UNIVERSIDADE DE LISBOA FACULDADE
DE CIÊNCIAS
DEPARTAMENTO DE MATEMÁTICA

INSTITUTO SUPERIOR DE CIÊNCIAS DO
TRABALHO E DA EMPRESA
DEPARTAMENTO DE FINANÇAS

EMPRÉSTIMOS BANCÁRIOS: VARIÁVEIS DETERMINANTES NA SUA AQUISIÇÃO

Carolina Ramos Estevam

Mestrado em Matemática Financeira

Dissertação orientada por:
Professora Doutora Diana Mendes

2021

AGRADECIMENTOS

Primeiramente e antes de qualquer outro agradecimento, quero agradecer à minha orientadora Diana Mendes. Agradeço-lhe por toda a disponibilidade, pela orientação e compreensão, por todo o apoio imediato e pela força e coragem que me prestou para realizar esta dissertação.

Em seguida, quero agradecer a todas as pessoas que se cruzaram na minha vida e que de direta ou indiretamente, fizeram parte de toda esta minha caminhada.

Quero agradecer aos meus avós, Arciolinda e Alfredo Ramos por todo o apoio incondicional, pela força e pela compreensão nos momentos em que estive mais ausente. Agradecer ao meu avô Agostinho e à minha estrelinha Maria Estevão por me fazerem sentir, sempre, especial.

Ainda na família, quero agradecer à minha família de coração, são eles (a minha tentação) o Fernando, a Marília e a Tânia Ângelo e o Miguel e a Shelby Ferreira. Estiveram presentes em todo o meu percurso académico, limparam-me todas as lágrimas nos momentos mais desafiantes, aturaram as minhas quebras de humor e celebraram comigo todas as vitórias, sempre com um sorriso no rosto e um abraço reconfortante.

Agradecer à família Marafuz, Carlos, Fernanda, Luís, Ricardo, Teresinha e Ana por acompanharem o meu percurso académico, oferecendo-me sempre força pela continuação. Obrigada Carlos por todos os, muitos, quilómetros percorridos pelos montes. És um cota maluco!

Nos amigos, tenho de agradecer a todos eles, mas fazer um agradecimento especial à Tatiana Ferreira, Daniela Pacheco, Adriana Henriques, Pedro Antunes e Bruno Esteves. Todos eles têm um lugar muito especial no meu coração. Estiveram presentes nos momentos mais difíceis e nunca me fizeram desistir. Um enorme obrigada!

No seguimento deste agradecimento, agradeço a todos os membros do grupo Cervejetarianos, Daniela Ferreira, Daniela Pacheco, David Lima, Diogo Calisto, Joana Gonçalves, João Francisco, João Grazina, João Leal, João Neto, Liliana Sobral, Mariane Moniz, Miguel Santos, Nelson Piedade, Pedro Pires e Tatiana Ferreira. Estes são os amigos que vou levar para a vida toda! Obrigada pelas saídas, cafés, conversas e festas; pelos treinos na garagem, na casa da escusa e nos montes; por estarmos sempre juntos quer na loucura, quer na tristeza; por acompanharem todo o meu crescimento a nível pessoal e por último, agradecer por aturarem as minhas birras de sono e o meu mau feitio. Um brinde a vocês!

De seguida agradecer aos grupos do mestrado, 104 e Powerpuff girls, constituídos pela Adriana Henriques, a Catarina Cardoso e o Rodrigo Falcão. Agradecer pelas viagens de comboio, de carro e de avião, por toda a dor e sofrimento partilhada - que ninguém merece - por acordar cedo a um sábado, pelos jantares, pelas risadas constantes, pelas aulas mais difíceis de passar, pelos trabalhos, pelo drama e pelas ajudas, que, sejam elas quais forem, ninguém precisa de saber. Quero ainda deixar uma palavra de agradecimento à Carolina Salazar por toda a ajuda em programação.

Agradecer e muito à Família Nogueira, Carlos, Helena, Rúben, Ricardo, Inês e Manuel, pois, sem eles, não teria adquirido as bases essenciais e fundamentais para chegar onde cheguei hoje, no meu percurso académico, mas também na minha vida pessoal. Um enorme obrigada!

Fazer um grande reconhecimento ao meu amigo sportinguista, Dr. Lima Natário. Obrigada por me deixar ficar no seu escritório tantas e tantas vezes. Obrigada pela companhia nos jogos de futebol do nosso grande amor. Obrigada por me tratar tão bem, a mim e à minha família.

Agradecer à família do meu namorado por me receberem tão bem, à avó Adosinda e ao avô Quim, ao pai Vítor e à mãe Carla, ao rapaz que adorava ser ribatejano para puder calçar uns merrell, mas que, não sabe que, primeiro se mete os cereais e só depois o leite, o Bernardo Ribeiro e também à sua namorada Catarina Romão. Obrigada por todo o apoio!

Agradecer à Ana Dias e ao Gonçalo Rato, à Maria do Céu, à Ana e ao Pedro Martins, ao João Medinas, ao Pedro Francisco, à Adelaide e ao João Teles, ao Carlos Pelixo, ao Fernando e à Fernanda Manuel, ao Sérgio Faria e à Família Fernandes. Também eles fizeram parte desta jornada.

Miguel, meu amor, já leste os agradecimentos todos? Ficaste de propósito para o fim. Assim sei que lês mais que uma folha da minha tese. Brincadeiras à parte, tenho de te agradecer e muito, pois fazes parte da conclusão desta etapa. Se calhar, se não fosse o "Se não acabares isto a tempo, não vamos de férias!", não estaria a concluir esta fase, que é tão minha quanto tua. Tenho de te agradecer pela paciência que tens comigo e pedir desculpa pela quantidade de vezes que ouviste "Isto é uma seca", "Nunca mais acabo", "Não vou conseguir", "Estive a fazer a tese", "Quero ir trabalhar" … ufa! Obrigada por toda a motivação e apoio que me deste. Começa agora uma nova etapa nas nossas vidas, seguimos juntos? Amo-te muito!

Por último e mais importante que tudo, estão os meus pais, Célia Estevam e Manuel Estevam, e ainda o homem da minha vida, o meu irmão, André Estevam. Sem todo o apoio, toda a ajuda, toda a força, toda a paciência, todo o amor, que me dão diariamente, não estaria a concluir mais uma etapa do meu percurso de vida. É a eles que dedico todo o meu trabalho! É incrível como continuam sempre do meu lado e nunca me deixam desistir, mesmo depois do meu mau feitio pré-exames e do meu mau humor quando algo não corre bem. Eles estão sempre ao meu lado, a celebrar todas as vitórias e eu sei, eu sei que sou a miúda mais sortuda, por tê-los na minha vida! Um enorme obrigada. Amo-vos muito!

RESUMO

O empréstimo é uma atividade que está presente desde o tempo dos nossos antepassados.

Outrora, se os camponeses necessitassem de um empréstimo, quer ele fosse de certos bens ou de produtos agrícolas, recorriam aos grandes proprietários, que lhos concediam em troca de certos benefícios. Caso houvesse qualquer incumprimento no pagamento do empréstimo por parte dos camponeses, poderiam estes, por exemplo, ver as suas terras confiscadas.

Assim, com o desenvolvimento do conceito de dinheiro, a criação da moeda e a criação dos bancos centrais, para a criação das políticas monetárias dos países, o crédito tornou-se algo essencial e imprescindível, nas instituições financeiras.

Os bancos tomaram consciência de que, se concedessem um crédito e o cobrassem aos devedores, pelo tempo de duração do empréstimo e ainda, se aceitassem depósitos de credores, compensando-os pelo tempo de permanência do dinheiro no banco, estes, conseguiriam obter lucros através da diferença entre as taxas de juro pagas e as taxas de juro cobradas.

No entanto, existe uma possibilidade de o devedor entrar em incumprimento no pagamento do empréstimo e o banco sofrer perdas financeiras.

Assim, torna-se importante e imperativo estudar o perfil do cliente e consoante o risco que o cliente apresente, o banco decide se concede ou não, o crédito bancário.

O atual cenário de crise económica que se faz sentir em Portugal, relativo ao alto nível de dívida e o seu incumprimento associado, conduzem a uma maior sensibilidade do setor bancário na concessão de crédito, o que leva os bancos a ficarem menos recetivos a atribuir crédito a clientes.

Posto isto, o principal objetivo desta dissertação é apurar quais são as variáveis determinantes na concessão de um crédito bancário, no setor privado. Para proceder a este estudo, esta dissertação irá usar a linguagem de programação *Python* para aplicar o modelo de regressão logística.

Palavras-chave:

Empréstimo bancário, crédito, regressão logística.

ABSTRACT

A loan is a process that is well known since ancient history.

In that time, if farmer needed a loan of an asset or certain agriculture related products, they would ask to the nobles for assistance and they would meet their request in change of some benefits. In case of the farmer did not follow this agreement, they could see their lands confiscated by the nobles.

Considering these loan agreements and adding the evolution of the concept of money, the creation of currencies and the creation of the bank, the credit concept became essential to financial institutions to create monetary policies for every country in the world.

On the other hand, the banks became aware that if they conceded a bank loan and then charge through a time period to the debtor, and if they were able to collect deposit from creditors with compensations over time, they can collect profits from the difference between the interest rates that were paid and the interest rates that were collected.

However, there is the possibility that the debtor isn't able to keep paying the loan and the bank suffer financial losses. Therefore, it is important and crucial to study the client's profile and evaluate the risk of the bank conceding, or not, the bank credit.

The current scenario of an economic crisis in Portugal, regarding the increasing public debt and it's associated non-compliance, it leads to a greater sensibility on the banking sector towards credit approval and makes banks less receptive to concede credit to new clients.

Hence, the main goal of this dissertation is to investigate which variables are determinant in the approval of a bank loan in the private sector. In order to substantiate this case study, this dissertation will seek to apply the logistic regression model by use of the programming language Python.

Keywords:

Bank loan, credit, logistic regression model.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

INTRODUÇÃO

Um empréstimo bancário não é mais do que um contrato, entre o cliente e o banco, no qual o banco empresta um montante solicitado pelo cliente. Este montante solicitado, que será pago pelo cliente, é acrescido de uma taxa de juro, a vigorar na data do contrato.

O banco deve reunir e analisar o maior número possível de informações do cliente, antes de proceder à concessão do crédito bancário, para amenizar os riscos financeiros e não enfrentar casos de incumprimento.

Em Portugal, a cedência de crédito bancário, tanto para habitação como para consumo e outros fins, tem sofrido algumas oscilações.

Sensivelmente entre 2001 e 2007, período que antecedeu a crise financeira, era concedido muito mais crédito em Portugal, do que é concedido nos dias atuais.

Entre 2008 e 2014 assinala-se um abrandamento na concessão de crédito, dado o impacto da crise financeira internacional e da crise das dívidas soberanas da zona euro, que levaram a uma forte quebra do PIB.

Consequentemente, os bancos ficaram pouco recetivos a emprestar capital e, além disso, a taxa de juro também ficou mais elevada.

Até aos dias de hoje, Portugal estaria a recuperar desta crise económica, não fosse, em 2020, aparecer uma crise pandémica, devido ao Covid-19, que deixará marcas profundas e de longo prazo na economia mundial.

Assim, dada a atual crise económica que se vive, as instituições financeiras foram obrigadas a adotar políticas de concessão de crédito mais rígidas e aplicá-las de forma mais rigorosa.

Nesta dissertação, o foco será procurar responder quais são as variáveis que os bancos consideram mais relevantes para procederem à concessão de crédito aos clientes, de maneira a reduzirem ao máximo a possibilidade de virem a sofrer de incumprimento.

Assim sendo, para proceder a este estudo, através do *Python*, foi usada a regressão logística, que prevê uma variável dependente, que é binária (tendo como resposta a concessão (1) ou não-concessão (0) do empréstimo bancário), dado um conjunto de variáveis independentes que causam a tomada de decisão.

Para se atingir o objetivo desta dissertação, o presente trabalho será dividido em três partes.

O capítulo 1 corresponde à revisão da literatura a respeito da concessão de crédito, recorrendo a alguns artigos empíricos considerados mais relevantes. São referidas algumas variáveis consideradas importantes, para haver lugar à concessão de empréstimos bancários e ainda é abordado a importância do *machine learning*, nas instituições de crédito.

O capítulo 2 diz respeito ao enquadramento teórico. Este capítulo elucida-nos da realidade pandémica e dos efeitos da mesma, nos dias que correm.

O capítulo 3, o último, refere-se ao estudo empírico, isto é, é neste capítulo que se tenta dar resposta ao objetivo principal desta dissertação.

A base de dados usada neste trabalho tem as observações simuladas e como tal, também é aqui explicado, como foi obtida/criada. É constituída por treze variáveis independentes – como por exemplo: *age*, *job* e *loan* – e uma variável dependente: *housing*. A variável dependente é binária e representa a

concessão ou não, do empréstimo bancário. Apesar de os dados serem sintéticos, o conteúdo desta dissertação é um exercício fundamental pois, este, é um dos problemas mais importantes que um banco tem de lidar diariamente.

Para dar resposta ao problema principal deste estudo é necessário ter em conta modelos econométricos adaptados aos algoritmos de aprendizagem, dirigidos por grandes bases de dados.

O objetivo deste estudo empírico, passa por entender qual é o impacto das variáveis no modelo. Assim sendo, primeiramente será analisada a matriz de correlação de *Pearson*, onde se obtém alguns resultados primários que serão essenciais para entender a correlação entre a variável alvo e as variáveis independentes. Posteriormente recorre-se ao *Recursive Feature Elimination* (RFE) de maneira a identificar as variáveis com menos impacto no modelo. Seguidamente, são analisados os *outputs* obtidos pela extensão *Jupiter Notebook*, do software *Anaconda Navigator*, antes e depois, da remoção das variáveis sugeridas pelo método *Recursive Feature Elimination* (RFE). Desta análise, consoante o melhor resultado obtido, fica a saber-se qual o modelo de regressão que pode ser validado, ficando necessariamente a saber, quais as variáveis determinantes na concessão de um empréstimo bancário.

Por último, segue-se a conclusão onde são sumarizados os principais resultados obtidos.

1 REVISÃO DA LITERATURA

Política monetária

A manutenção da estabilidade dos preços ou, por outras palavras, a manutenção do poder de compra da moeda, é o principal objetivo da política monetária de um país. A autoridade responsável pela sua definição e implementação, na área do euro, é o Eurosistema. (BdP, 2020)

Gameiro e Sousa (2010) estudaram qual o comportamento dos agentes económicos, a um choque na política monetária, na economia portuguesa. Neste estudo, concluíram que adotar uma política monetária contracionista, promove o aumento das necessidades de financiamento, por parte das sociedades não financeiras e por particulares. Este aumento de financiamento deve-se ao choque sentido pelos agentes económicos, ou seja, na sua dificuldade em se adaptar rapidamente a novas medidas económicas. Os autores defendem que, as empresas não financeiras e os particulares sentem alguma dificuldade em ajustar rapidamente as suas despesas face ao choque, o que leva a uma diminuição na aquisição de ativos financeiros e, consequentemente, um aumento dos seus passivos financeiros.

PIB e Taxas de Juro

Friedman e Kuttner (1993) mencionam que normalmente os estudos baseados na procura de crédito usam variáveis macroeconómicas tais como, o PIB e as taxas de juro. Referem que é habitual a utilização de modelos econométricos para o estudo, nos quais são utilizadas observações de series temporais longas.

O produto interno bruto, também designado como PIB, corresponde ao conjunto de todos os bens e serviços que geram valor, estejam eles relacionados com a produção, a compra, o investimento ou a exportação, ou seja, é a riqueza que um país consegue criar. É uma medida que avalia o desempenho da economia de um país e serve de comparação a outros países. (PORDATA, 2020)

Hofmann (2001) considera que as variáveis PIB real e a taxa de juro real não são suficientes para estudar o desenvolvimento do crédito, a longo prazo e, como tal, adiciona a variável preços dos imóveis.

Segundo o autor, o aumento no PIB real tem um efeito positivo no crédito. Em relação aos preços dos imóveis, o autor considera que existe uma forte ligação positiva e bidirecional com a procura de crédito, isto é, o aumento no preço dos imóveis impulsiona o crédito e vice-versa.

O mesmo não se pode dizer para a taxa de juro real, uma vez que esta tem efeitos significativamente negativos sobre o crédito.

Brzoza-Brzezina (2005) no seu artigo analisa a possível existência de *booms* de crédito e os desenvolvimentos adversos no setor bancário, que resultaram do processo de integração monetária de alguns Estados Membros na União Europeia.

Em países como a Irlanda e Portugal verificou-se aumentos substanciais nas cedências de crédito, nos anos imediatamente anteriores e posteriores, à adesão à UE. Em Portugal as taxas de crescimento dos empréstimos reais ultrapassaram os 25%.

Verificou-se que, em todos os países, o aumento do PIB fez aumentar o crédito. O autor explica que o aumento do crédito na economia, acontece normalmente após crises do setor bancário.

Relativamente à acentuada queda nas taxas de juro de referência do BCE, salienta ainda que, este fator faz aumentar o crédito e, consequentemente, impulsiona o crescimento económico.

Empréstimos

Estamos perante um empréstimo bancário, isto é, um contrato de crédito, quando uma instituição de crédito concede dinheiro a um cliente. Neste processo, a instituição de crédito assume o papel de credor/mutuante e o cliente adquire o papel de devedor/mutuário. O cliente tem a obrigação e o dever, de devolver a quantia total, acrescido de encargos, tais como juros e outros custos, numa data acordada na celebração do contrato. (BdP, 2020)

Segundo consta no site do Banco de Portugal, estão autorizadas a conceder crédito aos mutuários, todas as instituições de crédito e algumas sociedades financeiras, desde que estejam registadas no Banco de Portugal.

Moradi e Rafiei (2019) definem um banco como um tipo de instituição financeira, que presta serviços. Dentro destes serviços está compreendido a concessão de empréstimos bancários, a aceitação de depósitos e ainda a oferta de produtos de investimento básico.

Odegua (2020) considera que os empréstimos são concedidos, em geral, a indivíduos/consumidores, por bancos ou instituições financeiras, numa data acordada e normalmente concedidos para muitos propósitos, tais como negócios, saúde, estudos ou até mesmo para uso pessoal.

Risco de crédito

O risco de crédito é um dos riscos mais comuns, quando se fala de uma instituição financeira.

Corresponde ao risco associado à incapacidade de o cliente cumprir com as suas obrigações financeiras acordadas, perante a instituição financeira.

Maradi e Rafiei (2019) definem assim o risco de crédito, como sendo a probabilidade de não pagamento, de atraso no pagamento ou da incapacidade de reembolsar um empréstimo, por parte dos clientes. Consideram um risco muito importante de analisar e de ter em conta, visto que a sobrevivência de um banco está diretamente relacionada com risco.

Assim, as instituições financeiras controlando da melhor forma o seu risco de crédito, podem gerir e garantir a sua sustentabilidade e estabilidade financeira.

Como é de esperar, quanto maior for o risco de crédito associado a um cliente, maior é a probabilidade de o cliente não cumprir com as suas obrigações. Do mesmo modo que, quanto menor for o risco de crédito associado a um cliente, maior é a probabilidade de o cliente saldar as suas dívidas para com o banco.

Munkhdalai et al. (2019) consideram que um sistema de pontuação de crédito seria algo fundamental a adotar em instituições de crédito, de modo a satisfazer um princípio de perda mínima para a sua sustentabilidade. Seria um sistema de pontuação que apoiaria a tomada de decisão, gerindo potenciais riscos, de maneira a maximizar a estabilidade financeira da instituição.

Os avanços tecnológicos e a consequente automatização vieram ajudar as instituições de crédito, pois, tornar-se-ia necessário substituir agentes de crédito e modelos de pontuação de crédito - também feitos por especialistas de crédito - uma vez que iria evitar julgamentos, erros humanos e perdas de oportunidade ou perdas de crédito e ainda, reduzir os custos operacionais, num processo de tomada de decisão, sobre um empréstimo.

Munkhdalai et al. (2019) referem que, embora seja difícil executar um sistema eficiente e automatizado para estimar a credibilidade dos clientes, o *machine learning* veio desempenhar um papel fundamental na avaliação de crédito de clientes bancários. Assim, o que anteriormente era feito por especialista de crédito, passou a ser feito por *machine learning*, reduzindo tempo e dinheiro, nas tomadas de decisão.

Moradi e Rafiei (2019) consideram que os bancos devem garantir que os mutuários conseguem liquidar as prestações totais, antes de lhes conceder o empréstimo e para isso, a avaliação de risco de crédito é fundamental no processo de tomada de decisão.

Salientam que, houve imensa procura, por parte das instituições de crédito, de sistemas de pontuação de crédito capazes de analisar e modelar com precisão, o risco de cada mutuário. Assim, o objetivo é que estas técnicas de pontuação de crédito sirvam de suporte, à tomada de decisão, para uma vasta gama de clientes.

No entanto, apontam uma desvantagem importante, estes modelos são incapazes de funcionar de forma eficiente em crises, tanto financeiras, como económicas e também políticas. Apontam que os bancos usam estruturas de modelagem estáticas que não são capazes de dar resposta à evolução económica, tornando os modelos ineficientes.

Desse modo, os critérios do modelo precisam de ser constantemente atualizados, pois corre-se o risco, por exemplo, de clientes com alguma probabilidade de *default* continuarem classificados como "bons clientes" - clientes a quem o banco decide conceder o empréstimo. O risco destes "bons clientes" não cumprirem as obrigações acordadas, após uma crise, por exemplo, torna-se mais elevado, tornando-se prejudicial para instituições de crédito.

Contudo, Moradi e Rafiei (2019) consideram a regressão logística, as análises abrangentes de dados e a inteligência artificial, como modelos estáticos de qualidade. Estes modelos são obtidos através de dados demográficos, o que ressalva a ideia de seguirem um padrão estático.

Segundo Odegua (2020) dado o aumento de solicitações de empréstimos, o aumento de concorrência e a quantidade de dados disponíveis, as instituições financeiras devem criar modelos que possam ajudar a minimizar a probabilidade de *default*.

Reforça também a necessidade da existência de modelos eficazes de *machine learning*, que possam ajudar a capturar padrões importantes nos dados de crédito.

É de opinião que, através de vários estudos, as instituições financeiras conseguem identificar fatores importantes que podem estar ligados à falta do pagamento dos empréstimos, por parte dos mutuários, de modo a maximizar o seu lucro. Através destes estudos, conseguem obter informações sobre o comportamento, padrões de consumo e algumas caraterísticas comuns de indivíduos que solicitam crédito.

A escolha do modelo a usar, por parte das instituições de crédito, é algo extremamente importante na determinação da precisão, exatidão e eficiência de um sistema de previsão. Não existindo ainda o melhor modelo para sistema de previsão, dado que cada caso é um caso e cada banco tem a sua política de funcionamento diferente, então existem alguns melhores e mais usados que outros.

Alguns autores, identificam *neural network*, *random forest*, regressão logística, CART, entre outros, como os modelos mais usados.

Odegua (2020) neste estudo utiliza um *gradient boosting algorithm XGBoost* para estudar e analisar um conjunto de dados de empréstimos bancários, de maneira a prever o *default* de empréstimos bancários. Utilizando a linguagem de programação *Python*, afirma que os recursos mais importantes usados no modelo, para prever o *default* de um cliente, são a idade e a localização do cliente.

Existem várias teorias sobre quais as variáveis determinantes para a concessão do crédito, entre as quais, a idade do cliente, a sua atitude, o seu rendimento, se tem habitação, qual a sua ocupação, se tem contrato de trabalho, se possui conta poupança, o local onde vive, se paga uma renda mensal, a duração do empréstimo, entre outras.

Em concordância com vários casos práticos analisados na literatura científica, os resultados aqui encontrados referem que as variáveis com mais peso na decisão do *default* são: a idade do cliente, a sua localização e o seu rendimento.

Nota-se que, a regressão logística faz o seu papel de forma precisa e salienta-se ainda, a simplicidade e rapidez deste algoritmo.

2 ENQUADRAMENTO TEÓRICO

Como já foi referido anteriormente, 2020 é um ano desafiante e de certo modo crítico, em todos os setores.

O ano de 2020 fica marcado pelo princípio de uma pandemia mundial, o Covid-19.

Conforme mencionado no capítulo de revisão da literatura, a comunidade em geral tem alguma dificuldade em lidar com situações de choque, ou seja, tem dificuldade em adaptar-se e ajustar-se rapidamente a mudanças repentinas.

Este é então, um assunto preocupante no que toca à economia no mundo, isto é, o Covid-19 veio trazer novas realidades e muitas dificuldades em todo o mundo.

Todo este capítulo foi escrito com base em informações do Boletim Económico de Outubro de 2020, do Banco de Portugal.

PIB

Esta pandemia gerou um choque económico global e segundo o Fundo Monetário Internacional (FMI) terá atingido 90% das economias.

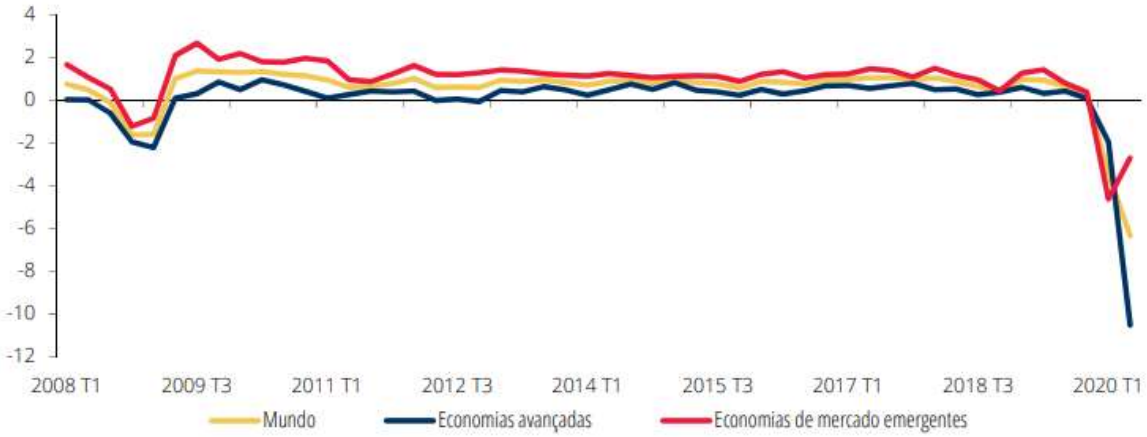

Figura 2.1 - PIB mundial

No período em analise, verifica-se uma forte quebra no PIB mundial, no primeiro trimestre de 2020.

O receio dos contágios, da incerteza/dúvida em torno do que realmente é este vírus, das medidas de confinamento impostas pelos governos e do distanciamento social, resultaram na queda da atividade económica mundial.

A quebra na atividade económica teve início no mês de março, atingindo o nível mínimo no mês de abril. Em maio regista-se uma melhoria, que se prolonga em junho, dado o alívio gradual das restrições impostas, devido ao abrandamento do contágio pelo Covid-19.

Apesar disso, a atividade económica manteve-se em níveis ainda abaixo do observado no período homólogo, do ano anterior.

No entanto, as medidas de política monetária, orçamental, prudencial e de supervisão impostas pelos governos, amenizaram o choque sobre a economia e criaram condições para a recuperação da atividade.

Fazendo também uma análise a nível da zona Euro, surge-nos o seguinte gráfico:

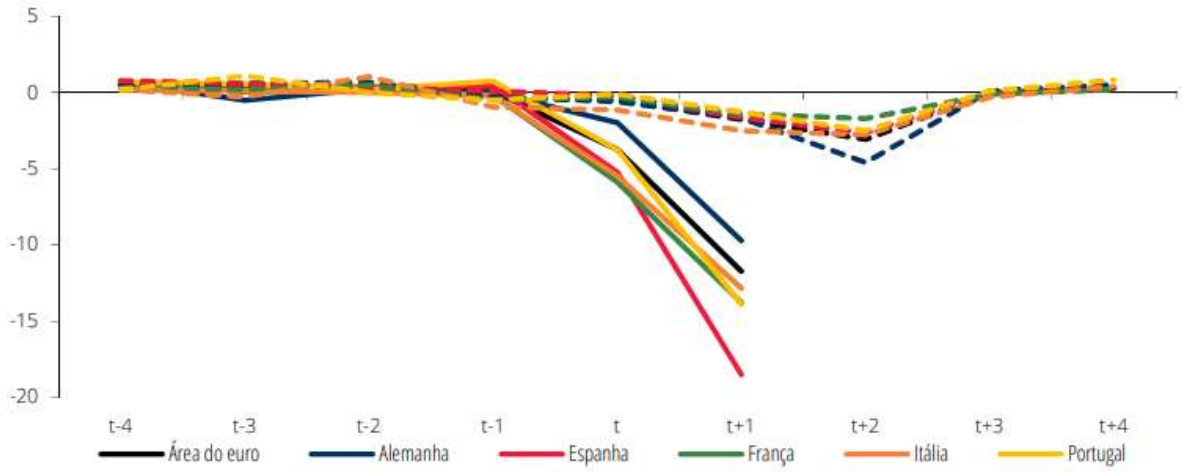

Figura 2.2 - Crescimento do PIB na área do euro

As linhas a cheio correspondem à crise pandémica e as linhas a tracejado correspondem à crise financeira global. O momento "t" refere-se ao trimestre em que o choque ocorre, isto é, nas linhas a tracejado corresponde ao terceiro trimestre de 2008 e nas linhas a cheio corresponde ao primeiro trimestre de 2020.

Observa-se, em ambos os gráficos, que a crise financeira global de 2008 teve um enorme choque na economia, mas é notório que, a crise pandémica de 2020 superou (de forma negativa) esse impacto.

As quebras na atividade e a deterioração da situação no mercado de trabalho levou à redução do investimento.

Emprego

Como seria de esperar houve uma forte redução nas horas de trabalho efetivo, devido a todas as medidas de contenção tomadas.

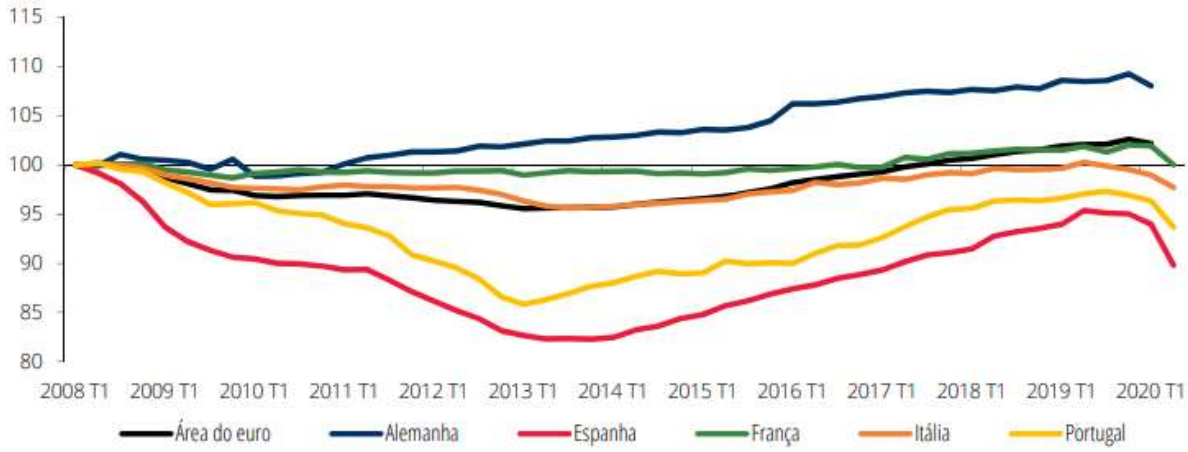

Figura 2.3 - Emprego na área do euro

Segundo a Organização Internacional do Trabalho a queda das horas trabalhadas em todo o mundo, atingiu a ordem dos 5.4% no primeiro trimestre e de 14% no segundo trimestre deste ano.

Alguns governos adotaram medidas de apoio ao mercado de trabalho.

Taxa de Poupança, Consumo nominal e Rendimento disponível nominal

Dado as restrições impostas, houve uma forte quebra no consumo privado diferente de recessões passadas, mais concretamente, nas despesas de hotéis e restaurantes, em transportes e consequentemente nos combustíveis, atividades de lazer que implicam interação pessoal e ainda na aquisição de vestuário e calçado, por exemplo.

Isto traduz-se numa estabilização do rendimento disponível e no aumento significativo da taxa de poupança.

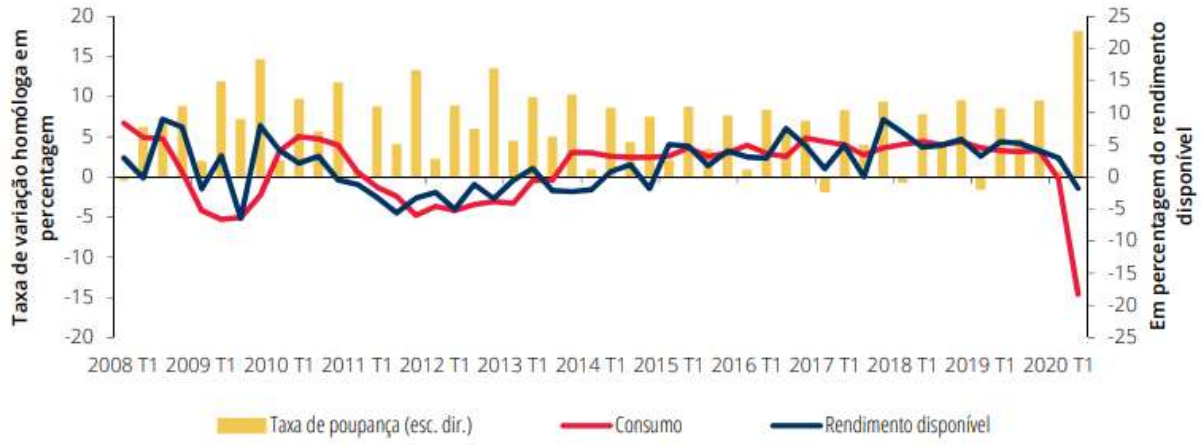

Figura 2.4 - Taxa de poupança, consumo nominal e rendimento disponível nominal

Uma das medidas impostas foi o distanciamento social, o que afetou grande parte das despesas habituais das famílias. Isto gera um comportamento de poupança forçada ou involuntária, por parte das mesmas.

Também é de esperar que num ambiente de incerteza e pessimismo dos consumidores, as famílias estejam mais cautelosas para momentos de grandes dificuldades e aumentem as suas poupanças.

Taxas de juro

Apesar das taxas de juro atingirem um mínimo histórico, houve um aumento dos depósitos dos particulares, justificado com o que foi dito anteriormente sobre o aumento significativo da poupança das famílias e da incerteza do amanhã.

Contudo este aceleramento dos depósitos, por parte dos particulares, é temporário, uma vez que esta poupança é "forçada" pelas medidas de confinamento e de distanciamento social.

Este aceleramento dos depósitos também aconteceu por parte das empresas, todavia pode ser justificado pelos empréstimos contratados nesse período.

Em Portugal, entre março e junho de 2020, os meses do choque inicial e talvez os mais críticos, as taxas de juro dos empréstimos, tanto para a habitação, como para o consumo permaneceram baixas.

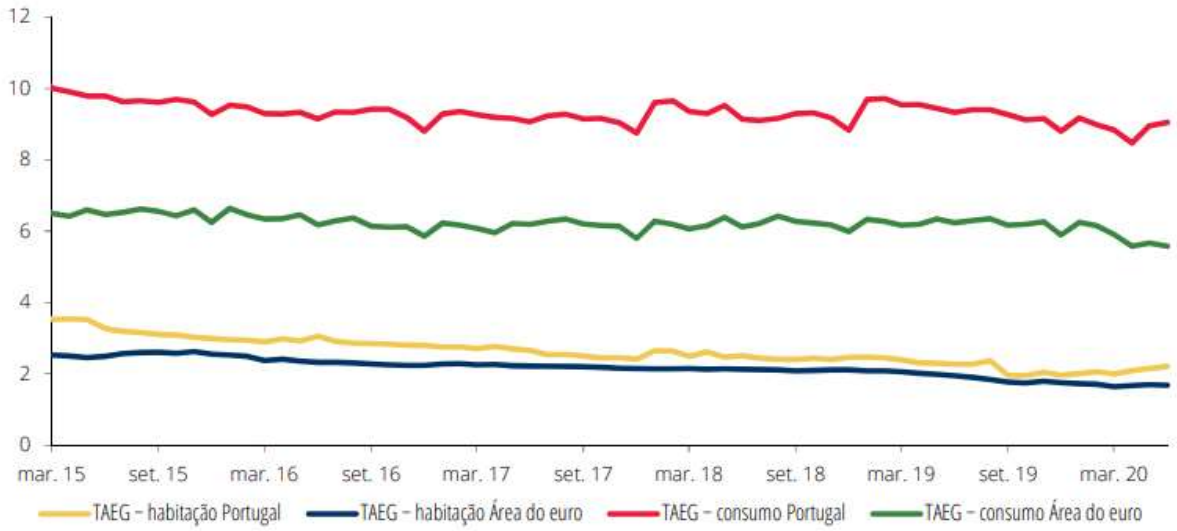

Figura 2.5 - TAEG dos novos empréstimos a particulares para habitação e consumo

TAEG diz respeito à taxa anual de encargos efetiva e global, isto é, representa o custo total do crédito para o consumidor, incluindo os juros e outros encargos que o consumidor tem de pagar pelo crédito.

Crédito

Como seria de esperar, dada toda a situação que se vive, os bancos portugueses tomaram uma atitude mais cautelosa/defensiva, para se protegerem de possíveis perdas financeiras. Assim, tornaram-se ainda mais criteriosos na concessão de crédito a particulares, no segundo trimestre do ano.

Tudo isto resultou do aumento do risco associado à situação pandémica e consequente redução da tolerância ao risco e ainda, das perspetivas económicas abaladas pela falta de credibilidade dos mutuários.

Segundo *o inquérito aos bancos sobre o Mercado de Crédito* na área do euro (*Bank Lending Survey* – BLS) o choque pandémico causou a redução da oferta de crédito, mas também da procura de crédito, justificada pela diminuição da confiança dos próprios consumidores, pela degradação das perspetivas para o mercado imobiliário, pelo fraco investimento em consumo de bens duradouros e pelo distanciamento social e restrições de circulação, que levaram a uma grande quebra no comercio.

Aglomerando o aumento da restritividade e simultânea redução da procura e da oferta, isto conduzirá à diminuição no volume de empréstimos concedidos.

Crédito à habitação e ao consumo pós 2020

No *Bank Lending Survey* (BLS) de julho de 2020, os bancos portugueses reportaram que relativamente aos empréstimos à habitação, se tornaram muito mais exigentes no rácio entre o valor do empréstimo e o valor da garantia (*loan-to-value*). Mencionam que existiu uma maior fração de pedidos de empréstimos rejeitados no segundo trimestre do ano, do que no primeiro trimestre.

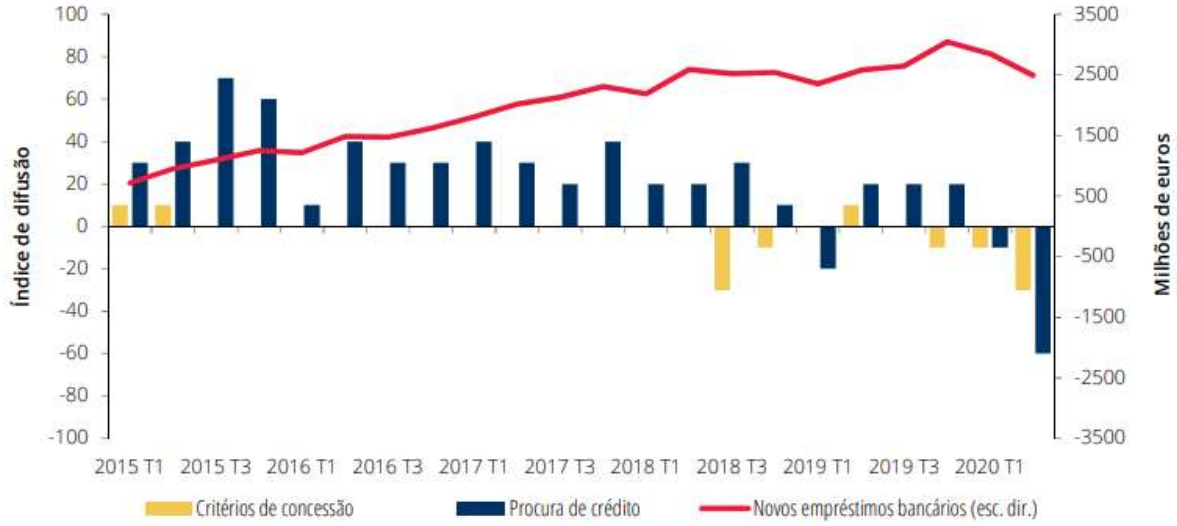

Figura 2.6 - Novos empréstimos e procura e oferta de crédito por bancos residentes a particulares para habitação

O crédito ao consumo e outros fins reduziu de forma mais significativa do que o crédito à habitação. Esta diminuição aconteceu tanto no crédito pessoal, como no crédito automóvel.

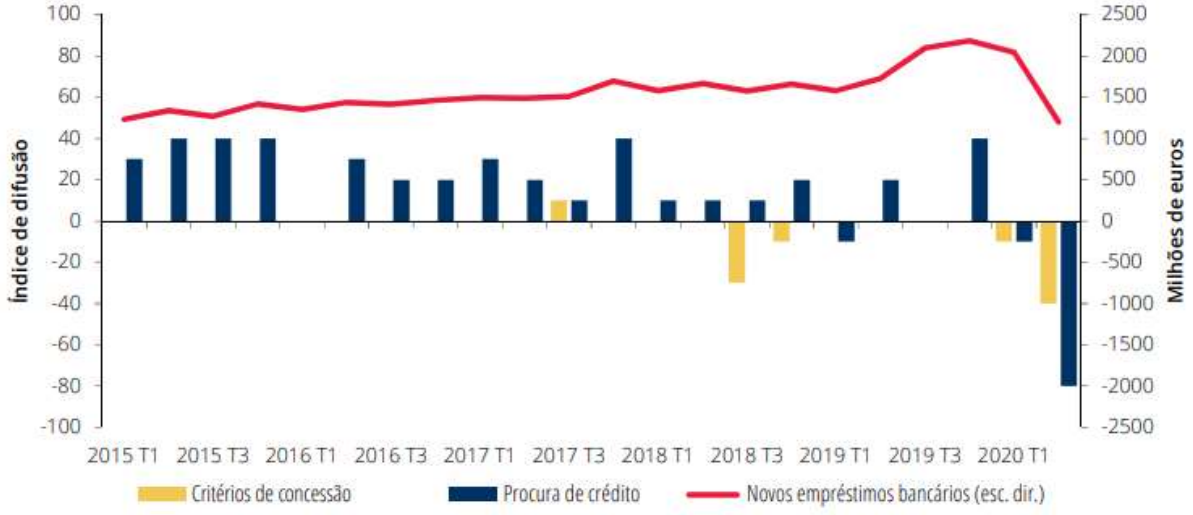

Figura 2.7 - Novos empréstimos e procura e oferta de crédito por bancos residentes a particulares para consumo e outros fins

Em termos de ordens de grandeza, no primeiro trimestre houve uma queda no crédito ao consumo na ordem dos 6.3% e de 41.4% no segundo trimestre.

A quebra foi de tal ordem que, a redução dos novos empréstimos ao consumo foi mais acentuada do que a diminuição do consumo privado (excluindo bem alimentares).

3 ESTUDO EMPÍRICO

3.1 Recolha de dados

A base de dados desta dissertação é uma simulação do mundo real.

Através do contacto direto com alguns bancos - nomeadamente a Caixa Geral de Depósitos, o Millenium BCP e o Novo Banco - é possível questionar alguns especialistas em crédito, obtendo algumas informações relativas à concessão de crédito no setor privado.

Consolidando a informação adquirida junto dos bancos, através da revisão da literatura e pesquisando/investigando dados reais disponibilizados no site oficial do Banco de Portugal e no PORDATA, surge uma base de dados constituída por 11162 observações e 14 variáveis.

O estudo labora numa base de dados, que contém informações fictícias relativas a clientes de um banco imaginário XYZ. O objetivo principal é prever quais são as variáveis determinantes para a concessão de um empréstimo bancário.

3.2 Caracterização da amostra

Antes de avançar para o processamento de dados, é importante referir quais as variáveis utilizadas neste estudo, o que ajudará na interpretação matemática que for feita.

A tabela 3.1 descreve resumidamente o conjunto de variáveis utilizadas.

Tabela 3.1 - Variáveis do modelo

Informações	Variável	Unidade	Descrição
Bank details	*Contact*	*cellular*	Tipo de comunicação utilizada para entrar em contacto com o cliente
		telephone	
		email	
	Month	*jan*	Mês em que a chamada foi efetuada
		...	
		dec	
	Client	*yes*	Se é cliente do banco
		no	
Client's data	*Age*	(numérico)	Idade do cliente
	Job	*admin.*	Emprego do cliente
		blue-collar	
		entrepreneur	
		housemaid	
		management	
		Retired	

		self-employed	
		services	
		student	
		technician	
		unemployed	
		unknown	
	Marital	*married*	Estado Civil do cliente
		divorced	
		single	
	Education	*primary*	Grau de ensinamento do cliente
		secondary	
		tertiary	
		unknown	
House conditions	*Household*	*1*	Agregado familiar do cliente
		...	
		8	
	Leasedhouse	*yes*	Vive numa casa arrendada?
		no	
Job information	*Salary*	(numérico)	Salário do cliente
	Effective	*yes*	Tem contrato sem termo?
		no	
		unknown	
Loan	*Default*	*yes*	Encontra-se em *default*?
		no	
	Loan	*yes*	Tem crédito pessoal?
		no	
Output variable (desired target)	*Housing*	*yes*	Tem crédito habitacional?
		no	

Na revisão de literatura são enunciadas algumas variáveis que os autores consideram determinantes para a concessão de crédito bancário, que foram tidas em conta, para a elaboração da base de dados.

3.2.1 *Client*

Conforme apurado junto de alguns especialistas em crédito, não é imperativo ou obrigatório que o mutuário seja cliente no próprio banco.

No caso de o mutuário ser cliente, toda a sua informação é de fácil acesso. Para além disso, consegue-se analisar com mais facilidade todo o seu histórico, o que poderá dar conforto nas decisões de concessão de novo crédito.

Contrariamente, se o mutuário não for cliente, então este terá a obrigação de trazer e prestar informações junto do banco. Isto é pouco creditício, uma vez que, poderá haver pouca informação do histórico do cliente, o que dificulta a sua classificação ao nível do incumprimento.

Fazem parte desta amostra 7477 indivíduos que são clientes do banco.

3.2.2 *Age*

Após a leitura de alguns artigos, alguns autores consideram que a faixa etária que mais solicita crédito é dos 20 aos 40 anos.

Ao questionar alguns especialistas em crédito, chega-se à conclusão de que, o crédito hipotecário é mais solicitado por pessoas entre os 30 e os 50 anos e o crédito pessoal é solicitado em maior quantidade, por pessoas entre os 20 e os 30 anos.

Na figura 3.1 encontra-se dados relativos à categoria *age*, da amostra de dados:

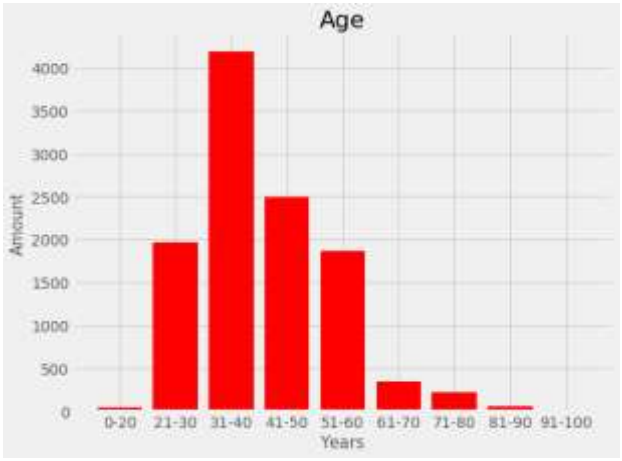

Figura 3.1 - Idade

Como se vê na figura acima, os dados estão de acordo com a informação adquirida.

A faixa etária dos 0 aos 20 e dos 61 aos 100 anos são idades nas quais são solicitados menos empréstimos bancários. Analisando esta informação, é algo que faz sentido, uma vez que os jovens dos 0 aos 20 anos, vivem normalmente com os seus pais e, portanto, não necessitam de crédito. Para além disso, entre os 61 anos e, impondo um limite de 100 anos, não é tão comum a solicitação de crédito, isto porque, depois de uma vida inteira de trabalho, estes indivíduos têm o seu "pé-de-meia" criado, não necessitando de financiamento extra.

3.2.3 *Marital*

De acordo com os especialistas de crédito bancário, o estado civil do cliente é importante na cedência de crédito, com especial incidência no crédito à habitação e no crédito pessoal. Consideram importante pois, de acordo com o tipo de crédito, poderá ser necessário a intervenção do cônjuge por questões patrimoniais e de garantias intrínsecas dos financiamentos, isto é, possibilidade de os bancos executarem o património dos clientes em caso de incumprimento.

A categoria *marital* é dividida em três classes: *married*, *divorced* e *single*. A classe *divorced* diz respeito a indivíduos divorciados ou viúvos.

Na amostra existem 56.9% de indivíduos que estão casados, 11.58% de indivíduos que se encontram divorciados ou viúvos e, por último 31.52% de indivíduos solteiros.

3.2.4 *Education*

De acordo com o verificado junto dos bancos, o nível de escolaridade/literacia é uma das variáveis dos motores de *scoring*, isso porque, o nível de escolaridade influencia o nível de risco e, consequentemente, os montantes a emprestar e o preço/*spread* a praticar. Conforme justificado, mais habilitações está estatisticamente associado a melhores empregos e, logo, melhor capacidade para obter crédito.

Fazem parte dos dados 49.06% de indivíduos que têm o ensino secundário concluído, 33.05% de indivíduos que chegaram mais longe e têm o ensino universitário concluído e 13.44% dos indivíduos tem apenas o ensino básico concluído.

3.2.5 *Household*

Os especialistas de crédito dos bancos também consideram a variável *household* essencial na cedência de crédito a particulares.

Esta variável influencia a Avaliação da Solvabilidade dos Consumidores (ASC) e o cálculo das Despesas Regulares do Consumidor (DRC).

A ASC (Avaliação da Solvabilidade dos Consumidores) nas operações de Crédito Pessoal e especializado, concretiza os procedimentos e critérios a observar na avaliação da capacidade e propensão de o consumidor cumprir as obrigações decorrentes do contrato de crédito.

Para efeitos da determinação da ASC são considerados dois parâmetros relevantes que entram na sua elaboração:

- DSTI – *Debt Service to Income*
- DRC – Despesas Regulares do Consumidor

A DRC (Despesas Regulares do Consumidor) diz respeito a despesas suportadas pelo RAL – rendimento anual líquido de impostos sobre o rendimento e prestações socias – e que se enquadram nas seguintes categorias:

- Despesas Gerais Familiares
- Saúde
- Educação
- Habitação (inclui impostos)
- Restauração e alojamento
- Automóvel (inclui impostos) e transportes.

Posto isto, esta é uma variável muito importante na tomada de decisão de cedência de crédito.

Para a constituição desta variável, usa-se ainda dados provenientes do site PORDATA, que nos elucida que a dimensão média dos agregados domésticos privados ronda os 2.5, no ano de 2019.

Ainda através dos dados do site, sobre o tipo de agregado doméstico privado, consegue-se saber, em média, qual a quantidade de indivíduos que vivem sozinhos, a quantidade de casais que vivem com e sem filhos, entre outros dados.

Por conseguinte, nos dados existem até 8 classes de agregado familiar sendo que, 34% diz respeito a um agregado de 3 indivíduos, 25% é relativo a um agregado de 2 indivíduos, 23% dos indivíduos vivem sozinhos, entre outros.

3.2.6 *Leasedhouse*

Contrariamente ao que acontecia em meados da década de 60, existem mais ocupantes proprietários do que inquilinos.

Segundo o site PORDATA, no ano de 2011 - que diz respeito ao último dado exibido - a diferença percentual entre viver numa casa própria e viver numa casa arrendada é quase o triplo.

Como tal, a base de dados é constituída por 75.20% de indivíduos que têm alojamento próprio e de 24.80% de indivíduos que vivem em casas arrendadas, como se pode ver na figura 3.2:

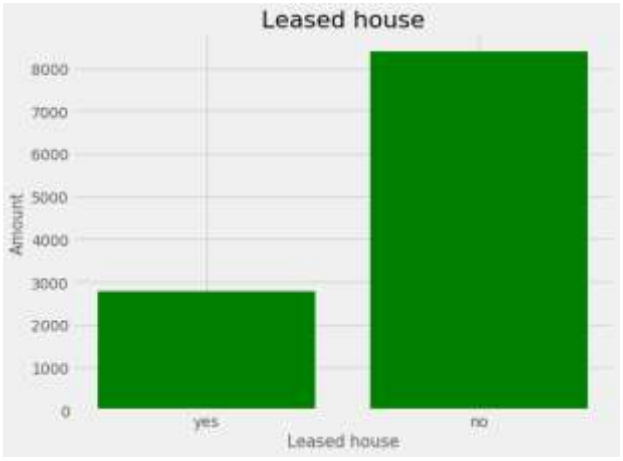

Figura 3.2 - Casa arrendada

3.2.6.1 *Leasedhouse* vs. *Housing*

Há nesta amostra 8394 indivíduos que não declaram que vivem numa casa arrendada. Portanto se existem 5281 proprietários de um empréstimo habitacional, então 3113 dos indivíduos deste estudo vivem em casa própria, mas não têm empréstimo bancário. Isto acontece por inúmeros motivos, tais como, estes indivíduos não precisaram de pedir empréstimo à habitação para comprarem uma casa ou até, nas melhores das hipóteses, já têm a casa totalmente paga, podem também ser indivíduos que vivem com alguém que possui o empréstimo, pode também ser uma casa herdada, como muitas outras possibilidades. Nisto, ao somar os 3113 indivíduos com as 2768 pessoas que declaram que vivem numa casa arrendada, obtém-se as 5881 pessoas que não contraem nenhum crédito habitacional.

3.2.7 *Salary* e *Job*

De acordo com os especialistas em crédito, o rendimento atual do cliente é uma das variáveis decisivas na tomada de decisão de cedência de crédito.

Esta variável entra para o cálculo do RAL (Rendimento Anual Líquido), que é uma das variáveis que vai estar na construção da DSTI (*Debit Service to Income*).

O *Debit Service to Income* é a relação entre o valor da primeira prestação mensal do período de reembolso, adicionada ao valor mensal dos seguros obrigatórios e de todas as responsabilidades mensais dos intervenientes em operações de crédito (a particulares), e um duodécimo do (RAL) rendimento anual líquido (de impostos associados ao rendimento e de prestações sociais obrigatórias) do agregado familiar.

A base de dados criada tem em conta a variável *job*, pois cada emprego tem a sua remuneração especifica. Assim temos:

Tabela 3.2 - Salário de cada trabalhador

Cargo	Remuneração mínima	Remuneração máxima
Administration	3000€	6000€
Blue-collar	635€	1000€
Entrepreneur	635€	3000€
Housemaid	635€	800€
Management	1000€	4000€
Retired	250€	2000€
Self-employed	500€	1500€
Services	1000€	2000€
Student	0€	0€
Technician	635€	1300€
Unemployed	0€	0€
Unknown	*unknown*	*unknown*

Também no site PORDATA, existe a informação de que o ganho médio mensal, do ano de 2018, é de €1170.3, o que se encontra de acordo com a base de dados deste estudo, como é observável na figura seguinte 3.3:

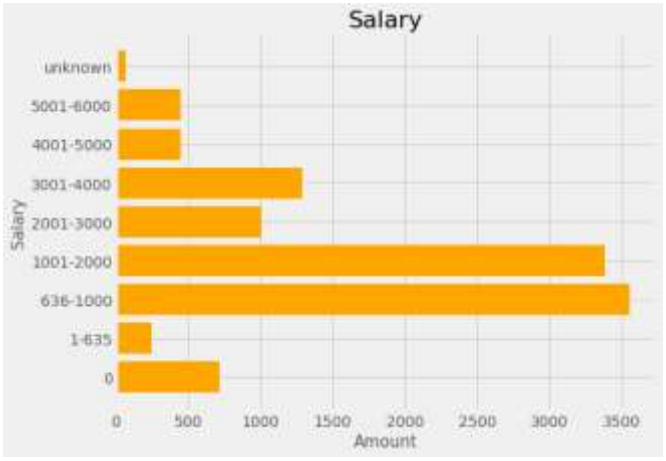

Figura 3.3 - Salário

3.2.8 *Effective*

O tipo de contrato do individuo, isto é, se tem um contrato permanente/sem termo ou um contrato a prazo/a termo, é um fator importante na cedência de crédito a clientes bancários.

Segundo o PORDATA, desde meados da década de 80 até à atualidade, sempre houve mais trabalhadores efetivos do que trabalhadores a prazo. A proporção de trabalhadores com contrato sem termo ronda em grosso modo os 70/80%.

A amostra de dados é constituída por 70% de indivíduos com contratos sem termo, trabalhadores efetivos, como é observável na figura 3.4:

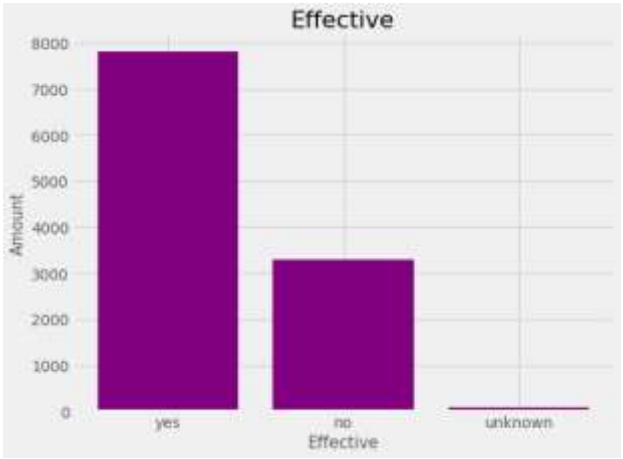

Figura 3.4 - Trabalhador efetivo

3.2.9 *Loan* e *Housing*

Segundo o Banco de Portugal, o crédito à habitação e o crédito aos consumidores são as principais categorias de crédito a que recorrem os clientes bancários e, como tal, foram estas as categorias de empréstimo escolhidas para fazer parte do estudo nesta dissertação.

Também ainda junto de especialistas em crédito chega-se à conclusão de que o crédito mais solicitado e também mais concedido, em termos de quantidade de operações, é o cartão de crédito, posteriormente o crédito à habitação e por último o crédito pessoal.

Já em termos de montante de operações, o crédito mais solicitado e também mais concedido é o crédito à habitação, de seguida o crédito pessoal e por fim os cartões de crédito.

Dado todas as variáveis que constituem a amostra, a variável *housing*, que diz respeito ao crédito à habitação, foi a escolhida como a variável *target* do modelo, descrita na figura 3.5:

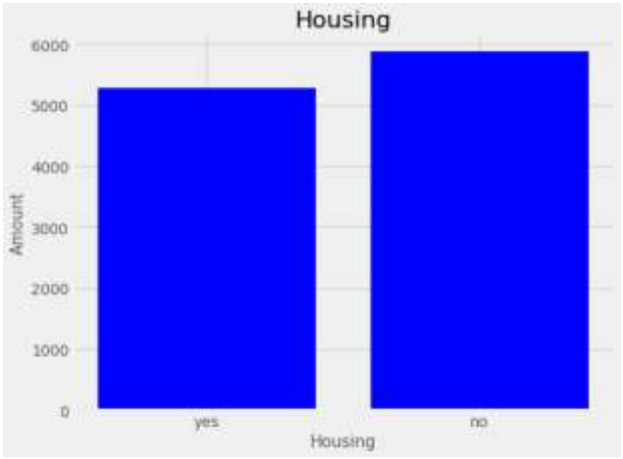

Figura 3.5 - Empréstimo à habitação

Neste estudo existem 5281 indivíduos a quem lhes foi concedido um crédito habitacional, o que perfaz um total de 47.31% dos dados, como se constata na figura 3.5.

3.3 Modelos econométricos e algoritmos de *Machine Learning*

É nesta secção que se irá falar sobre o modelo econométrico escolhido para fazer face aos objetivos desta dissertação, nomeadamente a regressão logística numa abordagem *data-driven*. A escolha do algoritmo baseia-se na sua simplicidade, performance e rapidez de execução e atualização.

Existem vários modelos econométricos e algoritmos de *machine learning* que abordam os problemas de classificação, contudo serão apresentados os mais adequados, para este caso. Com auxílio da revisão da literatura do Capítulo 1, serão apresentados alguns modelos.

3.3.1 *Random forest*

Um modelo de *machine learning* que poderia ser utilizado é o algoritmo *Random forest* (ou floresta aleatória) que, como o próprio nome o sugere, cria combinações (*emsemble*) de árvores de decisão.

É um modelo muito utilizado para problemas de classificação e também de regressão.

As árvores de decisão assumem a forma de um fluxograma com "nós", isto é, no topo existe uma decisão principal verificada e dessa decisão irão nascer ramos, formando novos "nós" na árvore de decisão. Posteriormente, destes novos nós, irão crescer outros ramos, e assim sucessivamente.

Ao utilizar *Random Forest*, o algoritmo irá selecionar aleatoriamente amostras do conjunto de treino.

Seguidamente, o algoritmo irá escolher, de forma também aleatória, a variável que irá constituir o primeiro nó da árvore de decisão.

Posteriormente, a partir dessa variável escolhida para o primeiro "nó", o algoritmo irá escolher - novamente de forma aleatória - quais as variáveis que farão parte do segundo "nó", excluindo as que já foram selecionadas anteriormente.

Este processo irá repetir-se sucessivamente até ser construído o último "nó" da árvore de decisão.

Por fim, após criada cada árvore do modelo, é feita a tomada de decisão a partir dos dados apresentados. A resposta mais frequente, é a resposta do algoritmo, como sugere a figura 3.6:

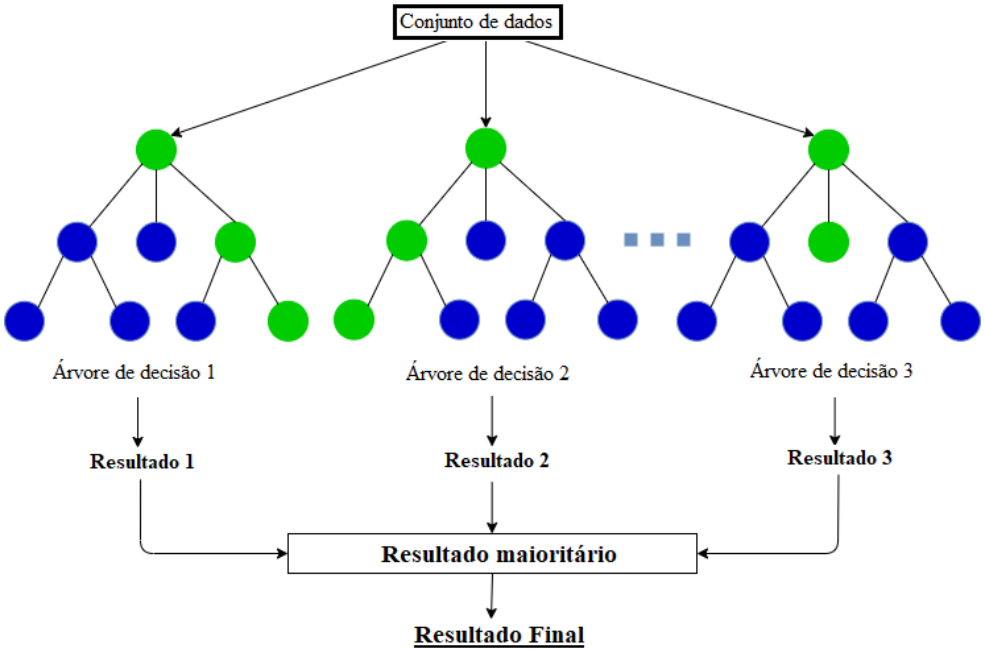

Figura 3.6 - *Random Forest*

Visto que todas as decisões são feitas aleatoriamente, existe o risco de o algoritmo escolher a pior variável para o primeiro "nó" da árvore de decisão e assim, comprometer o sucesso do método.

Para isto não acontecer, é aconselhável que sejam construídas várias árvores de decisão, para se atingir os melhores resultados.

Este método trabalha com variáveis binárias/categóricas e também numéricas e labora perfeitamente, mesmo com grandes volumes de dados, sendo estas as duas grandes vantagens da utilização da *Random Forest*.

3.3.2 Regressão Linear *vs.* Regressão Logística

Entenda-se a regressão logística como uma versão da regressão linear, que é utilizado quando queremos categorizar a variável *target*.

Para se entender melhor as diferenças da utilização da regressão logística, ao invés de regressão linear, vamos atentar o seguinte exemplo.

Supondo, como acontece neste estudo, que se está a decidir se, se deve ou não conceder crédito a um cliente, com base nas informações gerais que um banco possui do mesmo.

Para simplificar, considera-se que a probabilidade de o cliente realizar *default* só depende do tempo de permanência, em anos, no banco. Existe o acesso a essa informação de tempo de permanência, mas não se sabe como se relaciona com a probabilidade de o cliente praticar *default*.

Se, se retirar uma amostra de 10 clientes do banco, que contém a informação do tempo de permanência em anos, no banco e a informação de, se realizou ou não *default* e, posteriormente, se colocar os dados num gráfico (x,y), ele tomaria a forma apresentada na figura 3.7:

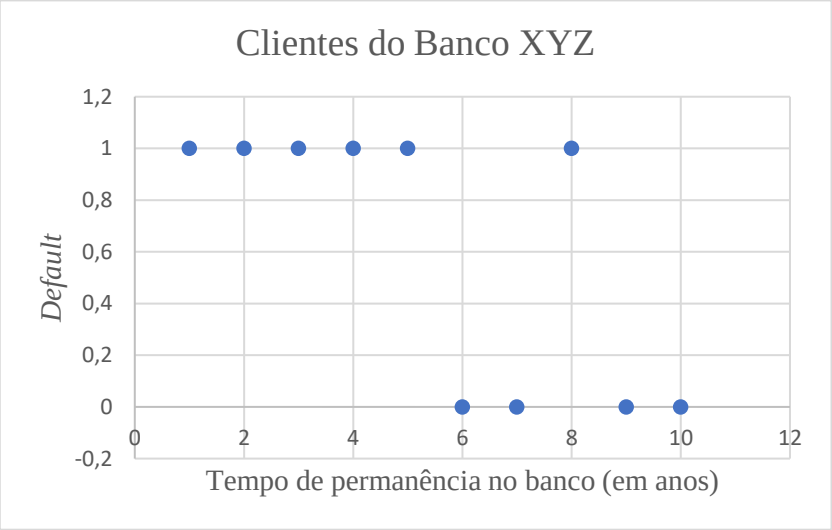

Figura 3.7 - Clientes do Banco XYZ

No eixo y, assume-se que o valor 1 condiz a "o cliente cometeu *default*" e, como contrapartida, o valor 0 diz respeito a "o cliente não cometeu *default*". O eixo do x corresponde ao tempo, em anos, em que o mutuário faz parte do leque de clientes do banco.

Ao observar o gráfico pode concluir-se que, a maioria dos mutuários que cometeu *default* é cliente do banco há relativamente pouco tempo.

No entanto, o objetivo é encontrar um modelo capaz de prever a probabilidade de o cliente vir a praticar *default*, com base no tempo de permanência no banco.

Supondo que existe um limiar de 0.5 então, os clientes cuja previsão de *default* for maior que 0.5 serão considerados de alto risco e como consequência, não será concedido o crédito.

A linha que melhor se ajusta a este problema, utilizando o método de regressão linear é:

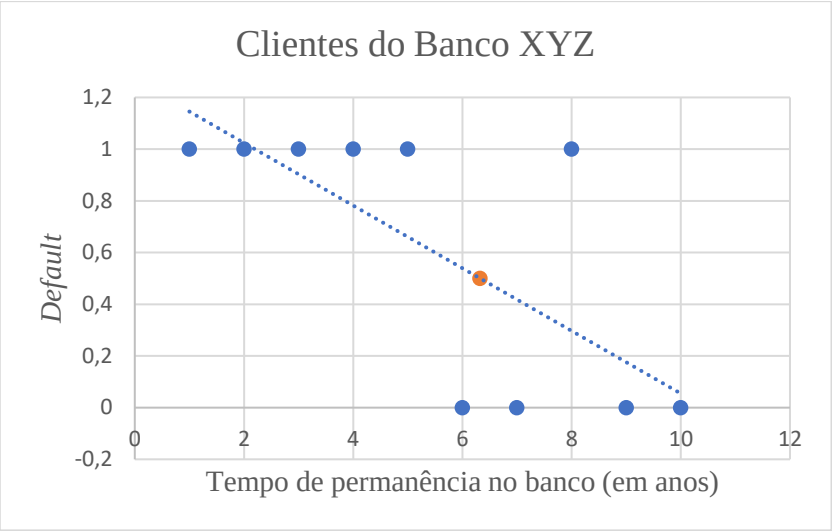

Figura 3.8 - Regressão Linear

Como se observa, utilizar a regressão linear e estabelecer um limiar de 0.5 parece dar um ótimo resultado ao problema.

No entanto, imagine-se agora que na nossa amostra de 10 pessoas, existe um cliente que tem conta no banco desde que nasceu e tem neste momento 50 anos.

Esse cliente à partida, pelo tempo de permanência no banco, não cometeu *default*. Ter este cliente na amostra, não traz muita informação nova, pois dado que é cliente ativo há 50 anos, é muito provável que cumpra os seus deveres de pagamento. Assim, em outras palavras, esta observação não gera incerteza.

Deste modo, a situação ideal passa pelo algoritmo de previsão não se centrar neste cliente, mas sim nas regiões de fronteira, isto é, onde é realmente difícil avaliar se o cliente é exatamente de alto ou baixo risco. Infelizmente isso não acontece usando o modelo de regressão linear:

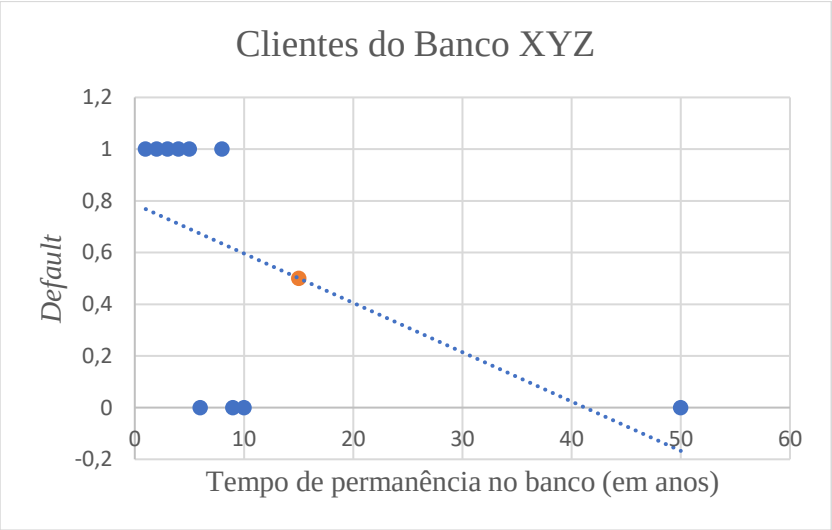

Figura 3.9 - Cliente extra

A presença, na amostra, do cliente do banco há 50 anos, faz com que a linha de regressão seja forçada a ir para a direita. Este efeito, fará com que, com o limiar de 0.5, se corra o risco de classificar clientes de baixo risco, como tendo alta probabilidade de cometer *default*.

Para além disso, como se verifica no gráfico acima, para clientes com mais de 40 anos de permanência no banco, está prevista uma probabilidade negativa de *default*, o que simplesmente não faz sentido algum.

Pode concluir-se que, neste caso, a utilização do modelo de regressão linear não faz sentido de ser utilizado. Faz sentido sim, utilizar a regressão logística para estes casos de classificação.

A regressão logística constrói sempre uma previsão no intervalo [0,1], estando sempre a trabalhar com probabilidades válidas, não negativas.

Outra situação que será resolvida pela utilização da regressão logística será o facto desta não ser influenciada por *outliers*, que não acrescentam informações novas ao modelo, como é o caso do cliente que permanece no banco há 50 anos.

Deste modo, o ideal será encontrar a curva em formato "S" que melhor se ajusta aos dados, usando assim, o método de regressão logística.

3.3.3 Regressão Logística

A regressão logística (RL) é um algoritmo de aprendizagem supervisionada que se usa para classificação. E será este o método escolhido para fazer face aos objetivos da dissertação.

É aplicado em situações em que a variável dependente é de natureza dicotómica ou binária, isto é, tem como *outcome* 1/0, Sim/Não, *True*/*False*, dado um conjunto de variáveis independentes, que tanto podem ser categóricas ou não. Para representar uma variável binária (categórica) temos de usar as variáveis *dummy*.

Dado que a regressão logística é uma técnica que permite estimar a probabilidade associada à ocorrência de um determinado evento, os resultados ficam compreendidos no intervalo [0,1].

A regressão logística tem uma expressão matemática muito semelhante à da regressão linear, sendo que, obtém-se a expressão da regressão logística através da expressão da regressão linear, pela aplicação da função sigmoide.

Definição 3.1:
Seja y uma variável aleatória (*target*) e sejam x_1,…, x_n, n variáveis independentes. A expressão seguinte representa a equação da regressão linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

Onde β_0, β_1,… são os coeficientes não-conhecidos a determinar pelo método de estimação dos mínimos quadrados. Finalmente, ε é um processo de ruído branco gaussiano.

Definição 3.2:
A função sigmoide é uma função matemática de amplo uso, em campos como a economia e a computação. O nome sigmoide vem da forma em S do seu gráfico e é definida como:

$$\rho(y) = \frac{1}{1 + e^{-y}}$$

Logo, substituindo a variável y pela equação da regressão linear, na expressão da função sigmoide, iremos obter a regressão logística, isto é:

$$\rho(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon)}}$$

Onde ρ é a probabilidade do resultado (*outcome*) e y o *output* predito.

O modelo de regressão logística segue uma distribuição binomial, e os coeficientes de regressão β (estimativas de parâmetros) são estimados usando o método de estimativa de máxima verossimilhança. Assim, através dos valores X e de uma combinação de coeficientes β, que maximiza a probabilidade dos valores preditos serem iguais aos valores atuais, chega-se à fórmula da regressão logística. (Hosmer, Jovanovic e Lemeshow (1989))

O gráfico assumirá uma forma em "S", que é característico do modelo de regressão logística, tal como aparece na figura 3.10:

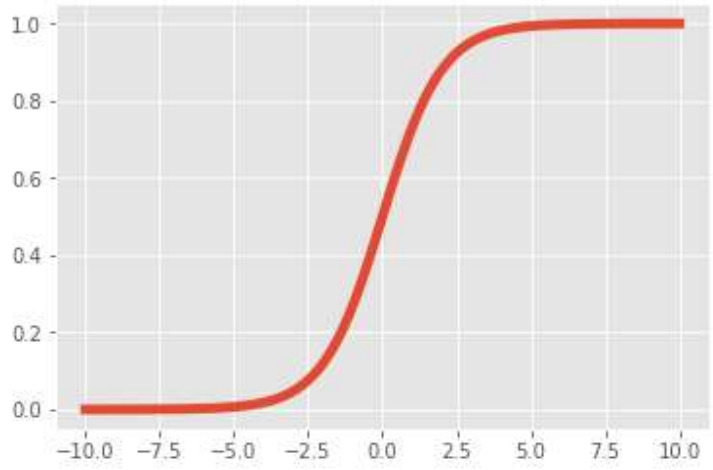

Figura 3.10 - Curva da regressão logística (função $\rho(y)$)

O gráfico em forma de "S" representa a relação entre as variáveis independentes e as probabilidades previstas.

Assim, quando:

- $y \longrightarrow +\infty, \quad então\ \rho(Y = 1) \longrightarrow 1$
- $y \longrightarrow -\infty, \quad então\ \rho(Y = 1) \longrightarrow 0$

Para prever em qual classe pertence um certo dado, precisamos escolher um valor de fronteira (*threshold*). Com base neste valor, a probabilidade estimada é classificada em classes. Esta fronteira é chamada fronteira de decisão (*Decision Boundary*).

Nesta dissertação o valor de fronteira é 0.5, e assim:

- $\rho(Y = 1) > 0.5\ então\ Y = 1$
- $\rho(Y = 1) < 0.5\ então\ Y = 0$

A regressão logística não tem requisitos de normalidade e variância como a regressão linear, mas precisa, no entanto, de seguir alguns pressupostos. Vejamos de forma sucinta alguns pressupostos e também, as principais vantagens e desvantagens do modelo de regressão logística, sintetizadas na tabela 3.3:

Tabela 3.3 - Pressupostos, Vantagens e Desvantagens da RL

Pressupostos da RL	Vantagens da RL	Desvantagens da RL
A variável dependente é binária.	Fornece resultados no intervalo [0,1].	Não trabalha com variáveis contínuas.
As variáveis independentes não são correlacionadas entre si (não são multicolineares).	Facilidade em trabalhar com variáveis independentes categóricas.	A regressão logística não funciona se as variáveis independentes não se correlacionam com a variável dependente (*target*).
Não tem *outliers* influentes		As bases de dados (amostras) têm de ser muito grandes, de maneira a garantir a estabilidade do resultado.

3.3.3.1 *Confusion Matrix*

Segundo Odegua (2020) a *confusion matrix* (ou matriz de confusão) é uma matriz bidimensional que contém informações sobre as classes reais e as classes previstas.

Vidhur Kumar (2019) refere que a matriz de confusão é amplamente usada para avaliar o desempenho de um algoritmo de classificação, como é o caso da regressão logística.

Em linhas encontra-se informações relativas ao valor atual e em colunas conta-se com as informações relativas ao valor predito. Assim, esta matriz terá tantas linhas e colunas quantas classes de classificação existam.

Nesta dissertação existem 2 classes (0,1), logo estamos perante uma matriz 2x2.

A figura 3.11 é uma simulação da matriz de confusão do modelo:

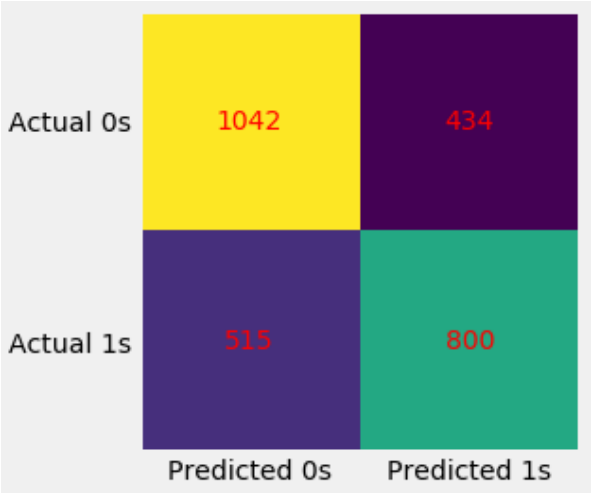

Figura 3.11 - *Confusion matrix*

Por exemplo, na matriz de confusão a posição correspondente à 1ª linha e 1ª coluna, diz respeito ao número de vezes em que o valor atual foi igual ao valor predito para classes com rótulo (*label*) igual a zero e na posição correspondente à 2º linha e 2º coluna, para classes com *label* igual a um.

Deste modo, o ideal é ter o máximo de valores na diagonal principal e poucos valores nas outras duas entradas. Desta maneira, no exemplo ilustrado na figura 3.11, a classificação foi predita corretamente na maior parte dos exemplos do conjunto de dados.

Formalmente, a matriz de confusão é lida da seguinte forma:

Tabela 3.4 - Explicação da *Confusion matrix*

		Valor predito	
		0	1
Valor atual	0	TP	FN
	1	FP	TN

Onde:

- TP (*true positive*): diz respeito ao número de casos em que o modelo prevê corretamente a classe positiva.
- TN (*true negative*): diz respeito ao número de casos em que o modelo prevê corretamente a classe negativa.
- FP (*false positive*): diz respeito ao número de casos em que o modelo prevê de forma incorreta a classe positiva.
- FN (*false negative*): diz respeito ao número de casos em que o modelo prevê de forma incorreta a classe negativa.

Com base na matriz de confusão pode definir-se as seguintes métricas para a performance da classificação:

Accuracy

A *accuracy* ou também chamada de exatidão, mede a proporção de previsões onde o valor atual é igual ao valor predito, ou seja, diz respeito ao número de dados classificados corretamente sobre o número total de dados, isto é:

$$Accuracy = \frac{TP\ +\ TN}{TP\ + TN + FP + FN}$$

Recall

A *recall* ou também chamada de sensibilidade, mede a capacidade de o método prever corretamente casos de uma determinada classe. Este valor, deve ser o mais alto possível. A sua fórmula é:

$$Recall = \frac{TP}{TP + FN}$$

Precision

De todas as classes positivas que se previu corretamente, a *precision* mede, quantas são realmente positivas. Este valor, também deve ser o mais alto possível. A sua fórmula é:

$$Precision = \frac{TP}{TP + FP}$$

f1-score

F1-score é a média harmónica entre a *precision* e a *recall* do modelo, anteriormente calculadas, sendo uma forma destes parâmetros, a *precision* e a *recall*, serem comparáveis. A fórmula é a seguinte:

$$f1 - score = \frac{2\ *\ Precision\ *\ Recall}{Precision\ +\ Recall}$$

ROC e AUC

Para além da matriz de confusão, a curva ROC (*Receiver Operating Characteristic*s) e a área por baixo da curva ROC, representam métricas importantes para mensurar a performance do modelo de classificação, em particular, de regressão logística.

A curva ROC é uma curva de probabilidade. Para a simplificar, foi criada a AUC (*Area Under the Curve*) e diz respeito à área que se encontra limitada pela curva ROC e o eixo do x.

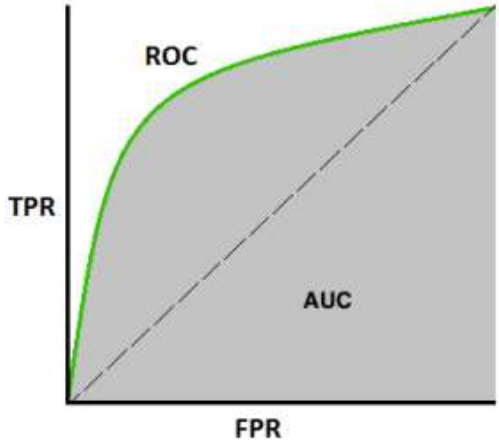

Figura 3.12 - Curva AUC-ROC

Conforme sugere a figura 3.12, no eixo x encontra-se a FPR (*false positive rate*) e no eixo y a TPR (*true positive rate*), ou seja, o número de vezes em que o modelo de classificação acertou na previsão, contra o número de vezes que o modelo errou a previsão.

$$TPR\ (true\ positive\ rate) = \frac{TP}{TP\ +\ FN}$$
$$FPR\ (false\ positive\ rate) = \frac{FP}{TN + FP}$$

Dado se tratar de taxas, ambos os eixos assumem valores no intervalo [0,1].

A AUC resume a curva ROC num único valor, tornando-se mais fácil a comparação entre modelos. Quanto maior o AUC, melhor o modelo prevê 0s como 0s e 1s como 1s.

Deste modo, um modelo excelente tem um AUC próximo de 1 e um modelo fraco tem um AUC próximo de 0. Quando o AUC assume um valor intermédio de 0,5, isto é, TPR=FPR, significa que o modelo não tem capacidade de separação de classes.

3.4 Modelação empírica

3.4.1 Algoritmos

Foram implementados alguns modelos de *machine learning*, como é o caso do modelo *Random Forest*. No entanto, é computacionalmente mais complexo e a performance não tem valores de destaque. Deste modo, levando em consideração o *trade-off* entre o tempo de execução/complexidade do algoritmo e performance na classificação, decidiu-se abandonar este algoritmo e dar mais destaque à regressão logística.

Possivelmente, para bases de dados de maior tamanho, o desempenho do *Random Forest* pode tornar-se mais competitivo, sendo necessária uma revisão dos algoritmos.

O objetivo era assim, encontrar um modelo acessível e prático, que se ajustasse da melhor forma aos dados.

Desta forma, o algoritmo escolhido nesta dissertação, como foi dito na secção anterior, é a regressão logística. É um dos algoritmos de *machine learning* mais simples, mas visto por vários autores como muito eficaz, rápido, de fácil interpretação e económico do ponto de vista da sua implementação (em termos de tempo e custos).

3.4.2 Variável independente *vs.* variável dependente

Uma das desvantagens do modelo de regressão logística, é que, se as variáveis independentes não são correlacionadas com a variável dependente (*target*), então a regressão logística não tem um desempenho considerável, levando para o enviesamento dos resultados.

Numa primeira fase da análise dos dados, a amostra é separada em dois conjuntos: de treino – para aprendizagem; e de teste – para avaliar a performance do modelo.

Antes de se passar aos resultados que foram obtidos no conjunto teste, pode observar-se a partir da figura 3.13, qual a relação que existe entre a variável *housing* (*target*) e algumas das variáveis independentes.

Desta forma, é possível tirar algumas primeiras informações acerca de quais serão as variáveis determinantes na concessão de um empréstimo, no entanto, não será nada conclusivo.

***Job*:** Estima-se que a cedência de crédito por parte dos bancos possa ser muito depende do cargo do cliente. A figura 3.13 relaciona a variável *job* com a variável *housing* (*target*):

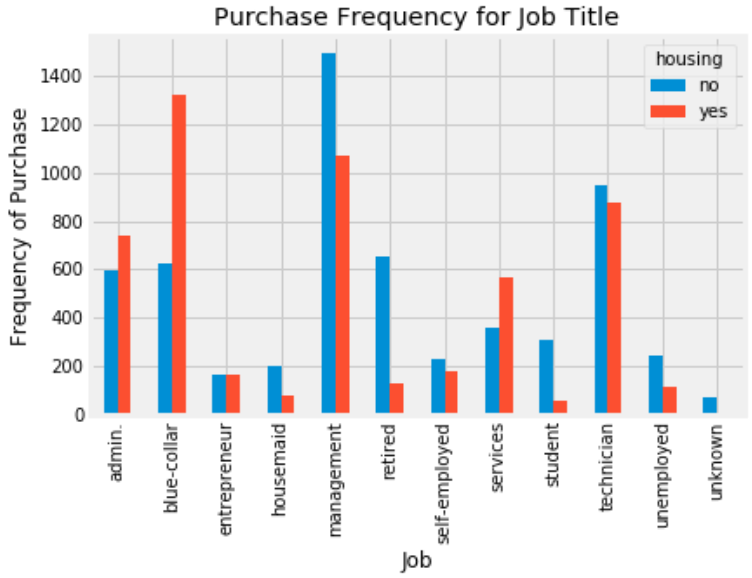

Figura 3.13 - *Housing vs. Job*

Como é observável na figura, esta variável poderá ser um bom preditor da variável *target*.

Marital: Acredita-se que, a variável *marital* será uma variável decisiva na concessão de cedência de crédito. A figura 3.14 relaciona a variável *marital* com a variável dependente *housing* (*target*):

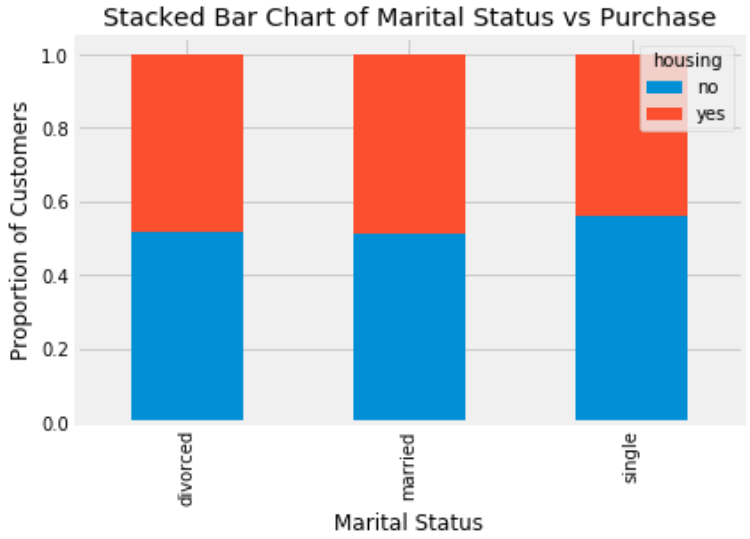

Figura 3.14 - *Housing vs. Marital*

Como se pode observar na figura acima, a variável *marital* não parece ser um forte preditor para a variável *target*. Da observação direta da figura, é ainda possível concluir que, nesta base de dados, o crédito bancário é cedido relativamente em menor proporção a pessoas solteiras.

Effective: Em relação à variável *effective*, pensa-se que o facto de o cliente ser efetivo no cargo que desempenha, é um fator decisivo na cedência de crédito. A figura 3.15 relaciona a variável *effective* com a variável *housing* (*target*):

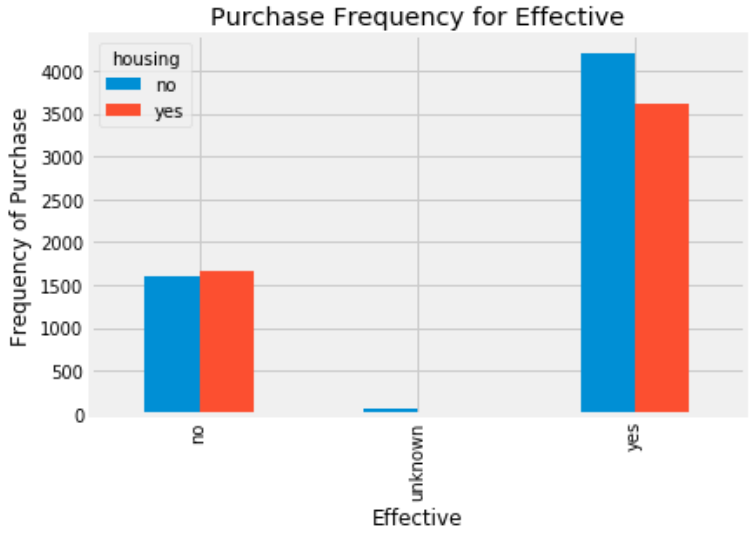

Figura 3.15 - *Housing vs. Effective*

A figura 3.15 sugere que a variável *effective* não é um bom preditor do resultado.

Analisando a figura acima pode concluir-se também que, nesta base de dados, das pessoas que são efetivas no seu trabalho, há mais pessoas a quem não foi atribuído crédito, do que pessoas a quem foi atribuído crédito.

Leasedhouse: Por último, é analisada na figura 3.16 a variável *leasedhouse*, em função da variável dependente *housing* (*target*):

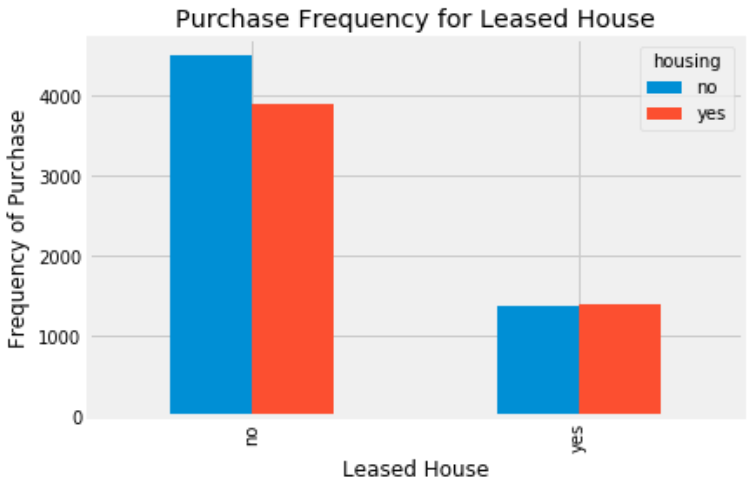

Figura 3.16 - *Housing vs. Leasedhouse*

No entanto, a figura acima dá a entender que este não será um bom preditor da variável dependente.

3.4.3 Base de dados

Dado que a base de dados é constituída por variáveis de texto, como por exemplo a variável *month*, é necessário transformar estas variáveis em variáveis binárias/categóricas, pois só assim é possível a execução do modelo.

Os primeiros cinco valores de cada categoria, depois de se obter apenas variáveis binárias/categóricas e numéricas, é representado na figura seguinte:

	contact	month	age	job	marital	education	household	client	leasedhouse	salary	effective	default	housing	loan
0	0	5	25	0	1	2	3	1	0	1954	2	0	1	0
1	0	5	14	0	0	1	3	1	0	1916	2	0	0	1
2	0	5	32	0	2	1	3	1	1	2590	2	0	0	0
3	0	1	32	0	1	1	3	1	1	3027	2	0	1	0
4	0	1	29	0	1	1	3	1	1	2474	2	0	0	0

Figura 3.17 - Variáveis do modelo

Nesta figura são visíveis as 13 variáveis independentes e a variável dependente, *housing*.

Seguindo a metodologia de *machine learning*, a base de dados dividir-se-á em dois conjuntos: de treino e de teste. O algoritmo faz a sua aprendizagem sobre o conjunto de treino e verifica a sua performance sobre o conjunto de teste.

O conjunto de teste é constituído por 25% da totalidade dos dados, o que perfaz 75% para o conjunto de treino.

A implementação do modelo de regressão logística faz-se sobre o conjunto de treino considerando, numa primeira fase, todas as variáveis independentes.

Passo a passo, as variáveis que não trazem contributo na classificação, são eliminadas. As variáveis independentes são permutadas e eliminadas com base nos resultados da performance da classificação.

Finalmente e, em concorrência com o que foi descrito até agora, as variáveis que têm mais contributo na classificação são consideradas no modelo final, modelo a partir do qual será validada a performance sobre o conjunto de treino.

No que se segue são apenas apresentados os resultados obtidos sobre o conjunto de teste, ou seja, o que o modelo de regressão logística nos garante quando é confrontado com dados novos.

Dado que existem 11162 dados na amostra, do conjunto de teste fazem parte 2791 observações.

3.4.4 Métricas

No que se prossegue, a regressão logística considera todas as variáveis independentes como parte da sua equação. A matriz de confusão que se obtém do conjunto de teste é:

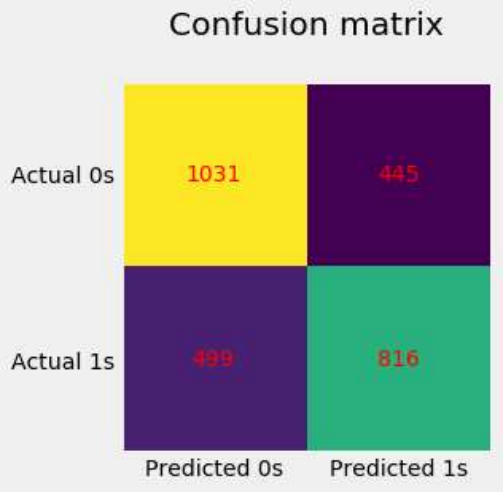

Figura 3.18 - Matriz de confusão

Da matriz de confusão, da figura 3.18, retiramos a informação de que, 1031+816 são previsões corretas e que 499+445 são previsões incorretas.

Somando os valores da diagonal da matriz de confusão, conclui-se que existem 1847 casos onde o valor predito é igual ao valor atual/real, ou seja, pode concluir-se que a classe foi predita corretamente na maior parte dos exemplos de teste.

Visto que, existem 2791 observações totais no conjunto de teste e que em 1847 observações os dados foram classificados corretamente então, atinge-se uma percentagem de *accuracy* na ordem dos 66%, o que é considerado um valor prazeroso.

A tabela 3.5 mostra os valores para as outras métricas de avaliação do modelo, referidas na secção anterior:

Tabela 3.5 - *Precision*, *recall* e *f1-score*

	Precision	*Recall*	*F1-score*	*Support*
0	0.67	0.70	0.69	1476
1	0.65	0.62	0.63	1315

Como se verifica na tabela 3.5, a *precision*, a *recall* e o *f1-score* assumem valores superiores a 60%.

A precisão do classificador de regressão logística no conjunto de teste é, em média, de 66%, ou seja, o classificador tem a capacidade de não rotular uma observação como positiva, se ela for negativa em 66% do conjunto teste.

Tendo em conta a *recall*, o classificador tem 66% de capacidade média de encontrar todas as observações positivas, no conjunto teste.

Por último, considerando que, no caso de uma pontuação de *f1-score* de 100%, esta métrica atinge o seu melhor valor e que, perto de 0% atinge a sua pior pontuação, então tendo como resultado um valor médio de 66% de *f1-score*, também aqui é notório que este modelo tem tudo para ser um bom modelo de classificação.

Como referido anteriormente, podemos resumir a previsibilidade do modelo com base na área sob a curva ROC. Um modelo com maior AUC tem maior previsibilidade.

A ROC-AUC da regressão logística considerada é:

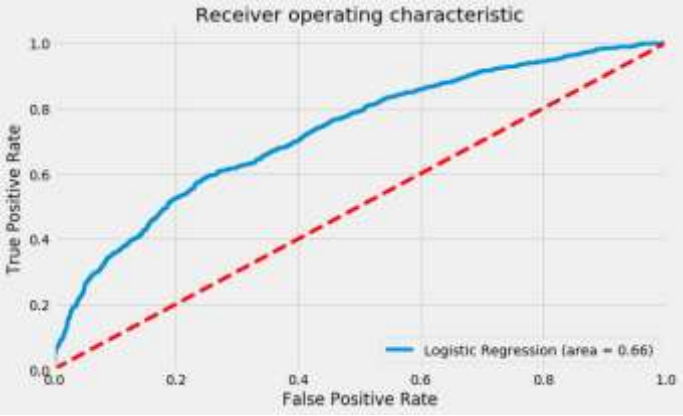

Figura 3.19 - ROC-AUC

O modelo tem um valor de AUC de 0.66, isto significa que 66% das previsões estão corretas.

Sendo esta uma métrica meramente comparativa, não existe muito mais a dizer sobre ela, sendo que uma área de 66% é significativamente aprazível.

Neste sentido, pode afirmar-se que de todo o conjunto de teste, o banco XYZ classifica que 66% dos clientes estão aptos para receber o empréstimo.

3.4.5 O impacto das variáveis no modelo

Para finalmente fazer face ao objetivo desta dissertação, é necessário responder à questão colocada: "Quais as variáveis que são determinantes na aquisição de um empréstimo bancário?".

Posto isto, tem de se analisar quais das variáveis, que foram consideradas no modelo, têm mais impacto na classificação.

3.4.5.1 Matriz de Correlação

Para se ter a certeza que se obtém a melhor previsão do resultado da regressão, é necessário incluir variáveis independentes que têm um contributo significativo sobre a variável alvo.

Assim sendo, é importante não existir multicolinearidade, ou seja, não existir variáveis altamente correlacionadas. Isto iria reduzir o desempenho do modelo.

A matriz correlação mostra todas as correlações de *Pearson* e tem como objetivo medir o grau de relação entre as variáveis, ou seja, avalia a força e a direção entre as mesmas.

Figura 3.20 - Matriz de correlação

Pela observação da figura 3.20 conclui-se, à partida, que não existem variáveis altamente correlacionadas com a variável *target*.

Da observação direta da figura, esta, ainda sugere que existe uma correlação positiva entre as variáveis independentes *contact, month, household, leasedhouse, salary, default* e *loan*, com a variável *target housing*. Enquanto que, as variáveis *age, job, marital, education, client* e *effective* têm uma correlação negativa com a variável *target*.

3.4.5.2 RFE

Ao incluir todas as variáveis no modelo de regressão, existe a possibilidade de não se alcançar os preditores mais significativos no modelo.

Assim sendo, foi necessário recorrer ao *Recursive Feature Elimination* (RFE), isto é, proceder à eliminação de um recurso/variável do modelo e repetir o processo com todas as outras variáveis, de maneira a descobrir o melhor ou pior recurso/variável. Este processo repete-se até que todas as variáveis do modelo tenham sido testadas.

Deste modo, trabalha-se apenas com as variáveis que mais contribuem para a previsão.

Ter muitas variáveis pode diminuir a precisão de um modelo. Esta técnica de eliminação de recursos, RFE, tem alguns benefícios, tais como:

- Reduzir o *overfitting*.
- Melhorar a precisão.
- Reduzir o tempo de treino.

Ao usar esta técnica no modelo de previsão, o top 5 das variáveis que foram consideradas menos relevantes foram as variáveis *contact*, *effective*, *month*, *default* e *leasedhouse*.

Resta agora, implementar o modelo e analisar quais os resultados obtidos na retirada destas variáveis.

3.4.5.3 *Output*

O *output* obtido antes da remoção das variáveis consideradas menos relevantes, pelo método RFE, isto é, com todas as variáveis do modelo de regressão, é o seguinte:

```
Optimization terminated successfully.
         Current function value: 0.612110
         Iterations 5
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.115
Dependent Variable: housing          AIC:              13690.7397
Date:               2021-02-18 16:03 BIC:              13785.9032
No. Observations:   11162            Log-Likelihood:   -6832.4
Df Model:           12               LL-Null:          -7720.8
Df Residuals:       11149            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
----------------------------------------------------------------
              Coef.   Std.Err.     z      P>|z|    [0.025   0.975]
----------------------------------------------------------------
contact       0.5816    0.0332   17.5164  0.0000   0.5165   0.6466
effective     0.2881    0.0293    9.8426  0.0000   0.2307   0.3455
month         0.1293    0.0066   19.4858  0.0000   0.1163   0.1423
default       0.1828    0.1678    1.0894  0.2760  -0.1461   0.5117
age          -0.0343    0.0018  -19.3740  0.0000  -0.0377  -0.0308
job          -0.0544    0.0068   -7.9433  0.0000  -0.0678  -0.0410
marital      -0.2604    0.0335   -7.7723  0.0000  -0.3261  -0.1947
education    -0.1973    0.0279   -7.0777  0.0000  -0.2519  -0.1427
household     0.0576    0.0199    2.9011  0.0037   0.0187   0.0966
client       -0.4512    0.0493   -9.1601  0.0000  -0.5477  -0.3546
leasedhouse   0.0275    0.0485    0.5684  0.5698  -0.0674   0.1225
salary        0.0001    0.0000    7.2590  0.0000   0.0001   0.0002
loan          0.2961    0.0604    4.9045  0.0000   0.1778   0.4145
================================================================
```

Figura 3.21 - Resultado Regressão Logística (*Output* 1)

Da observação da figura 3.21, com base nos valores de prova (*p-value*), conclui-se que todas as variáveis independentes, exceto a variável *default* e a *leasedhouse*, são estatisticamente significativas, pois o *p-value* é inferior ao nível de significância. A variável *default* tem um *p-value* no valor de 0.2760 e a variável *leasedhouse* tem um *p-value* de 0.5698. Assim sendo, é necessário proceder à remoção destas duas variáveis.

É importante relembrar que o *output* 1 comporta todas as variáveis independentes do modelo e como foi visto anteriormente, pelo método RFE, estas duas variáveis a excluir, fazem parte das variáveis menos relevantes no modelo.

Ainda da informação que é retirada do *output* 1, relativa aos coeficientes estimados, consta que as variáveis independentes *age*, *job*, *marital*, *education* e *client* são variáveis que têm um impacto negativo na variável *target housing*. Todas as outras têm um impacto positivo.

Pode-se ainda evidenciar outros parâmetros, tais como o Pseudo R-*squared*, AIC, BIC, *Log-Likelihood*, por exemplo. Contudo estes parâmetros fazem sentido de serem estudados, quando comparados com outros modelos, com conjuntos de dados semelhantes, que preveem o mesmo resultado.

Posteriormente, retirando apenas as variáveis *default* e *leasedhouse*, como sugerido pelo *p-value* das mesmas, no *output* anterior, conta-se com:

```
Optimization terminated successfully.
         Current function value: 0.612178
         Iterations 5
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.115
Dependent Variable: housing          AIC:              13688.2586
Date:               2021-02-18 18:39 BIC:              13768.7816
No. Observations:   11162            Log-Likelihood:   -6833.1
Df Model:           10               LL-Null:          -7720.8
Df Residuals:       11151            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
----------------------------------------------------------------
               Coef.   Std.Err.     z      P>|z|    [0.025   0.975]
----------------------------------------------------------------
contact        0.5851    0.0329   17.7762  0.0000   0.5206   0.6496
effective      0.2903    0.0292    9.9526  0.0000   0.2331   0.3475
month          0.1292    0.0066   19.4847  0.0000   0.1162   0.1422
age           -0.0341    0.0018  -19.3670  0.0000  -0.0376  -0.0307
job           -0.0543    0.0068   -7.9325  0.0000  -0.0677  -0.0409
marital       -0.2596    0.0335   -7.7530  0.0000  -0.3253  -0.1940
education     -0.1974    0.0279   -7.0820  0.0000  -0.2520  -0.1428
household      0.0591    0.0197    3.0012  0.0027   0.0205   0.0977
client        -0.4549    0.0490   -9.2758  0.0000  -0.5510  -0.3588
salary         0.0001    0.0000    7.3197  0.0000   0.0001   0.0002
loan           0.2962    0.0603    4.9080  0.0000   0.1779   0.4144
================================================================
```

Figura 3.22 - Resultado Regressão Logística (*Output* 2)

Como é evidente na figura 3.22, todos os *p-value* são inferiores ao nível de significância de 5%, portanto não será necessário eliminar mais nenhuma variável. O modelo de regressão pode ser validado.

Neste momento, existe um ponto de comparação para os parâmetros Pseudo R-*squared*, AIC, BIC e *Log-Likelihood*.

O *Pseudo R-squared* do *output* 2 manteve-se igual ao do *output* 1. O ideal seria ter um valor próximo de 1, pois significaria que se estava perante um modelo melhor ajustado.

Os parâmetros AIC, BIC e *Log-Likelihood* sofreram transformações. O AIC e o BIC desceram os seus valores, isto significa que, estamos perante um modelo melhor que o anterior (*output* 1), pois, quanto mais baixo for o valor assumido por estes critérios de informação, melhor é o modelo.

O mesmo não se pode dizer do parâmetro *Log-Likelihood*, pois o seu valor tornou-se ligeiramente mais negativo e, neste caso, quanto mais alto o valor assumido por este parâmetro, melhor será o modelo. No entanto, é mesmo uma ligeira descida.

Retirando agora do *output* 2, as restantes variáveis independentes sugeridas pelo método RFE – *contact*, *effective* e *month* – tem-se:

```
Optimization terminated successfully.
         Current function value: 0.661942
         Iterations 5
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.043
Dependent Variable: housing          AIC:              14793.1868
Date:               2021-02-18 18:41 BIC:              14851.7490
No. Observations:   11162            Log-Likelihood:   -7388.6
Df Model:           7                LL-Null:          -7720.8
Df Residuals:       11154            LLR p-value:      3.3154e-139
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
----------------------------------------------------------------
               Coef.   Std.Err.     z      P>|z|    [0.025   0.975]
----------------------------------------------------------------
age           -0.0210    0.0016  -13.3885  0.0000  -0.0241  -0.0179
job           -0.0265    0.0064   -4.1484  0.0000  -0.0391  -0.0140
marital       -0.0113    0.0306   -0.3685  0.7125  -0.0712   0.0486
education     -0.0840    0.0253   -3.3240  0.0009  -0.1335  -0.0345
household      0.1422    0.0181    7.8616  0.0000   0.1068   0.1777
client        -0.2019    0.0434   -4.6487  0.0000  -0.2870  -0.1168
salary         0.0002    0.0000   14.7247  0.0000   0.0002   0.0003
loan           0.4172    0.0577    7.2335  0.0000   0.3042   0.5303
================================================================
```

Figura 3.23 - Resultado Regressão Logística (*Output* 3)

Este *output* retratado na figura 3.23 sugere a remoção da variável *marital*, uma vez que esta possui um *p-value* superior a 0.05. Contudo, não será necessário, pois, os parâmetros de comparação Pseudo R-*squared*, AIC, BIC e *Log-Likelihood* assumiram piores valores neste *output* 3 e como tal, já não estamos perante o melhor resultado de regressão logística.

O facto de o *output* 3 não ser o melhor modelo encontrado, não contradiz os resultados obtidos pelo método RFE. Foi sugerido ao modelo identificar as cinco variáveis menos significativas, mas isso não quer dizer que se tenham de retirar as cinco do modelo.

Apesar disso, o método RFE identificou as mesmas variáveis que o modelo de regressão logística sugeria no *output* 1, por terem um *p-value* superior a 0.05.

Deste modo, o melhor resultado de regressão logística encontrado, foi o da figura 3.22 (*Output* 2), no qual foram retiradas as variáveis *default* e *leasedhouse*, do modelo original.

Assim sendo a matriz de confusão correspondente ao *output* 2, resultante da remoção das variáveis *default* e *leasedhouse* é representada na figura 3.24:

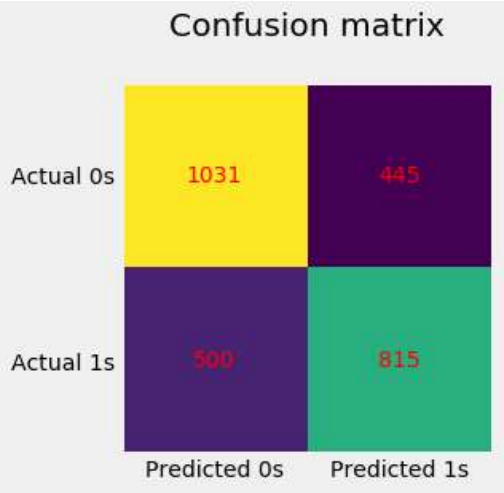

Figura 3.24 - Matriz de confusão do modelo de Regressão Logística

Observa-se que não houve uma grande alteração na matriz de confusão pois, os verdadeiros positivos mantém o mesmo número de acertos, enquanto que os verdadeiros negativos perdem 1 valor predito.

O modelo tem 1031+815 previsões corretas e 500+445 são previsões incorretas.

Fazendo novamente a soma dos valores da diagonal da matriz de confusão, conclui-se que existem 1846 casos onde o valor predito é igual ao valor atual/real – apenas menos um caso do que quando consideradas todas as variáveis do modelo.

Todavia a conclusão é a mesma, a classe foi predita corretamente na maior parte dos exemplos de teste.

Todos as outras métricas – *accuracy*, *recall*, *precision*, *f1-score* e a curva ROC-AUC – mantém os seus valores na ordem dos 66%.

Para concluir o estudo desta dissertação, no qual foi utilizado o modelo de regressão logística, as variáveis que são determinantes na concessão de empréstimos bancários são: *contact*, *effective*, *month*, *age*, *job*, *marital*, *education*, *household*, *client*, *salary*, *loan*.

Pode ainda concluir-se, que o modelo é muito estável na sua performance de classificação, não ultrapassando os 66% sobre os novos dados apresentados no conjunto de teste. Retirar e adicionar variáveis mantém a performance estável.

Observa-se também que, a variáveis *contact*, *effective*, *marital* e *client* são as que têm um maior peso/contributo na decisão de conceder ou não crédito ao proponente, como pode ser observado em todos os *outputs* apresentados.

Ultrapassar a meta dos 66% de performance deve ser realizada para um conjunto de dados maior e melhor balanceado.

CONCLUSÃO

O crédito bancário surge como uma importante fonte de financiamento na economia, seja para empresas ou particulares.

O crédito bancário influencia a política monetária, a atividade económica, o PIB e as taxas de juro.

Um choque de política monetária promove o aumento das necessidades de financiamento, uma vez que os agentes económicos têm alguma dificuldade em se ajustar rapidamente a novas medidas económicas, o que prejudica a evolução da atividade económica de um país.

Em relação ao PIB, segundo a revisão literária desta dissertação, este indicador tem uma relação positiva com o crédito bancário, uma vez que, quando o PIB aumenta, aumentam as necessidades de financiamento, tanto para empresas como para particulares. Já em relação às taxas de juro o mesmo não acontece, isto porque existe uma relação negativa entre a taxa de juro real e o crédito bancário.

No entanto, visto que foram usados dados simulados, não foi possível garantir tais afirmações.

O risco de crédito é dos riscos mais comuns na economia e diz respeito à incapacidade de o cliente cumprir com as suas obrigações financeiras acordadas, com uma instituição financeira. Para que estas instituições de crédito se protejam de eventuais perdas financeiras, é importante haver um sistema de pontuação de crédito, capaz de apoiar a tomada de decisão de concessão de empréstimos. Assim os agentes de crédito serão substituídos por modelos de pontuação de crédito.

Em Portugal, antes da crise de 2008 era concedido muito mais crédito, sendo este crédito concedido com mais facilidade e menos entraves.

Após a crise de 2008, Portugal estaria a recuperar da crise económica, não fosse no ano de 2020 surgir uma crise pandémica que afetou praticamente todos os setores económicos. Isto fez com que as instituições de crédito se tornassem muito mais exigentes na cedência de crédito, havendo assim uma maior fração de pedidos de empréstimo rejeitados.

Para fazer fase ao objetivo desta dissertação foi feita uma análise através da linguagem *Python* e foi usado o modelo de regressão logística.

Para o estudo foi construída uma base de dados com 11162 observações e 14 variáveis, sendo 13 destas, as variáveis independentes e uma variável dependente. Esta variável dependente, a variável *target* é a variável *housing* e diz respeito ao crédito à habitação. Todas as variáveis em estudo foram escolhidas, tendo em conta a leitura de alguns artigos e também, pela ajuda de especialistas em crédito.

Para avaliar se o modelo se adequa ao problema considerado, foram utilizadas algumas métricas como a matriz de confusão, a *accuracy*, a curva ROC-AUC, a *recall*, a *precision* e o *f1-score*.

Posteriormente para analisar se, se estava perante o melhor modelo, foi utilizada a técnica *Recursive Feature Elimination* (RFE), na qual se retira a informação de quais são as variáveis menos relevantes no modelo.

Ao confrontar o modelo com esta informação, chegou-se à conclusão que era necessário retirar duas das variáveis, a *default* e a *leasedhouse*. Desta maneira teríamos um modelo melhor ajustado do ponto de vista estatístico, mas sem efeito sobre a classificação.

Se, se procedesse à remoção de mais alguma variável de estudo, o modelo adquiria valores inferiores em todas as métricas.

Do modelo final considerado, pode concluir-se, através dos resultados obtidos nas métricas de desempenho, que este é um bom modelo, um modelo bastante funcional, simples, económico e apto a ser utilizado por instituições bancárias aquando da cedência de crédito aos seus clientes.

Os objetivos propostos foram parcialmente satisfeitos.

É verdade que não se chegou a patamares incríveis de 100%, mas, no entanto, existe sempre a lacuna de se estar a trabalhar com dados simulados. Talvez com a utilização de dados reais, e em maior quantidade, se consiga atingir resultados mais proeminentes.

Não obstante os resultados alcançados, é preciso reforçar que, quanto mais informação houver sobre o individuo que solicita o crédito, melhor informada está a instituição de crédito e menores serão as possibilidades de os sistemas de crédito falharem na pontuação de crédito de um cliente, diminuindo assim possíveis perdas financeiras para as instituições de crédito.

Assim conclui-se esta dissertação com a ideia de que, este projeto é um bom objeto de estudo para a concessão de crédito a particulares, a fim de identificar clientes de risco, numa amostra de clientes que solicitam crédito, usando algoritmos simples de *machine learning*.

Em resposta ao problema principal desta dissertação, as variáveis que são determinantes na concessão de empréstimos bancários são: *contact*, *effective*, *month*, *age*, *job*, *marital*, *education*, *household*, *client*, *salary* e *loan*.

REFERÊNCIAS BIBLIOGRÁFICAS

Avelar, A. (2019). "O que é AUC e ROC nos modelos de Machine Learning". *Medium*, 6.

Banco de Portugal (2020). "Boletim económico – outubro 2020". Departamento de Estudos Económicos.

Banco de Portugal (2020). "O que é e como funciona". https://www.bportugal.pt/page/o-que-e-e-como-funciona#o-que-e (consultado em outubro 2020)

Banco de Portugal (2020). "O que são e tipos de crédito". Portal do Cliente Bancário. https://clientebancario.bportugal.pt/pt-pt/o-que-sao-e-tipos-de-credito (consultado em outubro 2020)

Bedre, R. (2021). "Logistic regression in Python (feature selection, model fitting, and predition)". *Data Science blog*, 8.

Brzoza-Brzezina, M. (2005). "LENDING BOOMS IN THE NEW EU MEMBER STATES: WILL EURO ADOPTION MATTER?". *ECB Working Paper* (Vol. 543).

Friedman, B. M. & Kuttner, K. N. (1993). "Economic Activity and the Short-term Credit Markets: An Analysis of Prices and Quantities". *Brookings Papers on Economic Activity*, 2, 193-283. https://doi.org/10.2307/2534567

Gameiro, I. M. & Sousa, J. (2010). "O IMPACTO DA POLíTICA MONETÁRIA NAS TRANSAÇÕES FINANCEIRAS E DOS PARTICULARES EM PORTUGAL*". Banco de Portugal, *Boletim Económico*, Verão 2010.

Hofmann, B. (2001). "The determinants of private sector credit in industrialised countries: do property prices matter?". *BIS Working Papers*, 108. http://dx.doi.org/10.2139/ssrn.847404

Hosmer, D. W., Jovanovic, B. & Lemeshow, S. (1989). "Best Subsets Logistic Regression". *Biometrics*, 45, 4, 1265-1270. http://dx.doi.org/10.2307%2F2531779

Kulkarni, A., Chong, D. & Batarseh, F. A. (2020). "Foundations of data imbalance and solutions for a data democracy". *Data Democracy*, 83-106.

Kumar, V. (2019). "Applying Logistic Regression: Classifying Loans based on the risk of defaulting". *Towards Data Science*, 11.

Li, S. (2017). "Building A Logistic Regression in Python, Step by Step". *Towards Data Science*, 17.

Moradi, S. & Rafiei, F. M. (2019). "A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks". *Financial Innovation*, 5:15. https://doi.org/10.1186/s40854-019-0121-9

Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). "An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments". *Sustainability* (*Switzerland*), 11 (2019), 1–23. https://doi.org/10.3390/su11030699

Odegua, R. (2020). "Predicting Bank Loan Default with Extreme Gradient Boosting". ArXiv, 5.

PORDATA (2020). "Agregados domésticos privados: total e por tipo de composição".
https://www.pordata.pt/Portugal/Agregados+dom%c3%a9sticos+privados+total+e+por+tipo+de+composi%c3%a7%c3%a3o-19 (consultado em novembro 2020)

PORDATA (2015). "Alojamentos familiares clássicos de residência habitual segundo os Censos: total, por ocupantes proprietários e inquilinos".
https://www.pordata.pt/Portugal/Alojamentos+familiares+cl%c3%a1ssicos+de+resid%c3%aancia+habitual+segundo+os+Censos+total++por+ocupantes+propriet%c3%a1rios+e+inquilinos-145
(consultado em novembro 2020)

PORDATA (2020). "Dimensão média dos agregados domésticos privados".
https://www.pordata.pt/Portugal/Dimens%c3%a3o+m%c3%a9dia+dos+agregados+dom%c3%a9sticos+privados-511 (consultado em novembro 2020)

PORDATA (2019). "Salário médio mensal dos trabalhadores por conta de outrem: remuneração base e ganho".
https://www.pordata.pt/Portugal/Sal%C3%A1rio+m%C3%A9dio+mensal+dos+trabalhadores+por+conta+de+outrem+remunera%C3%A7%C3%A3o+base+e+ganho-857 (consultado em novembro 2020)

PORDATA (2020). "Trabalhadores por conta de outrem: total e por tipo de contrato".
https://www.pordata.pt/Portugal/Trabalhadores+por+conta+de+outrem+total+e+por+tipo+de+contrato+-844 (consultado em novembro 2020)

Rodrigues, V. (2018). "Entenda o que é AUC e ROC nos modelos de Machine Learning". *Medium*, 3.