLOps é uma parte essencial na estratégia de AI, assim sendo, afeta ou beneficia todos aqueles que estão ao redor do ciclo de vida de ML [77].

A Tabela 5 disponibiliza de forma sintetizada o papel e as obrigações de cada um dos envolvidos no negócio.

Tabela 5: *Personas* MLOps [77]

Papel no Ciclo de Vida do Modelo	**Obrigações MLOps**
Especialista no Assunto	
- Disponibilizar o negócio, objetivos ou *Key Performance Indicator (KPI)* - Assegurar que a performance do modelo se mantenha no plano	- Analisar o desempenho - Fornecer *feedback* cíclico dos resultados
Data Scientists	
- Construir modelos baseados nos requisitos delineados - Entregar modelos produtivos - Avaliar a qualidade do modelo	- Entregar modelos para *deploy* fácil - Efetuar testes e fazer melhorias - Analisar desempenho dos modelos
Data Engineers	
- Utilizar dados para alimentar o modelo	- Observar a performance dos modelos
Software Engineers	
- Integrar os modelos com os sistemas e as aplicações - Assegura que os modelos funcionam com as aplicações	- Gestão de versões e testes automáticos
DevOps	
- Construir sistema de controlo, testes de segurança, performance e disponibilidade - Gestão *Continuous Integration/Continuous Delivery (CI/CD)*	- Integrar MLOps com a estratégia DevOps da empresa - *Deploy* do modelo
Auditores	
- Minimizar o risco da empresa relativamente aos modelos em produção - Assegurar que os requisitos previamente definidos estão traçados antes dos modelos serem colocados em produção	- Relatórios dos modelos (em produção ou não)
ML Architect	
- Assegurar a escalabilidade e flexibilidade de todo o ambiente ML, desde o design, desenvolvimento e monitorização. - Disponibilizar novas ferramentas com o intuito de melhorar a performance dos modelos em produção	- Visão geral sobre os modelos e os recursos consumidos - Avaliar e ajustar as necessidades da infraestrutura

2.6.5 MLflow

MLflow é uma ferramenta *open-source* de MLOps gere o ciclo de vida de ML, desde a experimentação, desenvolvimento e registo do modelo *(end-to-end)*, veja-se a Figura 14. É uma ferramenta bastante diversificada e completamente agnóstica a qualquer biblioteca de ML ou a qualquer linguagem de programação acessível via REST API, ou *Command-Line Interface (CLI)* [59]. Este recurso disponibiliza 4 componentes bases no seu *workflow*: *MLflow Tracking, MLflow Projects, MLflow Models* e *MLflow Registry* [58].

- ***MLflow Tracking*:** Responsável pela monitorização dos parâmetros, versões de código e métricas. Capaz de ser usado em qualquer ambiente (como *notebook*) para registar os resultados localmente ou num servidor;
- ***MLflow Projects*:** Consiste num formato padrão para guardar e reutilizar o código, habilitado para partilhar com os outros *data scientists*;
- ***MLflow Models*:** Expressa-se numa convenção para empacotar os modelos ML e oferece múltiplas ferramentas para realizar o *deploy*. Cada modelo é guardado numa diretoria que contém ficheiros arbitrários e uma listagem de *"flavors"* em que o modelo poderá ser utilizado;
- ***MLflow Registry*:** Fornece uma *store* central para gestão colaborativa de todo o ciclo de vida do modelo MLflow, incluindo controlo de versões do modelo, transições (colocar em produção ou arquivar) e anotações, ou seja, foi projetado para permitir um fluxo de trabalho onde os modelos são testados e movidos entre blocos ao longo do tempo.

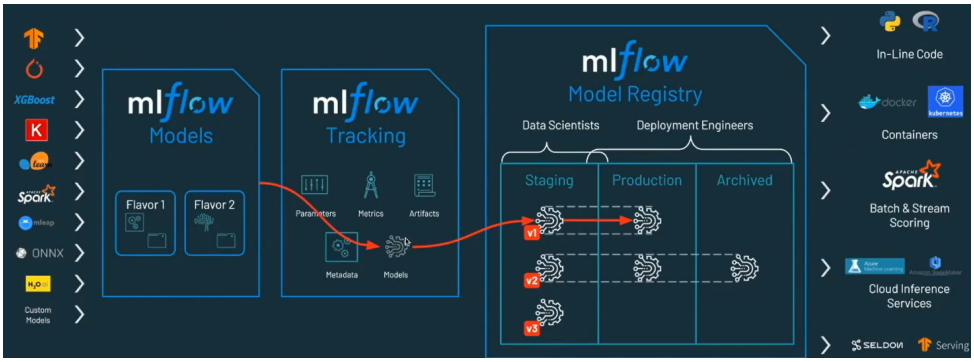


Figura 14: *Lifecycle* MLflow [23]

2.6.5.1 Alternativas

A capacidade do MLflow é imensa, contudo há aspetos que poderiam ser aperfeiçoados, nomeadamente ao trabalhar com grandes equipas ou o volume de testes/execuções é elevado. Então, esta secção explora e fundamenta algumas alternativas ao MLflow.

A documentação oferece uma infinidade de ferramentas disponíveis e catalogadas em diferentes secções, cada uma delas com as suas características únicas, são seguidamente dispostas 7 possibilidades sintetizadas, veja-se a Tabela 6 [18, 37, 47, 75].

1. **Neptune:** Contém um *User Interface (UI)* intuitiva e simplista, que simplifica o fluxo de trabalho, desde o armazenamento, organização, exibição e comparação de todos metadados gerados.

2. **Weights & Biases:** Também conhecido por WandB, diferindo do MLflow, este permite fazer *backup* de todos os testes num único local e permite trabalhar o projeto com uma equipa. Fornece recursos para *tracking*, controlo de versões e gestão de modelo, enquanto o MLflow cobre quase todo o ciclo de vida do ML.

3. **Comet:** Plataforma para *tracking*, comparação, teste e otimização de modelos. Contém uma abordagem híbrida relativamente a *host*, permitindo aos utilizadores terem *host* local ou na *cloud*.

4. **Valohai:** Semelhante ao fluxo de trabalho MLflow, porém contém adicionalmente a gestão da infraestrutura. Permite desenvolver em qualquer linguagem de programação.

5. **TensorBoard:** É uma ferramenta de visualização *open-source*, que permite analisar as execuções para o treino do modelo, permite visualizar as métricas, os gráficos dos modelos e os histogramas dos tensores.

6. **Metaflow:** Plataforma que possibilita construir e manipular projetos de ML, *end-to-end*, de nível empresarial em grande escala.

Tabela 6: Ferramentas alternativas ao MLflow [18, 47, 75]

	(a)	(b)	(c)	(d)
MLflow	✓	✓		Desenvolvido pelo Databricks
Neptune			✓	*Out-of-the-box*, permite integração com uma diversidade de ferramentas MLOps
Weights & Biases			✓	Regista o gráfico do modelo, para facilmente inspeccionar mais tarde
Comet	✓		✓	-
Valohai		✓	✓	-
TensorBoard	✓			Ótimo para trabalho com imagens
Metaflow	✓	✓		Desenvolvido pela Netflix

(a) *Open-Source*

(b) Engloba o Ciclo de Vida ML

(c) Colaboração e Gestão de Tarefas de Utilizadores

(d) Extra

Verifica-se na Tabela 6, que a utilização do MLflow é bastante viável e cumpre todos os requisitos necessários para a implementação do *use case* selecionado, sobretudo porque o desenvolvimento do modelo é realizado na plataforma Databricks, mantendo tudo incorporado no mesmo ecossistema.

2.7 Conclusões

Diversos conceitos, técnicas, ferramentas e algoritmos relacionados com a aplicabilidade de *Machine Learning* num sistema de gestão ERP foram analisados e sintetizados. Esta revisão do estado de arte foi complementada com o levantamento de literatura de diversas fontes fidedignas e devidamente mencionadas.

3

Planeamento

Este capítulo visa descrever o problema associado ao tema de trabalho e pormenorizar detalhes. A secção 3.1 detalha a importância do cálculo do CLV. A secção 3.2 descreve a proposta de desenvolvimento. A secção 3.3 descreve os desafios associados ao problema. A secção 3.4 disponibiliza algumas questões relacionadas. A secção 3.5 descreve genericamente os objetivos, pré-requisitos, dificuldades e desafios, potenciais falhas e planos de contingência para cada tarefa descrita. A secção 3.6 integra uma representação do diagrama de Gantt. Finalmente, a secção 3.7 sumariza o capítulo.

3.1 Descrição

Com o avanço das novas tecnologias, novas empresas surgem e a concorrência aumenta, consequentemente a aquisição de novos clientes e a retenção fica mais dispendiosa. Desta feita, o estudo detalhado sobre cada cliente devolve à empresa o valor de cada um deles, com a finalidade de realçar a análise de negócio e dando ênfase à informação relacionada com cada cliente. Posto isto, este processo manual seria inconcebível à medida que o volume de negócio aumenta. Portanto, o desenvolvimento de um modelo inteligente capaz de diferenciar e atribuir valor individual a cada cliente, de uma forma simples e fácil de interpretar, representa o principal objetivo da tese.

3.2 Proposta

O modelo RFM é limitado e não fornece explicitamente um valor monetário por cliente. No entanto, em 2005, Fader, Hardie e Lee mostraram que a sua integração com modelos probabilísticos superam algumas limitações [30].

Baseado no objetivo da dissertação, que visa extrair valor de cada cliente, com o auxílio de MLOps, esta secção detalha a proposta de implementação. Para a sua realização, o projeto será desenvolvido com base na classe de modelos estatísticos BTYD, que foi tipicamente desenhada para a captação da análise comportamental dos clientes não contratuais e produção de um modelo e previsão do cálculo CLV. Para isso, o desenvolvimento é realizado com a integração da biblioteca *Lifetimes* que sintetiza a

complexidade matemática dos modelos estatísticos. Posto isto, a listagem seguinte descreve a proposta de implementação:

- Modelos Pareto/NBD, Beta-Geometric/NBD e Modified Beta-Geometric/NBD para previsão do número expectável de compras por cliente num determinado espaço temporal e da probabilidade de *churn*;
- Modelo GammaGamma para previsão do valor médio gasto por transação e cálculo do CLV;
- MLflow como ferramenta para a gestão do ciclo de vida de ML;
- Segmentação dos clientes baseada no tempo, em dias, desde a primeira compra (métrica T);
- Segmentação dos clientes baseada nas métricas de RFM, seguindo uma convenção ligeiramente diferente daquela que é utilizada nas análises do CLV.

Portanto, o objetivo consiste em aplicar uma série de filtros, baseado nos requisitos, e extrair informação detalha para cada cliente. Como resultado, obtemos as previsões de compras futuras, valor médio por transação e do CLV, probabilidade de taxa de atividade (*churn*) e segmentação dos clientes. Observe-se o fluxograma genérico da aplicação do Lifetimes presente na Figura 15.

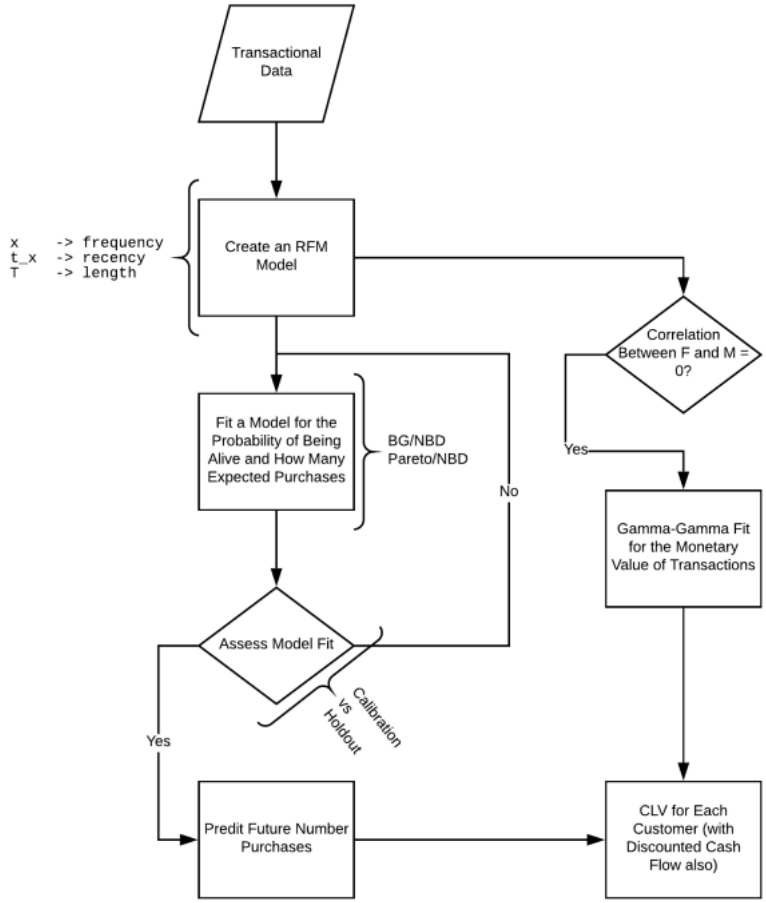


Figura 15: Fluxograma genérico da aplicação do Lifetimes [27]

3.3 Desafios

O problema mencionado está associado a diversos desafios que devem ser enfatizados. O desenvolvimento em escala conduz a um grau de incerteza maior e acentua a dificuldade do desenvolvimento. Portanto, o desenvolvimento deste projeto implica a interpretação de diversos conceitos e múltiplas decisões que precisam de ser tomadas, tais como:

1. Interpretação dos dados de cada instituição, de modo a associar o melhor modelo;
2. Trabalho detalhado sobre as empresas recentes ou com poucos dados;
3. Aplicação dos filtros lógicos para aprimorar o modelo e integrar o máximo de empresas possíveis;
4. Uma representação dos dados através de um formato simples e intuitivo, para uma fácil interpretação dos envolvidos;
5. Apresentação da solução com graus de incerteza baixos.

3.4 Questões relevantes

Perante isto, algumas questões emergem, tais como:

- **Questão 1:** Quais são os clientes que irão gerar mais receita no próximo mês?
- **Questão 2:** Quantas compras irá realizar o cliente X no próximo mês?
- **Questão 3:** Quais são os clientes em risco de *churn*?
- **Questão 4:** É possível categorizar e diferenciar os clientes relativamente ao volume de negócio?
- **Questão 5:** Quais são os clientes em ascensão?
- **Questão 6:** Qual o CLV previsto a 13 meses para o cliente X?

3.5 Descrição das Tarefas

Tabela 7: Descrição das Tarefas

	Nome	Início	Fim	Tarefas em Paralelo	Importância
1	Documentação MLOps e UCs	6 Out	31 Out	-	Média
2	Estado de Arte e casos de estudo	1 Nov	31 Dez	6	Alta
3	Análise do *Dataset*	1 Jan	31 Jan	6	Alta
4	Introdução do Databricks	1 Fev	28 Fev	6	Média
5	Desenvolvimento dos modelos	1 Mar	31 Mai	6	Alta
6	Revisão e documentação do trabalho	1 Nov	30 Jun	2, 3, 4, 5	Alta

São descritas abaixo as tarefas enumeradas na Tabela 7:

1. Objetivo(s)

    - **Tarefa 1:** Compreender MLOps, AutoML e a importância da AI aplicado a um ERP e investigar UCs de forma a definir a base do trabalho futuro.
    - **Tarefa 2:** Recolha e leitura de artigos científicos e revisão da literatura existente sobre o tema, de modo a redigir o capítulo 2.
    - **Tarefa 3:** Interpretar o *dataset*, interpretar anomalias e padrões existentes, visando apoio no processo de *data engineering* e *data science*.
    - **Tarefa 4:** Entender o fluxo de trabalho da plataforma de desenvolvimento (Databricks). Tarefa relevante para precaver dificuldades futuras durante o desenvolvimento.

- **Tarefa 5:** Base do projeto, consiste em produzir modelos inteligentes em paralelo com o MLflow.
- **Tarefa 6:** Tarefa importante que acompanha quase todas as tarefas antecedentes, nela é efetuada a revisão, melhoria e descrição escrita do trabalho desenvolvido.

2. Pré-requisito(s)

   - **Tarefa 1:** Recolha de documentação oficial e fidedigna.
   - **Tarefa 2:** Recolha de livros, artigos e *websites* de documentação oficial.
   - **Tarefa 3:** Seleção do *dataset* apropriado.
   - **Tarefa 4:** Criação de conta e associação à equipa de trabalho.
   - **Tarefa 5:** Leitura e interpretação da documentação dos modelos.
   - **Tarefa 6:** Seleção do *template* adequado.

3. Dificuldades e desafios

   - **Tarefa 1:** Documentação pouco fundamentada e duvidosa, podendo causar confusão e discórdia.
   - **Tarefa 2:** Interpretação e apuramento do conteúdo mais adequado para a redação do capítulo.
   - **Tarefa 3:** Um *dataset* complexo e pouco organizado poderá impedir uma boa progressão do trabalho.
   - **Tarefa 4:** Dificuldade em realizar tarefas básicas e conteúdo técnico pouco intuitivo.
   - **Tarefa 5:** Análise de possíveis bibliotecas para produzir modelos adequados e compreensão do *input* e o *output*.
   - **Tarefa 6:** Manter o fluxo de conteúdo coerente com capítulos reportados.

4. Potenciais falhas

   - **Tarefa 3:** Após uma análise detalhada sobre o *dataset*, concluí-se que não é suficientemente bom para o objetivo final.
   - **Tarefa 4:** Plataforma não corresponde às expetativas e não é capaz de realizar operações primárias.
   - **Tarefa 5:** Os modelos produzirem resultados pouco credíveis.

5. Planos de contingência

   - **Tarefa 3:** Pesquisar e estudar um novo *dataset* adequado ou recorrer ao uso de dados sintéticos, produzindo assim um *dataset* inexistente.

- **Tarefa 4:** Recorrer a *softwares* alternativo como: Cloudera ou Google BigQuery.
- **Tarefa 5:** Inovar e possuir a capacidade de tornar singular uma abordagem existente.

3.6 Diagrama de Gantt

De forma, a ilustrar o avanço das múltiplas fases do projeto, o Diagrama de Gantt descreve as tarefas e os intervalos temporais que representam o início e fim de cada tarefa (eixo horizontal do gráfico), observe-se a Figura 16.

Figura 16: Diagrama de Gantt

3.7 Conclusões

O objetivo deste projeto é apoiar os empresários na análise do valor de cada cliente e numa melhor distribuição de marketing, recorrendo ao uso de técnicas de *Machine Learning*. Como resultado, é expectável que cada empresa, nos limites delineados, tenha um modelo capaz de produzir previsões de compras futuras, o CLV por cliente, identificação de clientes com probabilidade de abandono e uma segmentação pormenorizada dos clientes.

4

Implementação

Este capítulo detalha a solução e a metodologia usada para a construção dos modelos. A secção 4.1 sintetiza a solução implementada. A secção 4.2 descreve a fase de preparação de dados, composto pelo *input* e o *output*. A secção 4.3 apresenta as decisões tomadas e os modelos aplicados. A secção 4.4 descreve o processo de segmentação dos clientes. A secção 4.5 demonstra a utilização do MLOps e a integração com o MLflow. A secção 4.6 especifica o formato do armazenamento dos dados concebidos pelos modelos. A secção 4.7 espelha os resultados dos modelos. Finalmente, a secção 4.8 extrai algumas conclusões da solução.

4.1 *Overview* da Solução

O primeiro passo na implementação no desenvolvimento do CLV em larga escala consiste em preparar o *dataset* e assegurar as *features* essenciais para a produção dos modelos. De forma genérica, a solução divide-se em 2 *notebooks*, em que o primeiro realiza as previsões relacionadas com as compras, enquanto o segundo está relacionado com as previsões de faturação (€).

Relativamente ao primeiro, começa pela aplicação do *package* Lifetimes, de forma a sumarizar os dados **diariamente** com base nas métricas do modelo RFM e T. Seguidamente, com base nisso, aplica-se a segmentação sobre os quantis das *features*: *recency, frequency, monetary value* e *T*. O próximo passo consiste na aplicação do *package* Hyperopt para a realização do *tuning* sobre os hiperparâmetros do modelo. Finalmente, baseado na melhor combinação extraída pelo *tuning*, é criado um modelo por empresa que produz previsões da probabilidade de *churn* e previsões de compras para os próximos 1, 3, 6 e 13 meses.

O segundo *notebook* surge com base no resultado do anterior, recolhendo os dados do primeiro *notebook* e gera um modelo que produz previsões sobre o valor médio monetário esperado e o cálculo do CLV para 1 e 13 meses.

Por último, estes dados são guardados no *data lake* e expostos nas interfaces dos clientes.

4.2 Preparação dos Dados

Os dados laborados são disponibilizados pela solução *Canonical Business Data* da PRIMAVERA (dados de produção anonimizados), que consiste numa solução que tem a capacidade de integrar dados em múltiplos sistemas e base de dados.

Posto isso e com a receção dos dados, o primeiro passo para a produção dos modelos depende da preparação dos dados, como mencionado anteriormente. Consequentemente, o processo de *data engineering* consistiu em:

- Renomear *features*;
- Tratar duplicados;
- Agregar *features* para identificar as empresas;
- Aplicar *casting* para a data da compra;
- Além disso, o total da compra está armazenada em euros (€), contudo o código está preparado para alterar a moeda com base nas taxas de câmbio atualizadas.

Após este tratamento, segue-se o filtro das *features* dominantes para os modelos: *InvoiceDate, Total, CustomerID* e *CompanyID*. A lista seguinte apresenta brevemente a descrição de cada uma delas:

- ***InvoiceDate***: Data final da fatura, formado por dia, mês e ano (Ex: 03-12-2022);
- ***Total***: Valor final despendido pelo cliente (Ex: 5.99);
- ***CustomerID***: Identificador único do cliente (Ex: ebc95f307);
- ***CompanyID***: Identificador único da empresa, composto pela junção de 2 *features* (*PhysicalKey* e *BusinessKey*) (Ex: 0fa144237_2rb142670).

Além destas, um identificador único incremental é requerido na utilização da *Feature Store*. Este tem como objetivo tornar cada linha do *dataset* única e apenas tem importância para guardar os dados processados no processo de *data engineering*.

Posteriormente a esta tarefa, processa-se o agrupamento dos dados por empresa. Este procedimento transforma o *dataset* em múltiplos *dataframes*, em que cada *dataframe* identifica as movimentações dos clientes de cada empresa.

Por fim, o objetivo resume-se em sumarizar os dados dos clientes de cada empresa com base na nomenclatura do *package* Lifetimes (descrito em 2.5.3.1), esta operação foi elaborada com o auxílio da função *summary-data-from-transaction-data()*. Observem-se as seguintes tabelas, nas quais, a Tabela 8 representa os dados de entrada sintetizados e a Tabela 9 representa os dados sintetizados para uma empresa.

Tabela 8: Dados de entrada - Vendas *(Canonical Business Data)*

	InvoiceDate	Total	CustomerID	CompanyID	...
1	12-01-2022	5.99	0fa144237_2rb142670	ebc95f307	...
2	03-02-2021	56.49	2rb142670_6cctd4806	ebc95f307	...
3	02-03-2022	17.96	0fa144237_2rb142670	ebc95f306	...
4	01-01-2022	1.96	6cctd4806_f1478jev5	ebc95f304	...
...	...	...	...	...	...

Tabela 9: *Output* por empresa

	Recency	Frequency	Monetary Value	T
1	234	195	152,83	236
2	154	12	544,98	154
3	121	3	57,33	138
4	150	5	10	150
...	...	...	...	...

4.3 Descrição dos Modelos

A presente secção pormenoriza a lógica envolvente para a produção dos modelos. Detalha o processamento sobre os dados, as decisões tomadas para aperfeiçoar performance dos modelos, e consequentemente dos resultados finais. Como especificado em 4.1, a implementação divide-se pelos modelos de previsões de compras e previsões de faturação.

4.3.1 Previsão de compras

Não sendo possível o desenvolvimento de um modelo *one-size-fits-all*, optou-se por produzir um modelo ajustado a cada empresa. Portanto, e por força do desenvolvimento em larga escala, antes da produção dos modelos, cada empresa necessita de pelo menos um mês de faturação para gerar um modelo. Caso isso não aconteça, são previstas exatamente o mesmo número de compras com base nesse pequeno histórico, ou seja, se um cliente efetuou 1 compra em 10 dias ($F = 1$ e $T = 10$) então, $((30 + 10) * 1/10) - 1) = 3$ compras para o próximo mês.

Em seguida, o segundo filtro consiste na limitação das empresas que tenham pelo menos 5 meses de faturação, isto para seguir a lógica da literatura, de forma a garantir um período de *calibration* (treino) de pelo menos 80% e 20% para *holdout* (teste), ou seja, o modelo é criado com base nos 4 meses de vendas e efetua a previsão para o próximo mês (método *holdout* - tipo de *cross-validation*, para evitar *overfitting* e testar o modelo). Assim dizendo, com o auxílio da função *calibration-and-holdout-data()* do Lifetimes, este procedimento divide o *dataframe* até uma data definida (*calibration period end*), originando o aumento de

features, veja-se o *output* na Tabela 10, dessa ação realizada sobre o cliente 4 da Tabela 9. A Figura 17 expressa esta lógica de forma diagramática.

Tabela 10: *Output* após divisão para validação - método *holdout*

	Recency Cal	Frequency Cal	T Cal	F Hol	Duration Hol
4	120	4	30	1	30

Em que:

- ***Recency Cal***: Diferença temporal entre a primeira e a última compra no período de ***calibration***;
- ***Frequency Cal***: Número de compras repetidas no período de ***calibration***;
- ***T Cal***: Diferença temporal entre a data da primeira compra e a última data registada no período de ***calibration***;
- ***F Hol***: Número de compras realizadas no período de ***holdout***:
- ***Duration Hol***: Quantidade de dias para efetuar a previsão.

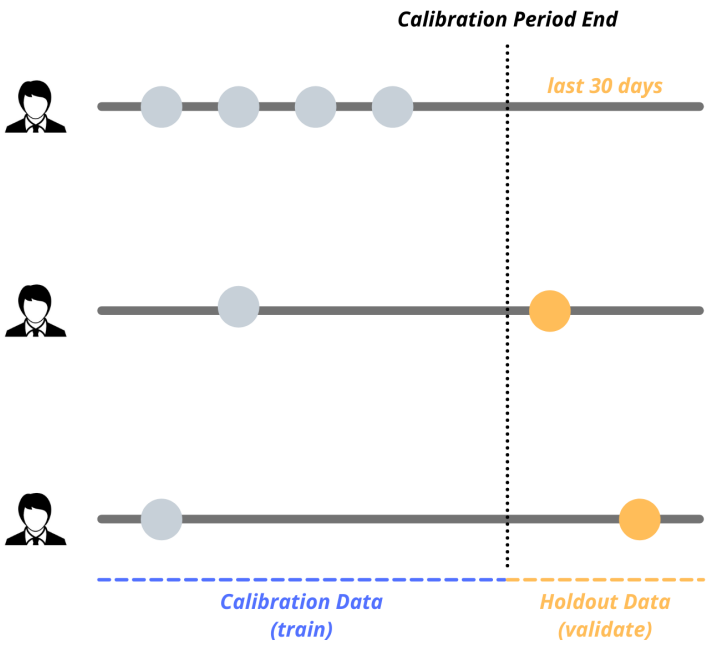


Figura 17: *Fragmentação do histórico* - método *holdout*

Por fim, o último requisito corresponde ao limite mínimo de clientes por empresa, em que é requerido que a empresa possua pelo menos 8 clientes para fazer *fit* do modelo, de forma a evitar erros de convergência. Diversos mínimos foram testados, contudo, o dígito 8 foi o que apresentou mais fiabilidade.

Posteriormente, com os *dataframes* filtrados, aplica-se o Hyperopt sobre os modelos: Pareto/NBD, BG/NBD e MBG/NBD. Neste processo iterativo, é calculado o MSE de forma a atingir a melhor combinação entre os modelos e o hiperparâmetro, o coeficiente de penalização. Para *dataframes* pequenos, os parâmetros podem ficar muito grandes, então, com a utilização de um penalizador da normalização L2 agregado à probabilidade, é possível controlar o tamanho desses parâmetros e, no que lhe concerne, torna o modelo menos complexo. Então, essa componente é implementada como configuração do modelo. Em aplicações típicas, penalizadores entre 0.001 e 0.1 são eficazes. Contudo, poderão existir números demasiado elevados, portanto o penalizador foi definido entre 0.0 e 1.0 evitando erros de convergência e com isto obtém-se um espaço de pesquisa bastante amplo. Adicionalmente, o parâmetro *TOL* foi alargado para $1e-02$ nos modelos Pareto/NBD, BG/NBD e MBG/NBD e $1e-06$ no modelo GG (valor padrão $= 1e-07$). Este consiste num critério de paragem, em que, quando a função erro (MSE) não melhora pelo menos 2x consecutivas o valor do *TOL* (salvo exceções), então a convergência é considerada alcançada e o treino do modelo termina.

A seguinte listagem detalha o processo anteriormente descrito:

- Agregação diária do histórico de cada cliente, recolhendo as métricas R, F, M e T;
- Exclusão de clientes sem compras repetidas (*frequency* $= 0$);
- Fragmentação do histórico: *calibration* equivalente a todo o histórico exceto o último mês e *holdout* equivalente ao último mês;
- Definição do espaço de procura (três modelos e penalizador entre 0.0 e 1.0), seguindo a nomenclatura do Hyperopt;
- Produção e ajuste (*fit*) do modelo com as *features* de calibração;
- Efetuar a previsão para o próximo mês;
- Calcular o erro, MSE, entre as compras realizadas no período de *holdout* e as previsões;
- Repetir o processo 100x e guardar a melhor combinação para cada empresa, observe-se a Figura 18.

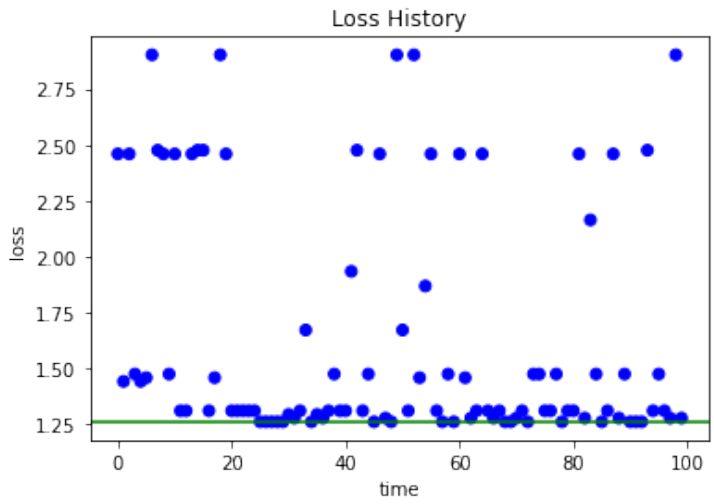


Figura 18: *Output* do Hyperopt - 100 execuções

De forma a abranger o máximo de empresas, a melhor combinação de resultados origina um modelo com configurações genéricas, partilhado pelas empresas que têm pelo menos 1 mês de faturação, proporcionando assim previsões para instituições recentes.

Por fim, um **novo modelo é gerado** por empresa, englobando o histórico total, em que cada um origina 5 novos *outputs* por cliente. São eles as previsões de compras para 1, 3, 6 e 13 meses (valor decimal) e a probabilidade da taxa de atividade ($[0-1]$) *(churn)*, observe-se a Tabela 11. O fluxograma A.1 exibe o processo descrito.

Tabela 11: *Output* típico de previsão de compras

	...	Expected 1M	Expected 3M	Expected 6M	Expected 13M	Prob Alive
1	...	0.467	1.234	4.625	7.212	0.978
2	...	0.004	0.010	0.020	0.120	0.218
3	...	3.972	6.142	10.123	20.645	1.0
4	...	6.091	18.312	36.621	72.351	1.0
...	...	...	...	...	...	...

4.3.2 Previsão de faturação

No sentido de implementar a previsão do cálculo do Customer Lifetime Value, o presente conteúdo pormenoriza a aplicação do modelo GammaGamma, baseado na literatura documentada em 2.5.3.2.

Como mencionado, este modelo tem a capacidade de prever um valor médio expectável por transação e calcular o CLV. Para isso, e assente no seus pressupostos, a frequência e o valor monetário têm de ser totalmente independentes, isto quer dizer que a correlação (de Pearson) entre estas *features* terá de ser inferior a 0.3 (considerada extremamente baixa), caso contrário as previsões poderão ser pouco credíveis. É notável que, caso haja uma frequência elevada notamos que a frequência e o valor monetário

são totalmente independentes, contudo com uma frequência baixa a tendência é criarem a ilusão de eles serem correlacionados [33].

Posto isto, no seguimento do modelo anterior, as empresas com modelos treinados e otimizados passam igualmente pelo processo de *tuning*. Neste caso, o espaço de procura está limitado apenas ao modelo GG, no entanto, também é acompanhado pelo penalizador da normalização L2. A seguinte listagem detalha o processo anteriormente descrito:

- Recolher os dados armazenados do modelo anterior;
- Agregação diária do histórico de cada cliente, recolhendo as métricas R, F, M e T;
- Exclusão de clientes sem gastos (*moneraty value* $= 0$);
- Fragmentação do histórico: *calibration* equivalente a todo o histórico exceto o último mês e *holdout* equivalente ao último mês (método *holdout*);
- Definição do espaço de procura (modelo GG e penalizador entre 0.0 e 1.0), seguindo a nomenclatura do Hyperopt;
- Produção e ajuste (*fit*) do modelo com as *features* de calibração;
- Efetuar a previsão do valor monetário para o próximo mês;
- Calcular o erro, MSE, entre o valor monetário no período de *holdout* e as previsões;
- Repetir o processo 150x e guardar a melhor combinação para cada empresa.

Da mesmo forma do modelo anterior, a melhor combinação de resultados produz um modelo com configurações genéricas para as restantes empresas com pelo menos 1 mês de vendas.

Em conclusão, um **novo modelo é gerado** por empresa, englobando o histórico total, em que cada um origina 3 novos *outputs* por cliente. São eles as previsões de CLV para 1 e 13 meses e a probabilidade do valor médio por transação, observe-se a Tabela 12. O fluxograma A.2 expressa o processo proferido.

Tabela 12: *Output* típico de previsão de faturação

	...	**Expected Monetary Value**	**Expected 1M CLV**	**Expected 13M CLV**
1	...	157	50,10	150
2	...	530	10,23	100
3	...	65	455	5750,45
4	...	11	60	659,80
...	...	...	...	...

De forma a avaliar os resultados, no início de cada mês, com um aumento do histórico, um novo modelo é concebido e aprimorado. Em paralelo, é efetuado o cálculo do erro com base nos resultados reais do mês antecedente.

4.4 Segmentação dos Dados

Ao contrário da secção anterior, independentemente do volume de negócio ou da dimensão, todas as empresas possuem os clientes segmentados assentes nas métricas de RFM e T.

Com base na documentação de marketing, é possível aplicar a segmentação dos clientes por níveis de importância. Os níveis são representados numa escala crescente entre 1 (menor) e 5 (maior). Esta transformação é realizada sobre os *quantis*[1], de forma individualizada, sobre a *recency*, *frequency* e *monetary value*. Assim sendo, para as métricas mencionadas a divisão aplicada consiste em:

- $<= 0.2 \Leftrightarrow 20\% \longrightarrow 1$
- $> 0.2 \;\&\; <= 0.4 \Leftrightarrow [20 - 40]\% \longrightarrow 2$
- $> 0.4 \;\&\; <= 0.6 \Leftrightarrow [40 - 60]\% \longrightarrow 3$
- $> 0.6 \;\&\; <= 0.8 \Leftrightarrow [60 - 80]\% \longrightarrow 4$
- $> 0.8 \Leftrightarrow 80\% \longrightarrow 5$

Veja-se a Figura 19, como a sua representação:

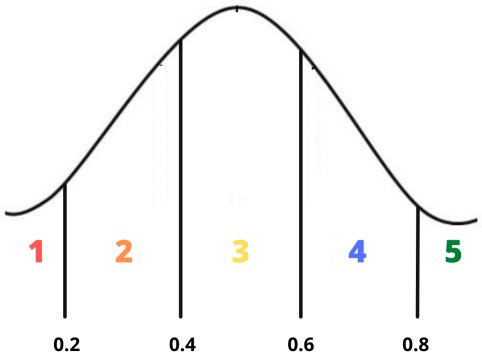


Figura 19: Segmentação para RFM

Portanto, caso a frequência máxima seja 50 então a divisão seria: $[0 - 10] = 1$, $[10 - 20] = 2$, $[20 - 30] = 3$, $[30 - 40] = 4$ e $[40 - 50] = 5$. Sendo assim, nesta analogia, um cliente com uma frequência equivalente a 23 corresponderia ao nível 3.

Logo após a esta divisão, cada cliente é associado a um grupo composto por 3 algarismos (entre 111 e 555), resultante da junção das 3 métricas. Consequentemente esse grupo é acompanhado por uma etiqueta relacionada com o histórico e capaz de identificar a categoria do cliente. Resultante disso, surgem 12 etiquetas caracterizadas na Tabela 13.

[1] Quantil divide os dados ordenados em subconjuntos de dados.

Tabela 13: Grupos segmentados

Intervalo	Nome	Cliente tipo
$11[1-5]$	*new customer or no repeat*	Novos ou antigos com poucas compras.
$[1-2][1-2][1-5]$	*hibernating*	Vieram poucas vezes e estão ausentes.
$[1-2][3-4][1-5]$	*at risk*	Frequente, mas ausente há algum tempo.
$[1-2]5[1-5]$	*cant loose*	Bastante frequente, mas ausente há algum tempo.
$3[1-2][1-5]$	*about to sleep*	Mediano em queda.
$33[1-5]$	*need attention*	Mediado.
$[3-4][4-5][1-2]$	*loyal spared*	Leal e poupado.
$[3-4][4-5][3-5]$	*loyal spender*	Leal e dispendioso.
$[4-5]1[1-5]$	*promising*	Comprou recentemente e com baixa frequência.
$[4-5][2-3][1-5]$	*potential loyal*	Esporádico com compra recente.
$5[4-5][1-3]$	*potential champion*	Leal, frequente e poupado.
$5[4-5][4-5]$	*champion*	Leal, frequente e dispendioso.

Relativamente à métrica T, e fundamentada na lógica de negócio, a segmentação compõe-se em 3 grupos: os clientes mais antigos, os clientes neutros e os clientes recentes. A lista seguinte detalha o método do agrupamento:

- $<= 0.33 \Leftrightarrow 33\% \longrightarrow$ 'C': Clientes mais recentes;
- > 0.33 & $<= 0.66 \Leftrightarrow [33-66]\% \longrightarrow$ 'B': Clientes neutros;
- $> 0.66 \Leftrightarrow 66\% \longrightarrow$ 'A': Clientes mais antigos.

Veja-se a Figura 20, como a sua representação:

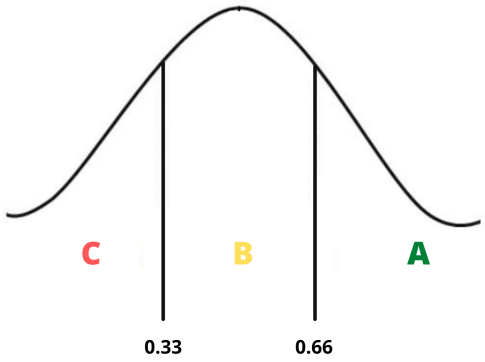


Figura 20: Segmentação para T

Portanto, caso o T máximo seja 30 então a divisão seria: $[0-10] = C$, $[10-20] = B$ e $[20-30] = A$. Nesse caso, um cliente com um T igual a 4 seria atribuído a letra 'C'.

Resumidamente, cada empresa terá agregado 6 novos *outputs* por cliente. São eles a representação escalar da *recency*, *frequency*, *monetary value* e *T*, grupo numérico (junção dos algarismos de RFM) e etiqueta associada, observe-se a Tabela 14. O fluxograma A.3 detalha o processo apresentado.

Tabela 14: *Output* típico de segmentação

	...	R Seg	F Seg	M. Value Seg	Group	Tag
1	...	1	2	3	123	hibernating
2	...	2	4	1	241	at risk
3	...	5	1	1	511	promising
4	...	5	5	5	555	champion
...	...	...	...	...	...	...

4.5 MLOps

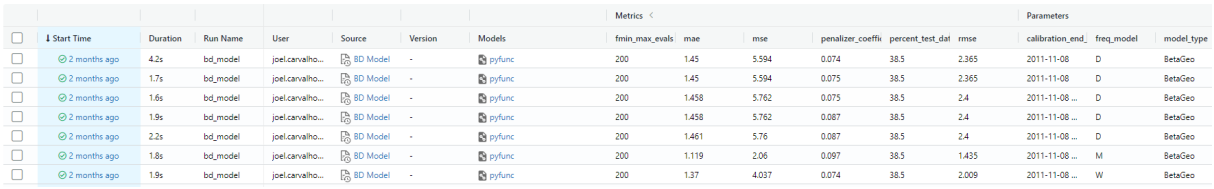
Como citado em 2.6, o MLOps simplifica da gestão do ciclo de vida de ML, visto que o desenvolvimento é realizado em larga escala, então a sua utilização evidencia-se.

Seguidamente segue-se a análise sobre as 4 componentes do MLflow, *Tracking*, *Projects*, *Models* e *Registry*, durante a implementação.

No que se refere à componente de ***Tracking***, que se destina essencialmente à manutenção do histórico dos parâmetros, esta tornou-se crucial para a interpretação dos resultados e afinação de algumas estratégias na implementação. Inicialmente, de forma a explorar todas as possibilidades associadas à agregação o histórico dos clientes e aos modelos, o conteúdo analisado dividiu-se em:

- ***Metrics***: Métricas numéricas
  - ***fmin_max_evals***: Número máximo de avaliações para o *tuning*;
  - ***MAE***: *Mean absolute error*
  - ***MSE***: *Mean squared error*, métrica selecionada pela popularidade e simplicidade.
  - ***RMSE***: *Root-mean-square error*
  - ***penalizer_coefficient***: Penalizador resultante do *tuning*;
  - ***percent_test_data***: Percentagem de data usada para *holdout*.
- ***Parameters***: Parâmetros textuais
  - ***calibration_end_date***: Data limite (inclusive) para a *calibration*;
  - ***freq_model***: Frequência para agregação do histórico de compras dos clientes;
  - ***model_type***: Nome do modelo utilizado.

Esta análise deu origem à definição da melhor combinação entre as métricas e os parâmetros descritos, veja-se o *output* típico na Figura 21.



								Metrics						Parameters		
	Start Time	Duration	Run Name	User	Source	Version	Models	fmin_max_evals	mae	mse	penalizer_coeffi	percent_test_dat	rmse	calibration_end_	freq_model	model_type
	2 months ago	4.2s	bd_model	joel.carvalho...	BD Model	-	pyfunc	200	1.45	5.594	0.074	38.5	2.365	2011-11-08	D	BetaGeo
	2 months ago	1.7s	bd_model	joel.carvalho...	BD Model	-	pyfunc	200	1.45	5.594	0.075	38.5	2.365	2011-11-08	D	BetaGeo
	2 months ago	1.6s	bd_model	joel.carvalho...	BD Model	-	pyfunc	200	1.458	5.762	0.075	38.5	2.4	2011-11-08 ...	D	BetaGeo
	2 months ago	1.9s	bd_model	joel.carvalho...	BD Model	-	pyfunc	200	1.458	5.762	0.087	38.5	2.4	2011-11-08 ...	D	BetaGeo
	2 months ago	2.2s	bd_model	joel.carvalho...	BD Model	-	pyfunc	200	1.461	5.76	0.087	38.5	2.4	2011-11-08 ...	D	BetaGeo
	2 months ago	1.8s	bd_model	joel.carvalho...	BD Model	-	pyfunc	200	1.119	2.06	0.097	38.5	1.435	2011-11-08 ...	M	BetaGeo
	2 months ago	1.9s	bd_model	joel.carvalho...	BD Model	-	pyfunc	200	1.37	4.037	0.074	38.5	2.009	2011-11-08 ...	W	BetaGeo

Figura 21: MLflow *Tracking*

Posteriormente, após a análise e a definição das métricas e dos parâmetros listados anteriormente, as métricas registadas foram suprimidas apenas para o MSE e o *penalizer_coefficient*, assim como os parâmetros registam apenas o ID da empresa (*company_id*) e o tipo de modelo associado (*model_type*).

Relativamente aos ***Models***, que consiste numa diretoria com a capacidade de fornecer ferramentas de *deploy* e armazenar *artifacts*[2] exclusivos a cada modelo, contribuiu principalmente para a análise dos *outputs* gráficos do modelo. Para além disto, cada modelo define um *schema* de *input* e *output* com o objetivo de identificar os dados de entrada e saída para outros *data scientists*. Vejam-se as Figuras 22 e 23, em representação da visão geral de um modelo e um exemplo de um *artifact*, respetivamente.

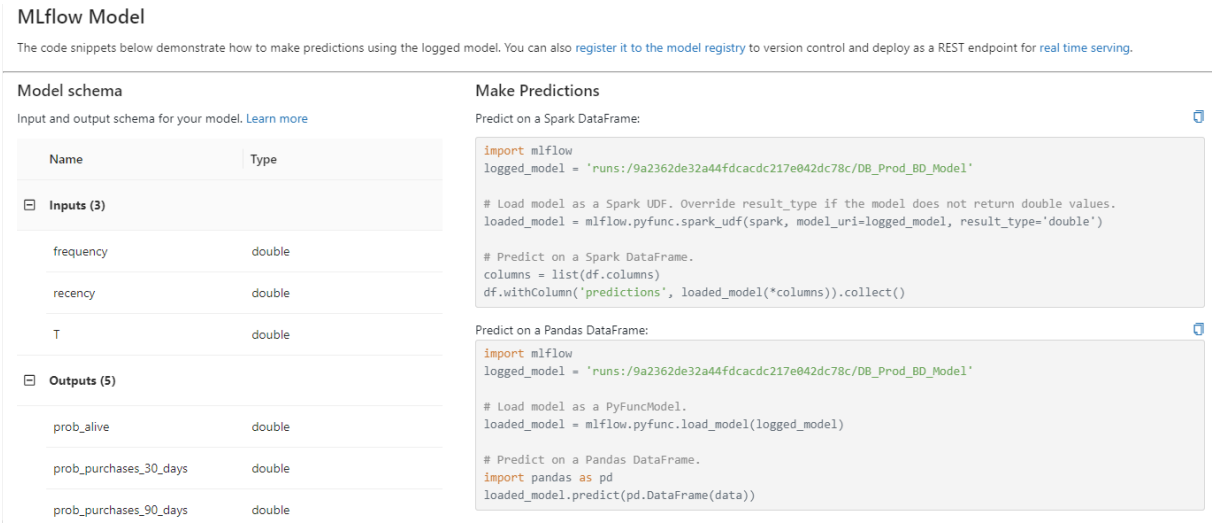


Figura 22: MLflow *Model* - Visão Geral

[2]Ficheiros de *output* associados ao modelo, como imagens ou ficheiros *.txt*.

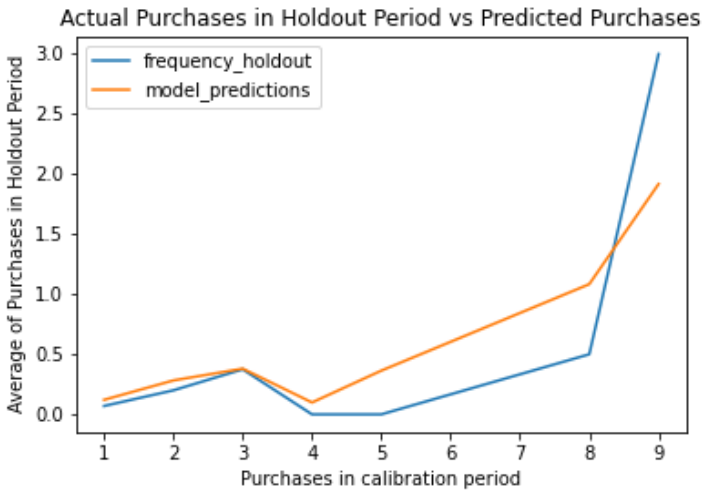


Figura 23: MLflow *Model* - Exemplo de um *artifact*

Quanto à componente ***Projects***, habilitado principalmente para a reutilização de código, composto por diversos ficheiros de configurações (como conda.yaml, train.py, MLmodel, requirements.txt), não releva vantagens de utilização visto que ainda se trata de um projeto individual.

Por fim o ***Registry***, tem a capacidade de gerir versões do modelo, como alterar o estado do modelo (*staging*, *production* e *archived*), e fornecer um *endpoint* para consumir o modelo via *Application Programming Interface (API)*. Contudo, atualmente, este componente não oferece nenhuma vantagem, visto que, o modelo em produção será sempre o último a ser treinado. Futuramente, a sua utilização poderá ser benéfica, como citado em 6.2.

4.6 Armazenamento dos Dados

O armazenamento dos dados representa grande importância, pelo motivo de que através dele somos capazes de manter o histórico das previsões e entender temporalmente a performance dos modelos mediante o cálculo de uma taxa de erro. Com esse armazenamento compulsivo e recorrente, o volume de dados torna-se abundante. À vista disso, o esquema está implementado com suporte à *Big Data*.

Face às exigências do volume de dados, a informação é armazenada no *Data Lake* através das *delta tables* assente nos princípios do *Delta Lake*. Assim sendo, o armazenamento dos dados é composto por 5 *delta tables*, representadas de forma diagramática na Figura 24.

Figura 24: Modelo Entidade-Relação

Em que:

- *ModelDetails* - Armazena os dados relacionados com os modelos - Veja-se a Tabela 15 como a sua representação;

Tabela 15: *Delta Table ModelDetails*

Coluna	Conteúdo
PhysicalKey	0fa144237
BusinessKey	8029a35fc
UpdateDate	01-01-2022
ModelName	BetaGeo
PenalizerCoefficient	0.01512
Tuning	true
MSE	0.484375

- *Binominal Distribution (BDPredicts)* - Armazena os dados relacionados com as previsões disponibilizadas pelos modelos de previsão de compras e *churn* - Veja-se a Tabela 16 como a sua representação;

Tabela 16: *Delta Table BDPredicts*

Coluna	Conteúdo
PhysicalKey	0fa144237
BusinessKey	8029a35fc
CustomerID	ebc95f307
UpdateDate	01-01-2022
Expected1MPurchases	0.21512
Expected3MPurchases	0.71512
Expected6MPurchases	1.484375
Expected13MPurchases	3.01512
ProbAlive	0.98

- *GammaGamma (GGPredicts)* - Armazena os dados relacionados com as previsões disponibilizadas pelos modelos de previsão faturação - Veja-se a Tabela 17 como a sua representação;

Tabela 17: *Delta Table GGPredicts*

Coluna	Conteúdo
PhysicalKey	0fa144237
BusinessKey	8029a35fc
CustomerID	ebc95f307
UpdateDate	01-01-2022
ExpectedAverageProfit	80.52894
Expected1MCLV	20.14982
Expected13MCLV	260.484375

- *RealResults* - Armazena os dados relacionados com o erro (avalia as previsões do mês passado e compara com as transações reais) - Veja-se a Tabela 18 como a sua representação;

Tabela 18: *Delta Table RealResults*

Coluna	Conteúdo
PhysicalKey	0fa144237
BusinessKey	8029a35fc
UpdateDate	01-02-2022
ModelName	BD
OriginalDate	01-01-2022
MSE	0.484375

- *RFMMetrics* - Armazena os dados relacionados com a segmentação - Veja-se a Tabela 19 como a sua representação;

Tabela 19: *Delta Table RFMMetrics*

Coluna	Conteúdo
PhysicalKey	0fa144237
BusinessKey	8029a35fc
CustomerID	ebc95f307
UpdateDate	01-01-2022
Recency	704
Frequency	171
T	705
MonetaryValue	185.40142
RecencySegmentation	5
FrequencySegmentation	5
TSegmentation	A
MonetaryValueSegmentation	5
Group	555
Tag	champion

Observável na Figura 24, todas as *delta tables* são compostas por 2 partições («*part*») que armazenam a informação pertencente às organizações, a *PhysicalKey* e a *BusinessKey*. A sua utilização é vantajosa na orquestração de um grande volume de dados porque:

- Divide tabelas grandes em subconjuntos de dados e mantém a integridade dos mesmos;
- As operações são mais eficientes devido ao *target* definido pelos subconjuntos, ao invés de analisar a tabela global;
- Acelera a gestão de dados e as consultas nas tabelas.

4.7 Representação dos Dados

A representação dos dados apresenta-se por 2 formatos viáveis: por *dashboards* individualizados no produto ou integrado nos *business report*. Na primeira solução a comunicação seria realizada diretamente aos empresários e o detalhe da informação seria diário e mais pormenorizado, enquanto que na segunda alternativa a comunicação seria efetuada aos contabilistas que reportam os resultados mensais do negócio aos empresários.

Seguindo as propostas da dissertação do Leandro Rocha [70], cujo objetivo é implementar uma interface para os meus resultados, vejam-se as Figuras 25, 26 e 27, a primeira forma de representação dos dados evidenciou-se. Visto que oferece a capacidade aos clientes de manipular e explorar *dashboards* de vendas e faturação, filtrar clientes, análise pormenorizada do histórico e das previsões de compras para 1 mês, 3 meses, 6 meses, 13 meses, previsões do valor monetário futuro e do CLV e a probabilidade de *churn*.

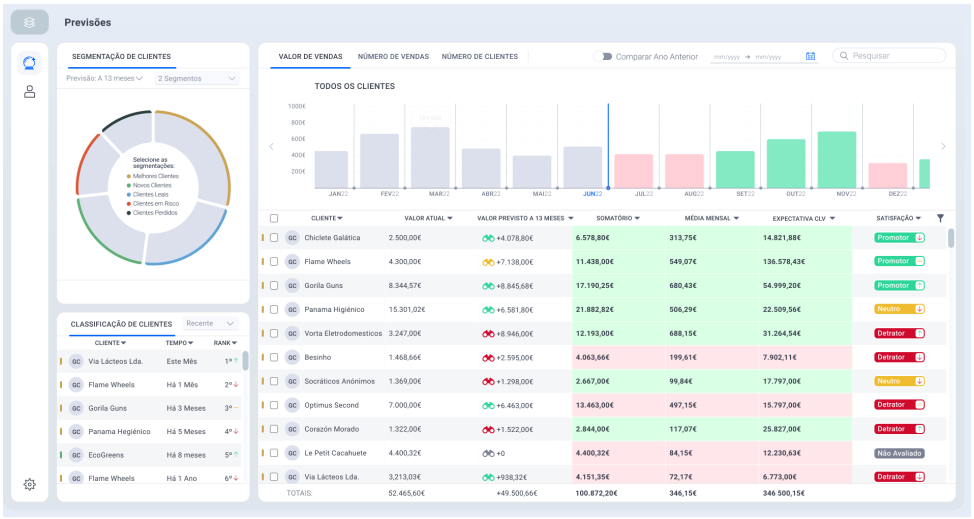


Figura 25: Representação dos dados - Layouts - Homepage

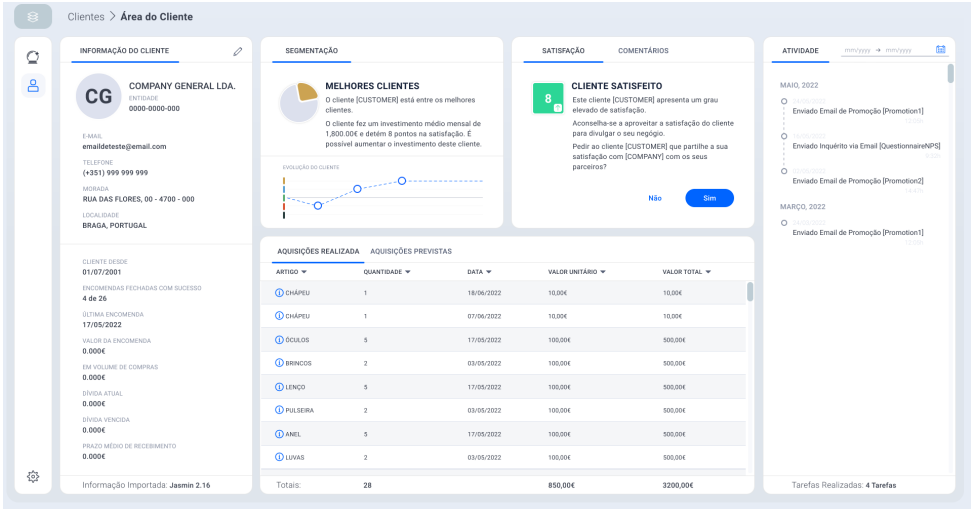


Figura 26: Representação dos dados - Layouts - Área de cliente individualizada

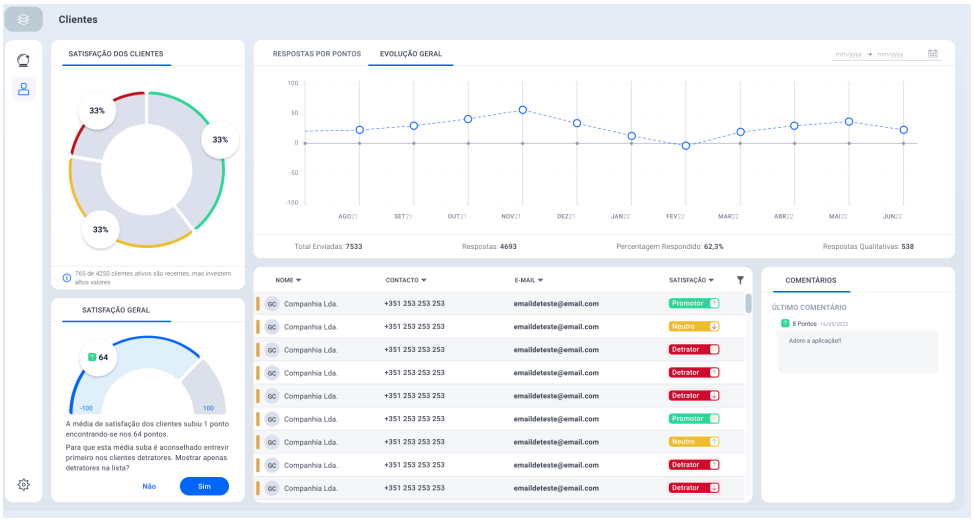


Figura 27: Representação dos dados - Layouts - Área de clientes

O seguinte conjunto de imagens, Figuras 28, 29, 30, 31 e 32 apresentam um protótipo desenvolvido que exibe alguns do *outputs* seguidamente delineados.

- Listagem do histórico de previsões passadas;
- Listagem de futuras previsões;
- Segmentação dos clientes seguindo a descrição da secção 4.4;
- Volume de vendas nos últimos X meses;
- Faturação gerada nos últimos X meses;
- Identificação dos clientes com maior e menor taxa de atividade;
- Listagem de tops gráficos:
  - Clientes mais frequentes;
  - Clientes mais recentes;
  - Clientes mais promissores (*promising*);
  - Melhores clientes (*champions*);
  - Clientes em risco (*at_risk*);
  - Entre outros.

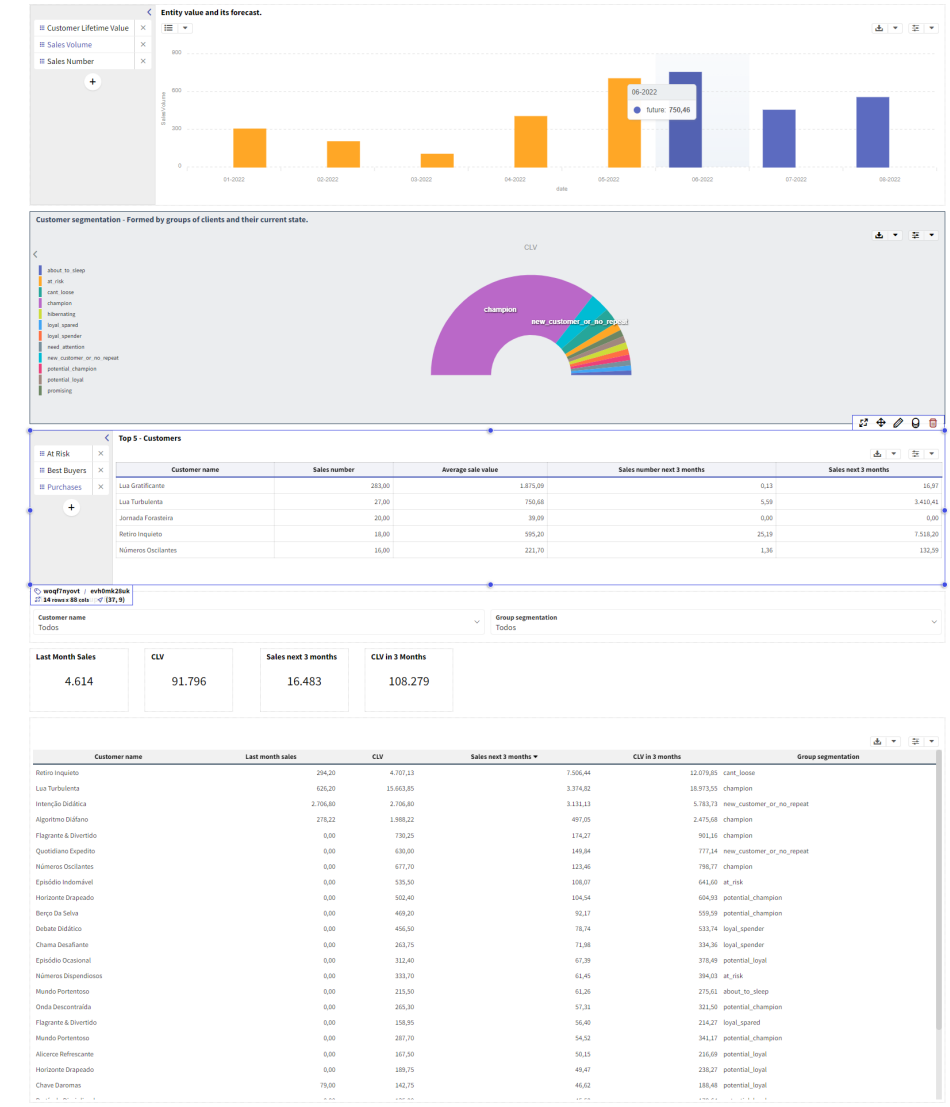


Figura 28: Representação dos dados - Homepage

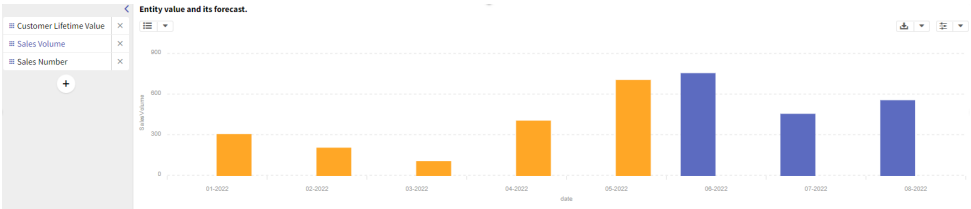


Figura 29: Representação dos dados - Volume de vendas (histórico e previsões)

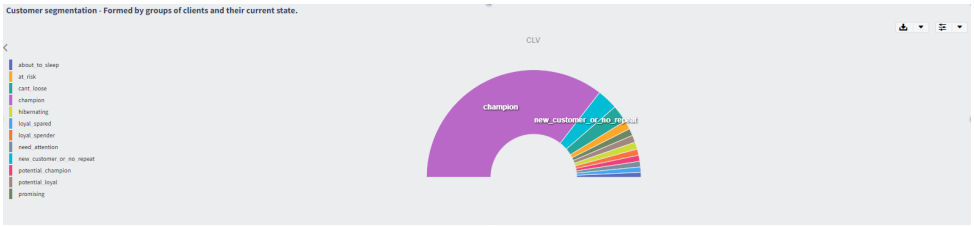


Figura 30: Representação dos dados - Segmentação

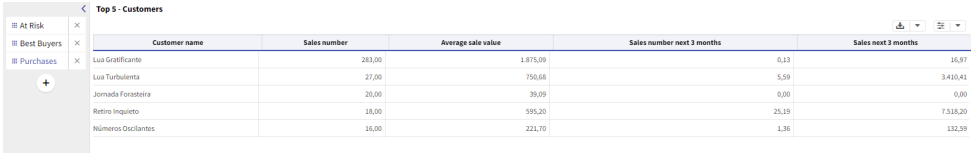


Figura 31: Representação dos dados - Tops

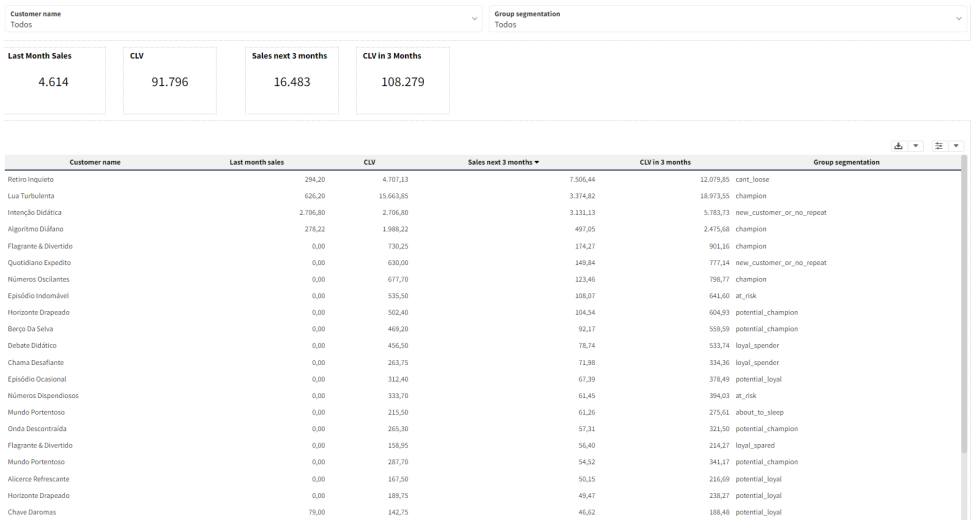


Figura 32: Representação dos dados - Lista de clientes segmentados e previsões

4.8 Conclusões

Este capítulo pormenoriza os passos da implementação. Detalha a etapa crucial da receção e da preparação dos dados. Além disso, identifica as fases cruciais que caracterizam e justificam a solução, aborda a integração com a ferramenta MLflow e de como os dados são armazenados. Por fim, considerou-se este formato como o modo de representação dos dados, mais conveniente para os empresários.

5 Resultados

O presente capítulo 5, irá descrever os resultados e abordagens tentadas relacionadas com a implementação dos modelos. Posto isto, a secção 5.1 descreve a validação dos modelos seguindo múltiplas abordagens e diferentes *datasets*. A secção 5.2 pormenoriza as diferentes tentativas abordadas e consequentemente os seus obstáculos. Finalmente a secção 5.3 sumariza este capítulo.

5.1 Validação dos Modelos

Com a finalidade de validar os resultados dos modelos, diversas abordagens foram produzidas. A presente secção divide-se em 3 subsecções, em que, particulariza o trabalho realizado sobre diferentes *datasets*.

5.1.1 *Dataset Online Retail*

Inicialmente, o trabalho efetuado debruçou-se sobre a exploração do *dataset* disponibilizado, pelo repositório online do centro de *Machine Learning* e Sistemas Inteligentes. Este *dataset* é composto pelas transações de uma loja online, efetuadas entre 01-12-2010 e 09-12-2011. Cada transação é composta pela seguinte informação:

- **InvoiceNo**: Número da fatura;
- **StockCode**: Código do produto;
- **Description**: Nome do produto;
- **Quantity**: Quantidade de cada produto;
- **InvoiceDate**: Data em que a compra foi realizada;
- **UnitPrice**: Preço unitário de cada produto;
- **CustomerID**: Identificador único do cliente;

- **Country**: País onde ocorreu a compra.

Posteriormente ao tratamento de dados, além da métrica MSE, para avaliação dos resultados foi introduzida outra para obtermos um erro em forma de percentagem, intitulada de *Manual Score*. O cálculo é efetuado seguindo a próxima listagem:

- Valor Real == Valor Previsto então $+1$;
- Valor Real e Valor Previsto aproximado com *threshold* $= 1$ (variável), então:
  - Valor Real $= 0$ e Valor Previsto $= 1$ então $+0$;
  - (Valor Previsto / Valor Real) < 1 então $+1$;
  - (Valor Previsto / Valor Real) > 1 então $+0.5$;

  No final, é devolvida percentagem entre o valor somado divido pelo total de amostras e multiplicado por 100, $soma/total * 100$.

Posto isto, a Tabela 20 resume a performance obtida sobre os modelos de previsão de compras. Evidencia-se, que quando maior for o histórico e menor espaço de previsão o erro é bastante aceitável.

Tabela 20: Resumo da performance do modelo de previsão de compras com o *dataset* online

			1 ANO TREINO E 1 ANO TESTE		
	7 dias	**1 mês**	**3 meses**	**6 meses**	**1 ano**
MSE	-	-	1.546	5.609	26.566
Manual Score	-	-	53.465%	41.848%	33.418%
			1.5 ANO TREINO E 6 MESES TESTE		
MSE	0.052	0.231	-	5.572	-
Manual Score	99.980%	92.289%	-	52.481%	-

De maneira a explorar ainda mais os modelos, aumentou-se, através de dados sintéticos, o tamanho do *dataset* em mais 1 ano (duplicação das transações do ano de 2011), isso permitiu avaliar as previsões requisitadas (1, 3, 6 e 13 meses). A Tabela 21 descreve o aumento expectável da precisão e consequentemente a descida do erro, mediante o aumento dos dados de treino.

Tabela 21: Resumo da performance do modelo de previsão de compras com o *dataset* online com dados sintéticos

		23 MESES TREINO E 13 MESES TESTE		
	1 mês	**3 meses**	**6 meses**	**13 meses**
MSE	0.592	0.881	1.651	5.238
Manual Score	79.295%	69.648%	67.697%	59.318%

Relativamente à performance do modelo GammaGamma, a avaliação com as métricas mencionadas torna-se complexa. Como este modelo assenta na previsão de lucro, ao calcular o MSE, caso o valor previsto for 10€ acima das expectativas traduzira um erro bastante elevado (como 10758.9). Outra adversidade é que para um empresário em que o modelo preveja 10€ abaixo das expetativas pode ser drástico, porém para outro não têm grande impacto. Apesar disso, as previsões do valor monetário médio e do CLV são bastante positivas.

5.1.2 *Dataset* de Produção - *Canonical Business Data*

Analisando a performance dos dados de produção, os resultados foram bastante confiáveis. Contudo, no total de 40 empresas disponibilizadas pelo *Canonical Business Data* apenas foram produzidos 22 modelos, dos quais, exclusivamente 2 cumpriram os requisitos definidos em 4.3, as 18 restantes produziram previsões baseado no cálculo descrito em 4.3.1. Veja-se a Tabela 22 que exibe detalhes sobre 40 as instituições.

Tabela 22: Sumário da performance dos modelos de previsão de compras com dados de produção

	N.º Clientes	N.º Dias	Cumpre Requisitos	Modelo	MSE
1	73	236	✓	BetaGeo	0.48438
2	64	269	✓	ModifiedBetaGeo	1.26230
3	8	489	-	BetaGeo	×
4	1	180	-	BetaGeo	×
5	2	271	-	BetaGeo	×
6	3	88	-	BetaGeo	×
7	2	185	-	BetaGeo	×
8	5	295	-	BetaGeo	×
9	2	121	-	BetaGeo	×
10	1	30	-	BetaGeo	×
11	16	60	-	BetaGeo	×
12	1	52	-	BetaGeo	×
13	6	346	-	BetaGeo	×
14	5	272	-	BetaGeo	×
15	10	314	-	BetaGeo	×
16	52	58	-	BetaGeo	×
17	5	329	-	BetaGeo	×
18	14	62	-	BetaGeo	×
19	3	50	-	BetaGeo	×
20	3	103	-	BetaGeo	×
21	1	124	-	BetaGeo	×
22	3	268	-	BetaGeo	×
23	25	29	×	×	×
24	23	25	×	×	×
25	23	25	×	×	×
26	2	29	×	×	×
27	1	10	×	×	×
28	1	8	×	×	×
29	1	23	×	×	×
30	1	1	×	×	×
31	1	23	×	×	×
32	0	0	×	×-	×
33	0	0	×	×	×
34	0	0	×	×	×
35	0	0	×	×	×
36	0	0	×	×	×
37	0	0	×	×	×
38	0	0	×	×	×
39	0	0	×	×	×
40	0	0	×	×	×

(✓) Sim
(-) Alguns
(×) Não

Relativamente às empresas que contêm um modelo sem *tuning*, apesar das advertências, os *outputs* são positivos e seguem a lógica do histórico. Isto significa que um cliente com um histórico de 5 meses tem a capacidade de produzir previsões 100% assertivas. Contudo, para clientes com histórico equivalente a 1 ou 2 meses os resultados podem ser incertos.

De forma a avaliar o processo incremental do negócio, fazendo um paralelismo com séries temporais, aplicou-se o *forward-chaining cross-validation*, também designado por *rolling-origin cross-validation*, para empresas com um histórico superior a 2 meses. Este método, permite treinar os últimos n meses de histórico e validar com a previsão do próximo mês (m), a janela de treino aumenta $n + m$ após cada avaliação, veja-se a Figura 33 como a sua representação.

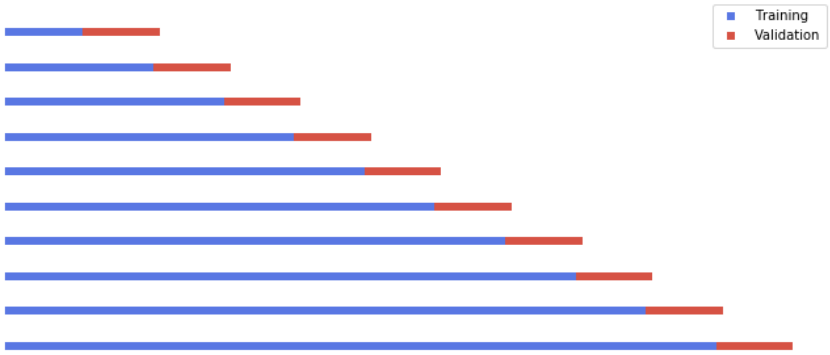


Figura 33: *Forward-chaining cross-validation*

Não obstante a sua aplicação ser mais objetiva com históricos superiores, esta validação oferece uma perceção da regularidade dos clientes por empresa, mitigação de possíveis erros inesperados e/ou previsões imprecisas (Figura 35), análise mensal do volume de clientes e expectavelmente, uma quebra gradual ou um ligeiro aumento (devido aos novos clientes ou clientes irregulares) do MSE (Figura 34). Observe-se esse processo aplicado à empresa 1 da Tabela 22 na Tabela 23.

Tabela 23: *Forward-chaining cross-validation* sobre a empresa 1 (data inicial:2021/04/08) (22)

Data de Calibração	Data Final	MSE	Manual Score (th=1)	Num. Clientes
2021-06-07	2021-07-07	13.00	50.00%	2
2021-07-07	2021-08-06	4.06	64.71%	17
2021-08-06	2021-09-05	0.59	70.37%	27
2021-09-05	2021-10-05	0.75	76.25%	40
2021-10-05	2021-11-04	0.29	79.81%	52
2021-11-04	2021-12-04	1.03	81.97%	61

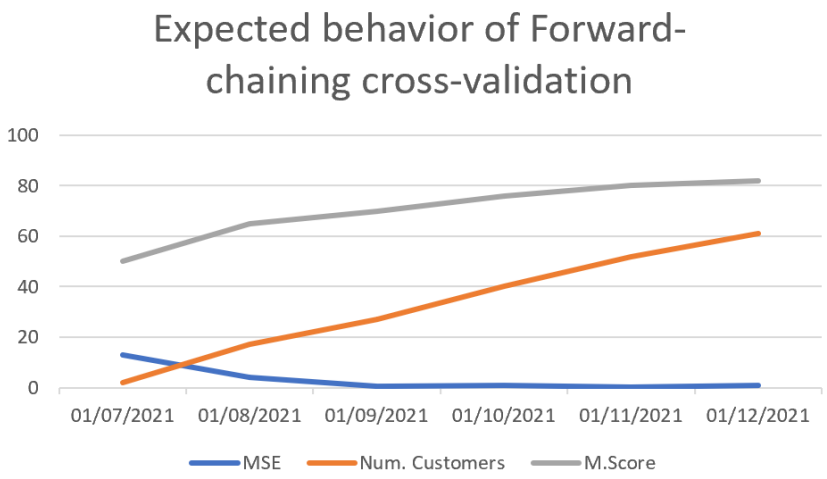


Figura 34: Comportamento esperado da aplicação do *forward-chaining cross-validation* relativo à Tabela 23

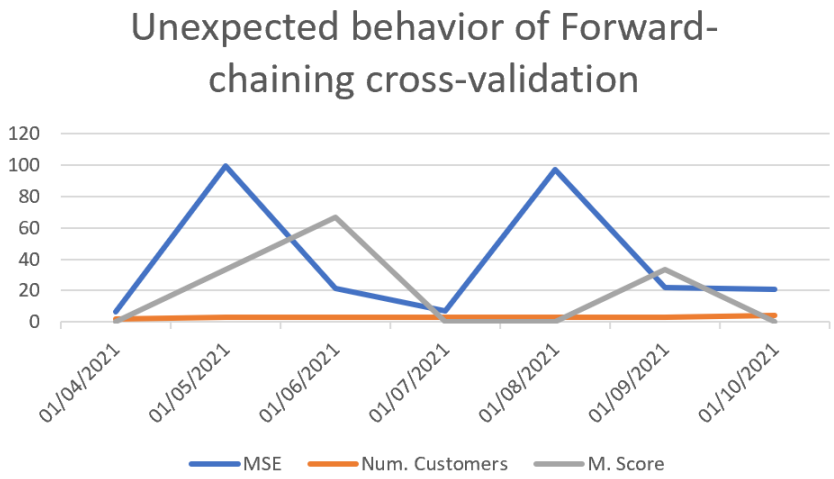


Figura 35: Comportamento inesperado da aplicação do *forward-chaining cross-validation*

Com o decorrer do tempo, os dados do *Canonical Business Data* avolumaram, o que permitiu interpretar o crescendo do negócio, veja-se a Tabela 24, e analisar a performance dos modelos 36 e 37.

Tabela 24: Aumento do volume de dados do *Canonical Business Data*

	1 Versão de dados	**2 Versão de dados**
N° linhas	149 633	393 930
N° faturas	34 915	68 469
N° empresas	40	106
N° empresas com *tuning*	2	12
N° empresas com modelo (> 1 mês)	22	44
N° empresas com novas vendas: 10 de 22		

A Figura 36 sintetiza graficamente o cálculo do MSE entre as previsões de vendas da primeira versão dos dados com os dados reais (segunda versão), enquanto a Figura 37 sintetiza graficamente a taxa de assertividade através do *Manual Score (threshold = 1)*. Uma análise particular para as empresas 1 e 2, que superaram os filtros na primeira versão dos dados, evidencia-se a alta performance da empresa 1, todavia a empresa 2, apesar de possuir um MSE baixo, a taxa de assertividade não alcançou os 50%. As restantes empresas, que produziram previsões com base no modelo da empresa 1, obtiveram resultados díspares, como o exemplo da empresa 7 e da empresa 10, que apesar de possuírem o mesmo modelo, alcançaram as piores e as melhores previsões, respetivamente. Em conclusão e como expectável, as previsões mais assertivas e lineares ocorrem em empresas com dados suficientemente compostos e ultrapassem os filtros delineados.

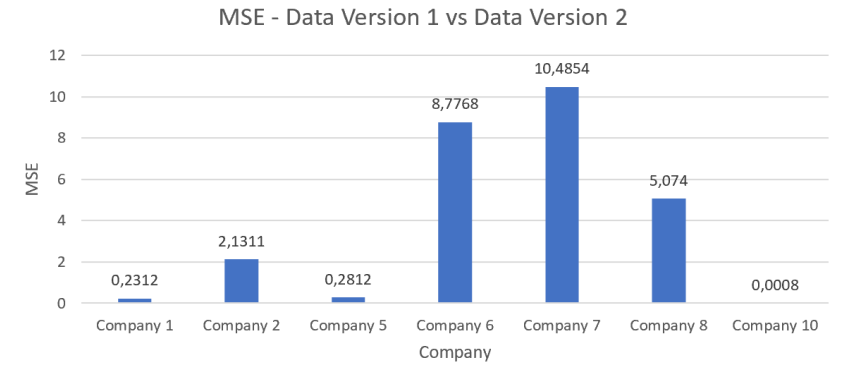


Figura 36: Comparação das previsões de vendas da versão 1 com a versão 2 - MSE

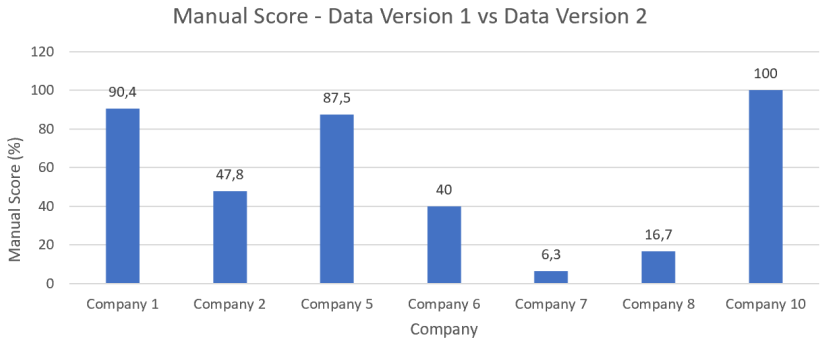
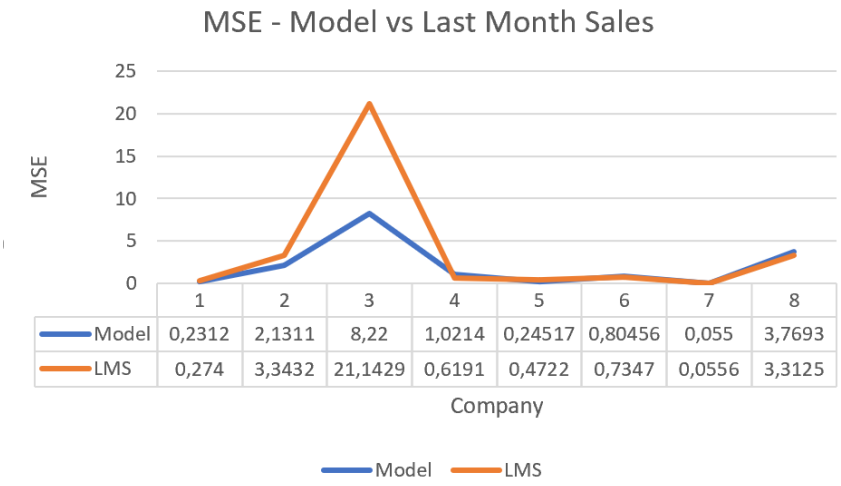


Figura 37: Comparação das previsões de vendas da versão 1 com a versão 2 - Manual Score

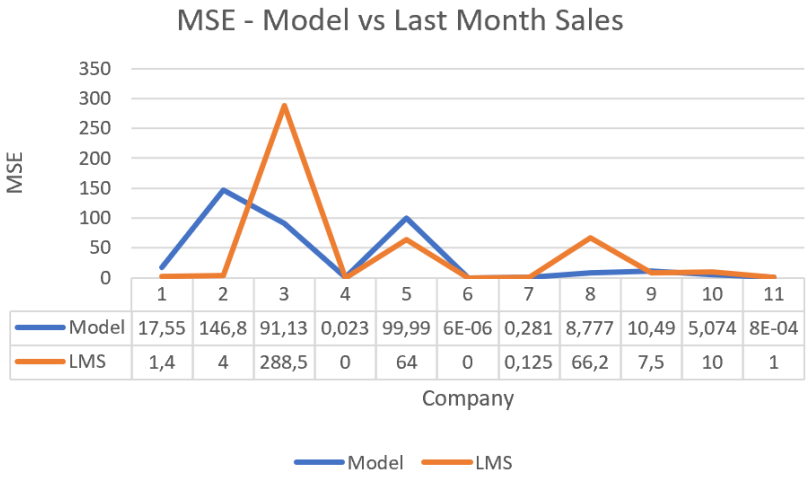
Representado na Figura 38, compara-se a performance do modelo das empresas que realizaram *tuning* (considerado os melhores modelos), com o histórico de vendas do último mês (*baseline*). Facilmente interpreta-se que o modelo prevê, maioritariamente, maior estabilidade e melhores resultados. Contudo, observando a Figura 39, concluí-se que as empresas que contêm o modelo de configurações genéricas (modelo com erro mais baixo - Empresa 1), na generalidade, apesar de conter mais estabilidade, tendem

a produzir previsões menos assertivas comparativamente com o último histórico mensal. Isto significa que, para empresas inferiores a 1 mês de faturação, a previsão baseada nesse ínfimo histórico é um bom presságio e tendem a ser eficientes.



	1	2	3	4	5	6	7	8
Model	0,2312	2,1311	8,22	1,0214	0,24517	0,80456	0,055	3,7693
LMS	0,274	3,3432	21,1429	0,6191	0,4722	0,7347	0,0556	3,3125

Figura 38: Comparação de resultados entre os melhores modelos e o histórico do último mês entre empresas que realizaram *tuning*



	1	2	3	4	5	6	7	8	9	10	11
Model	17,55	146,8	91,13	0,023	99,99	6E-06	0,281	8,777	10,49	5,074	8E-04
LMS	1,4	4	288,5	0	64	0	0,125	66,2	7,5	10	1

Figura 39: Comparação de resultados entre o modelo e o histórico do último mês entre empresas que realizaram *tuning*

5.1.3 *Datasets* Sintéticos

No seguimento da validação dos modelos foram produzidos 3 *datasets* simulados de forma a interpretar o desempenho em múltiplos cenários possíveis.

5.1.3.1 *Dataset 1*

Dataset de 12 meses, 01/01/2021 – 02/12/2021, composto por 1 cliente que efetua regularmente 2 compras mensais no dia 1 e 2 de cada mês. O objetivo passa por testar a capacidade de o modelo prever as duas compras mensais e a assertividade dos dias. O modelo utilizou 9 meses, 01/01/2021 – 02/09/2021, para treino e obteve os seguintes resultados:

- O modelo previu **corretamente** 2 compras para o primeiro mês (03/09/2021 – 03/10/2021);
- O modelo previu **corretamente** 4 compras para o segundo mês (03/09/2021 – 03/11/2021);
- O modelo previu **corretamente** 6 compras para o terceiro mês (03/09/2021 – 03/12/2021).

Assim sendo, o modelo consegue detetar as 2 compras mensais, contudo como o modelo calcula probabilidades não consegue interpretar o padrão dos dias de compras. Exemplificando, o modelo entre o 03/09/2021 – 12/09/2021 previu 0,687756 compras (0 compras na realidade), isto acontece porque o modelo dá primazia à assertividade mensal.

5.1.3.2 *Dataset 2*

Dataset de 11 meses, 01/01/2021 – 02/11/2021, composto por 1 cliente que efetua regularmente 2 compras mensais nos meses ímpares (janeiro, março, etc) no dia 1 e 2 de cada mês. O objetivo passa por testar a capacidade de o modelo prever as duas compras mensais nos meses ímpares e a assertividade dos dias. O modelo utilizou 7 meses, 01/01/2021 – 02/07/2021, para treino e obteve os seguintes resultados:

- O modelo previu **corretamente** 2 compras para os próximos 60 dias (03/07/2021 – 03/09/2021);
- O modelo previu **corretamente** 4 compras para os próximos 120 dias (03/07/2021 – 03/11/2021).

Assim sendo, o modelo consegue detetar as 2 compras a cada 2 meses, porém, da mesma forma do caso antecedente o modelo não deteta a regularidade e previu 1,077516 compras para o mês 08 (0 compras na realidade).

5.1.3.3 *Dataset 3*

Dataset de 16 meses, 01/01/2021 – 02/05/2022, composto por 1 cliente deixa de comprar 2 meses após 6 meses regulares de compras efetuadas no dia 1 e 2 de cada mês. O objetivo passa por testar a capacidade de o modelo prever as compras de um cliente em quebra. O modelo utilizou 13 meses, 01/01/2021 – 02/02/2022, para treino e obteve os seguintes resultados:

- O modelo previu **incorretamente** 1,55 compras para o primeiro mês (02/02/2022 – 02/03/2022);

- O modelo previu **incorretamente** 3,10 compras para o segundo mês (02/02/2022 – 02/04/2022);
- O modelo previu **incorretamente** 4,66 compras para o terceiro mês (02/02/2022 – 02/05/2022).

Assim sendo, o modelo não consegue detetar o intervalo das compras, esta contrariedade acontece devido ao histórico limitado e, possivelmente, há falta de utilizadores. A Tabela 25 resume o trabalho realizado sobre os *datasets* precedentes. Além disso, com a finalidade de sumariar as evidências, o apêndice A.4 descreve, em lista, algumas conclusões retiradas das validações dos modelos testados.

Tabela 25: Resumo dos *datasets* sintéticos

	Dias de previsão	**Real**	**Previsto**
Dataset 1	30	2	2
Dataset 1	60	4	4
Dataset 1	90	6	6
Dataset 2	60	2	2
Dataset 2	120	4	4
Dataset 3	30	1,55	0
Dataset 3	60	3,10	0
Dataset 3	90	4,66	2

Seguidamente, o próximo conjunto de imagens representa os *outputs* gráficos para a validação do modelo da empresa 1 da Tabela 22. A Figura 40 compara, através de um gráfico de linhas, o valor real e o valor previsto. Analisando-a, naturalmente identifica-se que o modelo foi capaz de realizar previsões bastante confiáveis.

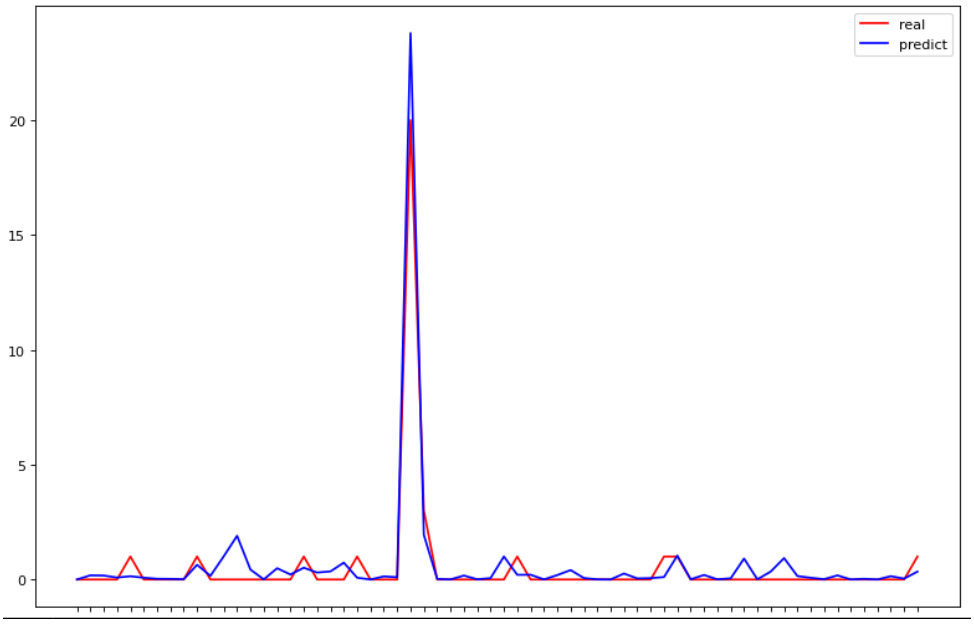


Figura 40: Comparação entre valores reais e previstos

A Figura 41 opõe valores reais com valores artificiais simulados com os parâmetros do modelo. Excluindo as previsões sobre os novos clientes (coluna 0.0), o modelo apresenta estabilidade pelo declínio constante e conseguiu produzir resultados aproximados da realidade, com a exceção de compras superiores a 5, onde se verifica um pequeno desvio entre os valores.

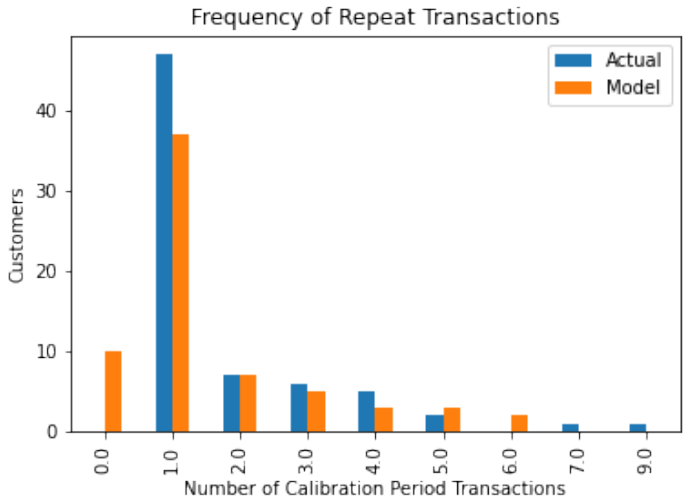


Figura 41: Comparação entre valores reais e dados sintéticos

A Figura 42 exibe as transações reais e previstas acumuladas dos clientes. Do mesmo modo, excluindo as previsões sobre os novos clientes (coluna 0.0), percebe-se que o modelo acompanha as expetativas, como evidenciado nas figuras anteriores.

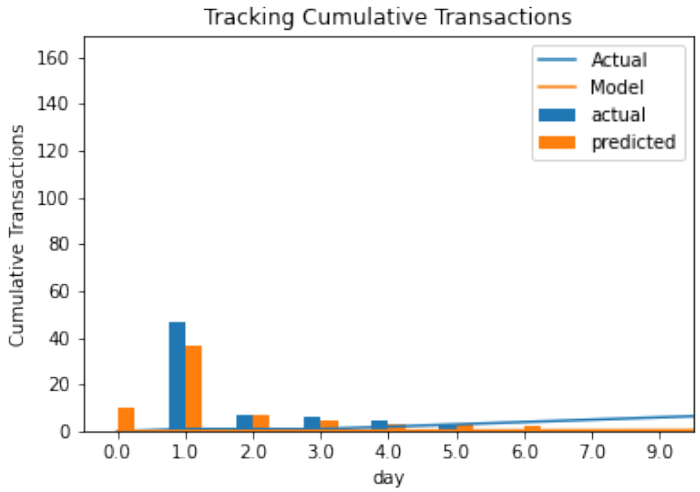


Figura 42: Comparação entre valores reais e previstos acumulados

A Figura 43 expõe graficamente a quantidade de compras realizadas no período de *holdout* e a quantidade de compras previstas desde a última compra efetuada. Bastante similar à Figura 40, porém, filtra as transações após a última compra realizada por cliente.

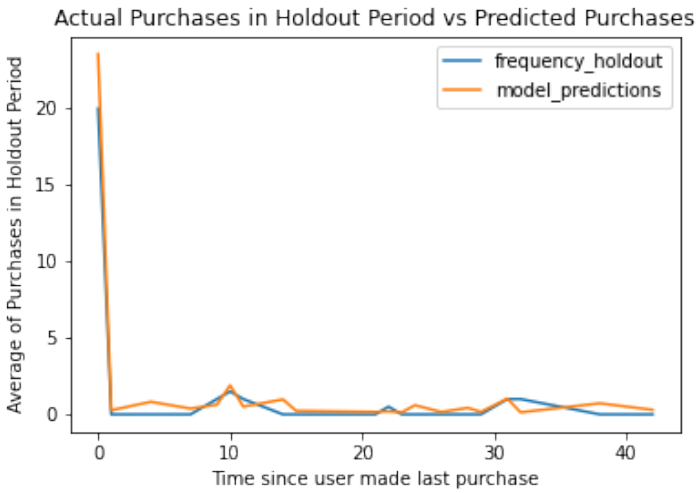


Figura 43: Comparação de compras realizadas no período de *holdout* e a quantidade de compras previstas desde a última compra efetuada

A Figura 44 apresenta uma matriz baseada no histórico da *frequency* e da *recency*. Os clientes que compraram muito, mas não recentemente, provavelmente desistiram, e quanto mais compraram no passado, maior a probabilidade de desistirem. Veja-se a matriz, no canto superior direito (zona escura mais larga) são representados os clientes débeis e no canto inferior direito (zona clara mais fina) são representados os clientes mais frequentes, isto significa que os clientes com mais de 150 compras e mais de 200 dias de atividade são os mais ativos no ecossistema da empresa 1.

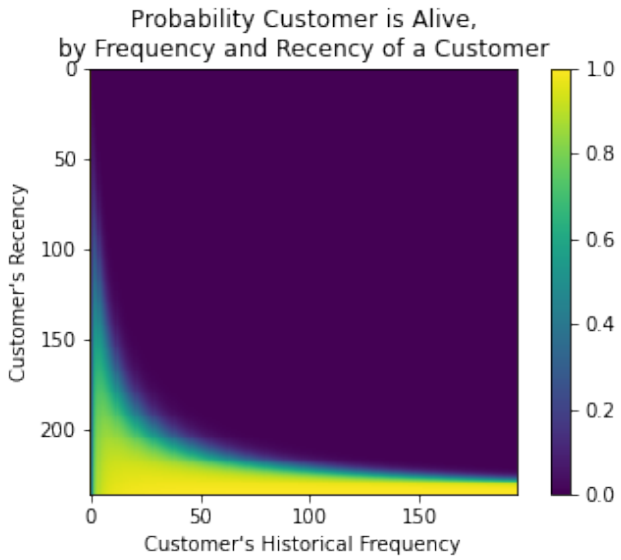


Figura 44: Matriz da taxa de atividade de um cliente

A Figura 45 apresenta, em formato matricial, a distribuição das probabilidades dos clientes da empresa 1, efetuarem compras no dia seguinte. É observável, que a probabilidade máxima seja aproximadamente 0.7, isso ocorre para os clientes que se encontram em torno dos 200 de *recency* e 175 de *frequency*, da mesma forma, concluímos que os clientes desta empresa tendem a não comprar diariamente.

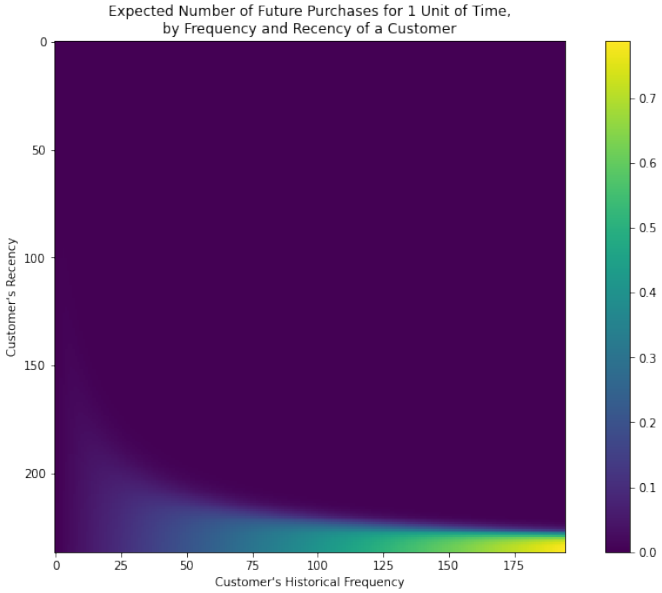


Figura 45: Matriz de probabilidade de compras para o dia seguinte

Por fim, a Figura 46 espelha a decadência de um cliente à medida que o tempo avança desde a última compra concebida. Como referido, o modelo tem a capacidade de calcular a taxa de atividade após cada compra, sendo assim, este cliente em particular, após 5 compras, provavelmente tornou-se inativo no início de agosto de 2021.

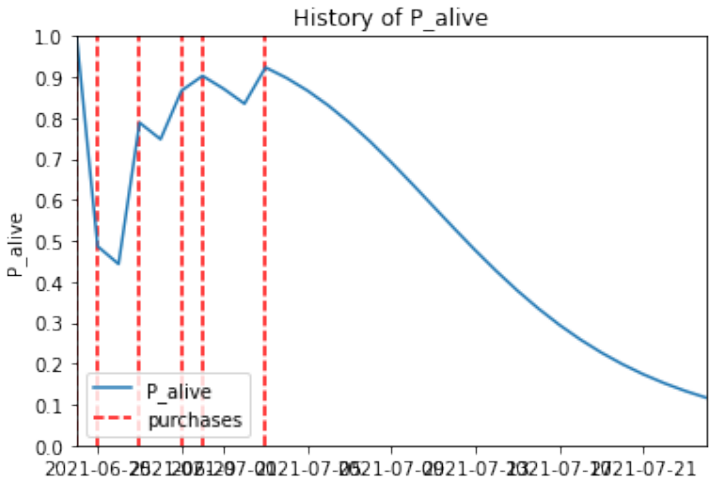


Figura 46: Probabilidade de atividade de um cliente assente na última compra

5.2 Diferentes Abordagens

No seguimento do desenvolvimento, inúmeras abordagens foram exploradas de maneira a suprimir o conjunto de possibilidades existentes. A princípio, estudou-se o caso de um modelo *one-size-fits-all*, que permitiria um modelo único para qualquer empresa, independentemente da sua dimensão. A vantagem da sua utilização era sobretudo a dedicação e o trabalho intensivo sobre si, alavancando o seu potencial. Contudo, essa solução não seria viável devido à panóplia de casos existentes no desenvolvimento em escala, a sua performance poderia variar e os resultados para algumas empresas seriam um fracasso. Além disso, limitaríamos a utilização dos 3 modelos apenas para 1.

Em relação ao sumário do histórico dos clientes, operada pela função *summary-data-from-transaction-data()* do Lifetimes, oferece a capacidade de agrupar os dados em diversas granularidades. Seguindo o desempenho dos modelos, os erros tendiam a reduzir com o aumento da granularidade (agrupamento semanal ou mensal), como visível na Figura 21. Todavia, com esse nível de detalhe reduzido, as previsões seriam menos tangíveis. Observemos a Tabela 26 que detalha a agregação do histórico de um utilizador em diferentes granularidades, enquanto a Tabela 27 especifica o número real de compras (9), o parâmetro de previsão e o valor previsto, facilmente detetamos que uma dificuldade na interpretação dos resultados para semanas e meses, isto porque o que modelo prevê são a quantidade de semanas (3 de 4) e a quantidade de meses que o cliente irá comprar (1 de 1). No final, optou-se pela diminuição de granularidade (agregação diária) devido à sua representação e desempenho igualmente eficiente.

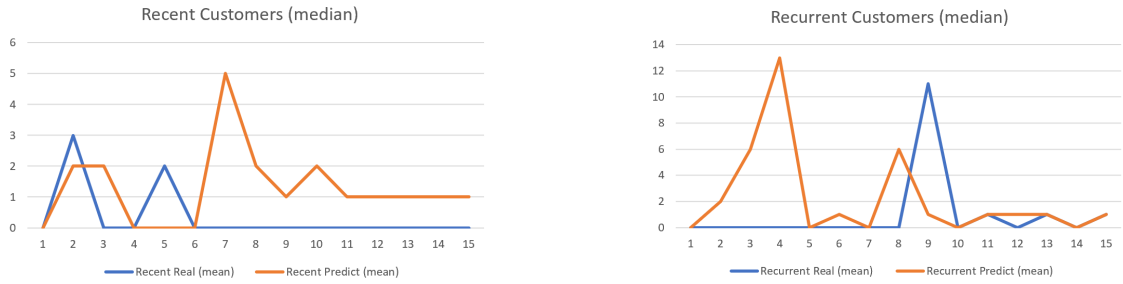
Tabela 26: Exemplo de agregação do histórico em diferentes granularidades

	R	**F**	**T**
Dias	171	704	705
Semanas	97	101	101
Meses	23	23	23

Tabela 27: Exemplo da previsão de compras para o próximo mês em diferentes granularidades

Nº Real: 9 compras	**Parâmetro de previsão**	**Valor previsto**
Dias	30	7
Semanas	4	3
Meses	1	1

De maneira a estudar o surgimento de novos clientes e a estabilidade dos recorrentes, foi tentada a seguinte experiência. O objetivo consistiu em criar 2 clientes fictícios, o cliente recente ($T < 30$) e o cliente recorrente ($T >= 90$ e $f >= 3$), em que ambos consistiam na agregação média mensal de diversos clientes. No final, estes eram concatenados aos clientes reais de forma a produzir previsões como um cliente comum. Veja-se nas Figuras 47 a divergência entre as previsões e os resultados reais, desta feita esta solução tornou-se prescindível.



(a) Clientes recentes

(b) Clientes recorrentes

Figura 47: Análise gráfica dos resultados do modelo

No que concerne à manipulação de hiperparâmetros, foi explorada a ferramenta Optuna. Da mesma forma que o Hyperopt, esta ferramenta tem a capacidade de otimizar os hiperparâmetros do modelo, com a finalidade de obter o erro menor. Com base no trabalho realizado pelo Jakub Czakon [17], o Optuna é superior em diversos aspetos, nomeadamente na visualização gráfica do *output* e na interpretação dos dados, porém, durante a implementação foi deduzido que o Hyperopt cumpre os requisitos e oferece uma performance superior e um *output* mais intuitivo.

A respeito da segmentação dos clientes, apresentaram-se soluções para a segmentação sobre RFM e T. Inicialmente, recorreu-se à segmentação com uso de técnicas de *clustering*, nomeadamente o *K-means* e a *Silhouette*. O *K-means* é um algoritmo não supervisionado com o intuito de agrupar os clientes através de padrões identificados nos dados, enquanto a *Silhouette* é um coeficiente que deteta a semelhança entre aquele ponto para outros no seu próprio *cluster* em comparação com pontos de outros *clusters*, estes 2 elementos integrados fornecem uma maior confiança nas decisões finais [54]. Apesar da sua eficácia, concluí-se que a sua integração levaria a alguma desordem na apresentação dos dados segmentados, caso um empresário tivesse 2 empresas e a segmentação diferisse entre elas poderia causar diferentes *outputs*, além disso, a sua aplicabilidade em T faria mais sentido agregada a pelo menos outra *feature*. No final, alguns filtros aplicados ao número de clientes, para evitar erros, limitariam a segmentação em algumas empresas.

Em conclusão, cada modelo, com a exceção do Pareto/NBD, disponibiliza os parâmetros, r, $alpha$, a e b, estimados pelo *Maximum Likelihood Estimation (MLE)*, que fornecem intervalos de confiança indicando alguma confiabilidade, veja-se a Figura 28 como exemplo. Com isso, o objetivo seria associar às previsões algum intervalo de confiança, contudo, após diferentes tentativas é percetível que esse *output* não traduz apenas imprevisibilidade de resultados e torna-se complexo relacioná-lo com as previsões.

Tabela 28: *Output* intervalos de confiança

	lower 95% bound	**upper 95% bound**
r	0,484091	0,855049
alpha	6,821513	16,695876
a	-0,005765	0,013972
b	-0,062195	0,468660

Em suma, a lista seguinte caracteriza algumas evidências e dificuldades, consequentes do caso precedente:

- Clientes com pouca frequência aumentam o desvio padrão;
- Clientes *churn* aumentam o desvio padrão - Isto não significa que as previsões são fracas;
- Empresas pequenas, com 1 cliente regular, tendem a aumentar o desvio padrão.

5.3 Conclusões

Esta secção detalha o processo de trabalho processado para a validação dos modelos e as principais diferentes abordagens tentadas. Ainda são discriminados os diferentes *datasets* manipulados e as evidências finais.

6 Conclusões

O presente capítulo 6 finaliza o documento, refletindo as conclusões do trabalho. A secção 6.1 descreve as principais dificuldades e explicações durante o desenvolvimento. Finalmente a secção 6.2 pormenoriza alguns tópicos relevantes com o propósito de potenciar o desenvolvimento futuro da solução e finda o documento.

6.1 Dificuldades

Diversas dificuldades foram surgindo ao longo do desenvolvimento. Inicialmente, a primeira dificuldade assentou na análise e interpretação dos modelos BTYD. Devido à panóplia de recursos e documentação existente a seleção inicial foi custosa, contudo sustentado nas avaliações efetuadas ao longo do relatório, concluí-se que os modelos são bastantes fiáveis.

Outra adversidade identificada, foi resultante da falta de dados de produção. Este problema está associado ao facto da PRIMAVERA disponibilizar um *On-premises software*, consequentemente, a princípio afetou o desenvolvimento e a análise dos modelos. Para colmatar isso, como citado em 5.1, recorreu-se ao uso de *datasets* online e sintéticos.

Relativamente a *datasets* pequenos, a capacidade do modelo convergir despoleta problemas associados à inconsistência dos dados, isto tornou-se evidente na realização do *tuning* com o Hyperopt. Para isso, juntamente com os filtros listados em 4.3, o parâmetro *TOL* foi manipulado diversas vezes com a finalidade de dissipar os erros na criação dos modelos e superar a descrita adversidade.

Por fim, a maior dificuldade passou pela interpretação e validação dos modelos, em que o objetivo seria associar uma taxa de imprevisibilidade às previsões. Contudo, após diversas pesquisas e conversas com profissionais da área, concluí-se que a melhor forma de avaliar os modelos, passa por calcular o erro através do MSE sob o *dataset* fracionado e que por se tratarem de modelos probabilísticos já se encontram associados a alguma incerteza, proporcionando uma ideia de incerteza associada às previsões. Apesar disso, ainda há uma ligeira incerteza de resultados para clientes novos, para esses, o objetivo passa por enfrentar o problema do Cone da Incerteza, isto é, à medida que o tempo decorre, o histórico aumenta, assim o erro e a incerteza tendem a afunilar e a desvanecer gradualmente. Inicialmente, caso as previsões

sejam totalmente opostas à realidade, as previsões são ocultadas até os modelos produzirem resultados aceitáveis.

6.2 Trabalho Futuro

Futuramente, como qualquer projeto, melhorias e funcionalidades adicionais poderão ser implementadas na solução.

Com o crescente aumento do histórico, uma das *features* que poderá ser relevante seria a extrapolação de clientes similares. De outra forma, traduzir-se-ia na partilha de um modelo por diversos clientes, de diferentes empresas, com características comuns (semelhante ao *cohort* aplicado em estatística), isto conduziria a uma redução da produção de modelos.

No seguimento da proposta anterior, outra melhoria consistia em tirar proveito do *MLflow Registry*. Assim sendo, o objetivo seria partilhar o *endpoint* da API do melhor modelo com as empresas com clientes similares.

Com base na dificuldade em 6.1, relacionado com os *datasets* pequenos. Este, faz um paralelismo ao problema típico dos Sistemas de Recomendação, designado por *Cold Start*. Posteriormente, a seguinte listagem apresenta algumas formas de mitigação que poderão ser aplicadas:

- Questionar aos clientes as tendências de compras e valores futuros;
- Analisar o histórico inicial das empresas e perquirir similaridades futuras;
- Aplicar o modelo mais consistente (MSE menor).

Em conclusão, todos os objetivos delineados foram alcançados com sucesso. Este projeto foi testado e avaliado por vários membros da comunidade da PRIMAVERA e será incorporado no sistema no próximo semestre.

Bibliografia

[1] "A modified Pareto/NBD approach for predicting customer lifetime value". Em: *Expert Systems with Applications* 36.2, Part 1 (2009), pp. 2062–2071. issn: 0957-4174. doi: https://doi.org/10.1016/j.eswa.2007.12.049. url: https://www.sciencedirect.com/science/article/pii/S0957417407006689 (ver p. 23).

[2] C. C. Aggarwal. "An introduction to outlier analysis". Em: *Outlier analysis*. Springer, 2017, pp. 1–34 (ver p. 9).

[3] E. Alecrim. *O que é o ERP e porque precisa dele?* https://www.infowester.com/erp.php. [Online; acedido 13-Nov-2021]. 2010 (ver pp. 2, 3).

[4] M. António. *Customer Lifetime Value Estimation via Probabilistic Modeling*. https://towardsdatascience.com/customer-lifetime-value-estimation-via-probabilistic-modeling-d5111cb52dd. [Online; acedido 24-Mar-2022]. 2022 (ver p. 20).

[5] J. Bergstra et al. "Hyperopt: a python library for model selection and hyperparameter optimization". Em: *Computational Science & Discovery* 8.1 (2015), p. 014008 (ver p. xviii).

[6] Y. D. Bergstra J. e C. D. D. *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures*. http://hyperopt.github.io/hyperopt/. [Online; acedido 04-Dez-2021]. 2013 (ver p. xviii).

[7] P. BSS. *Brochura Institucional*. https://pt.primaverabss.com/fotos/editor2/brochurainstitucional_2017_web2.pdf. [Online; acedido 12-Nov-2021]. 2017 (ver p. 2).

[8] P. BSS. *Relatório Anual Consolidado*. https://pt.primaverabss.com/fotos/editor2/Logos/rcp_2019.pdf. [Online; acedido 12-Nov-2021]. 2019 (ver p. 2).

[9] P. BSS. *Sobre a PRIMAVERA BSS*. https://pt.primaverabss.com/pt/primavera/. [Online; acedido 09-Nov-2021]. 2021 (ver p. 2).

[10] J. Bult e T. Wansbeek. "Optimal Selection for Direct Mail". Em: *Marketing Science* 14 (nov. de 1995), pp. 378–394. doi: 10.1287/mksc.14.4.378 (ver p. 18).

[11] S. Chen. "Estimating Customer Lifetime Value Using Machine Learning Techniques". Em: *Data Mining* (2018). doi: http://dx.doi.org/10.5772/intechopen.76990 (ver pp. 15, 16, 18, 19, 23).

[12] H.-M. Chuang e C.-C. Shen. "A study on the applications of data mining techniques to enhance customer lifetime value — based on the department store industry". Em: *2008 International Conference on Machine Learning and Cybernetics*. Vol. 1. 2008, pp. 168–173. doi: 10.1109/ICMLC.2008.4620398 (ver p. 15).

[13] G. Cloud. *MLOps: pipelines de entrega contínua e automação no aprendizado de máquina*. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning. [Online; acedido 03-Nov-2021]. 2020 (ver p. 23).

[14] P. C. Consul e G. C. Jain. "A generalization of the Poisson distribution". Em: *Technometrics* 15.4 (1973), pp. 791–799 (ver p. xviii).

[15] J. Couto. *How Machine Learning is reshaping Price Optimization*. https://tryolabs.com/blog/price-optimization-machine-learning. [Online; acedido 02-Nov-2021]. 2020 (ver p. 14).

[16] "Customers as assets". Em: *Journal of Interactive Marketing* 17.1 (2003), pp. 9–24. issn: 1094-9968. doi: https://doi.org/10.1002/dir.10045 (ver p. 16).

[17] J. Czakon. *Optuna vs Hyperopt: Which Hyperparameter Optimization Library Should You Choose?* https://neptune.ai/blog/optuna-vs-hyperopt. [Online; acedido 20-Abr-2022]. 2021 (ver p. 70).

[18] J. Czakon. *The Best MLOps Tools and How to Evaluate Them*. https://neptune.ai/blog/best-mlops-tools. [Online; acedido 06-Nov-2021]. 2021 (ver p. 29).

[19] S. Das e U. M. Cakmak. *Hands-On Automated Machine Learning: A Beginner's Guide to Building Automated Machine Learning Systems Using AutoML and Python*. Packt Publishing, 2018. isbn: 9781788629898 (ver pp. 6, 8–10).

[20] Databricks. *Databricks*. https://databricks.com/. [Online; acedido 18-Nov-2021]. 2021 (ver p. xvii).

[21] Databricks. *Databricks Solution Accelerators*. https://databricks.com/solutions/accelerators. [Online; acedido 02-Nov-2021]. 2021 (ver p. 13).

[22] Databricks. *Delta Lake*. https://databricks.com/product/delta-lake-on-databricks. [Online; acedido 13-Abr-2022]. 2022 (ver p. xvii).

[23] Databricks. *Managed MLflow*. https://databricks.com/product/managed-mlflow. [Online; acedido 31-Mar-2022]. 2022 (ver p. 28).

[24] Databricks. *Work with feature tables*. https://docs.databricks.com/applications/machine-learning/feature-store/feature-tables.html. [Online; acedido 01-Abr-2022]. 2022 (ver p. xviii).

[25] Datarobot. *What is a Feature Variable in Machine Learning?* https://www.datarobot.com/wiki/feature/. [Online; acedido 02-Dez-2021]. 2021 (ver p. 9).

[26] C. Davidson-Pilon. *Measuring users is hard. Lifetimes makes it easy.* https://lifetimes.readthedocs.io/en/latest/index.html. [Online; acedido 04-Dez-2021]. 2015 (ver p. xviii).

[27] C. Davidson-Pilon. *Measuring users is hard. Lifetimes makes it easy.* https://lifetimes.readthedocs.io/_/downloads/en/latest/pdf/. [Online; acedido 04-Dez-2021]. 2015 (ver p. 33).

[28] A. Dong. "ERP and Artificial Intelligence based Smart Financial Information System Data Analysis Framework". Em: *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. 2021, pp. 845–848. doi: 10.1109/ICICT50816.2021.9358659 (ver pp. 11, 13).

[29] "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study". Em: *Procedia Computer Science* 3 (2011). World Conference on Information Technology, pp. 57–63. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2010.12.011. url: https://www.sciencedirect.com/science/article/pii/S1877050910003868 (ver pp. xvii, 15, 18).

[30] P. Fader, B. Hardie e K. Lee. ""Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model". Em: *Marketing Science* 24 (mai. de 2005), pp. 275–284. doi: 10.1287/mksc.1040.0098 (ver pp. 19–21, 31).

[31] P. Fader, B. Hardie e K. Lee. "Customer-Base Analysis in a Discrete-Time Noncontractual Setting". Em: *Marketing Science* (ago. de 2010), pp. 1086–1108. doi: https://doi.org/10.1287/mksc.1100.0580 (ver p. 21).

[32] P. Fader, B. Hardie e K. Lee. "More than meets the eye". Em: (2006). url: https://faculty.wharton.upenn.edu/wp-content/uploads/2013/08/fader_et_al_mr_06.pdf (ver p. 23).

[33] P. Fader, B. Hardie e K. Lee. "RFM and CLV: Using iso-value curves for customer base analysis". Em: *Journal of Marketing Research American Marketing Association ISSN* XLII (dez. de 2005), pp. 415–430. doi: 10.1509/jmkr.2005.42.4.415 (ver pp. 19, 22, 43).

[34] B. G. e S. Hardie. "The Gamma-Gamma Model of Monetary Value". Em: fev. de 2013. url: http://www.brucehardie.com/notes/025/gamma_gamma.pdf (ver p. 20).

[35] M. Gadallah e H. Elmaraghy. "A Concurrent Engineering Approach To Robust Product Design". Em: *Concurrent Engineering* 1.4 (1993), pp. 237–251. doi: 10.1177/1063293X9300100407 (ver p. 11).

[36] Gartner. *Data Lake*. https://www.gartner.com/en/information-technology/glossary/data-lake. [Online; acedido 13-Abr-2022]. 2022 (ver p. xvii).

[37] Y. Gavrilova. *The Best Open-Source MLOps Tools You Should Know*. https://neptune.ai/blog/best-open-source-mlops-tools. [Online; acedido 06-Nov-2021]. 2021 (ver p. 29).

[38] R. Godinho. *Customer Lifetime Value*. https://goodi.pt/customer-lifetime-value/. [Online; acedido 13-Abr-2022]. 2022 (ver p. 15).

[39] S. Goundar. *Enterprise Systems and Technological Convergence: Research and Practice*. Information Age Pub Inc, 2021, pp. 85–96. isbn: 1648023428 (ver p. 11).

[40] A. Goyal. "Machine Learning Operations". Em: *International Journal of Information Technology Insights & Transformations*. Vol. 4, Issue 2. Scholar, Department of Computer Science, MDS University, Ajmer, India, 2020, p. 15. isbn: 2581-5172 (ver pp. xvii, 23).

[41] B. Groza et al. "Neural Network Based Framework for Optimization of Enterprise Resource Planning". Em: *2006 Canadian Conference on Electrical and Computer Engineering*. 2006, pp. 1889–1892. doi: 10.1109/CCECE.2006.277422 (ver pp. 12, 13).

[42] S. Gupta et al. "Modeling Customer Lifetime Value". Em: *Journal of Service Research - J SERV RES* 9 (nov. de 2006), pp. 139–155. doi: 10.1177/1094670506293810 (ver p. 16).

[43] A. Haponik. *What are Product Recommendation Engines? And the various versions of them?* https://towardsdatascience.com/what-are-product-recommendation-engines-and-the-various-versions-of-them-9dcab4ee26d5. [Online; acedido 02-Nov-2021]. 2017 (ver p. 14).

[44] A. Haponik. *What is Text Mining, Text Analytics and Natural Language Processing?* https://addepto.com/up-selling-cross-selling-5-reasons-use-machine-learning/. [Online; acedido 02-Nov-2021]. 2019 (ver p. 14).

[45] M. Hartmann. *Customer lifetime value (CLV) prediction*. https://docs.microsoft.com/en-us/dynamics365/customer-insights/predict-customer-lifetime-value. [Online; acedido 04-Dez-2021]. 2022 (ver p. 15).

[46] P. Jašek et al. "Comparative analysis of selected probabilistic customer lifetime value models in online shopping". Em: *Journal of Business Economics and Management* 20 (abr. de 2019), pp. 398–423. doi: 10.3846/jbem.2019.9597 (ver p. 22).

[47] P. Jenkner. *The Best MLflow Alternatives (2021 Update)*. https://neptune.ai/blog/the-best-mlflow-alternatives. [Online; acedido 06-Nov-2021]. 2021 (ver p. 29).

[48] P. S. Julian Soh. *Data Science Solutions on Azure: Tools and Techniques Using Databricks and MLOps*. Apress, 2020. isbn: 9781484264058 (ver p. 8).

[49] R. Kahan. "Using database marketing techniques to enhance your one‐to‐one marketing initiatives". Em: *Journal of Consumer Marketing* 15.5 (1998), pp. 491–493. doi: https://doi.org/10.1108/07363769810235965 (ver p. 15).

[50] H. Kang. "The prevention and handling of the missing data". Em: *Korean journal of anesthesiology* 64 (mai. de 2013), pp. 402–6. doi: 10.4097/kjae.2013.64.5.402 (ver p. 9).

[51] E. Korkmaz, R. Kuik e D. Fok. "Counting Your Customers, When will they buy next? An empirical validation of probabilistic customer base analysis models based on purchase timing". Em: (2013). url: https://core.ac.uk/download/pdf/18510752.pdf (ver pp. 15, 20, 21).

[52] Linguamatics. *What is Text Mining, Text Analytics and Natural Language Processing?* https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing. [Online; acedido 02-Nov-2021]. 2021 (ver p. 14).

[53] D.-R. Liu e Y.-Y. Shih. "Integrating AHP and data mining for product recommendation based on customer lifetime value". Em: *Information & Management* 42.3 (2005), pp. 387–400. issn: 0378-7206. doi: https://doi.org/10.1016/j.im.2004.01.008. url: https://www.sciencedirect.com/science/article/pii/S0378720604000394 (ver p. 15).

[54] R. Lletí et al. "Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes". Em: *Analytica Chimica Acta* 515.1 (2004). Papers presented at the 5th COLLOQUIUM CHEMIOMETRICUM MEDITERRANEUM, pp. 87–100. issn: 0003-2670. doi: https://doi.org/10.1016/j.aca.2003.12.020. url: https://www.sciencedirect.com/science/article/pii/S0003267003016246 (ver p. 70).

[55] I. Martinez. *Apply AI and ML to Your Predictive Next Best Actions and Meet Your Customers Where They Are*. https://www.credera.com/insights/apply-ai-ml-to-your-predictive-next-best-actions-and-meet-your-customers-where-they-are. [Online; acedido 02-Nov-2021]. 2020 (ver p. 14).

[56] D. McCarthy e E. Wadsworth. *Buy 'Til You Die*. 2014 (ver pp. 19, 20).

[57] Microsoft. *O que é o ERP e porque precisa dele?* https://dynamics.microsoft.com/pt-pt/erp/what-is-erp/. [Online; acedido 12-Nov-2021]. 2021 (ver p. 2).

[58] L. MLflow Project a Series of LF Projects. *MLflow Components*. https://www.mlflow.org/docs/latest/concepts.html. [Online; acedido 04-Nov-2021]. 2021 (ver p. 28).

[59] L. MLflow Project a Series of LF Projects. *MLflow Documentation*. https://www.mlflow.org/docs/latest/index.html. [Online; acedido 04-Nov-2021]. 2021 (ver p. 28).

[60] NICE. *What is Cognitive Robotic Process Automation?* https://www.nice.com/rpa/rpa-guide/what-is-cognitive-rpa/. [Online; acedido 02-Nov-2021]. 2021 (ver p. 13).

[61] N. J. Nilsson. *Principles of artificial intelligence*. Elsevier Inc, Morgan Kaufmann, 1982. isbn: 0934613109 (ver p. 5).

[62] Ş. Ozan. "A Case Study on Customer Segmentation by using Machine Learning Methods". Em: *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. 2018, pp. 1–6. doi: 10.1109/IDAP.2018.8620892 (ver p. 14).

[63] A. A. Patel. *Hands-On Unsupervised Learning Using Python*. O'Reilly Media, Inc., 2019. isbn: 9781492035640 (ver p. 8).

[64] G. C. Platform. *Predicting Customer Lifetime Value with AI Platform: introduction*. https://cloud.google.com/architecture/clv-prediction-with-offline-training-intro. [Online; acedido 13-Dez-2021]. 2021 (ver pp. 14–16, 19).

[65] E. Raj. *Engineering MLOps: Rapidly build, test, and manage production-ready machine learning life cycles at scale*. Packt Publishing, 2021, p. 29. isbn: 9781800562882 (ver pp. 7, 24).

[66] I. RapidMiner. *Industries and Use Cases*. https://rapidminer.com/solutions/. [Online; acedido 02-Nov-2021]. 2021 (ver p. 13).

[67] M. A. Rashid. "The Evolution of ERP Systems: A Historical Perspective". Em: (2002), p. 4 (ver pp. 3, 12).

[68] W. J. Reinartz e V. Kumar. "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing". Em: *Journal of Marketing* 64.4 (2000), pp. 17–35. doi: 10.1509/jmkg.64.4.17.18077. eprint: https://doi.org/10.1509/jmkg.64.4.17.18077. url: https://doi.org/10.1509/jmkg.64.4.17.18077 (ver p. 16).

[69] G. G. Robert C. Blattberg e J. S. Thomas. *Customer equity*. Boston : Harvard Business School Press, 2001. isbn: 0875847641 (ver p. 21).

[70] L. Rocha. *Anomalias detetadas por Inteligência Artificial e comunicadas através de padrões UI inteligentes*. Thesis. In Review. IPCA, jul. de 2022 (ver p. 51).

[71] S. Roger. *What are the branches of Artificial Intelligence?* https://www.h2kinfosys.com/blog/what-are-the-branches-of-artificial-intelligence/. [Online; acedido 21-Dez-2021]. 2021 (ver p. 6).

[72] O. K. e Roman Chuprina. *How to Use AI and Machine Learning in Fraud Detection*. https://spd.group/machine-learning/fraud-detection-with-machine-learning/. [Online; acedido 02-Nov-2021]. 2021 (ver p. 14).

[73] H. S. *A Definitive Guide for predicting Customer Lifetime Value (CLV)*. https://www.analyticsvidhya.com/blog/2020/10/a-definitive-guide-for-predicting-customer-lifetime-value-clv/. [Online; acedido 04-Mar-2022]. 2020 (ver p. 17).

[74] R. W. Stone e D. J. Good. “The assimilation of computer-aided marketing activities”. Em: *Inf. Manag.* 38 (2001), pp. 437–447 (ver p. 15).

[75] G. Symeonidis et al. “MLOps - Definitions, Tools and Challenges”. Em: *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. 2022, pp. 0453–0460. doi: 10.1109/CCWC54503.2022.9720902 (ver p. 29).

[76] T. Taulli. *Artificial Intelligence Basics: A Non-Technical Introduction*. Apress, 2019. isbn: 978-1-4842-5027-3 (ver pp. 5, 6).

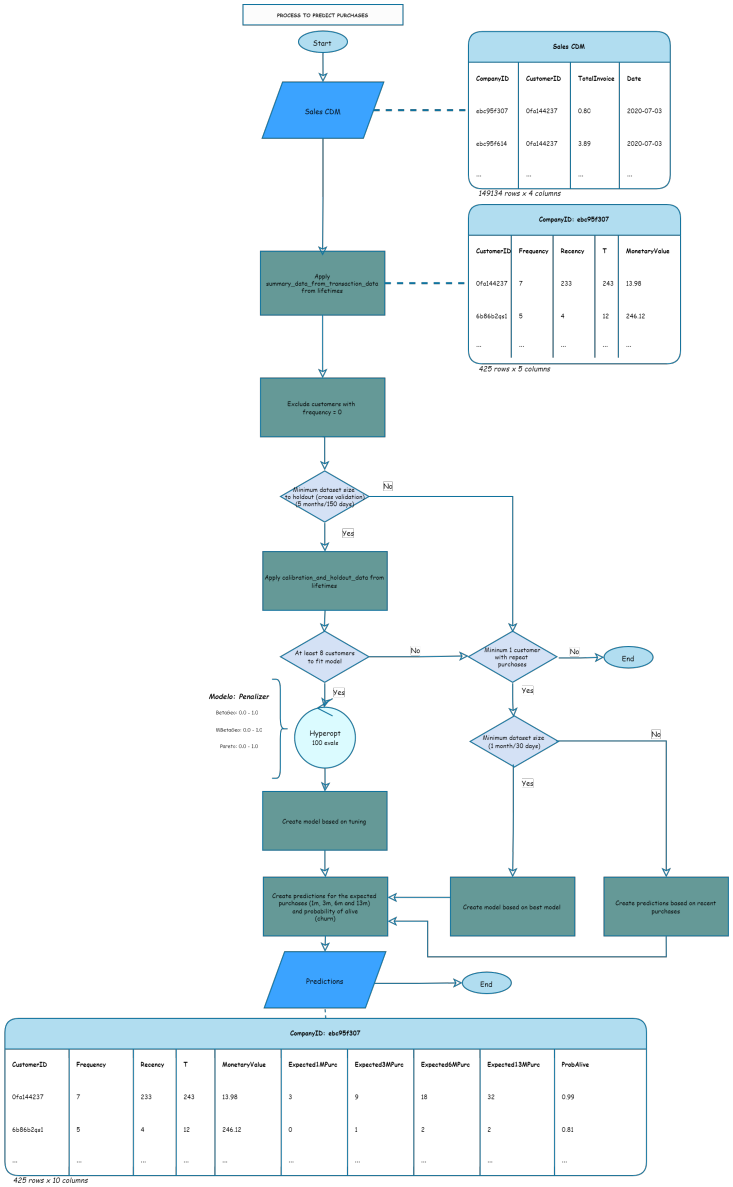
[77] M. Treveil. *Introducing MLOps: How to Scale Machine Learning in the Enterprise*. O’Reilly Media, Incorporated, 2020. isbn: 1492083291 (ver pp. 23–27).

[78] C.-F. Tsai et al. “A comparative study of hybrid machine learning techniques for customer lifetime value prediction”. Em: *Kybernetes: The International Journal of Systems & Cybernetics* 42 (mar. de 2013). doi: 10.1108/03684921311323626 (ver p. 15).

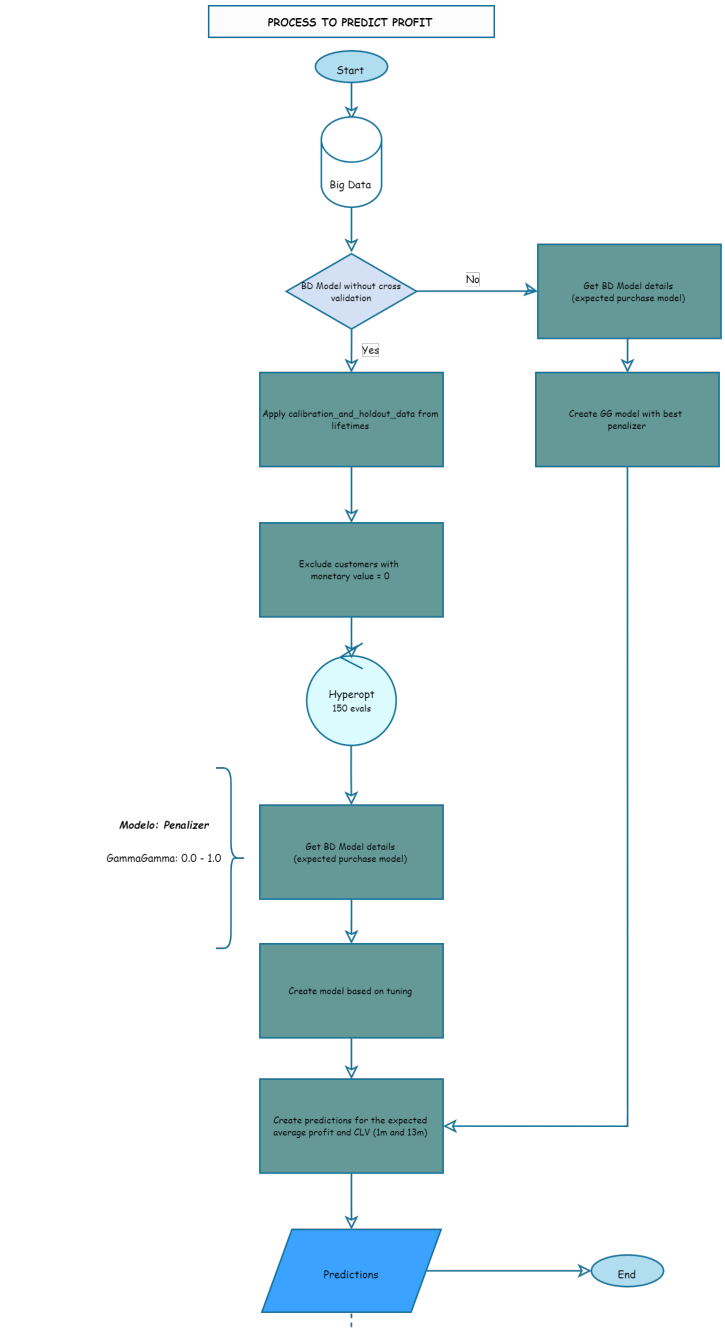
[79] Wikipédia. *Add-on*. https://pt.wikipedia.org/wiki/Add-on. [Online; acedido 16-Nov-2021]. 2018 (ver p. xvii).

A Fluxogramas

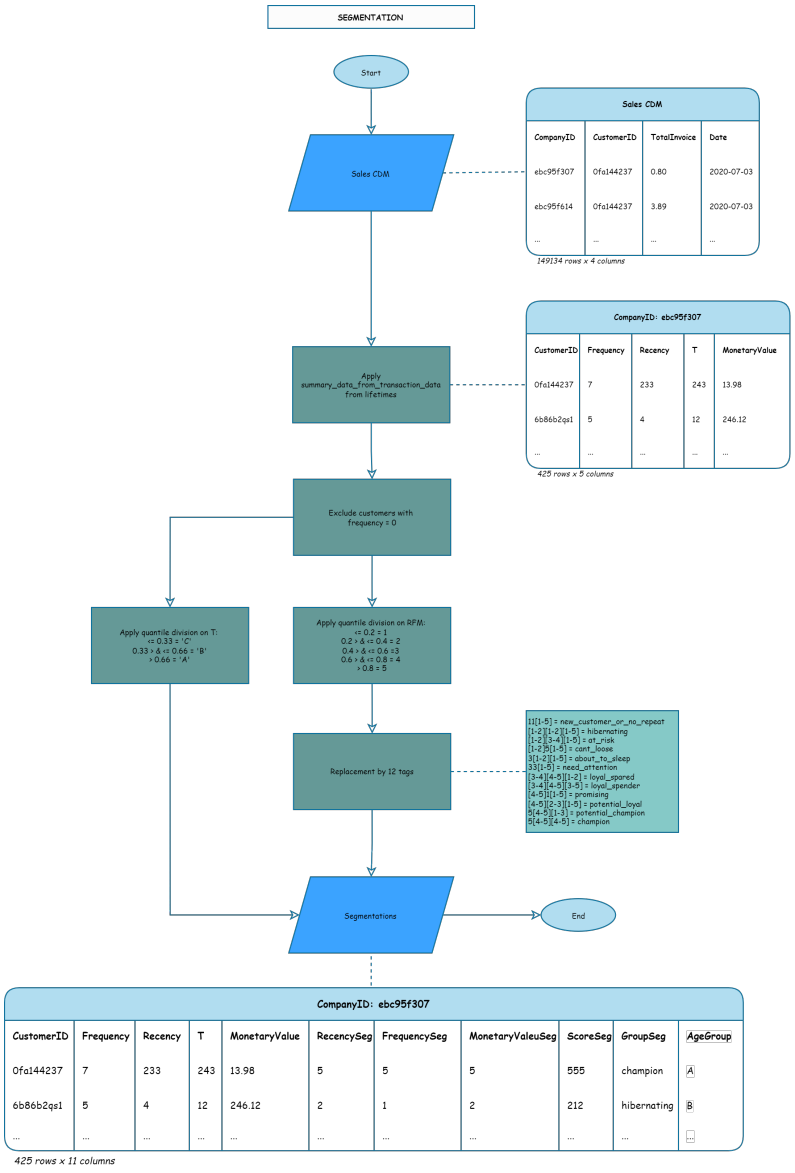
A.1 Fluxograma - Previsão de Compras



A.2 Fluxograma - Previsão de Faturação



A.3 Fluxograma - Segmentação



A.4 Evidências

A lista seguinte descrimina algumas evidências, consequentes do processo de validação dos modelos em 5.1:

- Os modelos realizam previsões mais assertivas com clientes recentes, clientes pouco frequentes e clientes regulares;
- Os modelos realizam previsões mais assertivas para espaços temporais mais curtos, portanto os modelos produzem melhores resultados para 1 mês do que para 3 meses;
- A taxa de atividade (*churn*) observa-se mais evidenciada em clientes com a frequência em queda;
- Os clientes com um histórico grande (superior a 1 ano), produz ótimas resultados (sem nunca exceder exageradamente o T);
- Os clientes com um histórico grande, o modelo é ótimo e realiza previsões não excessivas, ou seja, subestima o número real de compras;
- Os clientes com um histórico pequeno, o modelo poderá efetuar previsões excessivas;
- As previsões superiores a 1 mês são pouco assertivas para clientes com um histórico pequeno (inferior a 3 mês);
- O arredondamento das previsões deverá ser para baixo porque tipicamente os modelos subestimam os resultados reais;
- Com a manipulação de clientes recentes existe alguma dificuldade em prever compras futuras, devido à imprevisibilidade;
- O modelo, para clientes regulares, tem a aptidão de interpretar que se caso o tempo de previsão dobre ou triplique então provavelmente número de compras dobrará ou triplicará;
- Todos os clientes devem conter a *recency* inferior a T;
- Todos os clientes com a *frequency* igual a 0 então a sua *recency* é igualmente 0;
- Os modelos têm a perícia de variar mediante o histórico individual e da população;
  - Caso um cliente com $R = 181$, $F = 1$, $M = 10$ e $T = 730$ significa que presumivelmente não comprará mais, então a sua taxa de *churn* é igual a $0,000821$. Porém, agregando a este cliente outro com $R = 730$, $F = 2$, $M = 10$ e $T = 730$ a taxa de *churn* aumenta para $0,266558$, isto resulta do paralelismo estabelecido entre os clientes;

  - Caso um cliente com $R = 2$, $F = 2$, $M = 2$ e $T = 2$ significa que presumivelmente irá comprar diariamente, então a previsão para o próximo mês será 30 compras. Contudo, agregando um cliente com $R = 2$, $F = 2$, $M = 2$ e $T = 5$ a previsão para o próximo mês descerá para 11 compras.

- Um modelo sem realizar *tuning* e com dados recentes, como 1 mês de compras para prever o próximo mês, prevê exatamente o que foi comprado no mês passado;