NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

# Modelação Preditiva para a Distribuição de objetivos Comerciais no Setor Bancário

Rita Maria de Almeida Franco

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# Modelação Preditiva para a Distribuição de Objetivos Comerciais no Setor Bancário

by

Rita Maria de Almeida Franco

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

**Advisor:** Flávio Luís Portas Pinheiro

December 2020

# AGRADECIMENTOS

Em primeiro lugar agradecer aos meus pais, que sempre trabalharam horas e horas extra para nos poder proporcionar a melhor educação que podíamos pedir, sem eles nunca teria feito este mestrado. Depois agradecer ao João por todos os desabafos que ouviu, por toda a motivação que me deu e por nunca deixar de acreditar em mim.

Expressar também o meu agradecimento ao Professor Flávio Pinheiro pelo auxílio e preocupação demonstrados ao longo do estágio, e pelo contributo que deu na realização deste relatório.

Queria agradecer ainda aos meus amigos, Rodrigo Umbelino e Vitor Manita, por serem os meus parceiros no primeiro ano deste Mestrado. Muitas horas, muitos fins de semanas, muitas linhas de código, muitos projetos realizados em conjunto que contribuíram para o nosso crescimento enquanto *Data Scientists*.

Ao Millennium bcp agradeço a confiança em mim depositada e o investimento efetuado em formação ao longo do ano. Um agradecimento especial ao Hugo Resende por ser um Diretor presente, preocupado e exigente e ao Sérgio Moura, por sempre acreditar em mim e no meu trabalho. Por último, agradecer a toda a equipa que me acolheu e que auxiliou ao longo do projeto transmitindo o seu conhecimento sobre o negócio.

# SUMÁRIO

Ao longo dos anos assistiu-se a uma transformação no setor bancário conduzindo a um alargamento do leque de produtos e serviços oferecidos aos clientes. Surgindo a necessidade de informar a Rede Comercial de quais os produtos prioritários em cada período de tempo e de quais os seus objetivos individuais para cada produto. No Millennium bcp estas necessidades são respondidas através da atribuição de objetivos específicos e concretos de cada produto para cada unidade de negócio. A dinâmica de objetivos funciona então como forma de motivação individual do trabalho, como uniformização das prioridades na Rede e como base para a distribuição de incentivos monetários aos colaboradores. Para que os objetivos atribuídos tragam um acréscimo de motivação e produtividade é crucial a sua correta distribuição pelas unidades de negócio tendo em conta diversos fatores como o potencial das suas carteiras. No presente relatório é proposto um modelo de distribuição de objetivos de Crédito de Habitação pela Rede Comercial do BCP. Os resultados foram obtidos através do treino de diferentes algoritmos, nomeadamente *Gradient Boost*, Árvore de Decisão e Regressão Linear após a realização da recolha e processamento dos dados. Comparativamente à abordagem utilizada anteriormente, os resultados obtidos neste projeto demonstraram uma distribuição mais precisa dos objetivos com valores de erro menores.

# PALAVRAS-CHAVE

Modelo Preditivo; Data Mining; Aprendizagem Automática Supervisionada; Objetivos; Crédito Habitação; Banca

# ABSTRACT

Throughout the past years a transformation in the banking industry has been taking place leading to an increase in the range of products offered to clients. Thus, arose the necessity of informing the Retail Network which products are crucial and how much they should sell of each product was created. At Millennium bcp this necessity was answered through assigning specific and concrete goals of each product to each business unit. The goals dynamic works as personal motivation, as a uniformization of priorities throughout the Retail Network and as a base for distributing monetary incentives to employees. So that goals assigned create the increment in motivation and productivity wanted, it is crucial to ensure its correct distribution by each business unit considering their potential and market. The present report proposes a goals distribution model for Housing Credit through Millennium bpc's Retail Network. The results were obtained by training different algorithms such as, Gradient Boost, Decision Tree and Linear Regression after collecting and prepossessing data. In comparison with the previous methodology used, this model provided a more accurate goals distribution with lower error.



# KEYWORDS

Predictive Modelling; Data Mining; Supervised Machine Learning; Goals; Housing Credit; Bank

# INDICE

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIAÇÕES E ACRÓNIMOS

**AA** – Aprendizagem Automática

**BCP** – Banco Comercial Português

**CC** – Centro de Custo

**CRISP-DM –** *Cross-Industry Standard Process for Data Mining* – Processo Padrão Intersectorial para *Data Mining*

**CRM** – *Costumer Relationship Managment*

**DC** – Direção Comercial

**DQAR** – Direção da Qualidade e Apoio à Rede

**GPN** – Gestão Personalizada Negócios

**GPP** – Gestão Personalizada Prestige

**kNN** – *k-Nearest Neighbor* – k Vizinhos Mais Próximos

**MAE** – *Mean Absolte Error* – Erro Absoluto Médio

**MBCP** – Millennium bcp – Millennium – Millennium Banco Comercial Português

**MA** – Modelo Anterior

**ML** – *Machine Learning* – Aprendizagem Automática

**MM** – Mass Market

**MP** – Modelo Proposto

**MSE** – *Mean Squared Error* – Erro ao Quadrado Médio

**MxAE** – *Maximum Absolte Error* – Máximo Erro Absoluto

**NUT** – Nomenclatura de Unidade Territorial

**Obj** – Objetivo

**SEMMA –** Sample, Explore, Modify, Model, Assess – Amostrar, Explorar, Modificar, Modelar, Avaliar

**SML –** *Supervised Machine Learning* – Aprendizagem Automática Supervisionada

**SQL** – *Structured Query Language* – Linguagem de Consulta Estruturada

**SSE** – *Sum of Squared Errors* – Soma do Quadrado dos Erros

**SSMS** – *SQL Server Management Studio*

**UOIC** – Unidade de Objetivos e Informação Comercial

**VBA** – *Visual Basic for Applications*

# 1. INTRODUÇÃO

Ao longo dos últimos anos tem-se assistido a uma alteração dos produtos e serviços comercializados pelo setor bancário tradicional de forma a aproximar-se cada vez mais das necessidades dos clientes atuais. Estas alterações derivam, por um lado, do aparecimento de novos competidores mais ágeis e tecnológicos que oferecem serviços mais personalizados e modernos aos clientes. Por outro lado, derivam do incremento de exigência na qualidade de serviço e do aumento de conhecimento da concorrência por parte dos clientes. [1, 2] Esta transformação conduziu ao alargamento do leque de produtos e serviços oferecidos aos clientes, o que consequentemente incrementa a complexidade do negócio. Num setor com cada vez mais produtos é importante auxiliar os colaboradores a canalizarem os seus esforços para os produtos de maior valor acrescentado para o Millennium bcp e para os seus clientes. Este auxílio à Rede de Comerciais do MBCP é efetuado através da definição de objetivos de venda específicos e concretos para cada produto, tornando sempre claro quais as prioridades do Banco a cada momento.

O mecanismo de atribuição de objetivos comerciais na empresa é utilizado não só como forma de motivação do trabalho, mas ainda como forma de manter toda a rede de Retalho focada nos produtos e serviços mais importantes daquele período de tempo. O maior desafio da dinâmica montada em torno dos objetivos comerciais e sistema de incentivos do Retalho está na atribuição adequada dos montantes de objetivos de cada produto a cada uma das unidades de negócio. O projeto apresentado neste relatório teve como objetivo responder a este desafio, criando uma nova abordagem para a distribuição de objetivos comerciais de Crédito Habitação.

Renovar o modelo de distribuição de objetivos previamente utilizado tinha como principal intuito criar um modelo mais justo, apoiado em técnicas estatísticas robustas com um maior foque nos dados. Desta forma, o projeto consistiu na criação de um modelo preditivo para a distribuição de objetivos comerciais de Crédito Habitação. O projeto foi realizado desde a fase de definição do problema e recolha dos dados até à fase de modelação, avaliação e apresentação de resultados, englobando todas as fases de um projeto de *Machine Learning*. A adequação dos objetivos ao potencial de cada carteira de clientes e de cada região é fundamental para a motivação dos recursos humanos e para a performance do Banco sendo esta a principal motivação para a criação do projeto apresentado neste relatório.

Adicionalmente, pequenos projetos e tarefas realizados ao longo do estágio estão brevemente descritos neste relatório. Entre os quais destaca-se a criação de um *dashboard* dinâmico em Power BI para o acompanhamento da Ativação Digital dos Novos Clientes ao longo do tempo.

O relatório está organizado em secções iniciando-se pela apresentação do Millennium com especial foco na Rede de Retalho e pela contextualização do projeto dentro da organização na secção *2. O Retalho no Millennium bcp*. Seguidamente, na secção *3. Enquadramento Teórico*, tem-se uma contextualização teórica do projeto quer no âmbito das metodologias utilizadas e algoritmos testados, quer no âmbito das teorias de utilização de objetivos nas organizações. Na secção *4. Ferramentas* realiza-se uma apresentação dos programas e ferramentas utilizados ao longo do estágio. A secção *5. Modelo de Distribuição de Objetivos: Crédito Habitação* expõe o conteúdo do projeto e detalha cada uma das suas fases. Inicia-se esta secção com uma apresentação do estágio como um todo e breve explicação dos projetos realizados, seguindo-se para uma descrição detalhada da metodologia

utilizada no projeto. Por último, na secção *6. Conclusão* expõem-se as conclusões, limitações sentidas ao longo do projeto e trabalho futuro a desenvolver.

## 2. O RETALHO NO MILLENNIUM BCP

O Banco Comercial Português (BCP) é um grupo de origem portuguesa que desenvolve um conjunto de atividades financeiras e presta serviços bancários em diferentes geografias como Portugal, Polónia, Moçambique, entre outras. Desde 2019 todas as operações bancárias do Grupo desenvolvem a sua atividade sob a marca Millennium. No mercado português, o Grupo tem a presença de dois bancos: o Millennium bcp, a maior instituição bancária privada no país, e o AtivoBank um banco vocacionado para uma camada de clientes mais jovem e que privilegia uma comunicação com o seu banco assente no canal digital.

A atividade do Millennium bcp está assente numa forte segmentação de clientes particulares e empresas de acordo principalmente com património financeiro e volume de negócios. A segmentação, visível na Figura 1, divide os mais de 2 milhões de clientes por três redes distintas que operam de forma independente: Rede Retalho, Rede Empresas, e Private Banking.

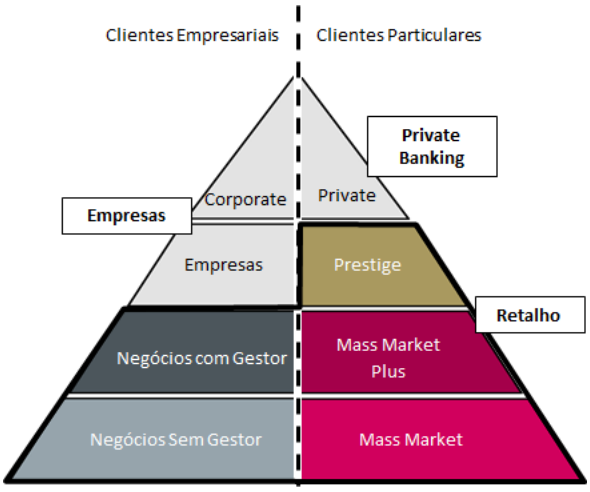



Figura 1. Segmentação dos Clientes no MBCP: Macro-segmentos

O Retalho é o conjunto da Rede de Retalho (RR) com as várias direções de apoio ao negócio, tais como a Direção de Marketing de Retalho, a Direção de Gestão de Segmentos, Direção de Crédito, a Direção de Qualidade e Apoio à Rede (DQAR), entre outras. Representa aproximadamente 90% dos clientes do Banco, o que se traduz em mais de 75% dos recursos e mais de 70% do produto bancário[1]. Os mais de 2 milhões de clientes da RR estão divididos em cinco macro segmentos, atribuídos através de critérios baseados em património financeiro, ordenado domiciliado e idade para clientes particulares e volume de negócios para clientes empresariais.

De forma a proporcionar a melhor experiência de cliente e extrair o maior valor acrescentado possível o Millennium bcp relaciona-se com estes cinco macro segmentos através de três plataformas distintas: Gestão Personalizada Prestige (GPP), Gestão Personalizada Negócios (GPN) e Mass Market (MM) conforme demonstra a Figura 2.

[1] Valores incluindo AtivoBank

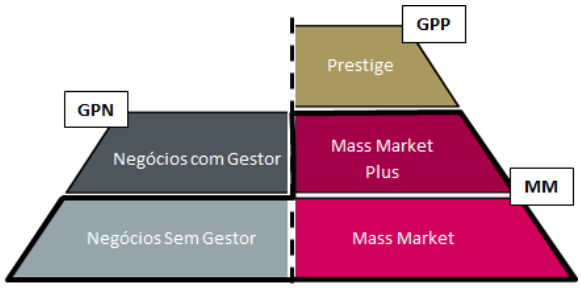



Figura 2. Distribuição dos macro-segmentos de clientes por tipologia de sucursais no Retalho

Nas sucursais de Gestão Personalizada a relação com o cliente é realizada através do seu Gestor de Conta. Existe uma comunicação dirigida e uma proposta de valor exclusiva e adaptável a cada subsegmento. Clientes Prestige (particulares com valores mais elevados de património financeiro) e Negócios Geridos (clientes empresa com volume de negócios considerável) têm uma relação mais personalizada derivada das necessidades de serviços bancários mais específicos. A diferença entre as duas plataformas prende-se assim com a tipologia de clientes e, consequentemente, com o tipo de produtos que comercializam. Para os gestores GPP existem produtos direcionados aos particulares como produtos de investimento, para gestores GPN existe um maior foco em produtos vocacionados aos negócios tais como crédito especializado e de soluções de gestão de tesouraria.

O MM caracteriza-se como uma plataforma de menor custo que serve os clientes com eficácia e conveniência. Na RR existem mais de 450 sucursais MM que operam através de diferentes modelos de negócio de forma a adequarem-se à realidade do local onde se encontram. Nesta plataforma a oferta de produtos e serviços é mais estandardizada abrangendo um maior número de clientes não deixando de existir uma subsegmentação e foco no conhecimento do cliente e na proximidade da relação. Tratando-se de um segmento com uma abordagem mais massificada existe um forte investimento nas plataformas digitais como potenciador da proximidade com o cliente.

Em suma, a Rede de Retalho é constituída por, aproximadamente, 450 sucursais MM, 70 sucursais GPN e 200 sucursais GPP que se dividem, de acordo com localização geográfica, em três Direções Coordenações: Norte, Centro, Sul e Ilhas, e subdivididas em sensivelmente 40 Direções Comerciais (DC). O modelo de segmentação dos clientes não é estático sendo que com alguma regularidade se promovem transferências de clientes entre segmentos. Adionalmente a própria Rede está em constante reavaliação, quer nos modelos de negócio utilizados nas sucursais quer na estrutura das Direções Comerciais.

### 2.1. Rede Comercial Focada em Objetivos

A equipa da Rede Comercial é composta por aproximadamente 3000 colaboradores que são extremamente motivados e focados não só em criar a melhor experiência de consumidor no mercado, como em proporcionar os melhores resultados para o MBCP. Durante os últimos seis anos o Retalho do Millennium orgulha-se de ter entregue resultados sempre acima dos objetivos anuais definidos.

Ao longo dos últimos anos o setor bancário tem sofrido alterações deixando de operar apenas como uma instituição que recebe depósitos e concede créditos para se tornar num ecossistema financeiro focado no cliente. Estas mudanças devem-se a variados fatores, por um lado, o aparecimento de novos concorrentes como as *fintech*. Estas novas empresas revelam-se mais pequenas e mais tecnológicas sendo, por isso, mais ágeis na adaptação às necessidades dos clientes. Por outro lado, os próprios

clientes tornaram-se cada vez mais informados e capazes de comparar serviços de diferentes instituições sem necessidade de sair de casa. [1, 2]

Ao aproximar-se cada vez mais das necessidades dos clientes o setor bancário criou um conjunto alargado de serviços e produtos que disponibiliza ao mercado, tornando mais complexa a atividade de negócio. Atualmente a banca tem uma grande variedade de produtos em que os comerciais podem focar a sua atividade sendo, assim, importante ajudar a focá-los nos produtos mais relevantes para a instituição em cada período de tempo.

No Millenium bpc a uniformização das prioridades do Banco para toda a Rede Comercial é feita através da Matriz de Campanhas: documento que explicita para cada produto qual sua importância relativa naquele determinado período. De forma a permitir flexibilidade na gestão de prioridades, a atividade comercial organiza-se em trimestres, denominados por ciclos comerciais, sendo a Matriz revista ao iniciar cada ciclo. Adicionalmente, os ponderadores atribuídos a cada campanha[2] diferem entre plataformas permitindo ajustar a estratégia do Banco a cada segmento de clientes.

Para cada campanha presente na Matriz são distribuídos objetivos comerciais concretos e específicos pelas unidades de negócio, denominadas de Centro de Custo (CC), ativos no ciclo comercial. Isto é, nas sucursais de gestão personalizada (plataformas GPP e GPN) os objetivos são atribuídos a nível individual, dado que cada gestor tem a sua própria carteira de clientes, cada gestor está alocado a um CC. Por outro lado, no MM a carteira de clientes é da sucursal desta forma os objetivos são atribuídos à sucursal e todos os seus colaboradores trabalham em conjunto para os atingir, a sucursal tem um CC único. A Figura 3 sumariza a estrutura do Retalho de forma a explicitar o seu funcionamento.

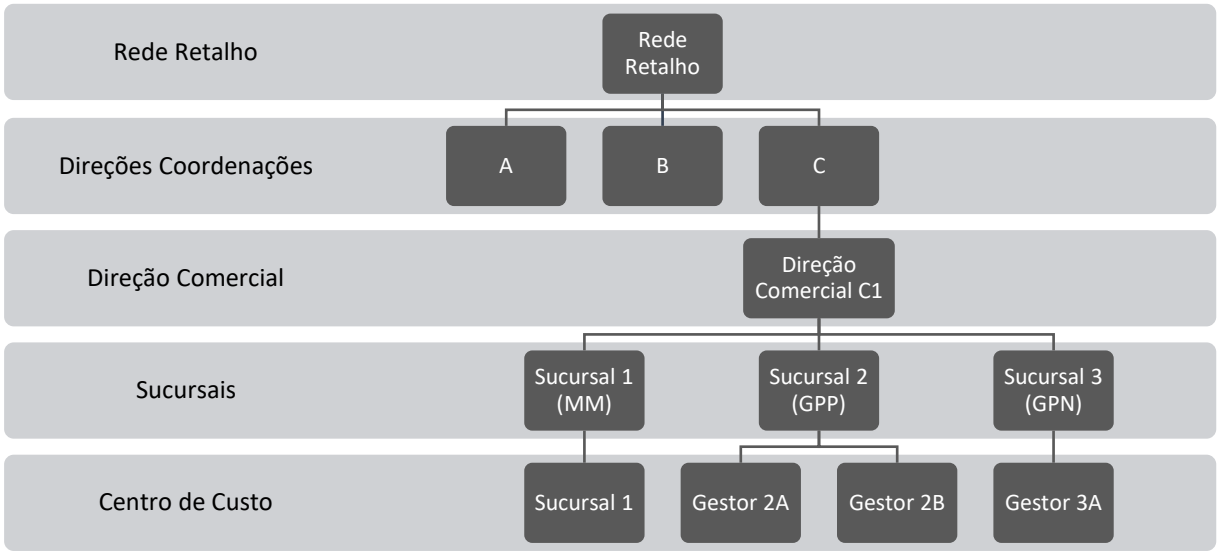



Figura 3. Estrutura Exemplo da Rede de Retalho

A avaliação de performance comercial de cada CC é efetuada através do Indicador de Performance (IP): um índice composto pelo desempenho dos colaboradores em cada um dos objetivos ponderado pelas importâncias definidas na Matriz e pela margem de lucro produzida pelas suas vendas. Por conseguinte atingir o objetivo em campanhas com maior relevância na Matriz é mais importante do que cumprir objetivos em produtos menos relevantes. Sendo por isso a Matriz tão importante para

[2] **Campanha**: denominação dada aos produtos, conjunto de produtos ou ações para os quais são definidos objetivos comerciais em cada ciclo. Como por exemplo: Crédito Habitação (produto), Cartões (conjunto de produtos) ou Captação de Clientes (ação).

focar a RR numa ou outra campanha consoante as necessidades e estratégia do Banco tem em determinado momento.

Em suma, a dinâmica de objetivos montada para a RR e seu o impacto tanto a nível individual na avaliação e compensação dos gestores como a nível global nos cumprimentos de metas do Banco justifica a necessidade de uma adequação correta dos objetivos distribuídos a cada CC.

## 2.2. DQAR

A Direção de Qualidade e Apoio à Rede (DQAR) é constituída por seis áreas ilustradas na Figura 4, as quais possuem posteriormente diversas equipas. A Direção divide-se entre duas responsabilidades principais que se traduzem no relacionamento constante quer com a RR quer com os clientes.

Por um lado, é responsável por controlar a qualidade do serviço prestado e gerir parte da relação com os clientes. Nesta vertente, a Direção atua em questões relacionadas com pequenos descobertos e incumprimentos (SAC); com a avaliação do serviço feita pelos clientes, recolhendo informação e elaborando os inquéritos e modelos de recolha do *feedback* (DEMS); e, por último, com o tratamento de queixas e reclamações de clientes do Retalho (CAC).

Por outro lado, é responsável por organizar a RR e prestar-lhe suporte na sua atividade diária. O Departamento de Suporte ao Negócio (DSN) é responsável pelo suporte técnico e operacional prestado à Rede. Neste departamento são geridas questões relacionadas com o rigor operacional (UGPO); com a transacionalidade e o parque de máquinas[3] (UOT); com o suporte técnico para os diversos aplicativos utilizados pela rede (USO); e com os objetivos e informação de gestão (UOIC).

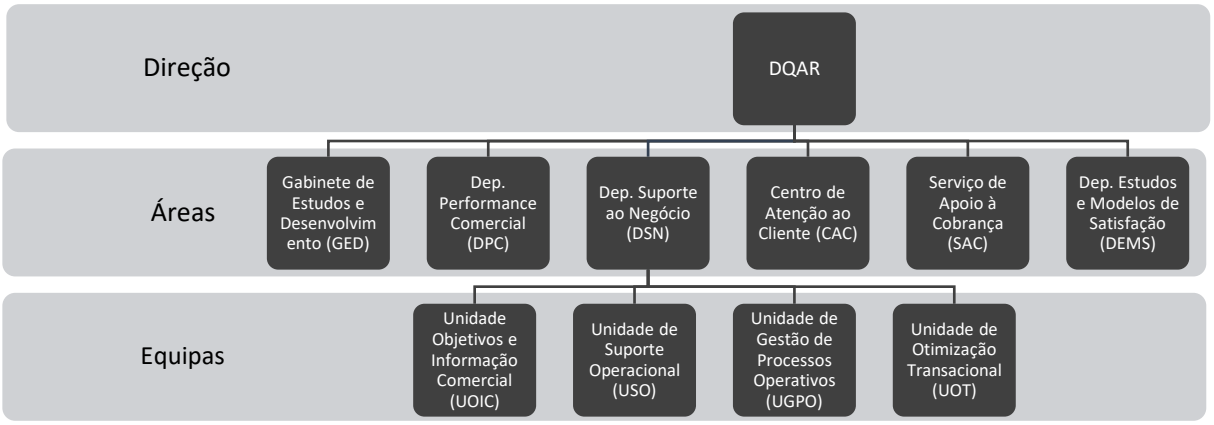



Figura 4. Organograma da Direção de Qualidade e Apoio à Rede

Adicionalmente a DQAR tem como responsabilidade a organização da rede para que esta se adeque da melhor forma às necessidades dos clientes em todo o território. Hoje em dia, as realidades estão em constante mudança e os clientes querem cada vez mais um serviço ágil e rápido. Desta forma, é importante existir um olhar cuidado à rede e uma preocupação na inovação do serviço e dos modelos de negócio. Para tal, surge o Gabinete de Estudos e Desenvolvimento (GED), uma área que avalia de

[3]O Parque de Máquinas consiste na rede de máquinas do BCP que permitem aos clientes realizar determinadas operações sem necessidade de se deslocarem aos balcões, tais como depósito e levantamento de numerário ou cheques, transferências, pagamentos etc. Existem vários tipos de máquinas que permitem realizar diferentes tarefas: multibanco/ATM, MTM entre outras.

forma constante as necessidades da rede proporcionando mudanças inovadoras nos modelos de negócio e restruturações das Direções Comerciais de forma a adequar ao máximo a estrutura às necessidades de cada geografia. Para além disto, são ainda responsáveis pelo Sistema de Incentivos da Rede.

Em suma, a Direção trabalha como um todo nas diferentes vertentes necessárias para apoiar a Rede Comercial. Tem como missão minimizar a carga operacional que existe nos colegas comerciais de forma a maximizar o seu tempo dedicado à venda e produtos e serviços de valor acrescentado para os clientes e para o Banco.

### 2.2.1. A Equipa e as Suas Tarefas

A Unidade de Objetivos e Informação Comercial (UOIC) é responsável por criar e monitorizar todos os processos relacionados com a dinâmica de objetivos da RR. De forma a manter esta dinâmica funcional é necessário, não só, mapas de acompanhamento que mostrem a cada *player* quão longe está da sua meta, mas também, uma infraestrutura de dados que permita atribuir cada venda ao gestor ou sucursal correta.

Ao iniciar cada ciclo, a equipa recebe informação de quais os objetivos globais para cada campanha, sendo estes distribuídos pela Rede de acordo com modelos pré-definidos. Ao longo do ciclo, a equipa utiliza processos de SAS e SQL para contabilizar as vendas realizadas e atribuí-las ao CC que a realizou. A UOIC tem então como responsabilidade manter atualizada, ao nível quase diário, a informação disponível no portal interno sobre a performance dos diversos CC em cada uma das campanhas.

| Sucursal | Campanha | Posição | Resultados | | |
|---|---|---|---|---|---|
| | | | Realizado | Objetivo | G.R.O. |
| | | 31Mar | 239 | 252 | 95% |
| | | 31Mar | 180 | 137 | 131% |
| | | 15Abr | 82 | 104 | 79% |
| | | 30Mar | 1.459.618 | 435.900 | 335% |
| | | 11Mai | 618.766 | 508.100 | 122% |
| | | 30Mar | 104.081 | 164.546 | 63% |
| | | 31Mar | 152.122 | 230.201 | 66% |
| | | 31Mar | 476.845 | 437.500 | 109% |
| | | 31Mar | 625.305 | 725.500 | 86% |
| | | 31Mar | 31.500 | 11.000 | 286% |
| | | 31Mar | 0 | 43.500 | 0% |
| | | 31Mar | 546.975 | 152.500 | 359% |
| | | 31Mar | 0 | 9.500 | 0% |
| | | 31Jan | 7.755 | 7.894 | 102% |
| | | 28Fev | 8.167 | 8.167 | 100% |
| | | 31Mar | 5.009 | 5.043 | 101% |

Figura 5. Secção do Mapa Global de Campanhas

Os mapas de apresentação de resultados são realizados pela equipa sendo por isso bastante versáteis e dinâmicos adequando-se às especifidades das campanhas de cada ciclo comercial. Os mapas encontram-se organizados hierarquicamente pela estrutura do Retalho o que permite a cada CC a localização rápida dos seus valores e uma vizualização agregada para cada nível hierarquico analisar os valores das suas equipas. Cada mapa é acompanhado de um mapa detalhado que apresenta todos os registos de venda ou acção que levaram a uma contabilização para o objetivo.

O exemplo demonstrado na Figura 5 representa o Mapa Global de Campanhas que possui para cada CC o resumo essencial de cada campanha: a data de atualização de cada resultado (**Posição**), qual o seu objetivo para o ciclo em questão (**Objetivo**) e qual o montante do produto que já efetuado

(**Realizado**). A avaliação do cumprimento de objetivos é realizada através da variável **GRO** que representa a percentagem de objetivo cumprida à data de atualização das campanhas, visível na Equação (1). O objetivo principal de cada CC é atingir pelo menos GRO 100 em todas as campanhas.

$$GRO = \frac{realizado}{objetivo} * 100 \quad (1)$$

Adicionalmente, a equipa é responsável por fornecer informação de gestão necessária para apoiar diversas tomadas de decisão, maioritariamente decisões relacionadas com mudanças nos modelos de distribuição de objetivos e nas regras de contabilização de cada venda.

Em suma, a equipa funciona como um relógio desportivo alertando a Rede Comercial quão longe está da meta, a que velocidade vai e qual o esforço final para atingir as metas pedidas para o ciclo.

# 3. ENQUADRAMENTO TEÓRICO

Neste capítulo realiza-se uma contextualização teórica do projeto, focando-se principalmente em literatura existente sobre atribuição de objetivos nas organizações e sobre as técnicas e metodologias utilizadas para o desenvolvimento do projeto.

## 3.1. DEFINIÇÃO DE OBJETIVOS NAS ORGANIZAÇÕES

A definição de objetivos influência a performance dos indivíduos e consequentemente das organizações através de quatro mecanismos. Em primeiro lugar, a definição de objetivos direciona a atenção e o esforço dos colaboradores das empresas para as tarefas mais relevantes e de maior valor acrescentado. Em segundo lugar, os objetivos têm uma função de motivação e de criação de energia, tendo-se demonstrado que objetivos maiores conduzem a um maior esforço do que objetivos mais pequenos. Em terceiro lugar, a existência de objetivos afeta a persistência ajudando os colaboradores a manterem o foco ao longo do tempo. Por último, os objetivos afetam indiretamente as ações por conduzirem à descoberta e utilização de conhecimento relevante para a realização das tarefas. [3]

Atualmente, a maioria das organizações possui uma gestão através de objetivos, quer pessoais, quer de equipa, sendo que a atribuição de objetivos é um dos fundamentos da maioria das teorias sobre motivação no emprego. [4] No BCP a distribuição de objetivos para a Rede Comercial funciona como um mecanismo para atingir os objetivos globais a que a instituição se propõe a cumprir. Inicia-se o processo com objetivos globais para a Rede que são posteriormente distribuídos da forma mais adequada possível por todas as unidades de negócio.

De forma a potenciar os efeitos dos mecanismos acionados pela atribuição de objetivos é importante que os objetivos cumpram algumas regras. A dinâmica montada no Retalho do MBCP para a repartição dos objetivos respeita as principais características que a literatura identifica como cruciais para que os objetivos conduzam à satisfação pessoal e ao incremento de performance nas organizações. De seguida são expostos estes aspetos, o porquê de serem importantes e como são aplicados na organização.

**Os objetivos devem ser específicos:** objetivos específicos, se possível quantificados, permitem aos colaboradores medir o seu próprio processo [4] e direcionar as suas ações de forma a atingir o que lhes é pedido. No MBCP são estabelecidos objetivos específicos para variados produtos sendo estes objetivos quantificados e ao nível individual. Os objetivos podem ser estabelecidos através de quantidades monetárias como é o caso do objetivo de Crédito Habitação, ou em número de vendas como o caso da venda de Cartões.

**Os objetivos devem ser difíceis, mas atingíveis:** se os objetivos forem demasiado acessíveis não existirá um aumento de esforço nem de motivação. Por outro lado, se os objetivos forem demasiado difíceis e inatingíveis os membros das organizações irão rejeitar os objetivos considerando-os demasiado irracionais. [4] É derivado deste princípio, que surge a necessidade de uma melhor adequação dos modelos de distribuição de objetivos do MBCP e a motivação para o projeto descrito neste relatório. A caracterização de um objetivo de difícil ou impossível depende fortemente da *self-efficacy* do indivíduo que o recebe. Isto é, depende da perceção e confiança de cada um nas suas próprias capacidades e competências, o que incrementa a dificuldade de adequação dos objetivos a cada pessoa. [4]

**Os objetivos devem ser aceites:** a participação dos membros das organizações no processo da atribuição de objetivos tende a incrementar o comprometimento para com os objetivos, conduzindo a níveis mais elevados de performance. O maior comprometimento surge de uma melhor compreensão da origem dos objetivos e da certificação por parte de cada um que o montante a si atribuído não é insensato. Não estando no processo de atribuição de objetivos a aceitação de cada indivíduo aos objetivos impostos será condicionada, mais uma vez, pela sua *self-efficacy*. Indivíduos com maior confiança nas suas capacidades tenderão a aceitar mais facilmente objetivos superiores do que indivíduos com menores níveis de *self-efficacy*. [4] No mecanismo de distribuição de objetivos montado para a RR existe participação dos membros da organização em diferentes passos consoante os níveis hierárquicos. Primeiro, os Coordenadores[4] participam no processo de alteração e definição dos modelos de distribuição de objetivos pelas suas equipas. Posteriormente, estando os objetivos calculados estes podem ser redistribuídos dentro de cada coordenação de acordo com o conhecimento de negócio de cada Coordenador. Os Diretores Comerciais, segundo nível hierárquico, podem transferir a maioria dos objetivos entre sucursais da sua Direção Comercial ajustando os objetivos aos seus conhecimentos sobre a realidade das equipas.

Outro fator que contribui para a aceitação dos objetivos propostos para cada sucursal do MBCP é o facto dos critérios de distribuição dos objetivos serem do conhecimento de toda a Rede. Para além destes critérios, também as formas de contabilização de cada venda são claras e do conhecimento de todos os colaboradores, ao iniciar cada ano.

**Deve existir feedback:** o feedback atua em duas frentes relevantes: primeiro permite aos membros da organização saber quão bem estão a agir em determinado momento, agindo como fator motivador e auxiliando a gerir esforços. Depois, permite aos colaboradores ajustar as suas estratégias para aumentar performance. [4] No Millennium, de forma a dar feedback constante às equipas existe um conjunto de mapas, atualizados diariamente, que mostram aos colaboradores da RR o seu objetivo e quanto já cumpriram desse objetivo e quanto tempo está disponível para o cumprimento do respetivo objetivo.

**Objetivos são mais eficazes quando são utilizados para avaliar a performance:** [4] Ao serem utilizados para avaliar a performance a importância dada aos mesmos sobe, subindo também o nível de compromisso. De acordo com Locke e Latham a relação objetivo-performance é maior quando as pessoas estão comprometidas com os seus objetivos. [3] Existem vários aspetos que podem contribuir para o aumento de compromisso para com os objetivos entre os quais, a *self-efficacy* dos indivíduos e a existência de incentivos monetários. No Retalho do MBCP uma grande componente da avaliação dos colaboradores e das equipas é efetuada através dos indicadores atingidos em cada objetivo. Adicionalmente, o Sistema de Incentivos está assente no mecanismo de objetivos para o Retalho.

**A aproximação de prazos aumenta a eficácia dos objetivos:** a aproximação de um prazo tende a levar os empregados a realizarem esforço adicional para o cumprimento das suas tarefas e objetivos. [4] Por isso, é importante estabelecer prazos pré-definidos que sejam do conhecimento dos colaboradores. Para o mecanismo de objetivos comerciais montado no MBCP os prazos são fixos uma vez que a atividade comercial se organiza em períodos de três meses, coincidentes com os trimestres do ano, estes períodos são denominados de ciclos comerciais.

[4] **Coordenador**: Responsável por uma Direção Coordenação da Rede de Retalho

**Objetivos de Grupo são tão importantes como objetivos individuais:** vários estudos demonstram que a combinação de objetivos individuais com os de equipa conduzem a maiores níveis de satisfação individual e rentabilidade. [4] Trabalhar em equipa pode aumentar produtividade das organizações, no BCP a atribuição de objetivos realiza-se tanto a nível individual como de equipa. Para além disto os resultados agregados das equipas condicionam os resultados dos Diretores Comerciais.

Concluindo, a distribuição de objetivos pode conduzir a um aumento de performance da organização quando implementada de forma eficaz e de acordo com alguns princípios estudados pela literatura. A dinâmica de objetivos montada para a RR do BCP respeita os princípios discutidos, necessários para maximizar os efeitos positivos na performance derivados da atribuição de objetivos. Existindo um esforço continuo para criar um sistema de objetivos justo e claro, procurando-se sempre distribuir objetivos ajustados à realidade de cada CC e do mercado em que está inserido através de modelos de distribuição adequados.

## 3.2. O PROCESSO DE DATA MINING

*Data Mining* define-se pelo processo de criar informação e revelar *insights* através de dados. As técnicas utilizadas podem ser aplicadas a uma variedade de indústrias e de problemas de negócio tais como segmentação de clientes, previsão e deteção de fraude, adequação de estratégias de marketing, entre outros.

No final do século XX as técnicas de *Data Mining* popularizaram-se como uma abordagem capaz para apoiar, melhorar e expandir processos existentes nas organizações. No entanto a crescente adoção de técnicas de *Data Mining* surgiu sem que existisse uma metodologia padrão de como as implementar, significando que nem sempre se conseguia recriar os projetos. A abordagem utilizada em determinado projeto dependia da equipa ou pessoa que o criasse. Desta forma, surgiu a necessidade de criar metodologias que normalizassem a implementação dos projetos dentro das organizações e entre elas, tornando os projetos mais confiáveis e recriáveis.

Nasceram então diversos *workflows*, ou metodologias, para o desenvolvimento de projetos de *Data Mining* que ganharam relevância especial pela adoção na indústria, tornando-se hoje *standards* no meio. Os exemplos mais populares incluem o *Knowledge Discovery Databases* (KDD); *Cross-Industry Standard Process for Data Mining* (CRISP-DM) e o *Sample, Explore, Modify, Model, Assess* (SEMMA). Estas metodologias representam um conjunto de etapas sequenciais que devem ser tomadas no desenvolvimento de um projeto, sendo no seu core semelhantes entre si. [5]

No fundo na realização de um projeto de *Data Mining* as etapas consideradas devem-se adequar ao problema a resolver e ao contexto em que se insere. Ao iniciar um projeto é importante realizar uma fase de planeamento e entendimento do negócio. Esta fase inicial foca-se na definição dos objetivos do projeto e nos requerimentos do ponto de vista de negócio, que devem ser traduzidos para a definição de um problema de *Data Mining*. [6]

Estando desenhado o problema procede-se para a fase de recolha e exploração dos dados. A exploração dos dados tem como intuito principal realizar uma análise aos dados que permita a descoberta de padrões escondidos nos dados, a avaliação da qualidade dos dados, a identificação e tratamento de valores omissos ou de valores anómalos, entre outros. De seguida, deve realizar-se a fase de preparação dos dados que tem como objetivo primário potenciar o poder preditivo das

variáveis conduzindo a um aumento da performance do modelo. Nesta fase realizam-se tarefas como a redução da dimensionalidade dos dados, a transformação de variáveis e a seleção das variáveis mais importantes. Resultando num *dataset* pronto para a modelação.

Os dados transformados resultantes das fases anteriores são então utilizados para a fase de modelação. Nesta etapa são aplicados diferentes modelos analíticos aos dados de forma a produzir-se o resultado pretendido. Para além da aplicação simples de um modelo é necessário efetuar a otimização os seus parâmetros de forma a obter o melhor resultado possível.

Por último, é necessário avaliar os resultados obtidos pelos diversos modelos testados de forma a escolher-se o modelo que produzirá os melhores resultados em dados novos e efetuar a sua implementação no contexto do negócio. A implementação varia de projeto para projeto podendo ser mais simples como por exemplo um relatório, ou mais complexa tal como a implementação de um processo regular e automatizado.

## 3.3. EXPLORAÇÃO DOS DADOS

A exploração dos dados é uma das fases do projeto apresentado neste relatório onde se incluem tarefas como a descoberta de padrões nos dados, a identificação de valores extremos ou de *outliers* e o tratamento de valores omissos, por exemplo. De seguida são apresentados diferentes métodos de tratamento dos valores omissos e de *outliers* presentes nos dados que devem ser aplicados antes de se prosseguir para a seleção de variáveis e modelação.

### 3.3.1. Valores Omissos

A existência de valores omissos nos dados recolhidos pode resultar de erros nos processos de recolha de dados ou de erros no armazenamento por exemplo. A ocorrência de falhas na informação pode ser detetada quer nos dados recolhidos para treino quer nos novos dados para os quais se pretende aplicar o modelo e conhecer o seu *output*. O tratamento destes valores omissos é de extrema importância uma vez que existem algoritmos que não conseguem suportar estes valores.

A eliminação de registos ou de variáveis pode ser uma solução a considerar para o tratamento de valores omissos. Se um registo apresentar valores omissos num alargado número de variáveis pode ser retirado da análise. Do mesmo modo, se uma variável apresentar valores omissos em grande parte dos registos pode esta ser excluída da análise. A principal vantagem da eliminação é a criação de um modelo robusto apenas com dados observados e recolhidos sem alteração. Contudo, este método apresenta desvantagens como a perda de informação.

Existindo um reduzido número de valores omissos e não se querendo excluir esses registos da análise pode realizar-se a imputação dos valores. A imputação dos valores, isto é, o preenchimento dos campos em falta, pode ser efetuada através de substituição por uma estatística descritiva ou através da previsão dos valores. A substituição de um valor por uma estatística descritiva da variável, tal como a média ou mediana para variáveis intervalares ou moda para varáveis categóricas, tem como principal vantagem a facilidade de implementação. No entanto, apresenta desvantagens como não ter em consideração a correlação entre variáveis, uma vez que é calculada ao nível da coluna, ou não ser um método preciso.

A previsão dos valores omissos para a sua imputação pode ser uma abordagem mais precisa e que tem em consideração as restantes colunas e a sua relação com a coluna em análise. Como desvantagens existem as desvantagens específicas de cada modelo utilizado quer seja a dificuldade computacional ou a dificuldade de otimização dos parâmetros. Os algoritmos utilizados para efetuar esta previsão de valores omissos podem variar entre algoritmos baseados em distância, em densidade ou mesmo uma regressão linear.

Um algoritmo baseado em distâncias entre observações que dada a sua simplicidade é ainda utilizado para a imputação de valores omissos nos dados é o *k-Nearest Neighbor* (kNN). [7] O algoritmo funciona através da *feature* similarity, ou semelhança entre variáveis, seleciona os *k* pontos do *dataset* mais perto da observação para a qual queremos estimar o *target*; posteriormente, calcula a média da variável *target* nesses *k* pontos retornando esse valor para a observação desconhecida. Sendo *k* o número de pontos a ser considerados para a recolha de informação.

Sendo um algoritmo relativamente simples a sua dificuldade encontra-se na adequação dos parâmetros: qual o número de vizinhos (*k*) a selecionar e qual a medida de distância a considerar. A medida de distância poderá ser simples tal como a distância Euclidiana ou mais complexa como a distância de Mahalanobis. [7] Apesar de ser um algoritmo simples possui algumas desvantagens visíveis na sua implementação sendo computacionalmente pesado. À medida que a dimensionalidade aumenta e são utilizadas mais variáveis e/ou observações a sua complexidade incrementa uma vez que o cálculo das distâncias será mais complexo. [8] Dito isto, se a dimensionalidade for reduzida esta desvantagem é mitigada.

Em suma, existem diversas formas de se lidar com dados em falta *datasets* utilizados para a modelação que apresentam algumas vantagens e desvantagens conforme a sua complexidade. Métodos mais complexos e avançados tendencialmente produzem resultados mais precisos sendo que têm uma dificuldade de implementação superior. Por outro lado, a simples imutação pela média ou mediana representa uma alternativa que consome menos tempo e menos poder computacional, produzindo resultados que poderão ser menos precisos.

### 3.3.2. Outliers

Hawkings define um *outlier* como uma observação que é tão desviada das restantes que levanta as suspeitas de ter sido criada por um mecanismo diferente. A identificação de *outliers* é importante pois a utilização destes valores na modelação pode conduzir a uma má interpretação de parâmetros e estimações enviesadas. [9, 10] Vejamos o exemplo na Figura 6, a estimação altera-se drasticamente com a eliminação dos pontos *outliers* aproximando-se mais da realidade.

A existência de *outliers* nos dados pode ser resultante de erros humanos na inserção de dados nas bases de dados, de erros nos processos de recolha ou tratamento de informação, de erros derivados de uma amostra não representativa da população deixando alguns pontos como *outliers* quando na realidade não o são, ou de uma variação natural dos dados.

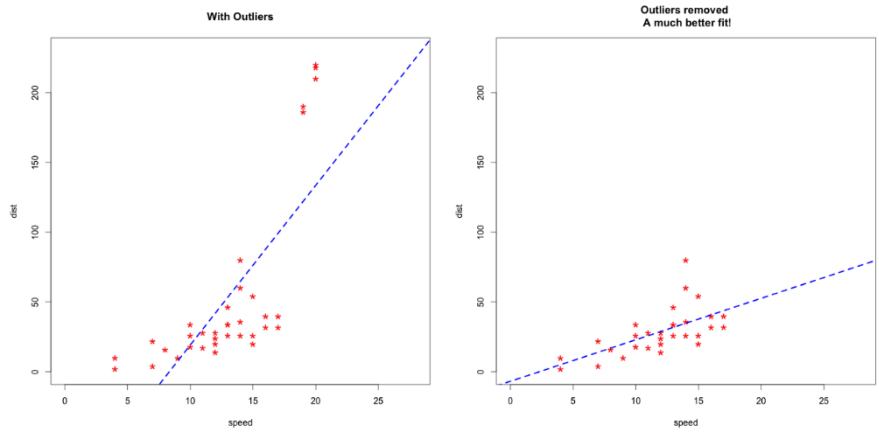



Figura 6. Efeito de *Outliers* na Estimação de um Modelo
Fonte: https://medium.com/@mehulved1503/effective-outlier-detection-techniques-in-machine-learning-ef609b6ade72

A identificação de *outliers* pode ser realizada através da análise de cada variável individualmente, sendo uma análise uni-variada, ou analisando as observações no espaço em que se encontram considerando duas ou mais variáveis, tratando-se então de uma análise multivariada. De seguida são sumarizados algumas das abordagens mais utilizadas para a identificação e *outliers.* [9]

- Análise de valores extremos: análise uni-variada à distribuição das variáveis, identificando valores extremos acima ou abaixo de um *treshold*. Usualmente os *tresholds* são definidos através da média ($\mu$) acrescentando-se ou diminuindo-se três vezes o desvio padrão ($\sigma$), resultando no intervalo $[\mu - 3\sigma, \mu + 3\sigma]$. Ou através do intervalo interquartil (IQR) que se obtém efetuando a diferença entre o valor do 3º quartil (Q3) e o valor do 1º quartil (Q1). Os *tresholds* utilizados nesta abordagem obtêm-se somando-se 1,5 vezes o IQR ao 3º quartil (Q3) e subtraindo-se 1,5 vezes esse valor ao 1º quartil (Q1), isto é, $[Q1 - 1{,}5IQR, Q3 + 1{,}5IQR]$. Sendo as observações que se encontram fora dos intervalos classificadas como outliers. A análise de valores extremos pode ainda ser realizada através da análise dos histogramas ou *boxplots* observando os valores de quebra na distribuição. Por um lado, a aplicação deste método é bastante simples e requer pouco poder computacional. Por outro lado, analisando as variáveis individualmente pode-se perder a identificação de padrões no espaço multidimensional em análise. Em adição, é importante a análise dos limites dos intervalos com risco de se cortar observações em demasia se a distribuição da variável não for normal.

- Modelos de *Clustering*: *clustering* e deteção de *outliers* são muitas vezes considerados problemas complementares uma vez que o primeiro procura identificar grupos de observações próximas e o segundo identificar observações isoladas. Assim, muitos dos algoritmos de *clustering* podem identificar *outliers* como um efeito secundário do seu objetivo principal.

- Modelos Baseados em Distâncias: calculando as distâncias entre todos os pontos no espaço multidimensional pode-se identificar *outliers* como sendo as observações cuja distância ao ponto mais próximo seja substancialmente maior do que os restantes pontos. Uma desvantagem da aplicação deste método é a dificuldade de definir o limite de corte de uma observação, isto é, quão grande deve ser a distância do ponto para ser considerado um *outlier*.

- Modelos Baseados em Densidade: a deteção de *outliers* é efetuada através do cálculo da densidade de cada ponto.

Os métodos que consistem em analises multivariadas podem produzir resultados mais precisos uma vez que analisam o espaço como um todo. No entanto estes métodos têm um nível de complexidade maior sendo mais dispendiosos quer a nível computacional quer a nível de tempo. A escolha da metodologia a utilizar deve ter em conta o contexto do problema em questão e as necessidades de negócio.

## 3.4. Aprendizagem Automática Supervisionada

Em 1959 Arthur Samuel definiu *Aprendizagem Automática* (AA), ou *Machine Learning* (ML), como um subcampo da Inteligência Artificial que pretende dar aos computadores a capacidade de aprenderem sem serem explicitamente programados para tal [11]. Com a utilização de algoritmos de *Machine Learning* conseguimos que os computadores aprendam através de um conjunto de *inputs* e que se adaptem para tomar decisões de forma automática de forma mais rápida, eficiente, e em muitos casos melhores do que um agente humano. A qualidade das decisões é avaliada de acordo com a sua proximidade às observações reais. [12]

As técnicas e algoritmos de Aprendizagem Automática têm vindo a ser aplicadas num vasto conjunto de problemas em áreas tão variadas como a astronomia, medicina e ciências sociais. Um dos aspetos em comum de todos os problemas onde é aplicado AA é serem problemas onde os dados têm padrões demasiado complexos para serem reconhecidos por humanos numa análise simples. [13] Os tipos de problemas resolvidos através de AA e análise de dados podem ser divididos em categorias [14]:

- Problemas de Classificação: em qual classe se integra cada observação, “é isto ou aquilo?”. Utilizados para diferentes fins tais como a identificação de clientes propensos a um produto, ou a deteção de anomalias;

- Problemas de *Forecast/*Previsão: prever uma variável intervalar, “Quanto? Ou Quantos?”. Problemas como a previsão do valor de venda de produtos, por exemplo;

- *Clustering*/Segmentação: detetar estruturas nos dados e grupos de observações. Problemas desta categoria são problemas como a segmentação de clientes.

Os diferentes problemas que podem ser resolvidos com recurso a algoritmos de *Machine Learning* têm características e métodos de avaliação diferentes. Por um lado, num problema de classificação pode-se avaliar quantos registos foram classificados de forma correta. Por outro, num problema de previsão quer-se avaliar o modelo através do cálculo do seu erro, do afastamento do valor previsto ao valor real. Assim, para os diferentes problemas existem tipos de algoritmos que, por sua vez, se dividem em categorias. As categorias mais comuns são a Aprendizagem Supervisionada, ou *Supervised ML*, e Não Supervisionada, ou *Unsupervised ML*.

Na Aprendizagem Supervisionada de forma aprender os algoritmos recebem um conjunto de observações que podem possuir variadas variáveis (*inputs*) para os quais existe um variável *target* conhecida. Isto é, para cada registo existe uma resposta correta. O algoritmo aprende através de exemplos. [12] O objetivo destes algoritmos é aprender com base no *dataset* de treino para posteriormente aplicar o conhecimento a novos dados para os quais ainda não existe a resposta

correta efetuando classificações ou previsões de valores. Em suma, o algoritmo é uma função que recebe *inputs* em forma de uma matriz e devolve uma variável *target*, com base no que analisou anteriormente. Os problemas de previsão e de classificação são problemas que se incluem no âmbito da Aprendizagem Supervisionada uma vez que à partida já se conhece o resultado desejado.

Por outro lado, na Aprendizagem Não Supervisionada, os algoritmos procuram por padrões e semelhanças nos dados sem que sejam atribuídas respostas corretas a cada observação. O algoritmo categoriza os *inputs* com maior semelhança entre si. [12] Problemas de *clustering* são problemas no âmbito da Aprendizagem Não Supervisionada uma vez que não se conhece, *a priori*, qual o cluster onde se insere cada registo.

O projeto apresentado neste relatório tem como intuito responder à pergunta "qual a proporção de objetivo que cada gestor/sucursal consegue realizar" sendo então classificado como um problema de previsão, encontrando-se no âmbito de Aprendizagem Supervisionada.

### 3.4.1. Regressão Linear

Uma regressão linear é um modelo que estabelece uma relação linear entre as variáveis de *input* e a variável de *output*/*target*. Este modelo é definido através do treino numa amostra representativa da população e os parâmetros estimados assumem-se válidos para futuras previsões na população. [15]

Uma Regressão Linear pode ser escrita como demonstra a Equação (2), para cada observação $i$ a variável *target* $y$ é dada através de um termo constante $\beta_0$, somando-se da combinação linear das variáveis de *input* $x_{1i}, x_{2i}, \dots, x_{ki}$ e o termo de erro $\varepsilon_i$. O método mais comum para a estimação dos coeficientes $\beta$ é *Ordinary Least Squares* (OLS) onde os parâmetros são estimados de forma a minimizar a soma do quadrado dos erros nas observações de treino.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad (2)$$

Após estimar os parâmetros para cada variável e o valor de $\beta_0$ é possível interpretar estes valores e afirmar, com um determinado nível de significância estatística o efeito que cada variável terá no *target*. Tendo como exemplo a regressão linear simples da Equação (3) onde se estima o valor do salário por hora através dos anos de educação, o coeficiente da variável de educação (0,54) implica que por cada ano de educação adicional o salário por hora de um indivíduo aumentará 0,54 unidades. [16] Ao interpretar os coeficientes numa regressão com mais que uma variável independente os efeitos de cada variável são apenas interpretáveis sozinhos. Isto é, cada variável terá o efeito do seu coeficiente no *target*, mantendo-se tudo o resto constante.

A interpretação da constante ($\beta_0$) deve ser efetuada de forma cuidadosa, neste exemplo a sua interpretação seria que para um indivíduo sem qualquer ano de estudo o seu salário por hora seria -0,90 o que em termos económicos não faria sentido. Wooldridge sugere que uma das causas que podem conduzir a uma constante negativa, como neste exemplo, é a pequena percentagem de casos no *dataset* com baixa escolaridade. [16]

$$salário_i = -0{,}9 + 0{,}54\, educ \quad (3)$$

As relações lineares entre as variáveis dependentes e a variável *target* nem sempre são o suficiente para todos os problemas existentes na economia ou ciências. Existem diversas formas de incorporar

relações não lineares nestes modelos tais como a transformação da variável dependente, das variáveis independentes ou de ambas no seu logaritmo natural.

A transformação da variável *target* no seu logaritmo permite impor um efeito das variáveis independentes de percentagem sobre o *target*. Analise-se o exemplo anterior com a transformação da variável **salário**, dado na Equação (4). Inicialmente, ao estimar os parâmetros com as variáveis originais o efeito de um ano a mais de educação para um indivíduo com 5 anos ou com 10 seria o mesmo. No entanto, ao utilizar o logaritmo da variável *target* a interpretação do coeficiente de **educ** seria por cada ano adicional de educação (por cada unidade adicional) o salário do indivíduo incrementará 0,083*100=8,3%. [16] O efeito do parâmetro $\beta_k$ na variável *target* pode ser escrito como: $\%\Delta y \approx (100 * B_k)\Delta x_k$, ou seja, por cada aumento de uma unidade de $x_k$, mantendo tudo o resto constante, a variável *target* incrementará $(100 * B_k)$ %.

$$\log(salário_i) = 0{,}584 + 0{,}083\, educ \tag{4}$$

### 3.4.2. Árvore de Decisão

As Árvores de Decisão, ou *Decision Trees*, são modelos construídos através da partição de um espaço em subespaços originado um conjunto de regras "Se-Então". As Árvores de Decisão são compostas por um conjunto de nós, ramos e folhas, cada nó representa a avaliação de uma regra numa variável de *input*, seguido por ramos que representam os caminhos a seguir consoante a resposta à regra. Todos os ramos terminam com uma folha: um nó final onde é retornado o valor previsto da variável *target*. [17]

A Figura 7, ilustra uma Árvore de Decisão onde a raiz, o primeiro nó, indica que a primeira variável a ser considerada é a idade. Cada ramo representa um dos caminhos: ou a idade é superior a 50 anos ou não é. Se, para uma determinada observação a variável idade tiver um valor superior a 50 então o valor de *target* previsto pelo modelo será 0, caso contrário as restantes regras serão avaliadas até se chegar a uma folha com um *output*.

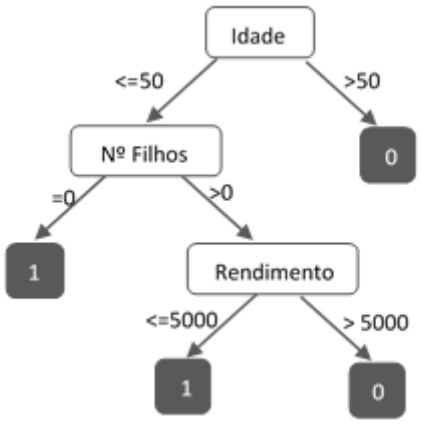



Figura 7. Exemplo de uma Árvore de Decisão criada com 3 variáveis: Idade, Número (Nº) de Filhos e Rendimento.

Assim, o processo de criar uma árvore é efetuado através da partição do espaço do *dataset* de treino em subespaços repetidamente. [18] A cada nova regra cria-se um novo subespaço de observações que são mais pequenos à medida que o comprimento da árvore aumenta. O comprimento da árvore denominado de *depth* é o equivalente ao número de nós existentes da raíz até à folha num

determinado caminho, sendo a *depth* da raíz igual a 0. Dependendo do algoritmo utilizado a partição em cada nó poderá ser binária, originando dois nós, ou não, originando 3 ou mais nós.

As Árvores de Decisão podem ser utilizadas quer para problemas de classificação quer para problemas de *forecast*. No primeiro, o valor retornado em cada folha representa a classe com maior número de observações naquele subespaço do *dataset* de treino. No caso de uma variável *target* intervalar, o valor da folha será a média dos valores de *target* das observações de treino que percorreram aquele caminho.

As principais questões na criação de uma Árvore de Decisão prendem-se com a definição da estrutura ótima: quantos nós deve ter, quando parar de acrescentar nós, que variável escolher e quando, que valor de separação deve ser escolhido para cada variável a cada nó, devem ser repetidas variáveis, entre outras questões. Testar todas as possibilidades torna-se computacionalmente dispendioso uma vez que existem infinitas combinações possíveis. Assim, a maioria dos algoritmos utilizados funcionam através de otimização *greedy*, ou gulosa. Isto é, a cada iteração as escolhas são efetuadas baseando-se no ótimo local.

A construção de uma árvore inicia-se com apenas um nó, que simboliza todo o espaço e aumenta-se a árvore adicionando um nó de cada vez. A cada passo é necessário escolher qual variável do espaço é escolhida para a partição e qual o valor de *treshold.* Num problema de *forecast* as escolhas são realizadas baseadas em procura exaustiva selecionando o par variável, *treshold* que origina menor erro, utilizando-se como medida de erro o *Sum of Squared Error* (SSE). [18]

Sendo um processo *greedy,* que tenta sempre minimizar o erro, poder-se-ia chegar ao ponto de criar folhas que contivessem apenas uma observação se o algoritmo criá-se a árvore sem nenhum critério de paragem. O que criaria uma árvore perfeita no *dataset* de treino mas com pouca capacidade de generalização para dados novos. Uma forma de resolver este problema é a criação de árvores grandes que posteriormente são "cortadas" através de *prunning*. O processo de *prunning* consiste em colapsar algumas das folhas juntando subespaços que anteriormente estavam divididos e é efetuado baseado num critério que balanceia o aumento do erro e a redução de complexidade do modelo. [18] Esta técnica ajuda a prevenir o *overfitting* do modelo. Pode-se ainda evitar *overffiting* através de outros critérios que param o crescimento da árvore. Entre eles destaca-se a definição de número mínimo de observações de treino por folha ou a definição da *depth* máxima da Árvore.

Em suma, as Árvores de Decisão são modelos de fácil interpretação devido à sua constituição por regras "Se-Então". Contudo existem muitos parâmetros a ter em consideração quando se implementa este tipo de algoritmos entre os quais *depth* máxima, número mínimo de observações por folha e número máximo de ramos criado em cada nó. Os resultados produzidos pelo modelo dependerão fortemente da especificação destes parâmetros e da sua otimização.

### 3.4.3. Gradient Boost

*Boosting* é uma técnica que permite a combinação de vários modelos individuais para a criação de um modelo que terá uma performance significativamente superior do que qualquer um dos seus modelos base por si só. Os modelos de base são treinados em sequência de forma a que o modelo final possa aprender com os modelos anteriores. Para cada novo treino é utilizado um *dataset* ponderado de

forma a que observações que tenham obtido erros maiores (*target* intervalar) ou que tenham sido mal classificadas (*target* binário) tenham um maior peso. [18]

Um modelo *Gradient Boost* aplica esta metodologia de *Boosting* utilizando o *gradient-descent* como algoritmo de minimização do erro no *ensemble*. [19] Este modelo utiliza, normalmente, *Decision Trees* como modelos base produzindo um algoritmo mais robusto do que as árvores individualmente sendo mais eficazes contra o *overfitting*.

À semelhança de uma Árvore de Decisão simples existe um inúmero conjunto de parâmetros que devem ser otimizados na implementação destes modelos tais como, o número de árvores a treinar, qual a sua *depth* máxima ou o número mínimo de observações numa folha.

### 3.4.4. Modelo *Ensemble*

*Ensemble* é uma técnica de ML que permite a combinação de dois ou mais modelos base num único modelo que é tendencialmente melhor que os modelos que o constituem individualmente. [20] A utilização de um *ensemble* contribui para a diminuição da variância e o enviesamento dos valores previstos para novos dados uma vez que se utiliza mais do que um modelo base. Adicionalmente, prever os valores com recurso a mais que um modelo de regressão ajuda a diminuir o efeito de *overffiting* de cada um dos modelos base utilizados.

As técnicas de *ensemble* podem ser categorizadas em *Bagging* e *Boosting*. Na subsecção anterior (3.4.3. Gradient Boost), explicou-se a técnica de *Boosting* dando o exemplo de um modelo *Gradient* Boost que aplica esta técnica para treinar vários modelos sequencialmente e por fim exportar um só resultado. Conforme exemplifica o esquema da Figura 8, através da técnica de *Boosting* o modelo final é apenas um, neste caso o Classifier-3.

A aplicação de *Bagging* permite a utilização de dois ou mais modelos base através da combinação dos seus dados preditos para cada observação. Nesta técnica os modelos base são treinados de forma independente sendo os resultados combinados através de uma média, ou média ponderada para *targets* intervalares ou da escolha da classe em maioria para modelos de classificação. Utilizando a técnica de *Bagging* os algoritmos dos modelos base podem ser diferentes, permitindo a combinação de Regressões com Árvores de Decisão ou outros algoritmos que se possa utilizar. Assim, os valores finais previstos derivam de dois ou mais modelos finais conforme ilustra a Figura 8.

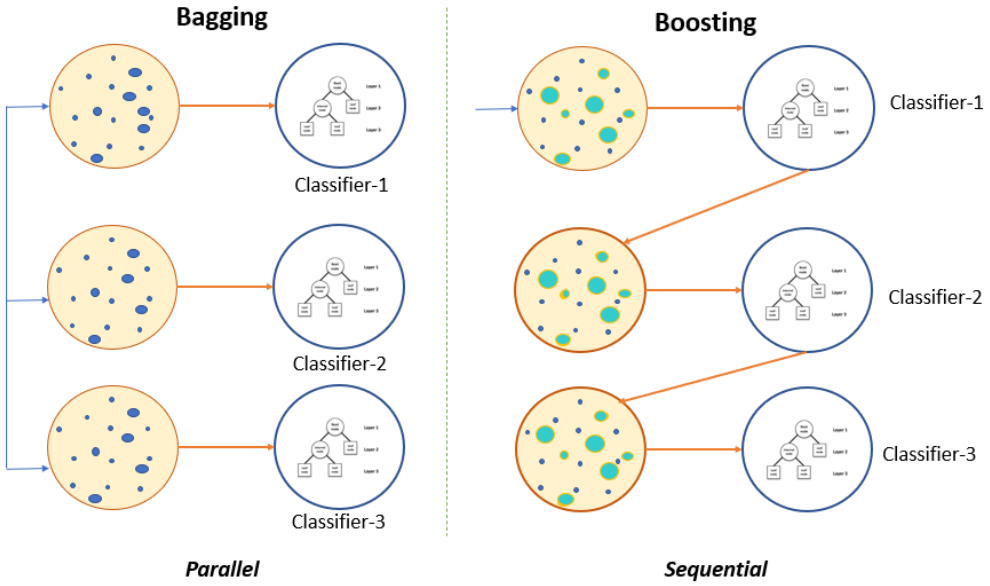



Figura 8. Esquematização das técnicas de *Ensemble*: Bagging e Boosting
Fonte: https://www.pluralsight.com/guides/ensemble-methods:-bagging-versus-boosting

### 3.4.5. Avaliação e Comparação de Modelos

O objetivo principal da modelação é obter uma função que recebendo um certo *input* e parâmetros produz uma previsão em novos dados com menor erro possível. Num problema de regressão o erro do modelo para cada registo *i* pode ser definido como a diferença entre o valor previsto pelo modelo ($\hat{y}$) e o valor original ($o$) do *target* naquela observação: $\varepsilon_i = \hat{y}_i - o_i$. As medidas de avaliação dos modelos são baseadas em sumários estatísticos dos erros de todas as observações. De seguida são expostas medidas de avaliação do erro tendo em conta um *target* intervalar. [21]

- **MAE** – *Mean Absolte Error* – Erro Absoluto Médio: Definido como $MAE = n^{-1}\sum_{i=1}^{n}|\varepsilon_i|$ representa a magnitude média de erro das observações em estudo na mesma escala do *target*.

- **MxAE** – *Maximum Absolute Error* – Erro Absoluto Máximo: Parte integrante do *output* do nó **Model Comparison**, devolve a magnitude máxima encontrada de afastamento do valor predito ao valor original.

- **SSE** – *Sum of Squared Errors* – Soma do Quadrado dos Erros: Definido como $SSE = \sum_{i=1}^{n}(\varepsilon_i)^2$ representa a soma total dos erros quadrados das observações em estudo. O recurso ao quadrado é justificado pela necessidade de eliminar o sentido do erro (sobrestimação versus subestimação). O cálculo do erro através do quadrado dos erros e não da sua magnitude (valores absolutos) conduz a uma influência relativa maior de erros maiores em comparação com influência de erros mais pequenos. O que significa que a SSE vai crescer à medida que o erro total se concentrar dentro de um número decrescente de erros individuais maiores.

- **MSE** – *Mean Squared Error* – Erro ao Quadrado Médio: Definido como $MSE = n^{-1}SSE = n^{-1}\sum_{i=1}^{n}(\varepsilon_i)^2$ sendo calculada através da SSE, esta medida não descreve o erro médio por si

só, dado que é mais influenciada por erros de maior magnitude, adicionalmente a MSE não está na mesma escala da variável *target*.

- **RMSE** – *Root Mean Squared Error* – Raíz Quadrada do Erro ao Quadrado Médio: Definido como $RMSE = MSE^{1/2} = [n^{-1}\sum_{i=1}^{n}(\varepsilon_i)^2]^{1/2}$, novamente, sendo esta medida calculada através da SSE, não descreve o erro médio por si só, dado que é mais influenciada por erros de maior magnitude.

Na Tabela 1 demonstra-se um exemplo fictício do impacto das magnitudes do erro nas diferentes medidas de avaliação. Estando as medidas a fornecer informação distinta é importante avaliar múltiplas medidas quando se compara diferentes modelos.

Nos diferentes casos apresentados observa-se que o MAE é igual não sendo representativo das diferenças de amplitudes dos erros. Para esta medida é indiferente se o erro ocorreu em apenas uma observação, mas com uma grande amplitude (Caso 5) ou se o erro é recorrente em múltiplas observações, mas com amplitudes menores (Caso 1). Por outro lado, a RMSE sendo sensível à amplitude dos erros vai aumentando à medida que existem erros com amplitudes maiores. Willmott e Matusuura defendem que sem a informação adicional não se consegue deferir quanto o RMSE e medidas semelhantes refletem a média do erro ou quanto representa as dos erros. [21] As medidas devem assim ser utilizadas em simultâneo sendo escolhidas de acordo com as regras e objetivos de negócios dependendo da importância da amplitude dos erros individuais no contexto do problema.

Em conclusão, existem diversas formas de avaliar o erro de um modelo que podem ser utilizadas para comparar abordagens diferentes e decidir qual a melhor. As medidas expostas revelam diferentes níveis de sensibilidade para a magnitude dos erros sendo por isso complementares e não exclusivas.

Tabela 1. Cinco casos hipotéticos de 4 erros e os seus totais correspondentes

| Variável | Caso 1 | Caso 2 | Caso 3 | Caso 4 | Caso 5 |
|---|---|---|---|---|---|
| $\varepsilon_1$ | 2 | 1 | 1 | 0 | 0 |
| $\varepsilon_2$ | 2 | 1 | 1 | 0 | 0 |
| $\varepsilon_3$ | 2 | 3 | 1 | 1 | 0 |
| $\varepsilon_4$ | 2 | 3 | 5 | 7 | 8 |
| $\sum\lvert\varepsilon_i\rvert$ | 8 | 8 | 8 | 8 | 8 |
| MAE | 2 | 2 | 2 | 2 | 2 |
| $\sum(\varepsilon_i)^2$ | 16 | 30 | 28 | 50 | 64 |
| RMSE | 2,0 | 2,2 | 2,6 | 3,5 | 4,0 |

Fonte: Willmott & Matusuura (2005) [21]

# 4. FERRAMENTAS

No contexto do estágio realizado no Millenium bcp foram usadas diversas ferramentas tais como o Microsoft Excel, o SAS Foundation, o SQL Server Management Studio, o Microsoft Access, o SAS Enterprise Miner e o Python.

De seguida apresenta-se uma descrição de cada ferramenta que não constou nos programas das unidades curriculares do mestrado, explicando-se qual o seu papel no dia a dia da equipa e qual a sua utilização no projeto apresentado neste relatório.

### Microsoft SQL Server Management Studio

O SQL Server Management Studio (SSMS) é um *software* que permite administrar bases de dados com elevados níveis de segurança. É uma ferramenta versátil que permite não só a criação de base de dados e a realização de modificações à sua estrutura, mas também, realizar pesquisas rápidas para a extração de dados. O SSMS permite facilmente integrar diferentes níveis de segurança para diferentes utilizadores, funcionalidade bastante relevante para empresas de elevada dimensão como é o MBCP.

Na equipa esta ferramenta é utilizada para a administração da base de dados que suporta todo o sistema de objetivos. Ao longo do projeto a ferramenta foi utilizada essencialmente para a verificação e controlo da qualidade dos dados bem como a criação de *scripts* para a atualização de tabelas de suporte ao *dashboard* realizado em Power BI.

### Microsoft Access

O Microsoft Access é outra ferramenta de gestão de base de dados que utiliza a linguagem SQL e VBA e possui uma *userface* mais simples e interativa. Através do Access é possível realizar diferentes operações como *queries* e criação de tabelas sem recurso a código. Este *software* permite ainda a criação de formulários que recebem *inputs* dos utilizadores e processam um determinado conjunto de *queries* pré-programadas apenas com recurso a um clique.

Todos os ciclos comerciais a equipa utiliza uma base de dados de Access, conectada à base de dados principal, de forma a gerir reduções e exceções necessárias nos objetivos das campanhas. A *interface userfrienly* do *software* permite realizar *queries* regulares com bastante rapidez e eficiência requerendo minimas alterações entre ciclos. Os principais processos realizados nesta base de dados são a redução de objetivos por ausências dos colaboradores ou encerramentos temporários de sucursais.

### SAS Foundation

O SAS Foundation é um *software* que permite aceder a dados de várias fontes, realizar análises estatísticas avançadas e reportar informação através de uma linguagem acessível que também pode incorporar linguagem SQL. O programa permite conexão a base de dados através de conexões ODBC, ou seja, permite aceder a toda a informação necessária da base de dados em SQL. Para além disto, através do SAS a equipa consegue visualizar e utilizar informações de bases de dados de outras direções tais como as informações de clientes guardadas no *datamart* do Marketing.

A linguagem utilizada é simples e versátil permitindo com grande rapidez a manipulação de dados. O *software* incorpora uma janela de log bastante detalhada onde se pode detetar erros nos *scripts* com bastante facilidade.

A maioria dos processos de contabilização de objetivos e alocação dos realizados a cada CC são realizados em SAS Foundation e executados de forma regular. Para além disto, todo o processamento de dados necessário para o cálculo de objetivos está implementado em *scripts* de SAS.

No projeto o SAS foi utilizado para a recolha de todas as variáveis internas, toda a manipulação de dados, pré-modelação e ainda na pós-modelação para gerar os objetivos para novos ciclos comerciais. Ao longo do estágio o SAS foi utilizado para a generalidade das tarefas de extração e manipulação de dados.

### SAS Enterprise Miner

O SAS Enterprise Miner (SAS Miner) é um programa que permite realizar um processo de *Data Mining* e produzir modelos preditivos através de uma *interface* amigável e visualizações gráficas detalhadas. A principal diferença entre os dois produtos da SAS referidos é a *interface*, no Miner a modelação é realizada através de *widgets* e numa perspetiva de fluxo. Os *widgets*, chamados de *nodes* ou nós, estão pré-programados para a maioria dos processos estatísticos que se precisa, se for necessário algo extra existe a possibilidade de adicionar **Code Editor Node** e programar livremente através da linguagem SAS.

No passo final da modelação o Miner permite exportar todo o fluxo como código SAS para ser executado facilmente no SAS Foundation.

O SAS Miner foi o *software* utilizado nas fases de modelação do projeto por permitir exportar um código SAS que facilmente pode ser executado por toda a equipa. Se o mesmo projeto fosse construído num *script* de Python tal não seria possível.

### Microsoft Power BI

O Microsoft Power BI (Power BI) é um conjunto de serviços que permitem criar relatórios e aplicações interativas de diversas fontes de dados tais como bases de dados, ficheiros Excel, dados na *cloud* entre outras. Os relatórios, ou *dahsboards*, são construídos de forma intuitiva com recurso a *widgets* pré-definidos e algum código de linguagem DAX. As visualizações são bastante customizáveis e interativas permitindo retirar conclusões dos dados em análise de forma simples e intuitiva.

O Power BI permite a partilha dos relatórios entre membros de uma organização com diferentes níveis de acesso aos dados, no entanto, para possibilitar a partilha é necessário a versão pro do *software*. Com esta versão existe ainda a funcionalidade de incluir relatórios em páginas web como por exemplo no portal interno da empresa.

Ao longo do estágio existiu a oportunidade de trabalhar com esta ferramenta para criar um *dahsboard* que permitisse de forma intuitiva e interativa analisar a utilização dos canais digitais por parte dos novos clientes do Retalho do Millennium bcp.

# 5. MODELO DE DISTRIBUIÇÃO DE OBJETIVOS: CRÉDITO HABITAÇÃO

Este capítulo contém uma descrição pormenorizada do projeto realizado durante o estágio, a metodologia e técnicas aplicadas, as fontes de informação utilizadas, bem como uma descrição dos dados. Adicionalmente, são explicadas outras tarefas realizadas ao longo do estágio no âmbito da extração, manipulação e análise de dados com recurso aos diferentes *softwares* referidos na secção anterior (4. Ferramentas ).

## 5.1. TIMELINE DO ESTÁGIO

O estágio teve duração de doze meses ao longo dos quais o principal objetivo foi rever os modelos de distribuição de objetivos, iniciando com as campanhas mais importantes: Captação de Clientes e Crédito de Habitação. Na Figura 9 encontra-se a cronologia detalhada dos projetos e tarefas explicados de seguida:

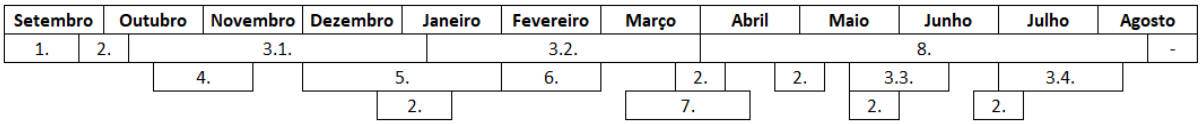



Figura 9. Cronologia do estágio

1. **Formação de integração**: formações com diversos responsáveis de área dentro das Direções Retalho com o objetivo de compreender como os diferentes departamentos se articulam e trabalham como um todo dentro de uma organização de grande dimensão. Existiu a oportunidade de receber formação de vários responsáveis e ainda a visita a uma sucursal de cada plataforma com vista em perceber a relação dos colaboradores com o sistema de objetivos e o trabalho realizado diariamente com os clientes.

2. **Cálculo dos Objetivos**: cálculo de objetos para diversas campanhas em cada ciclo comercial que consiste essencialmente na adaptação de *scripts* de SAS previamente desenvolvidos para adequação às regras e especificidades de cada ciclo. O *output* criado pelos programas é posteriormente utilizado e trabalhado no Excel de forma a calcular objetivos específicos para cada CC. Conforme referido anteriormente, o MBCP organiza-se em ciclos comerciais coincidentes com os trimestres do ano, contudo, devido ao Estado de Pandemia vivido no início de 2020, o 2º ciclo desse ano foi dividido em três ciclos mensais significando que a equipa teve de adaptar as infraestruturas de dados, os mapas de acompanhamento e o cálculo de objetivos a esta realidade. O cálculo de objetivos a nível mensal permitiu aos decisores uma maior adaptabilidade à realidade em meses de bastante incerteza.

3. **Modelo de Distribuição de Objetivos: Crédito Habitação:** o projeto de maior importância ao longo do estágio foi a realização de um modelo preditivo para a distribuição de objetivos da campanha de Crédito Habitação com recurso a técnicas de *Data Mining* e *Machine Learning* e utilizando-se várias fontes de informação quer internas como externas. Este projeto teve duas fases que serão explicadas com detalhe na subsecção seguinte.

    3.1. 1ª Fase do Modelo;
    3.2. 2º Fase do Modelo;
    3.3. Apresentação aos decisores;

### 3.4. Simplificação do Modelo e nova apresentação aos decisores.

4. **Ponderador Geográfico**: Em variadas campanhas o modelo de distribuição de objetivos inclui um ponderador geográfico atribuído a cada localidade ou cidade do país. Este projeto teve como propósito atualizar o ponderador existente, que possuía apenas três níveis, dotando o ponderador de uma maior granularidade. A classificação teve por base variados critérios tais como: população residente na localidade ou concelho, número de sucursais no mesmo lugar, índice de poder de compra do município, entre outros. O resultado final foi a segmentação das sucursais do país em seis grupos de acordo com critérios demográficos: Lisboa/Porto, Grande Cidade, Cidade, Pequena Cidade, Vila e Pequena Vila, que foram posteriormente combinados com cinco classes de acordo com critérios económicos. A cada classe atribuiu-se um ponderador entre 1,2 e 0,7, o ponderador final de cada sucursal resulta de uma combinação linear entre os ponderadores das duas classificações a ela atribuídos, podendo ser incrementado de acordo com a variação de população estrangeira no concelho. Este ponderador é utilizado em campanhas como a Captação de Clientes e Captação de Ordenados.

5. **Reporting e Análises para apoio à Decisão**: ao terminar cada ano existe uma revisão dos modelos de distribuição de objetivos para a generalidade das campanhas de forma a adequar os métodos de contabilização às necessidades e prioridades do ano seguinte e a adequar a distribuição ao comportamento que a Rede tem. Para tal é necessário extrair, transformar, analisar e apresentar informação aos decisores da melhor forma possivel. Usualmente estas análises incluem a simulação da distribuição de objetivos com diferentes cenários de forma a ser avaliado o impacto das alterações. Nestes casos a extração e modelação dos dados foi efetuada com recurso a SAS e SQL. A informação fornecida impactou em várias campanhas como por exemplo a Variação da Carteira de Crédito, onde se alterou por completo o modelo utilizado para a distribuição de objetivos.

6. **Mapa das Consistências**: para o Banco é importante não só o cumprimento dos objetivos ao final de cada ciclo ou de cada ano, mas também que exista um trabalho gradual ao longo dos períodos, que exista consistência. A consistência consiste na realização de um terço do objetivo do ciclo em cada um dos meses que o constituem. Esta variável é avaliada em parte das campanhas dado que é mais valioso que se cumpra um terço do objetivo a cada mês do ciclo do que cumpri-lo apenas nos últimos 15 dias do trimestre. Desta forma, a consistência já era uma parte integrante dos mapas de acompanhamento de algumas campanhas. Neste momento sentiu-se necessidade da existência de uma ferramenta que permitisse o acompanhamento das consistências para todas as campanhas simultaneamente. Criando-se então o Mapa de Consistências, Figura 10. Neste projeto, realizou-se toda a agregação de dados e a sua transformação criando-se um programa de SAS que alimenta a atualização do mapa no portal interno semanalmente.

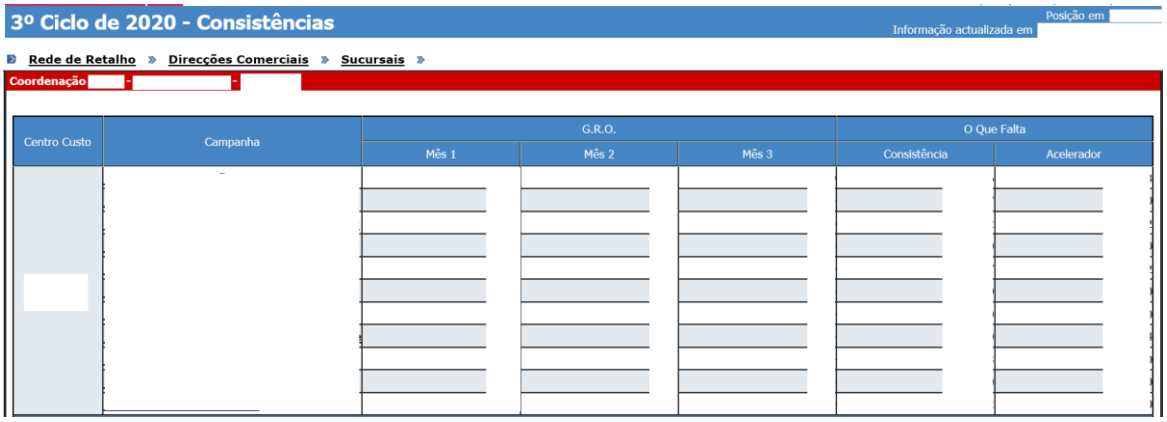



Figura 10. Mapa das Consistências

7. ***Dashboard* em Power BI: Ativação Digital de Novos Clientes:** Parte do *dashboard* pode ser consultado no Anexo I. Uma das prioridades para o MBCP é a utilização, por parte dos seus clientes dos canais digitais de comunicação com o banco. Com a situação de pandemia vivida na altura a comunicação com os clientes apoiada nos canais digitais recebeu uma importância acrescida. Assim, nasceu a necessidade de acompanhar a ativação de digital dos novos clientes a nível regular. Perceber quais os clientes que estão a ativar os canais digitais de comunicação com o banco, *app* e *site*, nos primeiros 30 dias desde que abriram conta para poder agir e melhorar a taxa de penetração nos canais digitais. Para tal, foi realizada a recolha de dados, a preparação e transformação dos dados, e a construção do relatório em Microsoft Power BI. O relatório permitiu aos decisores perceber quais as Direções Comerciais onde os canais digitais têm uma maior expressão, qual o ritmo de crescimento de novos clientes em situação de pandemia, quais as faixas etárias com maior adesão à *app* e ao *site* entre outros *insights* relevantes.

8. **Apoio na Gestão de Ajustes de Objetivos:** ao longo dos ciclos comerciais existe um conjunto variado de situações que requerem ajustes nos objetivos dos CC. Estes ajustes devem-se a mudanças de pessoal, ausências de gestores, a encerramentos das sucursais para obras, à existência de sucursais com períodos de funcionamento menor, entre outras situações. Com a situação de pandemia vivida a partir de março de 2020 estas situações aumentaram significativamente sendo necessário auxiliar a equipa no tratamento destes pedidos. De forma a realizar os ajustes recorre-se ao Microsoft Access e SQL para a manipulação dos dados nas bases de dados. Adicionalmente, ao iniciar cada ciclo é necessário carregar os objetivos nas bases de dados e realizar um conjunto de procedimentos que permite à hierarquia nos primeiros níveis (Coordenação) e segundos níveis (Diretores Comerciais) transferir objetivos entre CC dentro das suas direções para melhor ajustar os modelos à realidade do negócio local.

9. **Semanas de Formação:** Ao longo do estágio o Millennium bcp proporcionou aos seus estagiários um conjunto alargado de formações em diversas áreas quer para o aprofundamento de conhecimentos do setor bancário, através de seminários com diferentes departamentos. Quer diversas formações de caráter técnico em ferramentas como o SAS e Power BI. Através deste programa de formação realizou-se um projeto com uma equipa multidisciplinar orientado pela Direção Banca Direta para o estudo da criação de um assistente virtual

Em suma, ao longo do estágio, apesar do foco principal ser o projeto de Distribuição de Objetivos para a campanha de Crédito Habitação, foram realizadas diferentes tarefas de extração e manipulação e de análise de dados com recurso a diversos *softwares*.

## 5.2. CONTEXTO DO PROJETO

O Crédito Habitação trata-se de uma campanha de bastante importância para o Banco. A contratação de crédito habitação traduz-se numa relação banco-cliente de longo prazo sendo uma oportunidade de fidelizar um cliente ao Banco. Adicionalmente, sendo um processo maioritariamente presencial e demorado permite ao agente comercial conhecer o seu cliente e realizar outras ações de venda tais como cartões, seguros, soluções integradas, entre outros.

A campanha está presente na Matriz de Campanhas da generalidade dos ciclos apenas nas plataformas de clientes particulares: MM e GPP.

O projeto teve duas fases distintas em que se utilizaram abordagens diferentes devido à disponibilidade dos dados existentes em cada fase. Ao longo das secções seguintes serão explicadas as diferenças entre as duas fases e todo o processo de criação do modelo na segunda fase, desde a recolha de dados à sua implementação.

## 5.3. O MODELO ANTERIOR

O processo de distribuição de objetivos inicia-se com a definição de um objetivo global para cada plataforma que será posteriormente distribuído pelos diferentes CC através de um modelo definido no início de cada ano civil. A cada trimestre o modelo é aplicado aos CC que estão ativos no ciclo. Para todas as campanhas existe um conjunto de regras para a distribuição dos objetivos que estão de acordo com as regras de contabilização das vendas efetuadas.

O modelo de distribuição dos objetivos de Crédito Habitação possuía quatro variáveis ($v_1, v_2, v_3, v_4$) internas sobre aspetos relevantes do produto tais como o histórico de venda e os clientes elegíveis. Para cada uma das varáveis atribuía-se uma percentagem do objetivo, $p_j$, sendo essa percentagem do objetivo total distribuído proporcionalmente pela plataforma através dessa variável. Na Equação (5) e na Tabela 2 ilustra-se a fórmula utilizada para calcular o objetivo de cada CC, $o_i$, através das quatro variáveis e das respetivas percentagens definidas no modelo. Sendo $obj$ o objetivo global definido para a plataforma do CC $i$. Apesar do modelo ser comum às duas plataformas, este é aplicado a cada uma separadamente, sendo por isso as proporções calculadas dentro da plataforma e não do total da rede.

$$z_i = \frac{v_{1i}}{\sum v_1}\, p_1 * obj + \frac{v_{2i}}{\sum v_2}\, p_2 * obj + \frac{v_{3i}}{\sum v_3}\, p_3 * obj + \frac{v_{4i}}{\sum v_4}\, p_4 * obj \qquad (5)$$

Aplicando o modelo de distribuição de objetivos num cenário de apenas três sucursais MM, tendo como objetivo total da plataforma 10.000€, os objetivos das sucursais CC1, CC2 e CC3 seriam os apresentados na Tabela 3. Na Equação (6), exemplifica-se a aplicação da fórmula para o cálculo do objetivo correspondente à sucursal CC1.

Tabela 2. Exemplificação Teórica da Aplicação do Modelo Anterior de Distribuição de Objetivos de Crédito Habitação

| CC | Var1 | Var2 | Var3 | Var4 | Objetivo |
|---|---|---|---|---|---|
| CC1 | $v_{11}$ | $v_{21}$ | $v_{31}$ | $v_{41}$ | $o_1$ |
| CC2 | $v_{12}$ | $v_{22}$ | $v_{32}$ | $v_{42}$ | $o_2$ |
| CC3 | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| CCi | $v_{1i}$ | $v_{2i}$ | $v_{3i}$ | $v_{4i}$ | $o_i$ |
| Total MM | $\sum v_1$ | $\sum v_2$ | $\sum v_3$ | $\sum v_4$ | $obj$ |

$$z_1 = \frac{100}{300} * 0{,}3 * 10000 + \frac{20}{45} * 0{,}2 * 10000 + \frac{300}{700} * 0{,}4 * 10000 + \frac{400}{1050} * 0{,}1 * 10000 \quad (6)$$

Tabela 3. Exemplificação Prática da Aplicação do Modelo Anterior de Distribuição de Objetivos de Crédito Habitação

| Sucursal | Var1 | Var2 | Var3 | Var4 | Objetivo |
|---|---|---|---|---|---|
| CC1 | 100 | 20 | 300 | 400 | 3984 |
| CC2 | 80 | 10 | 150 | 250 | 2340 |
| CC3 | 120 | 15 | 250 | 400 | 3676 |
| Total MM | 300 | 45 | 700 | 1050 | 10000 |

De forma a uniformizar o esforço pedido à rede, após o cálculo dos objetivos para cada um dos CC são aplicados valores máximos e mínimos dependendo do tipo de centro de custo e de plataforma. Por exemplo, um gestor Prestige que seja simultaneamente gestor e diretor de sucursal terá valores máximos de objetivo inferiores a um gestor Prestige que não acumule funções.

O modelo é aplicado ao início de cada ciclo, ou seja, distribuem-se os objetivos quatro vezes ao ano. Apesar de a maioria dos clientes permanecer alocado ao mesmo centro de custo ao longo do ano, existem bastantes mudanças na constituição da Rede de Retalho quer em número de balcões quer em número de gestores. O que provoca mudanças no potencial das carteiras de clientes de cada CC para cada produto na Matriz de Campanhas.

### 5.4. A VARIÁVEL *TARGET*

A criação deste modelo teve como intuito ajustar ao máximo o objetivo de cada CC ao seu potencial de carteira e ao potencial do mercado onde se insere. O modelo deve distribuir os objetivos de forma justa de forma a maximizar os efeitos positivos da definição de objetivos nas organizações discutidos no capítulo 3.1. Definição de Objetivos nas Organizações.

Ao longo do tempo o objetivo global da Rede varia em montante e, por isso, também o grau de exigência pedido aos colaboradores varia. Além disso, ao longo dos ciclos as regras de contabilização de objetivo e de majoração de operações com maior valor acrescentado mudam. A majoração de operações consiste em contabilizar uma operação a mais de 100% se esta cumprir algum critério

estratégico como por exemplo a utilização da aplicação do Millennium bcp por parte do cliente. Desta forma, não é possível comparar valores absolutos de realizado em diferentes ciclos, a mesma operação pode contabilizar com montantes diferentes em ciclos distintos.

Em conclusão, o *target* precisaria de ser uma variável comparável ao longo de diferentes ciclos: uma proporção do realizado de cada centro de custo na sua plataforma. Com o modelo estimar-se-á qual o contributo de um centro de custo específico para o global da sua plataforma.

### 5.4.1. A Primeira Fase

Numa primeira fase do projeto a disponibilidade de dados internos era reduzida. Por um lado, tinha-se ao dispor as informações sobre objetivos e realizados para variados ciclos, uma vez que o *owner* desses dados é a própria equipa. Por outro lado, faltavam informações sobre clientes e constituição das carteiras das sucursais ou gestores em cada ciclo anterior. Não era possível saber qual o grau de risco do cliente, qual a sua sucursal ou carteira, idade ou outras informações para diferentes períodos de tempo. O que significava não conseguir saber as variáveis independentes que levaram à obtenção de um resultado *target* específico.

Assim, a informação disponível permitia apenas a utilização de um ciclo para a modelação. Uma vez que o montante de crédito realizado está dependente do comportamento dos colaboradores, sabia-se que utilizar apenas os dados de um ciclo representava assumir que aquele ciclo era representativo do comportamento de todos os elementos da rede, o que não se consegue afirmar. Para minimizar o efeito de esforço dos colaboradores num ciclo específico em detrimento de outro, a variável *target* foi criada incluindo o comportamento dos CC ao longo de um ano, como demonstra a Equação (7). Para cada CC ($i$), definiu-se o *target* ($y_i$) como a proporção média dos montantes realizados ($r_{i,c}$) ao longo dos últimos quatro ciclos ($c$). Sendo $P$ a plataforma a que pertence o CC.

$$y_i = \frac{\frac{r_{i,201803}}{r_{P,201803}} + \frac{r_{i,201804}}{r_{P,201804}} \frac{r_{i,201901}}{r_{P,201901}} + \frac{r_{i,201902}}{r_{P,201902}}}{4} \quad (7)$$

Assim, existia um *target* composto com informação de quatro ciclos e as variáveis independentes datavam do ciclo 201902[5].

Em conclusão, a dispoibilidade dos dados conduziu à criação de uma variável *target* composta que não era representativa da realidade à data de recolha das variáveis de *input*. Adicionalmente, o número de registos para o treino e validação do modelo era reduzido sendo apenas um trimestre.

### 5.4.2. A Segunda Fase

Alguns meses após o início do estágio foi possível obter as informações de clientes com as datas de início de ciclo desde o segundo ciclo de 2018 (201802) o que permitiu reorganizar a informação e restruturar a abordagem ao problema. Possuía-se então a constituição das carteiras, o número de cliente elegíveis, entre outros dados, à data do início de sete ciclos distintos.

[5] Código indicativo do ciclo em questão no formato AAAATT, isto é, o ano seguido do número do trimestre do ciclo. Neste caso 2º trimestre de 2019.

A variável *target* foi reformulada, existindo um maior número de ciclos não foi necessário criar uma variável composta. Simplificou-se o *target* e assistiu-se a um incremento substancial no número de registos para a modelação.

Conforme demonstra a Equação (8), calculou-se o *target* como uma proporção do realizado ($r_{i,c}$) do CC ($i$) no total realizado da sua plataforma ($P$) para cada ciclo ($c$) em análise.

$$y_{i,c} = \frac{r_{i,c}}{r_{P,c}} \tag{8}$$

As subsecções seguintes vão abordar as próximas fases do projeto tendo em conta o *target* definido nesta segunda fase, uma vez que foi a abordagem utilizada no modelo final.

### 5.5. Recolha, Tratamento e Exploração de Dados

A recolha e tratamento dos dados foi efetuada com recurso ao SAS por ser a ferramenta com integração de todas as fontes de dados internas necessárias, desde dados de clientes a dados de sucursais bem como aos valores de objetivos históricos. Nesta subsecção apresenta-se quais as variáveis recolhidas bem como as respetivas fontes de informação.

A recolha de informação foi realizada em quatro etapas: a recolha de variáveis internas para cada CC em cada ciclo em análise; a recolha da informação necessária para a criação do *target* bem como quais os objetivos definidos pelo modelo anterior para comparação posterior; a recolha de variáveis externas; e, por último, a integração de todas as fontes.

A recolha de variáveis internas teve por base as variáveis utilizadas anteriormente nos cálculos bem como informações relevantes provenientes do Modelo de Propensão à Compra de Crédito Habitação realizado e atualizado pelo departamento de *Costumer Relationship Managment* (CRM), da Direção de Marketing do Retalho. Este modelo calcula, mensalmente, a probabilidade estimada de cada cliente vir a adquirir um Crédito Habitação.

A informação disponível sobre clientes datava do ciclo 201802 até ao presente, desta forma foram recolhidas as variáveis para estes sete ciclos comerciais. De forma a recolher a informação estruturada por ciclo e de forma o mais autónoma possível foram criados um conjunto de macros que alimentadas com a variável ciclo devolviam as variáveis pretendidas para cada CC ativo naquele período. Obtendo-se assim sete tabelas com as variáveis sobre a carteira de Crédito Habitação e características da carteira de clientes que podem ser consultadas na Tabela 4.

Na recolha de variáveis referentes a clientes foram apenas contabilizados os clientes elegíveis para um Crédito Habitação de acordo com as regras aplicadas no modelo anterior. Sendo assim, apenas se contabilizou clientes entre 25 e 55 anos, que não possuem montantes em contencioso, nem em insolvência, nem crédito vencido, cujo grau de risco é menor que 10 (escala de 1 a 13) e que têm um valor de plafom hipotecário superior a 250 €.

Tabela 4. Variáveis recolhidas em fontes internas

| Variável | Descrição | Papel |
|---|---|---|
| id | Concatenação do CC e codciclo | Chave Primária |
| cc | Código identificativo de cada Centro de custo | Id |
| codciclo | Código identificativo do ciclo em questão | Id |
| Localidade | Nome da localidade onde se localiza a sucursal | Id |
| Município | Nome do município onde se localiza a sucursal | Id |
| Num_balcao | Número identificativo da sucursal à qual pertence o CC | Id |
| tiposuc | Código da plataforma à qual pertence o CC | Id |
| avg_idad | Idade média dos clientes elegíveis | *Input* |
| avg_score | Score de propensão médio dos clientes elegíveis | *Input* |
| avg_triad | Plafom Hipotecário[6] médio dos clientes elegíveis | *Input* |
| cartNrCh | Número de operações de Crédito Habitação na carteira de crédito do CC | *Input* |
| cartVolCh | Volume da carteira de Crédito Habitação do CC | *Input* |
| histNrCh | Número de operações de Crédito Habitação realizados no ano civil anterior pelo CC | *Input* |
| histNrChDir | Número de operações de Crédito Habitação contabilizadas para a realização do objetivo da Direção Comercial na plataforma do CC, no ano civil anterior | *Input* |
| histVolCh | Volume de Crédito Habitação contabilizado para a realização do objetivo do CC no ano civil anterior | *Input* |
| histVolCHDir | Volume de Crédito Habitação contabilizado para a realização do objetivo da Direção Comercial na plataforma do CC, no ano civil anterior | *Input* |
| n_25_43 | Total de clientes elegíveis com idades entre 25 e 43 anos (inclusive) do CC | *Input* |
| n_alto_sup | Total de clientes elegíveis com score "alto" e "superior" no modelo de Propensão à Compra do CRM | *Input* |
| ncli_eleg | Total de clientes elegíveis do CC, sendo que cada cliente REX é contabilizado apenas como 0,9 cliente. (critério utilizado anteriormente) | *Input* |

[6] **Plafom Hipotecário**: valor estimado que o cliente consegue suportar mensalmente como prestação do seu Crédito Habitação.

| Variável | Descrição | Papel |
|---|---|---|
| ncolab | Número de colaboradores do CC (valor sempre 1 em CC de GPP) | *Input* |
| p_25_43 | Percentagem de clientes elegíveis com idades entre 25 e 43 (inclusive) no CC | *Input* |
| p_alto_sup | Percentagem de clientes elegíveis com score "alto" e "superior" no modelo de Propensão à Compra do CRM (n_alto_sup/totclieleg) | *Input* |
| totclieleg | Total de clientes elegíveis para um CH do CC | *Input* |
| valmed | Valor médio das operações realizadas no CC no ano civil anterior (histVolCh/histNrCh) | *Input* |
| valmeddir | Valor médio das operações realizadas nas sucursais da mesma plataforma e na Direção Comercial do CC, no ano civil anterior (histVolChDir/histNrChDir) | *Input* |
| respsuc | Indicativo se o CC associado é diretor de sucursal (apenas para GPP) | *Input* |

Recolhidas as variáveis internas consideradas relevantes iniciou-se o cálculo da variável *target* para cada CC em cada ciclo. Ao realizar este processo recolheu-se ainda informação de quais os valores de objetivo definidos pelo modelo em vigor em cada ciclo anterior e de qual o objetivo reafectado por parte do Coordenador com o objetivo de posteriormente avaliar o comportamento dos diferentes modelos e poder compará-los. Um descritivo das variáveis pode ser consultado na Tabela 5.

O mercado de Crédito Habitação envolve vários bancos e não representa em todos os locais uma proporção clara da cota de mercado existente por cada instituição financeira. De forma a incorporar a dinâmica do mercado como um todo e não apenas do mercado interno do MBCP foram recolhidas variáveis externas através do INE e PORDATA.

A recolha de variáveis de fontes externas pode apresentar algumas desvantagens, das quais se destacam a atualização da informação depender de terceiros e não ser tão frequente como desejado; a granularidade dos dados ser as divisões administrativas nos seus vários níveis (freguesias, municípios, NUT) e não ao nível do CC; e o período de observação das variáveis, na maioria dos casos, não ser trimestral, mas sim anual.

Por outro lado, existe a vontade de incorporar informação de mercado nos modelos de distribuição de objetivos, com o intuito de olhar para o mercado pelo seu potencial total e não apenas para a cota de clientes do Banco naquele local. Adicionalmente, pretende-se uniformizar a distribuição de objetivos para CC na mesma cidade ou localidade uma vez que partilham o mesmo mercado. Esta uniformização nem sempre é conseguida com os critérios utilizados no modelo anterior uma vez que incluem, maioritariamente, variáveis de com granularidade ao nível do CC. Utilizando variáveis ao nível do CC podemos ter situações em que duas sucursais A e B geograficamente próximas têm objetivos significativamente distintos, uma vez que, as carteiras de clientes das sucursais não seguem distribuições normais e apesar de partilharem o mesmo mercado a carteira da sucursal A pode ter significativamente mais potencial que a carteira da sucursal B, por exemplo.

As variáveis recolhidas através de fontes externas estão descritas na Tabela 6 e são relativas a informações demográficas e relacionadas com o mercado habitação em cada município.

Tabela 5. Variáveis recolhidas para futura comparação dos Modelos

| Variável | Descrição | Papel |
|---|---|---|
| Id | Concatenação do CC e codciclo | Chave Primária |
| y | Proporção de realizado no realizado total da plataforma num determinado ciclo | *Target* |
| ObjMapa | Objetivo de cada CC no final do ciclo comercial que ficou visível no mapa de resultados | Futura comparação dos modelos |
| ObjOriginal | Objetivo de cada CC definido pelo modelo anterior | Futura comparação dos modelos |
| ObjCoordenador | Objetivo de cada CC após a reafectação de objetivos realizada pelos coordenadores[7] | Futura comparação dos modelos |
| Realizado | Montante de Crédito Habitação contabilizado para a realização da campanha | Futura comparação dos modelos |

Tabela 6. Variáveis recolhidas de fontes externas

| Variável | Descrição | Fonte | Granularidade | Papel |
|---|---|---|---|---|
| codmun | Código oficial do município | - | - | Chave Primária |
| mun | Nome do município | - | - | Id |
| nut_ii | Código identificativo da NUT II | - | - | Id |
| nut_iii | Código identificativo da NUT III | - | - | Id |
| cidade | Classificação se um lugar é oficialmente uma cidade ou não. | INE | | *Input* |
| crdhab | O crédito à habitação inclui os empréstimos bancários concedidos às famílias para comprar casas novas e usadas, para comprar o terreno, fazer obras ou construir a habitação própria; anual | PORDATA | Município | *Input* |
| denspop | Densidade populacional (N.º/ km²); anual | INE | Município | *Input* |

[7] **Coordenador**: Responsável por uma Direção Coordenação da Rede de Retalho

| Variável | Descrição | Fonte | Granularidade | Papel |
|---|---|---|---|---|
| ganho | Ganho médio mensal[8] (€); anual | INE | Município | *Input* |
| nraloj | Nº médio de alojamentos familiares por Km²; anual | PORDATA | Município | *Input* |
| pc | Poder de compra per capita por; bienal | INE | Município | *Input* |
| pib | Produto interno bruto (B.1*g) por habitante a preços correntes (Base 2016 - €); anual | INE | NUT II | *Input* |
| pop | População residente (Nº); anual | INE | Município | *Input* |
| purb | Valor médio dos prédios urbanos[9] transacionados em Portugal e referentes a prédios situados em território nacional; anual | PORDATA | Município | *Input* |
| sm | O saldo migratório é a diferença entre o número de pessoas que imigram e o número de pessoas que emigram; anual | PORDATA | Município | *Input* |
| valorbanca | Valores médios de avaliação bancária dos alojamentos por m² (€); anual | PORDATA | Município | *Input* |
| valortrimestre | Valor mediano das vendas por m2 de alojamentos familiares (€); trimestral | INE | Município | *Input* |

Estando recolhidas todas as variáveis foi necessário integrar as diferentes fontes de dados. Uma vez que a ligação entre os dados internos e externos era o nome do Município este processo requereu algum tratamento de texto. Tendo sido desenvolvido um *script* que permite a conversão dos nomes dos municípios em versão oficial para os nomes utilizados internamente.

Antes de proceder para a exploração e modelação efetuou-se uma análise à qualidade dos dados, procurando por incoerências e lacunas na informação. As regras de incoerência foram aplicadas a todos os CC independentemente de qual a sua plataforma. De seguida são listadas as regras aplicadas e a motivação por detrás de cada uma.

1. Eliminaram-se registos onde o CC encerrou durante o ciclo, uma vez que nestas situações aquelas sucursais não dispuseram do mesmo período de tempo para atingir o objetivo das restantes sucursais. O seu resultado foi condicionado pelo tempo disponível. (77 observações)

[8] **Ganho Médio Mensal**: Montante ilíquido em dinheiro e/ou géneros, pago ao trabalhador, com carácter regular em relação ao período de referência, por tempo trabalhado ou trabalho fornecido no período normal e extraordinário. Inclui, ainda, o pagamento de horas remuneradas, mas não efetuadas (férias, feriados e outras ausências pagas).

[9] **Prédio Urbano**: Os prédios urbanos são propriedades que incluem os alojamentos das famílias e os terrenos para construção, as construções e os edifícios que se destinam a habitação, comércio, indústria ou serviços.

2. Eliminaram-se registos com *target,* mas que o CC tinha encerrado antes daquele ciclo. Em raros casos a atribuição de vendas aos CC não é revista e pode ser atribuída uma operação a um CC que já encerrou. (28 observações)

3. Eliminaram-se registos que possuíam um **ObjMapa** igual a 0 ou está omisso. No caso da plataforma GPP, se o gestor estiver ausente mais de metade do ciclo não se elimina o seu CC, mas elimina-se os objetivos desse ciclo. Desta forma, CC com objetivo no Mapa de 0 representam CC que apenas trabalharam 50% do tempo para aquele realizado. (23 observações) E CC com objetivo no Mapa omisso representam CC que não estiveram ativos durante o ciclo. (168 observações)

4. Eliminaram-se registos para os quais o valor do **ObjOriginal** era nulo, uma vez que representam CC que não estavam ativos quando o ciclo iniciou. Mais uma vez são CC cujo resultado foi condicionado por um menor tempo disponível. (84 observações)

5. Eliminaram-se registos para os quais se registou valores de **histNrCh** maiores que 0 mas que não existiam valores na **cartNrCh**. Uma vez que um CC não pode possuir histórico de ter realizado Crédito Habitação, mas não possuir operações em carteira. O mesmo é verdade para o par **histVolCh** e **cartVolCh**. (3 observações)

6. Eliminaram-se registos onde a **cartVolCH** ou o **histVolCH** possuíam valores maiores que 0, mas eram 0 na **cartNrCh** ou **histNrCH**, respetivamente. Não se pode possuir volume numa carteira de Crédito sem ter realizado pelo menos uma operação. (0 observações)

Possuindo um *dataset* com a amostra completa dividiu-se o *dataset* em dados de treino e teste. De forma a permitir comparação entre o novo modelo a ser criado e o modelo anterior seria necessário poder distribuir o objetivo pelas plataformas nas mesmas condições que foi aplicado anteriormente. Posto isto, os dados de teste teriam de representar a estrutura da rede num determinado ciclo, pelo que se escolheu o ciclo mais recente com informação para testar o modelo, guardou-se para dados de teste todas as ocorrências referentes ao ciclo 201904.

À *priori* existia o pressuposto de modelar separadamente os dados das duas plataformas por possuírem características bastante diferentes, tanto a nível de modelo de negócio com a nível da constituição das suas carteiras de clientes. Apesar de o Modelo Anterior ser comum às duas plataformas, era aplicado em separado por dois motivos: 1) os objetivos globais da plataforma são definidos separadamente; 2) baseando-se em proporções juntar as duas plataformas distorceria os resultados. Desta forma, iniciou-se a exploração dos dados procurando evidência que justificasse, ou não, este pressuposto.

Na Figura 11, ilustra-se a distribuição das variáveis **cartvol** e **potencial** comparativamente ao *target* e distinguindo-se a plataforma a que pertence a observação. Através da análise dos *scatterplots* pode-se perceber, não só, que a distribuição das variáveis nas duas plataformas é distinta como também a sua relação com o *target* é diferente. Analisando, por exemplo, o coeficiente de correlação do *target* com a variável **cartvol** observa-se no MM uma correlação de 0,59 e no GPP uma correlação de 0,42. Para comprovar a diferente distribuição das variáveis nos diferentes segmentos realizaram-se testes *t student* às médias de diferentes variáveis (**cartvol**, **potencial**, **totclieleg**, entre outras) tendo se sempre rejeitado a hipótese nula com níveis de significância acima dos 95%. Ou seja, existe evidencia

estatística com elevados níveis de significância de que as distribuições das variáveis não são iguais nas duas plataformas.

Adicionalmente, existe o conhecimento de que o comportamento dos CC na plataforma MM é mais consistente do que o comportamento dos gestores Prestige. Isto é, sucursais MM ao longo do tempo têm resultados de Crédito Habitação semelhantes e estáveis, enquanto os gestores, em muitos casos existem ciclos muito bons e depois ciclos onde os resultados são nulos existindo uma maior volatilidade no comportamento. Esta diferença deve-se à composição das carteiras dos gestores ser de tamanho mais reduzido e menos rotativa o que leva a taxas de penetração do produto maiores nas carteiras dos gestores quando comparadas com carteiras de MM. Para além disto, os clientes Prestige têm médias de idade mais avançadas sendo um segmento de clientes que tem menos propensão à contratação de Crédito Habitação.

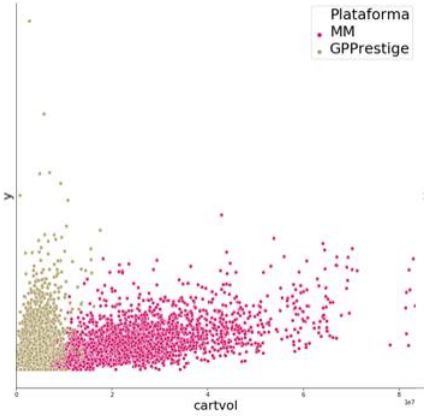



Figura 11 (a)

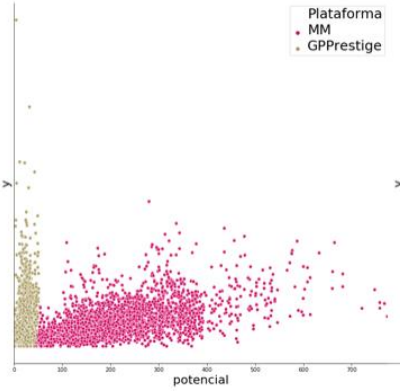



Figura 11 (b)

Figura 11. Distribuições das variáveis **cartvol** (a) e **potencial** (b) com o *target* por plataforma

Em suma, dada as distribuições distintas dos dados e os motivos derivados do funcionamento do negócio seguiu-se a exploração e modelação dos dados das duas plataformas em separado começando pelo MM.

Iniciou-se a exploração de dados do MM realizando a partição dos dados em dados de treino e de validação utilizando o nó **Data Partition** dos SAS Miner em 70 – 30, ou seja, 70% das observações ficaram para treino e 30% para validação. A partição dos dados foi realizada através de uma amostra aleatória. Após a partição dos dados, possuía-se um *dataset* de treino com 1894 registos e 29 variáveis.

Um dos intuitos da exploração dos dados é a identificação de possíveis *outliers*. No âmbito do projeto optou-se por realizar uma análise uni-variada às variáveis em estudo e consoante a sua distribuição identificar os pontos mais extremos com recurso ao nó **Filter** do SAS Miner. Na Figura 12 ilustra-se o processo de identificação de *outliers* na variável **p_alto_sup.** Este método consiste em analisar a distribuição dos dados e excluir os valores isolados desde que estes não contenham um elevado número de observações. A importância de avaliar cada variável individualmente ao invés de aplicar um intervalo pré-definido através da média ou do intervalo interquartil ficou visível ao avaliar as variáveis com granularidade municipal. Em muitos casos, como no **pc**, Figura 13, existiam valores extremos que seriam excluídos por esses limites. Contudo estes valores não podem ser considerados *outliers* uma vez que possuem um elevado número de observações e representam, na generalidade, os municípios de Lisboa e Porto. No total eliminaram-se 1,40% das observações.

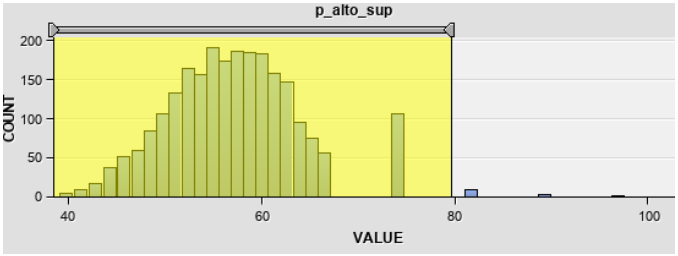



Figura 12. MM – Identificação de *Outliers*: Percentagem de Clientes com Score Alto e Superior

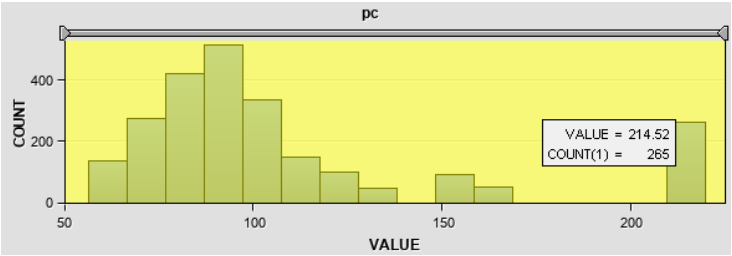



Figura 13. MM – Identificação de *Outliers*: Poder de Compra

Posteriormente, prosseguiu-se para o tratamento dos valores omissos nos dados. Na Tabela 7. Podemos observar que apenas três das variáveis em estudo possuíam valores omissos. Uma abordagem possível para esta situação poderia ser eliminar os registos com valores omissos, no entanto, tratando-se de variáveis com granularidade ao nível do município tal ação significaria eliminar um dos municípios da análise pelo menos para determinado período de tempo. Em adição, sendo informação externa é espectável que existam valores omissos em nos dados futuros onde se aplicará o modelo pelo que é importante decidir como os preencher.

Outra abordagem simples seria eliminar as variáveis onde existem valores omissos excluindo-as da análise, perdendo assim algum potencial preditivo que essa variável traga à análise. Esta abordagem foi utilizada para a variável **valorbanca** por possuir uma percentagem de valores omissos superior a 2% do *dataset* de treino. Para além disso, esta variável possuía coeficientes de correlação elevados com outras variáveis tais como **valortrimestre**: 0,975, **purb**: 0,933 entre outras, o que permitia manter a maior parte do seu poder explicativo na análise.

Tabela 7. MM – Valores Omissos

| Variável | Número de Registos Omissos | Percentagem do total de registos |
|---|---|---|
| crdhab | 6 | 0,32% |
| valorbanca | 42 | 2,22% |
| valortrimestre | 1 | 0,05% |

Para as restantes variáveis optou-se por fazer a imputação dos valores omissos tendo sido testados diferentes métodos de imputação para cada uma das variáveis: **crdhab** e **valortrimestre**. Os métodos considerados foram a substituição do valor pela média do *dataset*, substituição pela mediana e a imputação através do algoritmo kNN *(k-Nearest Neighbor*).

A implementação do kNN foi realizada através do **PROC KRIDGE2D** no SAS que foi desenhado para apenas receber duas variáveis de *input* trabalhando num espaço bidimensional. Desta forma diversas combinações de variáveis foram testadas de forma a escolher o par de variáveis que produzia resultados mais precisos.

Para a avaliação dos métodos foram utilizadas 35% das observações, escolhidas de forma aleatória, tendo-se substituído o seu valor pelo valor previsto através do método. No final, os diferentes métodos foram comparados através do seu erro absoluto máximo (MxAE) e do seu erro ao quadrado médio (MSE). Os métodos escolhidos foram distintos para as duas variáveis em questão: para o **valortrimestre** escolheu-se o kNN utilizando as variáveis **ganho** e **purb**, por outro lado na variável **crdhab** o melhor método foi a substituição pela mediana.

Terminada a fase de exploração no MM, repetiram-se os procedimentos para o *dataset* da plataforma Prestige. Após a separação em treino-validação nas proporções 70-30, possuíam-se 29 variáveis e 1903 registos de treino. Relativamente à identificação de *outliers* repetiu-se o processo referido anteriormente tendo-se eliminado 0,84% das observações. Para esta plataforma existia um menor número de valores omissos, apenas 21, e todos concentrados na mesma variável **valorbanca**. À semelhança da decisão tomada para o MM optou-se por eliminar a variável uma vez que possuía um coeficiente de correlação elevado com diversas outras variáveis tais como **valortrimestre**: 0,98 e **purb**: 0,94.

## 5.6. Seleção e Desenho de Variáveis

Estando completa a fase de exploração e tratamento dos dados procedeu-se para a seleção de variáveis. A seleção de variáveis pode trazer bastantes benefícios à modelação tais como a redução da complexidade do modelo, uma maior facilidade de interpretação dos resultados e um aumento da capacidade preditiva. Adicionalmente, com a seleção de variáveis os modelos tornam-se mais rápidos e eficientes. No sentido de potenciar o poder explicativo das variáveis, variáveis que mantenham uma correlação com o target devem ser incluídas, no entanto variáveis excessivamente correlacionadas entre si devem ser retiradas uma vez que trazem informação redundante ao modelo. [22, 23]

Inicialmente a seleção de variáveis foi efetuada tendo em conta a totalidade das variáveis recolhidas, no entanto, ao iniciar a segunda fase do projeto as variáveis de histórico (**histNrCh**, **histNrChDir**, **histVolCh**, **histVolCHDir**) foram eliminadas da análise por decisão de negócio. O princípio por de trás da eliminação destas quatro variáveis é tentar ao máximo atribuir objetivos pelo potencial da carteira e do mercado onde os CC se encontram e não pelos seus comportamentos anteriores, uma vez que esse critério tende a distribuir maior objetivo aos gestores que mais se esforçam por atingir os resultados. Sendo assim, os resultados da seleção de variáveis apresentados de seguida não incluem estas quatro variáveis na análise.

Para a seleção das variáveis neste projeto foram testados três métodos distintos de forma a perceber qual traria melhores resultados. O primeiro consiste em escolher as variáveis através do cálculo de um ranking de importância, o segundo consistiu na utilização do nó **Variable Selection** do SAS Miner e o terceiro método utilizado é um método de redução de dimensionalidade com recurso a *clustering* conseguido através do nó **Variable Clustering** do SAS Miner.

A criação de um ranking de importância das variáveis é conseguida através da ordenação da importância de cada variável de acordo com diferentes métodos de seleção implementados com recurso a diferentes bibliotecas de Python. De seguida introduz-se cada método utilizado que foi aplicado após uma normalização dos dados com recurso à função **MinMaxScaler**.

- **R-Quadrado:** Efetuou-se uma Regressão Linear simples apenas com um *input* e constante guardando-se para cada variável o seu valor de R-Quadrado. Atribuindo-se *rank* 1 à variável com maior valor de R-Quadrado.

- **Correlação com o *target*:** Calculou-se a correlação de cada variável com o *target* utilizando o coeficiente de Pearson para as variáveis intervalares e binárias e o coeficiente de Spearman para as variáveis ordinais (como a variável ciclo). Atribuindo-se *rank* 1 à variável com coeficiente de correlação com maior valor absoluto.

- ***Recursive Feature Elimination* (RFE) com Regressão Linear:** aplicado através de um método da biblioteca de Python Sklearn, o RFE permite treinar um modelo com todas as variáveis disponíveis e interactivamente retirar a variável que menos contribui para a estimação. O algoritmo para quando é atingido o número desejado de variáveis e devolve um *rank* que ordena as variáveis sendo o valor 1 atribuído à última variável a sair da modelação.

- ***Decision Tree Feature Importances*:** através da biblioteca Sklearn do Python é possível treinar uma Árvore de Decisão e posteriormente extrair a importância normalizada de cada variável para a construção do modelo. Os valores devolvidos pela propriedade *feature_importances_* são calculados como o decréscimo de erro de um nó ponderado pela probabilidade de atingir esse nó. A probabilidade de atingir um nó é a relação entre o número de observações que atingem esse nó e o número total de observações. No final, este método devolve um valor de importância para cada variável aos quais se atribuiu valores de *rank*: valor 1 para o maior valor de importância.

Cada método utilizado permite ordenar as variáveis da mais relevante para a menos relevante, resultando em quatro rankings diferentes. De forma a escolher as variáveis tendo por base os diferentes métodos foi necessário criar um proxy que juntasse todos os rankings num só. Para tal, calculou-se a média dos rankings para todas as variáveis obtendo-se o **RankFinal**.

Posteriormente, ordenando as variáveis de acordo com o seu **RankFinal** foram eliminadas as variáveis que estivessem altamente correlacionadas com uma variável de *rank* superior. Ou seja, para cada par de variáveis com um coeficiente de correlação igual ou superior a 0,8 manteve-se a variável de *rank* superior (maior importância) eliminando-se a outra de forma a eliminar problemas de multicolinearidade.

O segundo método aplicado foi o nó **Variable Selection**, do SAS Miner, com o método R- Quadrado. Este método seleciona as variáveis em dois passos: 1) computar o R-Quadrado entre o *target* e todas as variáveis de *input*; 2) com as variáveis que possuam um R-Quadrado superior ao limite (o valor pré-definido é 0,005) aplica-se um método de *foward selection* sequencial. No segundo passo, a primeira interação seleciona-se a variável com maior correlação com o *target*, de seguida, a cada iteração seleciona-se a variável que produz o maior incremento ao R-Quadrado do modelo sendo que se pode

definir o incremento mínimo desejado para a seleção de uma variável (o valor pré-definido é 0,0005). [24]

Por último, com recurso ao nó **Variable Clustering** foi testado um método de redução da dimensionalidade dos dados que permite incluir um maior número de variáveis e não eliminar variáveis correlacionadas entre si. Os clusters de variáveis são criados através de um processo divisivo e interativo, todas as variáveis iniciam agrupadas num só cluster sendo iterativamente escolhido um cluster para ser dividido. O cluster a ser dividido em cada interação é o cluster que tem a menor percentagem de variância explicada das variáveis (quando a propriedade **Variation Propotion** é selecionada). As variáveis são interactivamente transferidas entre clusters de forma a maximizar a variância explicada por cada cluster. O algoritmo para quando atinge o máximo número de clusters pré-definido ou quando todos os clusters cumprem os critérios de mínima variância explicada. Assim, com o nó **Variable Clustering** os clusters são todos criados com diferentes conjuntos de variáveis ao contrário do que acontece numa Análise de Componentes Principais. [25]

Ao aplicar uma análise de clusters às variáveis reduz-se a dimensionalidade dos dados agrupando as variáveis mais correlacionadas entre si e minimizando a correlação entre clusters. [25] Utilizando apenas variáveis intervalares e um mínimo de variância explicada pelo cluster de 0,85 em ambas as plataformas foram criados seis clusters com composições ligeiramente diferentes. A decisão do número de clusters a utilizar foi efetuada tendo por base a análise dos dendrogramas e a propriedade de mínima variância explicada. De seguida, aplicou-se um nó de **Variable Selection** ao conjunto de clusters criados, eliminando assim os que não cumprissem critérios de mínimo R-Quadrado com o *target*.

Concluindo, procedeu-se para a fase da modelação com quatro conjuntos de variáveis selecionadas por diferentes métodos para cada uma das plataformas.

A capacidade preditiva de uma variável pode ser incrementada se realizadas as transformações adequadas. O nó **Variable Transformation** pode efetuar transformações tanto a variáveis categóricas como a variáveis intervalares. Para o primeiro tipo de variável pode-se optar por criar variáveis *dummy* ou agrupar os grupos mais raros de classes numa única classe. Para variáveis intervalares pode-se escolher as transformações matemáticas simples tais como a raiz quadrada, o logaritmo ou a exponencial. Ou transformações mais complexas que maximizam a correlação com o *target* ou que maximizam a normalidade da distribuição. Por último, pode-se aplicar *optimal binning* que permite criar classes através da divisão da variável em diferentes intervalos. [24]

A transformação das variáveis iniciou-se pela transformação da variável *target* em que se aplicou uma transformação simples logarítmica em ambas as plataformas. Sendo que todas as transformações a variáveis dependentes foram aplicadas tendo por base ambas as versões da variável *target*: **y** e **log(y)**.

De seguida, utilizou-se três nós de **Variable Transformation** para testar diferentes métodos de transformação das variáveis, onde foram criadas variáveis *dummy* para a variável categórica ciclo, não tendo sido aplicada transformação às variáveis binárias. Relativamente às variáveis intervalares, no primeiro nó aplicou-se a transformação que maximiza a correlação com o *target*, no segundo as transformações que maximizam a normalidade dos dados e, no último, aplicou-se uma transformação logarítmica.

## 5.7. MODELAÇÃO

Estando realizadas as fases da pré-modelação – Recolha, Exploração, Transformação e Seleção – procedeu-se para a modelação com vista em prever a proporção de objetivo adequada a cada CC. Na fase de modelação era importante utilizar modelos que permitissem uma interpretação dos resultados de forma a explicar à Rede os critérios utilizados para a distribuição do objetivo, desta forma, os modelos testados foram Regressão Linear, Árvore de Decisão e *Gradient Boost*, explorados na secção 3.4 Aprendizagem Automática Supervisionada.

Os três modelos foram aplicados aos diversos conjuntos de variáveis que provieram da fase anterior: os clusters; clusters após aplicar o nó **Variable Selection**; variáveis provenientes do nó de **Variable Selection** sem transformação e com cada um dos três métodos de transformação; variáveis selecionadas através do ranking de importância sem transformação e com cada um dos três métodos de transformação. Todos estes *flows* foram criados para o *target* simples e para o *target* transformado em logaritmo.

Desta forma, todas as combinações testadas foram comparadas, através de um nó de **Model Comparison** que inclui no seu *output* as medidas MSE, RMSE, SSE e MxAE. Por predefinição para *targets* intervalares, o modelo escolhido é o modelo com menor MSE no *dataset* de validação. [26] Nesta fase do projeto escolheram-se os cinco melhores modelos em cada plataforma, baseado no valor do MSE e de MxAE no *dataset* de validação. Importante notar que a comparação entre modelos do *flow* com a variável *target* transformada e não transformada foi efetuada através do MxAE tendo-se transformado o erro para a escala original. Os modelos em que se utilizou a transformação logarítmica da variável **y** produziram melhores resultados na maioria das opções.

De seguida, com o intuito de adequar o melhor possível o algoritmo ao problema em mãos, sem fazer *overfitting* do modelo, foi aplicada a técnica de *grid search* (“procura em grelha”) aos parâmetros dos cinco melhores modelos. O *grid search* é utilizado para encontrar os valores para diferentes parâmetros dos modelos que produzem os resultados mais corretos. De forma a aplicar a técnica, cria-se uma grelha com os parâmetros que queremos avaliar e os valores para esses parâmetros e treina-se o modelo com todas as combinações de valores indicadas.

O *grid search* foi aplicado a diversos parâmetros dos modelos *Gradient Boost* como a *depth* máxima e o número de ramos a ser gerado por cada decisão, entre outros. Depois de treinados, os modelos são comparados escolhendo-se a combinação de valores que produzem melhores resultados. Esta comparação pode ser efetuada apenas através dos dados de validação ou de uma forma mais robusta utilizando-se *cross-validation* (validação cruzada).

A utilização de *cross-validation* prende-se com a divisão do *dataset* de treino num conjunto de partes, denominadas de *folds*, que serão utilizadas, à vez, como *dataset* de validação do modelo produzido através dos restantes dados. Os modelos são posteriormente avaliados através da sua prestação média nos diferentes *folds*. No caso de uma variável intervalar pode ser escolhido o modelo com melhor MSE médio. A aplicação desta técnica pode ter variadas formas desde um diferente número de *folds* à utilização de métodos de amostragem dos dados distintos. A utilização desta metodologia, pode reduzir o risco de criar um modelo com pouco poder preditivo em novos dados uma vez que estamos a avaliar os modelos num maior conjunto de dados não vistos. [27]

Tendo os modelos com os seus parâmetros otimizados, através da aplicação do *grid seacrh* com *cross-validation,* foram comparados através dos valores médios de MSE e MxAE em cinco *folds*. Adicionalmente, calculou-se o desvio padrão de cada métrica para cada modelo de forma a avaliar a capacidade de generalização do modelo. Ao utilizar *cross-validation* não se pretende escolher um modelo que se comportou excecionalmente bem num *fold* e excecionalmente mal noutro, mesmo tendo a melhor média. O objetivo é escolher os modelos que têm menores valores de erro e simultaneamente se comportam com coerência em diferentes conjuntos de dados.

Por último, foi utilizado o nó **Model Ensemble** de forma a combinar os resultados previstos pelos três melhores modelos através de *bagging*. Este nó cria novos valores através do cálculo da média dos valores previstos nos nós antecedentes, tornando os resultados mais robustos e estáveis uma vez que não dependem de um único algoritmo e conjunto de variáveis. É importante denotar que o modelo *ensemble* apenas produz melhores resultados que os modelos incluídos se estes forem divergentes, devendo sempre ser comparado com os modelos individuais. [28] Em ambas as plataformas o modelo *ensemble* revelou ser a abordagem que tinha menor erro nos dados de validação.

Em conclusão, para cada uma das plataformas possuía-se um modelo final: um *ensemble* de uma Regressão Linear e dois *Gradient Boost*.

## 5.8. RESULTADOS

Estando definido o modelo final é importante testá-lo com dados que o modelo nunca viu, de forma a averiguar a sua capacidade preditiva em novas observações. No contexto do negócio a única forma de avaliar o comportamento do modelo é através da utilização de uma população completa, ou seja, é importante avaliar a distribuição de objetivos na Rede como um todo. Dado que o objetivo definido tem de ser distribuído na sua totalidade pelos CC ativos num ciclo não é válida a avaliação numa amostra de sucursais. Considerando esta necessidade de negócio foram utilizados dois ciclos para testar o modelo em cada plataforma: 201904 e 202001[10].

Através do nó de **Score** do SAS Miner foram estimados os valores de *target* para dados de teste. Este nó permite ainda a criação de um *script* SAS com todo o processo desde as transformações de variáveis até à definição dos parâmetros dos modelos que pode, mais tarde, ser utilizado em dados novos.

Os resultados devolvidos pelo modelo representam a proporção prevista de quanto objetivo deveria ser alocado a cada CC por plataforma. No entanto, uma vez que o objetivo real alocado à plataforma é predefinido estes resultados previstos têm de ser recalculados de forma a dar uma base de 100% e distribuir-se a totalidade do objetivo. Em adição, é necessário controlar os valores máximos por cada colaborador que devem ser aplicados aquando da distribuição do objetivo. Os resultados apresentados aos decisores representam, portanto, a transformação do *target* previsto no valor de objetivo distribuído para cada CC.

As métricas de avaliação com interesse para o negócio não são as métricas de erro do modelo por si só, mas sim quais as melhorias que a implementação de um novo modelo traz face ao modelo utilizado anteriormente. A quantificação destas possíveis melhorias deve ser apresentada não na escala da variável **y** mas sim na escala do objetivo distribuído pelos CC, ou seja, em euros.

[10] Aquando a fase de teste do modelo já foi possível recolher o ciclo adicional: 202001.

Assim, na apresentação de resultados é efetuada uma comparação dos dois modelos em cada plataforma evidenciando as vantagens e desvantagens de cada uma das abordagens. Os modelos foram comparados através das métricas seguidamente expostas. Sendo o $r_i$ o montante contabilizado de Crédito Habitação em cada CC $i$ no ciclo em análise e $o_i$ o objetivo distribuído para o CC $i$ no mesmo ciclo. O erro de cada observação é calculado como $\varepsilon_i = r_i - o_i$.

- **$MAE_{obj}$**: Erro Médio Absoluto: $n^{-1}\sum_{i=1}^{n}|\varepsilon_i|$ que se traduz no afastamento médio em euros do objetivo ao realizado pelos CC.

- **Afastamento Médio de GRO 100 ($MAE_{GRO}$)**: representa o MAE calculado em função de um Erro Absoluto dado por $\varepsilon_i = |100 - GRO_i|$. Esta medida representa o afastamento médio de cada CC de ter realizado um GRO perfeito de 100. O intuito é distribuir os objetivos pela capacidade real de cada CC pelo que a existência de GRO muito superiores a 100 não é benéfica, simboliza que poderia ter sido distribuído mais objetivo para alguns CC em detrimento de outros.

- **Percentagem de CC com GRO>=100 (% GRO >=100)**: representa a percentagem de CC que atingiriam o objetivo proposto.

- **Percentagem de CC com GRO entre 90 e 110 (% GRO 90-110):** representa a percentagem de CC que se inserem num intervalo de resultados para os quais se considera os objetivos ajustados à realidade (variação do realizado face ao objetivo de 10 pontos percentuais).

- **Percentagem de CC que ficaram mais próximos de GRO 100 comparativamente ao Modelo Anterior (% + próximo 100)**: representa a percentagem de CC onde o afastamento de GRO 100 é menor com um modelo em relação com o outro.

Nas subsecções seguintes apresentam-se os resultados obtidos pelos modelos finais de cada plataforma isoladamente. Avaliando os resultados numa perspetiva de negócio e estabelecendo um paralelo entre o Modelo Proposto (MP), criado ao longo do projeto, e o Modelo Anterior (MA).

Os valores de objetivo apresentados sofreram uma transformação[11] de forma que as distribuições se mantenham, mas os valores sejam fictícios. Os montantes estão agrupados por Direção Coordenação e Direção Comercial no entanto, os nomes de cada zona estão anonimizados.

### 5.8.1. Mass Market

Na plataforma MM os resultados esperavam-se bons uma vez que na exploração dos dados se observou que a relação de diversas variáveis com o *target* era simples e existiam variáveis com elevados coeficientes de correlação. Adicionalmente as sucursais apresentam valores de realizado com

[11] A transformação efetuada foi a divisão por uma constante $k$ aplicada após a distribuição dos objetivos pelas plataformas às variáveis Realizado, Objetivo MA e Objetivo MP. A divisão por uma constante permite manter a relação de grandeza entre as variáveis e ainda os valores de GRO reais uma vez que $GRO = \frac{Realizado}{Objetivo}$*100, Aplicando a transformação das variáveis ficaria $GRO = \frac{\frac{Realizado}{k}}{\frac{Objetivo}{k}} * 100 = \frac{Realizado}{Objetivo} * \frac{k}{k} * 100 = \frac{Realizado}{Objetivo} * 100$. Ao mesmo tempo que mascara os valores de Crédito Habitação produzidos nos trimestres em causa.

reduzida variância ao longo dos ciclos. Sendo, por isso, mais simples de estimar e dando ao modelo maior facilidade de generalizar para novos ciclos comerciais.

A Tabela 8 demonstra os valores de erro do modelo nos dois ciclos utilizados como teste sendo o erro substancialmente menor no último ciclo de 201904. De notar que pelo menos nos últimos 3 anos ambas as plataformas superavam sempre o objetivo proposto por pelo menos 15% (GRO mínimo na plataforma de 115%) e pela primeira vez em 202001 assistiu-se a um GRO menor que 100 em ambas as plataformas. A redução na performance deveu-se sobretudo à pandemia vivida no final deste trimestre e à necessidade de adaptar um modelo de negócio ao teletrabalho bem como às consequências sentidas no mercado pela incerteza vivida. Desta forma, é natural que os resultados de 202001 do modelo não sejam tão positivos como os observados em 201904. Os valores de erro MAE refletem que em média um valor previsto erra por cerca de 0,05 no ciclo 201904.

Para a empresa e para os decisores o mais relevante não é o comportamento do modelo na variável *target*, mas sim, com a variável transformada no objetivo que receberia cada CC. A comparação dos dois modelos nas diferentes métricas, Tabela 9, demonstrou uma clara diminuição no erro absoluto médio ($MAE_{obj}$) que se traduz numa melhor adequação dos objetivos do MP ao realizado das sucursais de MM face ao modelo utilizado anteriormente. O erro absoluto médio no objetivo ($MAE_{obj}$) diminuiu 36% e a mesma métrica no GRO ($MAE_{GRO}$) teve uma redução de 43%. Apesar do número de CC que atinge um GRO 100 ou superior ser semelhante com os dois modelos em 201904 ou mesmo menor em 202001. Existe, na generalidade das sucursais, uma aproximação ao resultado ideal ao aplicar-se o Modelo Proposto. Isto é, em 201904 aproximadamente 72% das sucursais ficariam mais próximas de GRO 100 com o Modelo Proposto. O que se traduz num erro absoluto médio da variável GRO menor quando utilizado o Modelo Proposto.

Tabela 8. MM – Valores de erro do Modelo Proposto no *dataset* de Teste

| Medida de Erro | Ciclo 201904 | Ciclo 202001 |
|---|---|---|
| Nº de Observações | 453 | 457 |
| SSE | 2,0007 | 7,2228 |
| MSE | 0,0044 | 0,0200 |
| RMSE | 0,0665 | 0,1257 |
| MAE | 0,0475 | 0,0962 |
| MxAE | 0,3394 | 0,5779 |

Tabela 9. MM – Comparação do Modelo Anterior e Modelo Proposto no *dataset* de Teste

| Ciclo | 201904 | | 202001 | |
|---|---|---|---|---|
| Modelo | MA | MP | MA | MP |
| Nº de Observações | 453 | | 457 | |
| $MAE_{obj}$ | 864,28 | 566,64 | 928,00 | 864,42 |
| $MAE_{GRO}$ | 52,99 | 29,98 | 51,29 | 44,32 |
| % GRO >=100 | 62,91% | 66,00% | 45,30% | 39,61% |
| % GRO 90-100 | 16,78% | 25,17% | 10,72% | 11,60% |
| % + prox. 100 | 27,37% | 71,74% | 45,95% | 50,55% |

Tabela 10. MM – Resultados do Modelo Anterior e Modelo Proposto no *dataset* de Teste (ciclo 201904) por Coordenação

| DC | Realizado | Obj MA | Obj MP | $MAE_{obj}$ MA | $MAE_{obj}$ MP | GRO MA | GRO MP |
|---|---|---|---|---|---|---|---|
| A | 349.065 | 266.722 | 283.215 | 918 | 599 | 131 | 123 |
| B | 354.047 | 317.601 | 300.364 | 696 | 493 | 111 | 118 |
| C | 301.987 | 267.753 | 268.498 | 1.018 | 625 | 113 | 112 |
| MM | 1.005.099 | 852.077 | 852.077 | 864 | 567 | 118 | 118 |

Na Tabela 10, pode-se ver os resultados (transformados) dos dois modelos em análise, para cada Coordenação no último trimestre de 2019 (ciclo 201904). Observando os resultados das Coordenações (A, B, C) percebemos que a nível macro o modelo comporta-se como o esperado: aumentar objetivo em quem possuía GROs bastante elevados (A) e diminuir onde a performance era mais baixa (B).

Analisando os dados das Direções Comerciais a tendência de crescer o objetivo em locais onde existiam maiores níveis de performance manteve-se. No entanto, denotam-se algumas assimetrias consoante a região do país. Os valores médios de erro são mais elevados nas Direções Comerciais da Área Metropolitana de Lisboa, onde apesar do aumento de objetivo proposto pelo modelo os valores de GRO ainda são superiores a 110%. Por contraste as regiões onde os erros médios são menores são as Direções Comerciais localizadas no interior.

Relevante realçar que nas 10 Direções onde o objetivo distribuído sobe 5% ou mais com a aplicação do Modelo Proposto os valores de GRO com o Modelo Anterior variavam entre 119 e 163. Por outro lado, nas 11 regiões em que o objetivo diminuía 5% ou mais os valores anteriores de GRO variavam entre 91 e 109. Na Figura 14 podemos observar como o modelo apresentado diminui a variância na variável GRO permitindo uma distribuição de objetivos mais ajustada aos valores realizados por cada sucursal, cumprindo assim o seu objetivo.

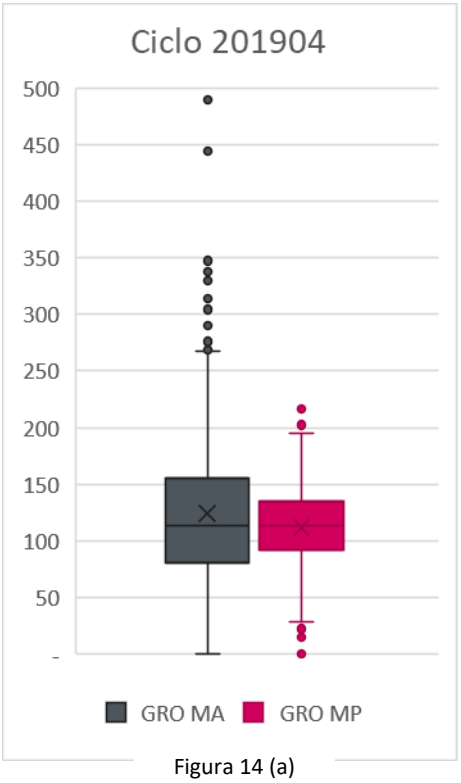



Figura 14 (a)

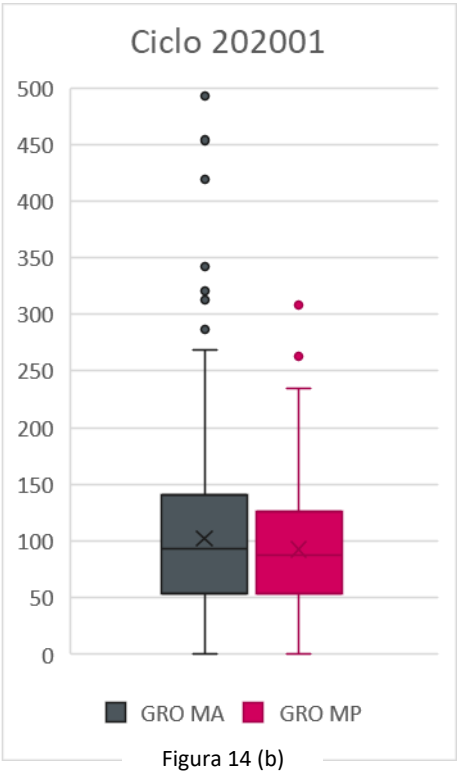



Figura 14 (b)

Figura 14. MM – Distribuição do GRO dos CC nos dados de teste em cada Modelo nos ciclos 201904 (a) e 202001 (b)

### 5.8.2. Gestão Personalizada Prestige

Na plataforma GPP os resultados são menos positivos relevando erros médios superiores. Observando-se a Tabela 11 vê-se que, por exemplo, o valor de $MAE_{obj}$ é de 0,12 em ambos os ciclos comparativamente com 0.05 no MM.

Ao comparar o Modelo Proposto com o Modelo Anterior nos valores de objetivo em euros e não em proporção, Tabela 12, percebe-se que as reduções de erro nesta plataforma não são tão expressivas como no MM. Um dos fatores que contribui para esta redução menos expressiva é o facto do erro do modelo inicial no GPP já ser notavelmente menor do que era a aplicação desse mesmo modelo ao MM. O que nos leva a concluir que o Modelo Anterior se adequava melhor às especificidades da plataforma de clientes geridos do que ao MM.

Adicionalmente, é importante mencionar que os maiores erros nesta plataforma podem resultar da volatilidade no comportamento dos gestores. Isto é, gestores com elevados valores de realizado num ciclo podem não fazer nenhuma operação no ciclo seguinte. Sendo comum existir CC com GRO 0 nesta plataforma, algo que não acontece no MM.

Tabela 11. GPP – Valores de erro do Modelo Proposto no *dataset* de Teste

| Medida de Erro | Ciclo 201904 | Ciclo 202001 |
|---|---|---|
| Nº de Observações | 453 | 576 |
| SSE | 14.9401 | 15.1687 |
| MSE | 0.0330 | 0.0253 |
| RMSE | 0.1816 | 0.1590 |
| MAE | 0.1238 | 0.1205 |
| MxAE | 0.7513 | 0.7918 |

Ainda assim, a implementação do Modelo Proposto traria uma redução de erro absoluto médio quer no objetivo quer no GRO. No global existe uma maior percentagem de CC a ficarem melhor ajustados com o MP, 45% versus 42% em 201904 e 43% versus 35% em 202001. Relevante denotar que o $MAE_{obj}$ mesmo sendo semelhante nos dois modelos em comparação é menor do que a mesma medida de erro na plataforma MM nos dois ciclos em teste.

Tabela 12. GPP – Comparação do Modelo Anterior e Modelo Proposto no *dataset* de Teste

| Ciclo | 201904 | | 202001 | |
|---|---|---|---|---|
| Modelo | MA | MP | MA | MP |
| Nº de Observações | 453 | | 576 | |
| $MAE_{obj}$ | 572,79 | 562,54 | 587,96 | 563,79 |
| $MAE_{GRO}$ | 73,85 | 71,72 | 73,49 | 68,34 |
| % GRO >=100 | 55,63% | 55,63% | 38,19% | 39,76% |
| % GRO 90-100 | 10,60% | 10,15% | 7,12% | 9,55% |
| % + próximo 100 | 41,94% | 45,03% | 35,24% | 43,40% |

Tabela 13, podemos ver os resultados (transformados) dos dois modelos em análise, para cada Coordenação no último trimestre de 2019 (ciclo 201904). Observando resultados das Coordenações (A, B, C) podemos concluir, tal como no MM, que a nível macro o modelo comporta-se como o esperado: aumenta objetivo em quem possuía GRO bastante elevados (A) e diminui onde a performance era mais baixa (B).

Tabela 13. GPP – Resultados do Modelo Anterior e Modelo Proposto no *dataset* de Teste (ciclo 201904) por Coordenação

| DC | Realizado | Obj MA | Obj MP | $MAE_{obj}$ MA | $MAE_{obj}$ MP | GRO MA | GRO MP |
|---|---|---|---|---|---|---|---|
| A | 188.582 | 134.721 | 137.279 | 637 | 624 | 140 | 137 |
| B | 134.196 | 116.858 | 111.548 | 421 | 426 | 115 | 120 |
| C | 130.453 | 111.156 | 113.908 | 719 | 689 | 117 | 115 |
| GPP | 453.231 | 362.735 | 362.735 | 573 | 563 | 125 | 125 |

Ao analisar no segundo nível hierárquico, as Direções Comerciais, observa-se resultados semelhantes ao MM as Direções Comerciais com erros médios mais elevados são as de Lisboa e circundantes, sendo os erros mais pequenos observados nas Direções mais interiores e do Norte. Com auxílio da Figura 15 podemos concluir que apesar da variância na distribuição do GRO pelos diferentes gestores ser menor com a aplicação do Modelo Proposto a sua redução não é tão significativa como na plataforma MM.

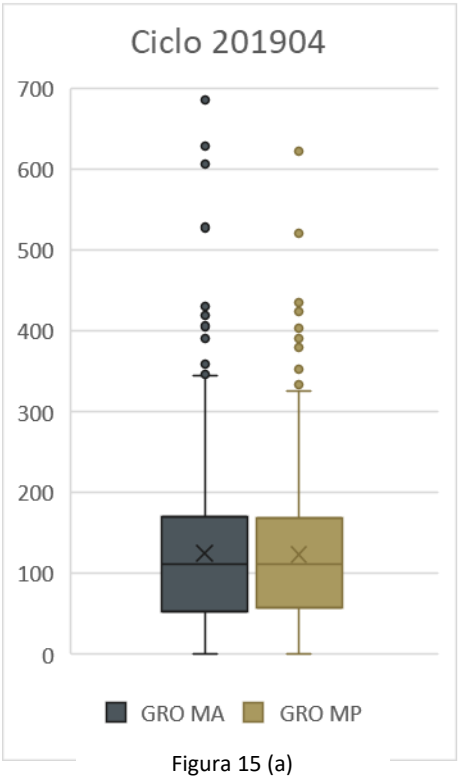



Figura 15 (a)

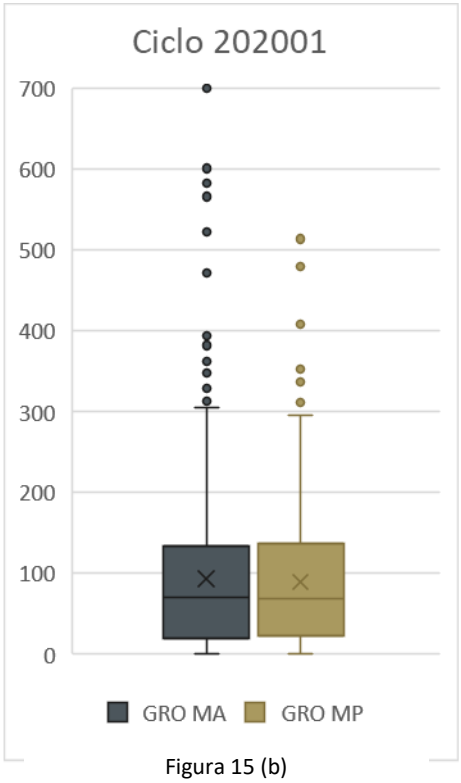



Figura 15 (b)

Figura 15. GPP – Distribuição do GRO dos CC no *dataset* de Teste em cada Modelo nos ciclos 201904 (a) e 202001 (b)

### 5.8.3. Rede Retalho

Por último é necessário avaliar a RR como um todo. Esta necessidade surge porque apesar do cálculo dos objetivos na maioria das campanhas ser realizado por plataforma, os objetivos são apresentados aos Coordenadores e Diretores Comerciais no seu global. Para além disso, conforme referido

anteriormente, os Coordenadores e Diretores Comerciais têm a possibilidade de ajustar os objetivos das suas Coordenações e Direções Comerciais, respetivamente, desde que o montante global de cada região seja igual. Na prática esta possibilidade permite às chefias aliviar uma plataforma em detrimento da outra por possuírem um maior conhecimento no negócio efetuado em cada local e saberem como melhor atingir o objetivo global da Coordenação/Direção Comercial.

Avaliando a Rede como um todo observa-se, Tabela 14, que através do Modelo Proposto quase 60% dos CC ficariam com resultados mais próximos de 100 do que se fosse aplicado o Modelo Anterior. No último trimestre de 2019 afastamento médio de GRO 100 diminuía em 20% e a mesma diminuição para o erro absoluto médio do objetivo. Já no ciclo seguinte as melhorias são uma redução de 9% e 6% respetivamente.

Uma análise aos valores de Realizado, de Objetivo, de GRO e valores de erro $MAE_{obj}$ e $RMSE_{obj}$ dos dois modelos para as Coordenações (A, B e C) e respetivas Direções Comerciais no ciclo 201904 demonstra na generalidade que o Modelo Proposto acresce objetivo face ao Modelo Anterior nas regiões onde o anterior produzia valores de GRO acima de 120, possuindo o efeito inverso em Direções onde o indicador de performance era menor.

Tabela 14. RR – Comparação do Modelo Anterior e Modelo Proposto no *dataset* de Teste

| Ciclo | 201904 | | 202001 | |
|---|---|---|---|---|
| Modelo | MA | MP | MA | MP |
| Nº de Observações | 906 | | 1033 | |
| $MAE_{obj}$ | 718,53 | 564,59 | 738,39 | 696,79 |
| $MAE_{GRO}$ | 63,42 | 50,85 | 63,67 | 57,71 |
| % GRO >=100 | 59,27% | 60,82% | 41,34% | 39,69% |
| % GRO 90-100 | 13,69% | 17,66% | 8,71% | 10,45% |
| % + próximo 100 | 34,66% | 58,39% | 39,98% | 46,56% |

## 5.9. Conclusões e Discussão

O modelo apresentado e criado através da metodologia exposta apresenta variadas vantagens, tais como, a inclusão de variáveis de mercado que permitem incluir o potencial da localidade da sucursal ou gestor sem o limitar aos clientes da instituição. A maior vantagem do Modelo Proposto é a redução considerável do erro na plataforma MM que se traduz numa melhor adequação dos objetivos à capacidade de realização dos colaboradores.

No entanto, o modelo apresenta uma desvantagem face ao Modelo Anterior uma vez que não permite uma explicação clara e fácil interpretação dos parâmetros de forma a explicar à RR porque cada unidade de negócio tem um objetivo específico. Adicionalmente, analisando a distribuição de objetivos por CC ao longo de 3 ciclos percebeu-se que variações consideradas pequenas na carteira de clientes poderiam resultar em variações significativas no objetivo. Em termos do contexto de negócio apesar

de se pretender um recálculo de objetivos trimestralmente a variação de objetivo em cada CC deve ser reduzida.

Face ao exposto decidiu-se simplificar o modelo utilizando apenas um dos algoritmos presentes no modelo *ensemble*. Assim, para cada uma das plataformas escolheu-se a Regressão Linear pela sua interpretabilidade. Esta abordagem representa uma redução de performance do modelo existindo níveis der erro superiores. No entanto, a capacidade de interpretação dos parâmetros e de explicação clara das variáveis que conduzem à atribuição de um objetivo a cada CC fazem com que seja a abordagem mais adequada ao negócio.

A redução do erro trazida pela adoção do Modelo Proposto é mais visível e significante na plataforma MM, conforme demonstrado na apresentação de resultados. A melhor performance no MM é justificada por diversas razões, das quais se destaca o facto do comportamento das unidades de negócio ser mais consistente nesta plataforma provocando relações das variáveis dependentes com o *target* mais evidentes e lineares.

Relativamente à plataforma de cliente geridos, observam-se um MAE na escala real do objetivo ($MAE_{obj}$) consideravelmente baixo, no entanto o afastamento médio de GRO igual a 100 ($MAE_{GRO}$) é bastante elevado, próximo de 70 em ambos os ciclos de teste. O que demonstra a volatilidade sentida nos resultados deste segmento. Os objetivos encontram-se na sua maioria ajustados, mas o elevado número de CC com resultados iguais a zero incrementa o erro médio no GRO.

Em suma, o Modelo Proposto apresenta melhores resultados que o Modelo Anterior nos dois ciclos de teste, comprovando a necessidade sentida para uma melhor adequação do modelo utilizado anteriormente. Não obstante, com a aplicação do Modelo Proposto a percentagem de unidades de negócio com GRO entre 90 e 100% ainda é consideravelmente baixa, apenas 18% da RR, existindo espaço para uma melhoria da abordagem.

Ao longo da realização do projeto apercebeu-se que existiam aspetos relacionados com o fator humano difíceis de incorporar no modelo que no contexto real influenciam fortemente os resultados. Por um lado, tem-se a definição do *target* dependente do comportamento das unidades de negócio nos ciclos anteriores. Este comportamento pode não ser um reflexo real do potencial da carteira desse CC. Primeiramente porque ao longo do tempo as equipas e gestores mudam provocando mudanças no potencial humano das unidades de negócio. Depois, os valores de produção de cada campanha, utilizados para calcular a *target*, estão condicionados pelos valores de objetivos distribuídos nesse ciclo para cada CC uma vez que as campanhas não funcionam isoladamente. Isto é, o esforço comercial será distribuído por vários produtos e assim que se atinge GRO 100 no Crédito Habitação, por exemplo, o foco será redirecionado para outras campanhas.

Para além deste aspeto, outro fator humano que influencia o resultado da rede são as chefias. A gestão da Direção Comercial condiciona os indicadores de performance e os resultados em cada campanha a dois níveis: na gestão de pessoas e na gestão de objetivos. Na primeira, fala-se de aspetos como a motivação das equipas, a satisfação pessoal dos colaboradores com a sua gestão de carreira, entre outros. A nível da gestão de objetivos, o principal impacto acontece na forma como procedem à redistribuição dos objetivos pelos seus CC.

Outro fator não quantificado no modelo são as operações de Crédito Habitação realizadas fora da sua zona de influência e/ou a clientes de outras carteiras que não a sua. Estas operações não são contabilizadas no potencial do CC, uma vez que as variáveis são especificas da carteira e/ou da região em que o CC está localizado. Se uma unidade de negócio tiver localizado no interior mas realizar operações no Algarve ou Lisboa, por exemplo, essas operações tendencialmente serão superiores ao que seria esperado para a sua carteira/região.

Em conclusão, apesar de existirem fatores que não foram considerados na análise pela sua complexidade ou incapacidade de quantificar, os resultados obtidos pelo modelo foram considerados bons e apresentam melhorias ao método anteriormente utilizado. Estas melhorias são visíveis quer pela melhor adequação dos valores de objetivos distribuídos à capacidade real dos CC, quer pela inclusão de variáveis externas ao MBCP. O Modelo Proposto com recurso à técnica *ensemble* não conseguiu cumprir todos os critérios do negócio de clareza e interpretação optando-se por isso por uma versão mais simples com apenas um algoritmo. Esta versão mantém a vantagem da inclusão de variáveis de potencial externas ao Banco, contudo representa melhorias menos significativas no erro.

# 6. CONCLUSÃO

Ao longo da realização do estágio foi possível aplicar as técnicas e conhecimentos adquiridos ao longo do ano curricular do mestrado num problema prático e real de previsão. O projeto permitiu aprofundar conhecimentos nas técnicas de *Data Mining* e *Machine Learning* e ainda potenciar o conhecimento sobre o setor bancário e de ferramentas como SAS e SAS Miner. Para além do projeto principal, as tarefas de extração, manipulação de dados contribuíram para a compreensão do negócio e do funcionamento do departamento. Por último, realçar a mais valia que foi desenvolver um projeto em Power BI que é uma ferramenta cada vez mais utilizada nas organizações.

## 6.1. LIMITAÇÕES E LIÇÕES APRENDIDAS

Ao longo do projeto a primeira limitação sentida foi a inicial disponibilidade reduzida de dados. Conduzindo à produção de um modelo ineficaz modelado a partir de dados insuficientes, onde a variável *target* não era o resultado real do conjunto de variáveis independentes.

Em termos técnicos, existiram algumas limitações com o *software* disponível. Inicialmente o projeto começou a ser implementado em Python. No entanto, sendo um programa de licença aberta, as permissões de atualização e instalação de novos pacotes necessários era demorada por motivos de segurança. Para além disto, a versão do Anaconda disponível, incluindo os diversos pacotes instalados, é comum aos colaboradores o que significava que alguns pacotes necessários não fossem possíveis de instalar por entrarem em conflito com determinadas versões de outros pacotes utilizados em outras áreas.

Para solucionar a limitação do Python pediu-se a licença de SAS Miner e criou-se o projeto num *software* desconhecido o que, apesar de ser uma limitação inicialmente, potenciou o alargamento dos conhecimentos tornando-se numa mais valia.

No decorrer do estágio foi possível aplicar os conhecimentos teóricos e teórico-práticos adquiridos ao longo no percurso académico num contexto real de negócio. A criação de um projeto de *Machine Learning* num contexto real permitiu a compreensão de que nem sempre a capacidade preditiva dos modelos e o menor erro possível são os aspetos mais importantes do projeto. Num contexto real a capacidade interpretativa dos algoritmos e a sua simplicidade podem ser mais valorizados.

Em suma, a realiação de um projeto no Millennium bcp demonstrou que para além das capacidades técnicas e teóricas o conhecimento do negócio e do funcionamento da empresa são fundamentais para desenvolver projetos que tragam valor à instituição e cumpram os requisitos de negócio.

## 6.2. TRABALHO FUTURO

Estando implementado o modelo é importante avaliar como será o resultado da Rede com o novo método de distribuição de objetivos e se o montante global de objetivo é atingido, sendo esse o principal objetivo do banco. Para além dos valores globais de objetivo e realizado é importante a avaliação do GRO dos CC percebendo se a distribuição convergiu para um GRO igual a 100.

Posteriormente, se, como indicam os dados de teste, a adequação dos objetivos à realidade da Rede for melhor que a obtida através do Modelo Anterior o próximo passo será a aplicação da metodologia às restantes campanhas tais como os Crédito de Negócios e o Crédito Pessoal.

Ao longo do tempo ser importante a adequação dos modelos às alterações regras de contratação e contabilização dos produtos. Por exemplo, com o clima de incerteza potenciado pela pandemia a partir do terceiro trimestre de 2020 o valor médio de *Loan to Value*[12] (LTV) aplicado em empréstimos a clientes estrangeiros sofreu uma redução passando de 80 para 50%. Esta alteração vai afetar de maneira diferenciada o valor das operações dos diversos CC. Especialmente em zonas como o Algarve, Lisboa e Porto no segmento Prestige, onde as operações contratadas por clientes estrangeiros tem uma expressão maior, o montante das operações e o seu valor médio poderá reduzir, o que afetará a realização de objetivos destes CC usualmente com valores elevados quando comparados com outras geografias.

Por último, seria interessante testar uma diferente abordagem ao problema olhando para os objetivos ao longo do ano como um único objetivo por cada CC e não como objetivos trimestrais. Para tal seria necessário a agregação dos objetivos e da produção anual de cada CC calculando-se apenas um registo por ano. A agregação dos objetivos ao longo do ano bem como das variáveis de *input* permitiria uniformizar o comportamento dos diversos CC com características semelhantes que demonstram uma elevada variância ao longo do ano conseguindo contratar crédito nuns ciclos e em outros não. Agregar a informação traria vantagens como o incremento de correlação de algumas variáveis com o *target* e o potencial crescimento de poder preditivo. No entanto, para utilizar esta abordagem teria de se garantir que apenas se utilizaria CC que não alteraram a sua constituição ao longo do ano em termos de constituição da carteira. Deveria ser garantido ainda que o gestor dos CC no segmento Prestige não teria sido alterado, uma vez que, mudando o colaborador encarregue da carteira, o comportamento no atingimento de objetivos será distinto. Esta metodologia poderá ser estudada na realização dos projetos para os restantes produtos.

[12] **LTV**: o valor de *Loan to Value* representa a relação do montante do empréstimo com o montante de avaliação do imóvel, em condições normais este valor não pode ultrapassar os 80%, ou seja, o valor do empréstimo só pode ter valores iguais ou inferiores a 80% do valor de avaliação do imóvel.

# 7. BIBLIOGRAFIA


[1] K. Heinonen, "Multiple Perspectives on Customer Relationships," *International Journal of Bank Marketing,* vol. 32, no. 6, 2014.

[2] Indra, "From Traditional Banking System to the Customer-Centric Financial Ecosystem," 2014.

[3] E. A. Locke and G. P. Latham, "Bulding a Practicakky Usefuk Theory of Goal Setting and Task Motivation," *American Psychologist,* vol. 57, no. 9, pp. 705-717, Setembro 2002.

[4] F. C. Lunenburg, "Goal-Setting Theory of Motivation," *International Journal of Management, Business, and Administration,* vol. 15, no. 1, 2011.

[5] U. Quaiser, "A comparative study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)," *International Journal of Inovation and Scientific Research,* vol. 12, no. 1, pp. 217-222, Novembro 2014.

[6] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000.

[7] P. Meesad and K. Hengpraprohm, "Combination of knn-based feature selection and knn-based missing-value imputation of microarray data," in *rd International Conference on Innovative Computing Information and Control*, 2008.

[8] X. Liang, "kNN Classification and Regression using SAS," in *NESUG*, 2012.

[9] C. Aggarwal, Outlier Analysis, in: Data Mining, Cham: Springer, 2015.

[10] I. Ben-Gal, "Outlier Detection," in *Data Mining and Knowledge Discovery Handbook*, L. M. Oded, Ed., Boston, MA, Springer US, 2005, pp. 131-146.

[11] A. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Reasearch and Development,* vol. 3, no. 3, pp. 210-229, Julho 1959.

[12] S. Marsland, Machine Learning: An Algorithmic Prespective, CRC Press, 2015.

[13] S. Shalev-Shwartz and S. Ben-David, Understanding machine learning: From theory to algorithms, Cambridge university press, 2014.

[14] J. Alzubi, A. Nayyar and A. Kumar, "Machine Learning from theory to algorithms: an overview," *Journal of physics: conference series,* vol. 1142, no. 1, 2018.

[15] M. Verbeek, "2.2 The Linear Regression Model," in *A Guide to Modern Econometrics*, 2ª ed., John Wiley & Sons, Ltd., pp. 14-16.

[16] J. M. Wooldridge, Introductory Econometrics A Modern Approach, 6ª ed., USA: CENGAGE Learning, 2015, pp. 20-59.

[17] D. Wei and J. Wei, "A MapReduce Implementation of C4.5 Decision Tree Algorithm," *International Journal of Database Theory and Application,* vol. 7, no. 1, pp. 49-60, 2014.

[18] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006, pp. 653-676.

[19] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics,* vol. 7, p. 21, 2013.

[20] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *International workshop on multiple classifier systems*, 2000.

[21] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research,* vol. 30, pp. 79-82, Dezembro 2005.

[22] M. A. Brookhart, S. Scnneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn and T. Sturmer, "Variable Selection for Propensity Score Models," *American Journal of Epidemiology,* vol. 163, no. 12, pp. 1149-1156, 2006.

[23] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* 2003.

[24] K. S. Sarma, "Variable Selection and Transformation of Variables in SAS Enterprise Miner 5.2," 2007.

[25] SAS Institute, "SAS Documentation: Variable Clustering Node," Agosto 2017. [Online]. Available: https://documentation.sas.com/?docsetId=emref&docsetTarget=p19e837tepmjz0n1hjt2gdk3sqfg.htm&docsetVersion=14.3&locale=en. [Accessed Julho 2020].

[26] SAS Institute, "SAS Documentation: Model Comparison Node," Agosto 2017. [Online]. Available: https://documentation.sas.com/?docsetId=emref&docsetTarget=p01jgc9rmzsg37n1lfncp67t0unm.htm&docsetVersion=14.3&locale=en. [Accessed Julho 2020].

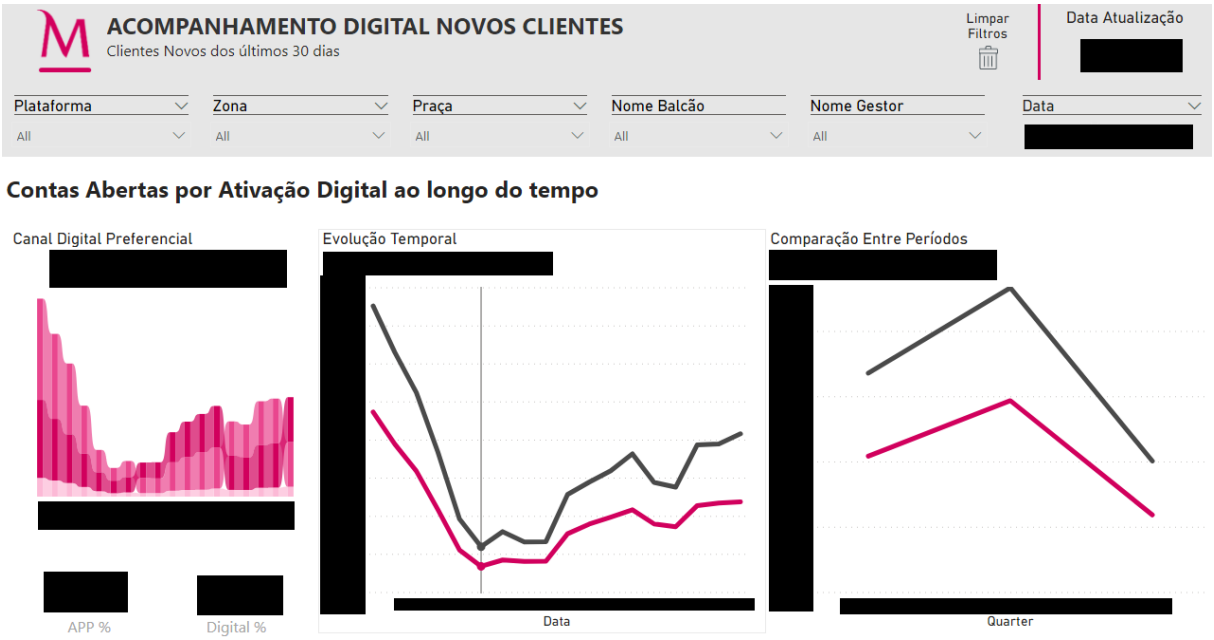
[27] C. Shaffer, "Selecting a Classification Method by Cross-Validation," *Machine Learning,* vol. 13, pp. 135-143, 1993.

[28] SAS Institute, "SAS Documentation: Ensemble Node," Agosto 2017. [Online]. Available: https://documentation.sas.com/?docsetId=emref&docsetTarget=n1upixvafylkyon1nfbokdnx3hpu.htm&docsetVersion=14.3&locale=en#p0fuh4t0ue5mgcn1s93p07jl1iq7. [Accessed Julho 2020].

# 8. ANEXOS

## ANEXO I. *DASHBOARD* EM POWER BI: ATIVAÇÃO DIGITAL DE NOVOS CLIENTES

*Dashboard* em Power BI: Ativação Digital de Novos Clientes – Evolução Temporal:




*Dashboard* em Power BI: Ativação Digital de Novos Clientes – Comparação Nacional: