FACULDADE DE CIÊNCIAS E TECNOLOGIA

Ano Lectivo: 2018/2019

Mestrado em Matemática e Aplicações

Dissertação para obtenção do Grau de
Mestre em Matemática e Aplicações
Especialização em Finanças

Estimação Da Probabilidade De Incumprimento De Uma Carteira De Crédito Ao Consumo

Nasolino Fernandes Varela

Praia, Março de 2019

FACULDADE DE CIÊNCIAS E TECNOLOGIA

Ano Lectivo: 2018/2019

Mestrado em Matemática e Aplicações

Dissertação para obtenção do Grau de Mestre em Matemática e Aplicações
Especialização em Finanças

Estimação Da Probabilidade De Incumprimento De Uma Carteira De Crédito Ao Consumo

Nasolino Fernandes Varela

Dissertação apresentada à Universidade de Cabo Verde para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica do *Prof. Doutor José Moniz Fernandes*, Faculdade de Ciências e Tecnologia, Universidade de Cabo Verde, Cabo Verde

Praia, Março de 2019

Estimação Da Probabilidade De Incumprimento
De Uma Carteira De Crédito Ao Consumo

Copyright © Nasolino Fernandes Varela, Faculdade de Ciências e Tecnologia, Universidade de Cabo Verde

A Faculdade de Ciências e Tecnologia e a Universidade de Cabo Verde têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel e/ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

It is not the knowlegde, but the act of learning,
not the posession but the act of getting there,
which grants the greatest enjoyment.
Carl Friedric Gauss

Dedicatória

Aos meus familiares, principalmente aos meus filhos Alyssa Nayara e Naylsson.

O júri :

Presidente	_____
Arguente	_____
Orientador	*Prof. Doutor José Moniz Fernandes*, Faculdade de Ciências e Tecnologia, Universidade de Cabo Verde, Cabo Verde.
Co-orientador	_____

Agradecimentos

Agradeço à UNICV (Universidade de Cabo Verde) pela oportunidade dada para fazer este curso e chegar a este nível.
A todos os que foram meus professores e que lidaram comigo durante este percurso, que muitas vezes tiveram paciência e calma.
Ao meu orientador pela sugestão do tema e principalmente pela paciência que teve desde o inicio da pesquisa.
À minha família por se dedicar com tanto empenho à minha formação, pelo apoio incessante e incentivo, pelo carinho, pelo incorajamento nos dias menos fáceis e pela compreensão pelas minhas poucas atenções.
Por fim, mas certamente não menos importante, aos meus amigos pela amizade, incentivo constante, companheirismo e apoio durante todos os momentos da minha vida e em especial durante este percurso.

Resumo

A análise de risco de crédito nas instituições bancárias é de extrema importância para as instituições, uma vez que o crédito é a sua principal atividade. A capacidade de distinguir clientes cumpridores e incumpridores é um processo decisivo na constituição do crédito, pelo que são aplicados modelos quantitativos que auxiliam os gestores na tomada de decisão.

O principal objetivo desta dissertação é estimar a probabilidade de incumprimento de cada cliente no momento da concessão do crédito e de uma carteira de crédito ao consumo ao longo de um determinado período temporal.

Utilizando o Modelo de Regressão Logística, e recorrendo a variáveis sociodemográficas e financeiras de cada cliente, estimamos a probabilidade de incumprimento para cada cliente.

O Modelo de Regressão Probit foi utilizado neste estudo como forma de efetuar a comparação com o Modelo de Regressão Logística, dado que estes são bem comparáveis.

A probabilidade de incumprimento não é constante ao longo do tempo, sendo assim, analisamos o incumprimento de uma carteira utilizando o Modelo Vórtices Estocásticos.

Relativamente ao Modelo Vórtices Estocásticos as populações são sujeitos a entradas e saídas de elementos, consideramos que o fluxo de entrada é modelada através da forma funcional sigmoidal $\lambda_i = \left(a + b.e^{-\theta i}\right)^{-1}$, estudada e aplicada nos estudos de Fernandes [12].

Com o Modelo de Regressão Logística obtemos resultados relativos a probabilidade de incumprimento de cada cliente no momento da concessão de crédito. E verificamos que o modelo obtido está a classificar corretamente cerca de 99.3% dos clientes cumpridores e 4.1% dos clientes incumpridores, se se tiver em conta que apenas se consideram clientes incumpridores os clientes com probabilidade de incumprimento estimada superior a 40%. Donde, em média, o modelo classifica corretamente 51.7% dos clientes da amostra de teste.

Os resultados obtidos através do Modelo Vórtices Estocásticos permitem-nos estimar o número e a proporção de clientes nas várias classes de risco, através de estimativas pontuais e intervalos de confiança. Verificou-se por exemplo que no mês 20 cerca de 93% de clientes estão em cumprimento, enquanto que 0.48% estão em incumprimento.

Através desses resultados se pode classificar cada cliente como bom ou mau pagador, dependendo da política da instituição, do ponto corte e da classe de risco que se encontra.

Palavras-chave: Crédito, Risco, Regressão Logística, Vórtices Estocásticos

Abstract

The analysis of credit risk in the bank institutions is of extreme importance for the institutions, once the credit is the main activity. The capacity to distinguish customers in default and non-default is a decisive process in the constitution of the credit, for the that are applied quantitative models that they aid the managers in the socket of decision.

The objective of this dissertation is estimate the default probability of the customer at the time of credit granting and of consumption portfolio credit, to in a given period.

The estimate of the default probability for each customer has been made by using the Logistic Regression form, taking into account the socioeconomic and financial variable.

The Model of Regression Probit was used in this study as form of making the comparison with the Model of Regression Logistics, given that these are very comparable.

The default probability is not regular in time. Therefore, the default of the consumption portfolio credit is analyzed using Stochastic Vortices Model.

Relatively to the Stochastic Vortices Model the people subjected to addition and subtraction of elements,we consider the entry flow is modelled through sigmoidal functional form $\lambda_i = \left(a + b.e^{-\theta i}\right)^{-1}$, studied and applied in Fernandes [12] studies.

With Logistic Regression Model we obtain relative results the probability of each customer's default in the moment of the credit concession. And we verified that the obtained model is to classify correctly about 99.3% of the customers non-default and 4.1% of the customers default, if it be had in bill that you/they are just considered customers default the customers with probability of superior dear default to 40%. From where, on average, the model classifies 51.7% of the customers of the test sample correctly

The results obtained through the Stochastic Vortices Model allow to be considered the number and the customers' proportion in the several risk classes, through punctual estimates and trust intervals. It was verified for instance that in the month 20 about 93% of customers they are in no default, while 0.48% are in default.

Through of those results if it can classify each customer as good or bad payer, depending on the politics of the institution, of the point cuts and of risk class that is.

Keyswords: Credit, Risk, Logistic Regression, Stochastic Vortices.

Índice

Lista de Figuras

Lista de Tabelas

Símbolos e Abreviaturas

NOTAÇÃO	DESCRIÇÃO
MLGs	Modelos Lineares Generalizados
f.d.p.	Função densidade de probabilidade
f.m.p.	Função massa de probabilidade
A.I.C	Akaike Information Criterion
PCC	Percentagem dos Corretamente Classificados
SENS	Sensibilidade
Esp	Especificidade
KS	Kolmogorov-Smirnov
AR	Racio de Precisão
ROC	Receiver Operating Characteristic
Ω	Espaço de resultados ou universo
$\mathbb{P}$	Uma medida natural
H_0	Hipotese nula
H_1	Hipotese alternativa
$\mathbb{N}$	Conj. de numeros naturais
$\mathbb{R}$	Conj. de numeros reais
P_c	Ponto de corte
S_i	Score atribuido aos clientes
d.e.n	Desvio equivalente normal

Introdução

Contextualização

O trabalho que ora se apresenta, constitui um trabalho final de curso cujo tema é *estimção da probabilidade de incumprimento de uma carteira de crédito ao consumo*, apresentada à Universidade de Cabo Verde como requisito parcial para obtensão do grau mestrado em matemática e aplicações, especialização em finanças.

O risco de crédito tornou-se nos últimos tempos, o tema destaque no dia-a-dia dos gestores de crédito e o desenvolvimento de modelos de *credit scoring*, tornou-se uma tarefa fundamental para as instituições financeiras, como via de minimizar a perda de créditos e auxiliar os gestores no processo de tomada de decisão.

Com a globalização, tornou-se necessário reestruturar os métodos de avaliação de risco, primeiramente era com base apenas nos modelos de análise subjetivos que por vários motivos pode levar a perdas muitos elevados.

As alterações macro-económicas dos países, o aumento da concorrência bancária, as alterações constantes da taxa de juro ou as margens de *spreads* mais competitivas fez com que ocasionaram uma necessidade urgente de gerir o risco de crédito de modo eficaz. Perante esta realidade, as instituições financeiras têm vindo a desenvolver modelos internos de gestão de risco de crédito com o intuito de oferecer ferramentas mais eficientes para a valorizaçao da carteira , medição de riscos, apreçamento de novos empréstimos, etc.

No entanto o risco que cada cliente representa para a instituição bancária deve ser avaliado e estimado a priori, aquando da concessão do crédito, usando técnicas adequadas. Por vezes, as condições socio-económicas do cliente, analisadas aquando da concessão do crédito, podem alterar-se durante o período do empréstimo, condicionando o cumprimento do pagamento das prestações, representando um risco acrescido ao estimado inicialmente.

Dado a importáncia do crédito e risco de crédito para a instituição bancária são vários os autores que desenvolveram estudos que auxiliam a tomada de decisão no momentos de concessão de crédito, destacamos aqui alguns:

Almeida [1], desnvolveu estudos sobre crédito e risco de crédito, onde afirma que o crédito é de longe o negócio mais importante da banca e que continuará a sê-lo por muitos anos no futuro. Sendo assim os modelos de análise de risco do crédito desempenha um papel fundamental no processo da concessão do crédito.

Araújo [3] em seu estudo, afirma que a concessão de crédito constitui uma das principais atividades das instituições financeiras sendo a principal causa de insolvência das mesmas e, uma vez que o risco de crédito é inerente a qualquer operação financeira, é essencial que estas instituições procedam a uma análise e gestão rigorosa dos créditos que concedem.

Caoutte [8], afirma que os primeiros banqueiros na europa mediaval frequentemente cobravam dos clientes pequenas tarifas em função dos custos associados a guarda dos seus recursos. Contudo, não demorou muito para que eles percebessem que, emprestando esses recursos a outros, poderiam fazer essa atividade rentável.

Fernandes [12], que utilizou os dados de uma carteira de crédito ao consumo de uma instituição bancária de Cabo verde para desenvolver estudos sobre uma carteira de crédito ao consumo. Neste estudo verificou que os fatores como a competitividade dos mercados, a evolução dos sistemas e gestão de informação, a diminuição do pedido de crédito e o aumento da probabilidade de incumprimento condicionam as intuições financeiras à mitigação de melhores técnicas para a análise e gestão do risco de crédito.

É de referir que o crédito é o principal negócio de uma instituição bancária, que por este motivo é necessário criar políticas, condições, meios e técnicas que aquando da concessão de crédito, garantem a totalidade do reembolso, com isso sim garantir sustentabilidade da instituição. Imaginemos uma isntituição bancária em que todos os créditos são malparados, talvez funcionaria por um curto período de tempo, não teria condições para sustentar as despesas da própria empresa. Mas graças aos estudos desenvolvidos por vários autores, qualquer instituição já tem ideia e planos para minimizar o risco aquando da concessão do crédito.

Escolha e justificação do tema

Uma das motivações é poder colocar em prática conhecimentos adquiridos em alguns pontos curiculares da minha maior preferência nesse mestrado, também escolha do tema justifica-se principalmente pela preocupação constante das instituições bancárias perante a análise do risco de crédito. Nota-se que no setor bancário o principal risco é o de crédito e nem sempre estas instituições possuem um controlo operacional adequado, gerando retornos indesejáveis. Posto isso torna-se necessário desenvolver estudos que permite aos gestores de crédito uma visão esclarecedora no que tange ao crédito, risco de crédito e seus prós e contra no momento de tomada de decisão.

Objetivos

O presente trabalho tem como objetivo geral:

- Estimar a probabilidade de incumprimento de cada cliente no momento da concessão do crédito e de uma carteira de crédito ao consumo ao longo de um determinado período temporal.

Para que o objetivo geral seja atingido, propomos os seguintes objetivos específicos:

- Mencionar os principais conceitos relacionados com crédito bancário.
- Conhecer um dos modelos de análise do risco do crédito bancário: o *credit scoring*.
- Desenvolver um modelo capaz de avaliar adequadamente o risco de crédito com base nos fatores chave do devedor que influência o incumprimento.
- Identificar fatores chave que influência a probabilidade de incumprimento.
- Estimar a probabilidade de incumprimento de um cliente aplicando o modelo de Regressão Logística.
- Relacionar os modelos logístico e probit.
- Estimar as probabilidades de transição entre classes de risco, assumindo uma carteira de crédito aberta (permitindo a subscrição de novos contratos e a anulação de contratos existentes).
- Analisar numa perspetiva temporal, a evolução da dimensão de cada uma das classes de risco.
- Estimar o peso relativo de cada uma das classes dentro da carteira de crédito.

Metodologia

Todo e qualquer trabalho científico implica a adopção de uma metodologia que consiste num conjunto de métodos e de técnicas, não só de recolha como também de tratamento de dados, para se alcançar os objetivos pretendidos numa investigação.
Neste sentido, o trabalho em questão representa uma investigação exploratória e bibliográfica, buscando decrever o processo de análise do risco de crédito, utilizando o modelo de Regressão Logística e o modelo Vórtices Estocásticos. Ou seja, fizemos pesquisa bibliográfica nas bases de dados, da web, das principais monografias, livros, artigos dedicadas ao tema em estudo, considerando as seguintes palavras chaves: Incumprimento, Regressão

Logística, Vórtices Estocásticos, Risco de Crédito.

O modelo de Regressão Logística foi utilizado para a estimar a probabilidade de incumprimento e a identificação dos determinantes das mesmas, admitindo que a carteira de crédito ao consumo é constituída pelas variáveis financeiras e sociodemográficas.
O modelo de Regrssão Probit foi utilizado neste estudo como forma de efetuar a comparação com o modelo de Regrssão Logística, dado que estes são bem comparáveis sendo que a principal diferença está no fato de a logística ter caudas ligeiramente mais achatadas como pode ser visto na figura 2.1 da seção 2.4.6.
A metodologia dos Vórtices Estocásticos, aplicada e desenvolvida nesta dissertação, baseia-se nos estudos de Fernandes [12], Guerreiro e Mexia [16] e Rodrigues [24].
Consideramos populações abertas, divididas num número finito de sub-populações. Os elementos da população são inicialmente colocados numa sub-população e periodicamente reclassificados, podendo, em cada reclassificação, ser colocados em qualquer das sub-populações existentes. Com este modelo podemos estimar a probabilidade de incumprimento de uma carteira de crédito ao longo de um determinado período de tempo e posteriormente classificar os clientes com bom ou mau pagador de acordo com a classe de risco a que pertence.
Para dar fim a esta dissertação utilizamos os seguintes softwares:

- SPSS para categorização das variáveis em estudo.
- R-studio para desenvolver o modelo e estimar a probabilidade de incumprimento.
- Wolfram Mathematica 9 para ajustamento da forma funcional sigmoidal aos dados da carteira e estimação dos parâmetros do fluxo de entrada.
- JabRef utilizado para enserir as referências bibliográficas.
- Latex-Winedit10.1 utilizado para digitar o testo, obtendo assim o testo em formato PDF.

Estrutura da dissertação

Esta dissertação está estruturada em três grandes capítulos, cada capítulo inicia-se com uma pequena introdução que nos ajuda a compreendê-los.
O primeiro capítulo onde fizemos um breve apanhado do crédito bancário, está dividido em quatro pequenas seções. Na primeira seção falamos um pouco sobre o risco do crédito, na segunda sobre política do crédito bancário, na terceira sobre análise do crédito bancário e na quarta seção o principal objetivo é analisar e conhecer os principais conceitos relacionados com um dos modelos de classificação de análise do risco do crédito bancário: o modelo *credit scoring*.
No segundo capítulo, onde o principal objetivo é estimar a probabilidade de incumprimento de um cliente, está dividido em quatro seções. Na primeira seção apresentamos o

modelo Regressão Logística como um caso particular dos modelos lineares generalizados, na segunda seção apresentamos a carteira de crédito ao consumo, na terceira seção, apresentamos os métodos para desnvolvimento e validação do modelo e na quarta seção fazemos aplicação do modelo de Regressão Logítica a uma carteira de crédito ao consumo e também comparamos este modelo com o modelo de Regressão Probit.
O terceiro capítulo onde o principal objetivo é estimar a probabilidade de incumprimento de uma carteira de crédito ao longo de um determinado período temporal, está dividido em duas seções. Na primeira seção descrevemos o modelo Vórtices Estocásticos e na segunda seção aplicamos o modelo Vórtices Estocásticos para estimação da evolução temporal da probabilidade de incumprimento de uma carteira de crédito ao consumo.
Por fim apresentamos as conclusões e algumas recomendações.

Capítulo 1

Crédito Bancário

Introdução

Neste capítulo apresenta-se uma breve revisão à literatura sobre assuntos que contribuem para a compreensão do crédito bancário.
Segundo Almeida [1], o crédito continua a ser, de longe, o negócio mais importante da banca e, sem qualquer dúvida, continuará a sê-lo por muitos anos no futuro. Hoje em dia afirma-se, e com plena razão, que os bancos são instituições fornecedoras de serviços financeiros aos agentes económicos.
Almeida [1], afirma que de entre os vários serviços prestados, o da disponibilização de fundos a crédito é aquele que continua a ter maior relevância.
De acordo com Chaia [9], existem um grande número de definições para o termo crédito ou operações de crédito, contudo é necessário conhecer o seu sentido etimológico. A palavra **crédito** vem do latin *creditu*, significando eu acredito ou confio. A confiança não representa uma atividade unilateral, ocorrendo tanto por parte do vendedor, que acredita na capacidade ou desejo do comprador de honrar os comprimissos assumidos, como do adquirente em acreditar na qualidade do produto comprado.
Para Silva [27], essa confiança representa um dos elementos necessário, porém não sufuciente, para uma decisão do crédito. Para ele, crédito representa entrega do bem presente mediante uma promessa de pagamento. Essa definição serve tanto ao chamado crédito comercial ou industrial, no qual o bem entregue é representado pelos recursos financeiros disponibilizados.
Este capítulo está dividida em quatro seções. Na primeira seção falamos um pouco sobre o risco do crédito, na segunda sobre política do crédito bancário, na terceira sobre análise do crédito bancário e na quarta o principal objetivo é analisar e conhecer os principais conceitos relacionados com um dos modelos de classificação de análise do risco do crédito bancário: o modelo *credit scoring*.

1.1 Risco

Segundo Almeida [1], o risco, dependendo do contexto em que é referido, terá definições diferentes. Na sua forma mais ampla ele é definido com sendo um acontecimento futuro incerto que possa influenciar o alcance dos objetivos estratégicos, operacionais e financeiros da organização.
Então pode-se dizer que o risco está associado à incerteza em obter ou perder algo. O risco de crédito não deixa de ser uma incerteza em obter ou perder um montante de dinheiro emprestado a outrem. Sendo assim podemos dizer que, o risco de crédito está associado ao grau de incerteza dos retornos esperados quer por incapacidade do tomador de um empréstimo, quer do emissor de um título ou da contraparte de um contrato, em cumprir com as suas obrigações.
Na análise da capacidade do devedor para assunção do crédito, tem grande importância a avaliação do risco.
Nessa avaliação, distingue-se quatro tipos de risco:

- Risco geral
- Risco do ramo de atividade ou profissional
- Risco particular ou pessoal
- Risco da operação

1.2 Política do Crédito Bancário

Segundo Almeida [1] os bancos têm necessidade de definir de forma clara e precisa a política de crédito que deve orientar a atuação de todos os seus empregados que lidam com a matéria.
Os recursos próprios de um banco constituiem, na maioria dos casos, uma pequeníssima parte dos seus meios de ação. A grande fatia dos seus meios de ação é constituída pelas poupanças dos seus clientes, que lhes são confiadas em depósito. Nestes termos, quando um banco empresta dinheiro a um cliente, é o dinheiro dos seus depositantes que está a emprestar.
De tempos a tempos, o banco tem de reembolsar os seus depositantes, pelo que, ao exercer a sua função creditícia, deverá assegurar que o crédito concedido está em segurança, e que a concessão gera uma mais valia (juro), com o qual remunera os seus e os capitais dos depositantes. De igual modo a política de aplicação deve ter em conta o eventual levantamento de fundos por parte dos seus depositantes, o que impõe aos bancos uma sã gestão de contratação de prazos para os depósitos e aplicação.

Almeida [1], defende que os bancos orientam a sua política de crédito com base em três princípios básicos: princípio de segurança, princípio da rendibilidade e princípio da liquidez.

Atendendo a estes princípios, o banco terá de definir a sua política de crédito, a qual vai mudando consoante as alterações que se vão surgindo: no mercado, nas preferências dos clientes, nas situações dos concorrentes e nas regras impostas pelas autoridades monetárias.

Em matéria de crédito, o fator risco está sempre presente em maior ou menor grau. Quando o banco, decide favoravelmente uma operação de crédito, isso significa que considerou que o risco associado a essa operação era comportável.

Uma vez acordado o crédito, o banco deverá acompanhar esse crédito, nomeadamente, vigiando os devedores e assegurar que as garantias oferecidas foram corretamente constituídas.

Para Chaia [9], os principais componentes da política de crédito em bancos são: definição estratégica do banco, forma de decisão e delegação de poderes, análise de crédito, limites de crédito e normas legais. O mesmo afirma que a política do crédito bancário não representa apenas avaliação dos clientes e aprovação de limites, devendo conter também regras de precificação em função das avaliações, formas de gestão de risco durante a vida da operação e instrumentos da recuperação das dívidas em atraso.

1.3 Análise de Crédito Bancário

Para muitos autores a atividade de crédito como temos hoje em dia teve início na revolução industrial.
No final do século XIX a revolução industrial colhia seus frutos desenvolvimentistas e os bancos já atuavam como intermediários financeiros. A massificação dos empréstimos levou os bancos a procurarem padronizar suas análises exigindo dos demandantes de empréstimos, demonstrativos financeiros que comprovassem sua capacidade de futuro pagamento. Os demonstrativos financeiros, analisados sistematicamente, levaram à criação dos índices que relacionam fatos administrativos de interesse contábil. Salienta-se a criação do índice de liquidez corrente, o qual faz sucesso ate hoje (Pereira [23]).

Araújo [3], afirma que uma boa gestão do risco de crédito por parte das instituições financeiras, é indispensável para que se evite a insolvência das mesmas. A análise de crédito é um processo que deve reunir informações a respeito do tomador de crédito, com o intuito de avaliar a sua capacidade de cumprir com as suas obrigações e definir quanto à concessão ou não do crédito.
Para Sousa [29], a análise de crédito é uma ferramenta fundamental para a decisão de crédito, e consiste num estudo da situação global do devedor. Ela possibilita a elaboração de um parecer que demonstra de maneira clara e objetiva o desempenho económico-financeiro do cliente.
Vale [32], afirma que a análise de crédito pode ser tratada tendo em conta duas metodo-

logias: a **qualitativa** e a **quantitativa**. A análise qualitativa remete para julgamentos subjetivos por parte do analista de crédito, em relação à capacidade de pagamento do tomador de crédito. Nesta abordagem, pessoas especializadas são encarregues de tomar a decisão sobre a concessão de crédito, utilizando critérios qualitativos e subjetivos. A análise quantitativa utiliza informações provenientes de modelos estatísticos e econométricos. Nesta última abordagem consideram-se os modelos de credit scoring e os modelos baseados na teoria de carteiras.
A principal vantagem da abordagem qualitativa é a especificidade com que é tratado cada caso. A principal desvantagem é a sua dependência na experiência do avaliador, o baixo volume de produção e o envolvimento pessoal do concedente de crédito.
As regras bem definidas em relação às características dos clientes e às operações de crédito, baseadas, em geral, em modelos estatísticos, são o fatores positivos da análise quantitativa.
Nas teorias de Vale [32], a análise de crédito clássica é mais antiga depende essencialmente da opinião de especialistas bem treinados, no entanto, atualmente, a necessidade de obter uma avaliação de risco mais aprofundada e mais correta obrigou ao desenvolvimento de técnicas estatísticas de modo a garantir uma diminuição do risco de crédito das instituições financeiras. As técnicas de pesquisa estatística, tais como a análise de sobrevivência, redes neuronais, programação matemática e simulação probabilística contribuíram para o avanço das técnicas de mensuração do risco de crédito.

1.4 Modelos de Análise de Risco de Crédito

Neste seção apresentamos os conceitos relacionados com os modelos de análise de risco de crédito. Existem vários modelos de análise do risco de crédito, mas neste estudo vamos destacar o modelo de *credit scoring*, dado que este tem mais haver com objetivo e o tema do trabalho a desenvolver.
Segundo Vale [32], modelo é, por definição, uma representação simplificada de algo real. Desta forma, algoritmos, fórmulas, sistemas ou regras que visam representar processos ou atributos reais relacionados com risco de crédito, podem ser considerados modelos de risco de crédito.
Devido às suas características, os modelos facilitam a compreensão de um fenómeno e, eventualmente, a sua exploração. Os modelos de risco de crédito não são excepção.
Caoutte et al. [7], fala da construção dos modelos de risco de crédito, afirmando que a construção de um modelo de risco de crédito exige, em primeiro lugar, a identificação das variáveis que podem provocar a ocorrência de incumprimento. Segue-se a utilização de um conjunto de ferramentas para construir um modelo formal, com base num conjunto de dados que representem a carteira de crédito. Finalmente, devem ser aplicados testes para determinar se o modelo tem o desempenho esperado.
Segundo Brito [6], os modelos de risco de crédito podem ser classificados em três grupos: modelos de classificação de risco, modelos estocásticos de risco de crédito e modelos de risco de portfolio. Os modelos de classificação de risco buscam avaliar o risco de um toma-

dor ou operação, atribuindo uma medida que representa a expetativa de risco de *default*[1], geralmente expressa na forma de uma classificação de risco (*rating*) ou pontuação (escore). Os modelos de classificação de risco são utilizados pelas instituições financeiras em seus processos de concessão de crédito.
Os modelos estocásticos de risco de crédito são aqueles que têm por objetivo avaliar o comportamento estocástico do risco de crédito ou das variáveis que o determinam. Esses modelos são utilizados pelas instituições financeiras principalmente para precificar títulos e derivativos de crédito. Os modelos de risco de portfolio visam a estimar a distribuição estatística das perdas ou do valor de uma carteira de crédito, a partir da qual são extraídas medidas que quantificam o risco do portfolio. Esses modelos constituem uma importante ferramenta no processo de gestão de riscos das instituições, pois permitem que o risco de crédito seja avaliado de forma agregada, considerando os efeitos da diversificação produzidos pelas correlações entre os ativos da carteira.

1.4.1 Modelos de credit scoring

Conforme Baptista [4], *o credit scoring* é um processo de avaliação da capacidade de crédito do cliente obtido através de informações registadas e cujos dados são convertidos em números, que depois combinados (normalmente adicionados) originam uma pontuação (score). Já Almeida [1], define *o credit scoring* como sistemas ou modelos de apoio à decisão, baseados no princípio de objetividade de critérios e que visa avaliação de operações de crédito a particulares.
Os modelos de *credit scoring* permitem, recorrendo a técnicas estatísticas, estabelecer um processo de atribuição de pontuações às variáveis de decisão de crédito. Este processo permite estimar a probabilidade de um dado cliente ser um cliente cumpridor ou incumpridor.(Vale [32]).
Na visão de Saunders [25], os sistemas de pontuação de crédito encontram-se em quase todos os tipos de análise de crédito. O objetivo é geralmente o mesmo, pré-identificar, através de técnicas estatísticas, fatores-chave que determinem a probabilidade de incumprimento, e a combinação ou ponderação dos mesmos de modo a produzir uma pontuação quantitativa. Ainda afirma que a metodologia básica para o desenvolvimento de um modelo de *credit scoring* deve ter em conta as seguintes etapas:

- Planeamento e definições
- Identificação das variáveis
- Planeamento amostral e coleta de dados
- Determinação da fórmula de classificação através de técnicas estatísticas

[1]Default é a incapacidade para cumprir as condições de uma obrigação ou acordo, ou seja, é não fazer um pagamento em dívida. A palavra Default pode ser simplesmente, Incumprimento.(Vale [32])

- Determinação do ponto de corte[2]

Nos modelos de *credit scoring* supõe-se que as características dos clientes que entrarão em incumprimento no futuro são semelhantes às características dos clientes que entraram em incumprimento no passado. Tendo em conta este pressuposto, é comum utilizar-se amostras de clientes cumpridores e incumpridores da instituição bancária e aplicar técnicas estatísticas apropriadas para inferir sobre os indícios de incumprimento de um cliente em particular, como refere Vale [32].
Para Chaia [9], apesar do *credit scoring* ser um processo matemático, não ilimina a possibilidade de se recusar um bom pagador ou de se aceitar um mau pagador. Isto devido á falta de sistema de avaliação que consegue capturar todas as informações relevantes que são necessárias para a precisa classificação dos devedores.
Lewis [20], refere que o primeiro modelo estatístico de análise de crédito remonta ao ano de 1945. Os primeiros modelos de *credit scoring* destinavam-se ao crédito ao consumo. A expansão do mercado de crédito massificado obrigou os analistas a uma maior rapidez e homogeneidade no tratamento dos seus clientes e, por isso, a um aumento da utilização destes modelos. A evolução dos sistemas informáticos possibilitou o tratamento estatístico adequado a esse aumento de dados.
Saunders [25], afirma que a diferença mais acentuada entre os modelos subjetivos e o modelo de *credit scoring* reside no fato de, nesse último, se valorizar a utilização de métodos estatísticos que permitem a seleção dos fatores-chave e dos respetivos pesos, a partir dos quais se obtém uma pontuação para cada cliente de acordo com as suas características. A pontuação gerada, fornece indicadores quantitativos das hipóteses de incumprimento desse cliente e representa o risco de perda. Esta pontuação pode ser comparada com um **ponto de corte** ou com uma pontuação mínima aceitável a partir da qual a instituição financeira aprova ou não a concessão de crédito.
Os modelos de *credit scoring* dividem-se em duas categorias: os modelos de aprovação de crédito (*credit scoring* propriamente dito) e os modelos de classificação comportamental, estes últimos conhecidos como *behavioural scoring*, conforme refere Caouette et al. [7].
Segundo Araújo [3], o *behavioural scoring* leva consideração os aspetos comportamentais e as atividades dos clientes da instituição, ou seja, auxiliam a gestão dos créditos dos clientes que já possuem créditos na instituição.
Almeida [1], entende que o modelo *behavioural scoring* avalia a forma como o cliente se comporta, quer perante o banco, quer na sua vida pessoal, e ultrapassa o âmbito da gestão do risco de crédito.
Os modelos de *scoring* de aprovação (*credit scoring* propriamente dito) são técnicas estatísticas, baseadas em análise discriminante de fatores de incumprimento, que utilizam a informação sobre os clientes para determinar o segmento a que pertencem e o correspondente risco, permitindo dividir o mercado de particulares em segmentos de risco semelhante.(Almeida [1]). Nestes modelos a instituição não tem conhecimento do comportamento do cliente e é muito utilizado quando se trabalha com os não clientes e ou clientes que solicitam pela primeira vez o crédito.

[2]ponto de corte é um patamar mínimo de risco no qual a instituição de crédito estaria disposta a assumir

A tabela 1.1 é um exemplo do modelo de *scoring* de aprovação com os principais fatores geralmente analisados pelas instituições bancárias na concessão de crédito.

Tabela 1.1: Fatores analisados na concessão de crédito

Fatores de avaliação	**Escalões**	**Pontuação**
Idade	18 a 25 anos	0 pontos
	25 a 35 anos	5 pontos
	35 a 50 anos	8 pontos
	50 a 65 anos	10 pontos
	mais de 65 anos	5 pontos .
Estado civil	Solteiro	3 pontos
	Casado	8 pontos
	Viúvo	5 pontos
	Divorciado	1 ponto
	separação de bens	0 ponto.
Número de filhos	0	3 pontos
	1 ou 2	6 pontos
	3 ou 4	4 pontos
	4 ou 5	2 pontos
	mais de 5	0 pontos.
Residência	Arrendada	0 pntos
	casa familiares	0 pontos
	Própia com ónus	4 pontos
	Própia sem ónus	8 pontos
	Casa dos vizinhos	0 pontos .
Número de anos na residência	até 1 ano	0 pontos
	1 a 3 anos	2 pontos
	3 a 5 anos	4 pontos
	5 a 6 anos	6 pontos
	mais de 6 anos	8 pontos .
Número de anos no emprego	até 1 ano	0 pontos
	1 a 3 anos	2 pontos
	3 a 5 anos	6 pontos
	5 a 6 anos	8 pontos
	mais de 6 anos	10 pontos.
Rendimento líquido mensal	até 50.000cve	0 pontos
	50.000cve a 75.000cve	2 pontos
	75.000cve a 125.000cve	5 pontos
	125.000cve a 250.000cve	10 pontos
	mais de 250.000cve	8 pontos.

Fonte:Almeida [1]

Com base nesta grelha, os clientes potenciais podem obter pontuações que variam entre 0 e 60 pontos.
O modelo é completo com a definição de três escalões que correspondem a:

- Créditos aprovados;
- Situação duvidosa;
- Créditos recusados.

Em geral, as situações duvidosas são novamente analisadas pelo processo tradicional de avaliação, por um analista, e podem ser objeto de decisão favorável.

1.4.2 Vantagens dos modelos de credit scoring

Chaia [9] descreve de forma resumida vantagens dos modelos de *credit scoring*:

- Consistência: são modelos bem elaborados, que utilizam a experiência da instituição, e ajudam na gestão dos créditos de clientes já existentes e de novos solicitantes.
- Facilidade: os modelos de credit scoring visam a simplicidade e a fácil interpretação, com instalação relativamente fácil.
- Melhor organização da informação de crédito: a sistematização e organização das informações contribuem para a melhoria do processo de concessão de crédito.
- Redução da metodologia subjetiva: o uso do método quantitativo com regras claras e bem definidas contribui para a diminuição da subjetividade da avaliação do risco de crédito.
- Maior eficiência do processo: o uso de modelos de credit scoring na concessão de crédito direciona os esforços dos analistas, trazendo redução de tempo e maior eficiência a este processo.

Almeida [1], destaca que uma das vantagens deste modelo é a homogeneidade, isto é, com sistema deste tipo, torna-se possível ter um processo de decisão homogéneo para todo o banco, seja qual for a sua dimensão. Ainda refere que este modelo pode ser programado e trabalhado por computador, o que torna o processo de decisão muito rápido e com custo muito baixo , pois não é necessário recurso a pessoal especializado.

1.4.3 Desvantagens dos modelos de credit scoring

Caouette et al. [7] referem o aspeto temporal como a principal limitação:
Um modelo de *credit scoring* pode degradar-se com tempo se a população em que ele é aplicado diverge da população original que foi usada para construir o modelo.
Chaia [9] fala das principais desvantagens dos modelos de *credit scoring*:

- Custo de desenvolvimento: o desenvolvimento de sistemas de *credit scoring* acarreta custos ao nível da sua construção e manutenção.
- Excesso de confiança nos modelos: algumas estatísticas podem estimar por valores superiores a eficácia dos modelos, provocando um excesso de confiança nos mesmos por parte dos utilizadores, pois os menos experientes consideram-nos perfeitos e não põem em causa o seu resultado.
- Falta de dados adequados: a necessidade de dados não facultados pode originar problemas na utilização dos modelos e gerar resultados diferentes dos esperados. É necessário analisar a qualidade das informações disponibilizadas.
- Interpretação equivocada das classificações: o uso inadequado do sistema, devido à falta de treino e falta de formação sobre a sua utilização, pode provocar problemas sérios à instituição.

Capítulo 2

Estimação da Probabilidade de Incumprimento

Introdução

A concessão de crédito traduz-se na disponibilização de um valor, geralmente por uma instituição financeira, mediante uma promessa de pagamento desse mesmo valor no futuro, que pressupõe a confiança que o mesmo irá honrar os seus compromissos nas datas acordadas previamente.
O risco de crédito é o risco de perda em que se incorre quando há incapacidade de pagamento numa operação de concessão de crédito.
Para avaliar este risco, e tentar discriminar bons de maus clientes, muitas vezes utiliza-se modelo estatístico para quantificar o risco.
Segundo Fernandes [12] o interresse em utilizar modelos estatísticos na gestão e avaliação de risco de crédito no setor bancário e no sistema financeiro em geral tem aumentado a cada dia. Num cenário de muita competitividade no sistema bancário, as técnicas estatísticas tornaram-se, atualmente, uma das ferramentas mais importantes utilizadas na avaliação do risco de crédito dos empréstimos bancários.
A estimação da probabilidade de incumprimento de cada cliente no momento da concessão do crédito ao consumo, utilizando Regressão Logística é o objetivo principal deste capítulo. O modelo ajustado têm como variáveis mais significativas o Taxa nominal, Valor das prestações, Valor do empréstimo, Idade, Agência, Actividade profissional, Genero, Entidade patronal e Habilitações.
Este capítulo, está dividido em quatro seções. Na primeira seção apresentamos o modelo de Regressão Logística como um caso particular dos modelos lineares generalizados, na segunda seção apresentamos a carteira de crédito ao consumo, na terceira seção apresentamos os métodos para desnvolvimento e validação do modelo e na quarta seção aplicamos o modelo de Regressão Logítica e do modelo de Regressão Probit(este de forma superficial) a uma carteira de crédito ao consumo.

2.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG's) vêm unificar modelos anteriores desenvolvidos para a modelação estatística e foram introduzidos por [Nelder e Wedderburn [22]]. Silva [27] afirma que os MLG's são constituidos por três componentes:

1. **Componente aleatória**
   Esta componente do modelo estabelece que as variáveis aleatórias Y_i, que se pretendem modelar, são independentes com distribuição pertencente á Família Exponencial, em que
   $E[y_i|x_i] = \mu_i = b'(\theta_i), i = 1, ..., n$

2. **Componente estrutural ou sistemática**
   A componente sistemática dos MLG's, também designada de preditor linear, consiste numa combinação linear das variáveis independentes dada por
   $\eta_i = \mathbf{x}_i^T \beta$
   onde $\mathbf{x}_i$ é um vetor de especificação de dimensão $p+1$ tal que $\mathbf{x}_i = \left(1, x_{i1}, x_{i2}, ..., x_{ip)}\right)^T$ e β é um vetor de parâmetros de dimensão $p+1$.

3. **Função de ligação**
   Outra componente destes modelos é a relação entre o valor esperado μ e o preditor linear η, que se estabelece através de
   $\mu_i = h(\eta_i) = h\left(\mathbf{x}_i^T \beta\right)$
   onde $h(.)$, designada por função de ligação, é uma função monótona e diferenciável, tal que $g(.) = h^{-1}(.)$.
   Quando o preditor linear coincide com o parâmetro canónico, isto é, $\theta_i = \eta_i$, então a função de ligação denomina-se de **função de ligação canónica**.

Para Turkman [31] os MLG's, correspondem a uma síntese de vários modelos estatísticos. O mesmo autor indica os modelos que são casos prticulares dos MLG's: a regressão linear, regressão logística, modelo probit para estudos de proporção e regressão de Poisson. Os modelos lineares generalizados pressupõem que a variável dependente tenha uma distribuição pertencente a uma família particular, a família exponencial.

Seja $\mathbf{Y}$ a variável aleatória, de interesse primário, também designada por variável dependente ou variável resposta, e um vetor $\mathbf{X} = (x_1, ..., x_p)^T$ de p variáveis independentes, também designadas por covariáveis , variáveis explicativas ou variáveis preditoras, que se crê explicarem parte da variabilidade inerente a $\mathbf{Y}$. A variável dependente $\mathbf{Y}$ pode ser contínua, discreta ou dicotómica.
As variáveis independente, determinísticas ou estocásticas, podem ser de qualquer natureza: contínuas, discretas, qualitativas de natureza ordinal ou dicotómicas. Assume-se que os dados têm a forma

$$(y_i, x_i), i = 1, 2, ..., n \tag{2.1}$$

resultantes da realização de (**Y**,**X**) em n indivíduos, sendo as componentes Y_i do vetor aleatório $\mathbf{Y} = (y_1, ..., y_n)^T$ independentes. Pode-se representar 2.1 na forma matricial,

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \tag{2.2}$$

Os modelos lineares generalizados são uma extensão do modelo linear clássico,

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{2.3}$$

ou

$$\mathbf{Y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \tag{2.4}$$

onde $\beta = (\beta_0, \ldots, \beta_p)^T$ de parâmetros e ε um vetor de erros aleatórios.
A escolha da função de ligação depende do tipo davariável dependente. Por exemplo, para dados binários utiliza-se a função de ligação Logit que será tratada no ponto seguinte, na exposição acerca do modelo de Regressão Logística.

Definição 2.1.1 (**Família Exponencial**)**.** *Diz-se que uma variável aleatória* y *tem distribuição pertencente à* família exponencial *se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma:*

$$f\left(y|\theta,\phi\right) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.5}$$

Onde ϕ e θ são parâmetros escalares, $a(.), b(.)$ e $c(.)$ são funções reais conhecidas. Neste trabalho considera-se $\phi = 1$. Para aplicar a metodologia dos MLG's a um conjunto de dados existe a necessidade de após a formulação do modelo que se pensa adequado, de se proceder à realização de inferências sobre esse modelo. A inferência em MLG's baseia-se essencialmente na verosimilhança.

2.1.1 Fases dos modelos lineares generalizados

Segundo Silva [28] existem três fases que se devem seguir para modelar dados através dos Modelos Lineares Generalizados:

- **Formulação dos modelos**
  Nesta primeira fase, a formulação do modelo, há a necessidade de explorar cuidadosamente os dados, para se determinar uma distribuição adequada que defina a variável dependente e que permita selecionar as variáveis independente que melhor explicitam o modelo em estudo. Deve-se ainda escolher a função de ligação, que depende do tipo de variável dependente e do estudo particular que se pretende efetuar.

- **Ajustamentos dos modelos**
  O ajustamento do modelo, consiste na estimação dos parâmetros do modelo, isto é, na estimação do vetor dos coeficientes β associados às variáveis independente e respetivos erros padrão. Determinam-se intervalos de confiança e realizam-se testes de ajustamento, que permitam avaliar a qualidade do mesmo ou seja aqui fazemos toda a inferêcia relativamente aos parâmetros.

- **Seleção e validação dos modelos**
  Nesta última fase, procura-se encontrar submodelos que ainda se adequem aos dados, bem como procurar divergências que possam existir entre os dados e os valores preditos, localizar resíduos excessivos e possíveis outliers.

2.1.2 Regressão logística

Segundo McCullagh [21], Regressão Logística pode ser utilizada quando se deseja perceber a natureza do relacionamento entre a resposta média (probabilidade de ocorrência de um evento) e uma ou mais variáveis independentes, ou então com o objetivo preditivo, quando se deseja prever se determinado evento ocorrerá num prazo pré-definido, dado um conjunto de variáveis independentes.
A Regressão Logística é amplamente usada em ciências médicas e sociais, e tem outras denominações, como **modelo logístico**, **modelo logit** e **classificador de máxima entropia**. A Regressão Logística é utilizada em áreas como as seguintes:

- Em medicina, permite por exemplo determinar os fatores que caracterizam um grupo de indivíduos doentes em relação a indivíduos sãos.

- No domínio dos seguros, permite encontrar fracções da clientela que sejam sensíveis a determinada política securitária em relação a um dado risco particular.

- Em instituições financeiras, pode detetar os grupos de risco para a subscrição de um crédito.

- Em econometria, permite explicar uma variável discreta, como por exemplo as intenções de voto em actos eleitorais.

Este modelo usa como função de ligação a função logit:

$$\theta_i = ln\left(\frac{p_i}{1-p_i}\right)$$

A Regressão Logística é um caso particular dos modelos lineares generalizados, onde cada variável dependente é binomialmente distribuída, isto é,$Y_i \sim B(1, p_i)$ com probabilidade de "sucesso"p_i e de "fracasso"$(1-p_i)$.
A função probabilidade é dada por

$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}, y_i = 0, 1 \tag{2.6}$$

A cada indivíduo i está associado um vetor de especificação $\mathbf{z}_i$, que resulta do vetor das variáveis independentes $\mathbf{x}_i, i = 1, ..., n$.
A distribuição binomial pertence à família exponencial donde temos que

$$E[y_i] = p_i = \frac{e^{\theta_i}}{1 + e^{\theta_i}}.$$

Fazendo $\theta_i = \eta_i = \mathbf{z}_i^T \beta$ conclui-se que a associação entre o valor esperado da variável dependente e as variáveis independentes é feita através da função de ligação canónica, **função logit**. Assim a probabilidade de sucesso, $p_i = P[Y_i = 1|X = x_i]$, está relacionada com o vetor $\mathbf{z}_i$ através de

$$p_i = \frac{exp\left(\mathbf{z}_i^T \beta\right)}{1 + exp\left(\mathbf{z}_i^T \beta\right)} \tag{2.7}$$

assim sendo,

$$logit\,(p_i) = ln\left(\frac{p_i}{1 - p_i}\right) = ln\left(e^{\theta_i}\right) = \theta_i \tag{2.8}$$

e

$$logit\,(p_i) = \theta_i = \mathbf{z}_i^T \beta = \beta_0 + \sum_{i=1}^{p} (\beta_i x_i), i = 1, ..., p$$

Segundo Silva [28], os valores possíveis de p_i se situam no intervalo $[0, 1]$, daí o valor de p_i é frequentemente interpretado como a probabilidade de incumprimento. A principal vantagem da Regressão Logística é a capacidade de estimar as probabilidades individuais de cada cliente entrar em incumprimento, sendo este um dos objetivos deste capítulo.

2.1.3 Método de estimação

Segundo Turkman e Silva [31], para poder aplicar a metodologia dos Modelos Lineares Generalizados a um conjunto de dados, há necessidade, após a formulação do modelo que se pensa adequado, de se proceder à realização de inferências sobre esse modelo. A inferência com MLG é, essencialmente baseada na verosimilhança. Com efeito, não só o método da máxima verosimilhança é o método de eleição para estimar os parâmetros de regressão, como também os testes de hipóteses sobre os parâmetros do modelo e de qualidade do seu ajustamento são, em geral, métodos baseados na verosimilhança.
Os procedimentos de estimação e inferência a serem utilizados em Regressão Logística são casos particulares da metodologia de MLG's já descritos. A função de verosimilhança para o modelo logístico é dada por:

$$L\left(\beta_0, \beta_1, ..., \beta_p\right) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \tag{2.9}$$

Em que

$$p_i = \frac{exp\left(\mathbf{z}_i^T \beta\right)}{1 + exp\left(\mathbf{z}_i^T \beta\right)}, i = 1, ..., n. \tag{2.10}$$

O logaritmo da função de verosimilhança (função log-verosimilhança) de um modelo linear generalizado é dado por:

$$l(\beta) = ln\left[L(\beta)\right] = ln\left[\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\right] = \sum_{i=1}^{n} y_i\left(ln\left(p_i\right)\right) + \sum_{i=1}^{n}\left[(1-y_i)\, ln\left(1-p_i\right)\right]$$

Os estimadores são obtidos através da matriz hessiana da função de log-verosimilhança.

$$\frac{\partial^2 l(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{i,j}^2 p_i\left(1-p_i\right) \tag{2.11}$$

e

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{i,j} x_{i,l} p_i\left(1-p_i\right) \tag{2.12}$$

Igualando as expressõs 2.11 e 2.12 a zero, o sistema de equações obtido não é linear. Em virtude disso, essas equações não têm, em geral, solução analítica. Portanto, a sua resolução implica o recurso a métodos numéricos. Como a verosimilhança depende da probabilidade de sucesso desconhecida p_i, que, por sua vez, depende dos parâmetros $\beta's$, a função de verosimilhança pode ser vista como função de β.

2.1.4 Teste de significância

Segundo Turkman e Silva [31], quando estamos perante um problema de seleção de variáveis independentes e queremos testar se um submodelo é melhor que o modelo original, é comum utilizar a estatística de *Wald*, a estatística de *Wilks* ou a estatística da Razão de Verosimilhança . Estas estatísticas são deduzidas a partir das distribuições assimtóticas dos estimadores de máxima verosimilhança e de funções adequadas desses estimadores. Considera-se o teste de hipóteses da forma:

$$H_0 : \mathbf{C}\beta = \xi \quad versus \quad H_1 : \mathbf{C}\beta \neq \xi \tag{2.13}$$

Onde $\mathbf{C}$ é uma matriz $q \times p$, com $q \leq p$, de característica completa q e ξ é um vetor de dimensão q , previamente especificado. Seja o caso particular:

$$H_0 : \mathbf{C}\beta_j = 0 \quad versus \quad H_1 : \mathbf{C}\beta_j \neq 0 \tag{2.14}$$

para algum j, sendo $q = 1$ e $\mathbf{C} = (0, ..., 0, 1, 0, ..., 0)$, um vetor com todas as componentes nulas excepto a j-ésima que será igual a 1 e $\xi = 0$.
No caso em que uma variável é policotómica[1] e que toma $r+1$ valores distintos, é aconselhável construir r variáveis dicotómicas para as representar, havendo, nesse caso, r parâmetros $\beta's$ que lhe estão associados. Então, para averiguar se essa variável deve ou não ser incluída no modelo, interessa testar se os r parâmetros são significativamente diferentes de zero.

2.1.4.1 Estatística de *Wald*

A estatística de *Wald*, baseia-se na normalidade assimptótica do estimador de máxima verosimilhança, $\widehat{\beta}$.
Supõe-se que hipótese nula estabelece que $\mathbf{C}\beta = \xi$, Onde $\mathbf{C}$ é uma matriz $q \times p$, com $q \leq p$, de característica completa q. Seja $\widehat{\beta}$ o estimador de máxima verosimilhança de β, o qual tem uma distribuição assimptótica $N_p\left(\beta, \mathfrak{I}^{-1}\left(\widehat{\beta}\right)\right)$ (aqui o vetor β já foi substituído pela sua estimativa), onde $\mathfrak{I}^{-1}\left(\widehat{\beta}\right)$ é matriz de covariâncias. Dado que o vetor $\mathbf{C}\widehat{\beta}$ é uma transformação linear de $\widehat{\beta}$ então, pelas propriedades da distribuição normal multivariada,
$\mathbf{C}\widehat{\beta} \sim N_q\left(\mathbf{C}\beta, \mathbf{C}\mathfrak{I}^{-1}\left(\widehat{\beta}\right)\mathbf{C}^T\right)$
e portanto, sob a hipótese nula, a estatística

$$W = \left(\mathbf{C}\widehat{\beta} - \xi\right)^T \left[\mathbf{C}\mathfrak{I}^{-1}\left(\widehat{\beta}\right)\mathbf{C}^T\right]^{-1}\left(\mathbf{C}\widehat{\beta} - \xi\right) \tag{2.15}$$

Tem uma distribuição assimptótica de um χ^2 com q graus de liberdade.
A estatística W em 2.15 designa-se por *estatística de Wald*.
Para o teste de hipóteses referido em 2.14, designando por σ_{ii} o i-ésimo elemento da diagonal de $\mathfrak{I}^{-1}\left(\widehat{\beta}\right)$, a estatística de *Wald* resume-se a:

$$W = \left(\widehat{\beta_j} - \beta_j\right)^T [\sigma_{ii}]\left(\widehat{\beta_j} - \beta_j\right)$$

logo , sob H_0,

$$W = \frac{\widehat{\beta_j}^2}{\sigma_{ii}} \sim \chi_1^2$$

A estatística de *Wald* , geralmente, é mais utilizada para testar hipóteses sobre coeficientes individuais, embora também se use para testar hipóteses nulas do tipo $\beta_r = 0$ quando o subvetor β_r representa o vetor correspondente a uma recodificação de uma variável policotómica. A estatística de *Wald* é muito conhecida e útil na seleção/exclusão das variáveis independentes como iremos ver no ponto seguinte e na subseção 2.1.5.

2.1.4.2 Estatística da razão de verosimilhança

A estatística da razão de verosimilhanças, também conhecida por estatística de *Wilks*, é definida por:

[1] Variáveis qualitativas com mais de duas categorias

$$\Lambda = -2ln\left[\frac{max_{H_0}L(\beta)}{max_{H_0\bigcup H_1}L(\beta)}\right] = -2\left\{l\left(\widetilde{\beta}\right) - l\left(\widehat{\beta}\right)\right\} \tag{2.16}$$

onde $\widetilde{\beta}$, o estimador de máxima verosimilhança restrito, é o valor de β que maximiza a verosimilhança sujeito às restrições impostas pela hipótese nula, $C\beta = \xi$ e $l(.)$ corresponde ao máximo da função log-verosimilhança.
O Teorema de *Wilks* estabelece que, sob certas condições de regularidades , a estatística Λ tem, sob H_0, uma distribuição assimptótica de um χ^2 onde o número de graus de liberdade é igual à diferença entre o número de parâmetros a estimar sob $H_0 \bigcup H_1$ (neste caso p) e o número de parâmetros a estimar sob H_0 (neste caso $p - q$).
Assim, sob H_0,

$$\Lambda = -2\left\{l\left(\widetilde{\beta}\right) - l\left(\widehat{\beta}\right)\right\} \sim \chi^2_q \tag{2.17}$$

Com base na estatística da razão de verosimilhanças rejeita-se a hipótese nula $H_0 : \mathbf{C}\beta = \xi$, a um nível de significância α, se o valor observado da estatística Λ for superior ao quantil de probabilidade $1 - \alpha$ de um χ^2_q.
A estatística da razão de verosimilhanças, é mais utilizada quando é preciso comparar modelos que estão encaixados, isto é, modelos em que um é submodelo do outro.

2.1.5 Método de seleção das variáveis explicativas

As variáveis escolhidas inicialmente para a construção da carteira de crédito são apenas potenciais variáveis do modelo final, pois na prática e em várias situações, nem todas fornecem informações relevantes para explicar a natureza de crédito. Tornou-se assim, necessário escolher dentro desse grupo de variáveis, as mais significativas para explicar a natureza de crédito, sendo que essas compõem o modelo final.
Os métodos utilizados para realizar esta tarefa são: o método *Stepwise* e o *AIC*(*Akaike Information Criterion*).

2.1.5.1 *Stepwise*

O método *stepwise* é utilizado para inclusão/exclusão das variáveis independentes, dependendo da metodologia de seleção que se está a seguir.
Segundo Fernandes [12] há duas versões da abordagem *stepwise* que são a seleção *forward* seguido do teste de eliminção *backward*, ou eliminção *backward* seguido por seleção *forward*.
A seleção *forward* inicia-se somente com o modelo nulo, ou seja, sem nenhuma variável independente, e de seguida seleciona a variável a incluir, com base no menor valor de *p-value* da estatística de *Wald*. A eliminção *backward* das variáveis inicia-se com o modelo saturado, ou seja, considere-se todas as variáveis e elimine-se a variável com maior valor de *p-value* da estatística de *Wald*.
Esta abordagem combina esses dois passos que incluem ou excluem variáveis em cada iteração.
Segundo Turkman e Silva [31], o método de seleção *stepwise*, baseia-se no valor dos *p-values*

relativos às estatística da razão de verosimilhanças de *Wilks* entre modelos com inclusão ou exclusão de variáveis independentes para decidir quais as variáveis independentes que devem ser incluídas no modelo final.
Começa por calcular o valor do *p-value* dado pela estatística de *Wald* e, com base neste, escolhe qual a variável que, em primeira análise, deve sair (ou entrar) no modelo final.
Após a escolha da variável independente, faz-se uma segunda análise ao seu grau de importância através do valor do *p-value* da estatística da razão de verosimilhanças entre os modelos que a incluem e a excluem, e assim toma a decisão final sobre a exclusão (ou inclusão) da variável no modelo final.

2.1.5.2 AIC (*Akaike Information Criterion*)

Segundo Vale [32] o critério de informação de *Akaike* foi desenvolvido em 1971 por *Hirotsugu Akaike* sob o nome "*Akaike Information Criterion*"(AIC) e foi proposto por Akaike em 1974.
Turkman e Silva [31] afirmam que AIC não é um teste ao modelo no sentido de testar hipóteses, mas sim um teste entre modelos, ou seja, é uma ferramenta para selecionar um modelo de entre um conjunto de modelos. Dado um conjunto de dados, e vários modelos para os mesmos, o AIC classifica-os e o que tiver o menor AIC deve ser considerado o melhor modelo.
Este critério de seleção baseia-se na função Log-verosimilhança, com a introdução de um fator de correção como modo de penalização da complexidade do modelo.
Ainda segundo os mesmos autores a estatística correspondente para o modelo em H_0 é

$$AIC = -2l\left(\widetilde{\beta_1}, 0, \widetilde{\phi}\right) + 2r,$$

sendo:

- $r = dim\,(\beta_1)$
- $\widetilde{\beta_1}$, um estimador de β_1, e β_1 um subvetor de β.
- $\widetilde{\phi}$, um estimador consistente do parâmetro desconhecido ϕ.

A metodologia AIC tem como objetivo encontrar o modelo que melhor explica os dados com um mínimo de parâmetros livres.
Mas é importante perceber que o valor de AIC atribuído a um modelo serve apenas para classificar os modelos concorrentes e dizer qual é o melhor entre as alternativas dadas.

2.2 Carteira de Crédito

Segundo Fernandes [12] uma carteira de crédito ao consumo são uma mistura de variáveis contínuas, semi-contínuas e categóricas. Muitas vezes, a carteira possui milhões de registos e centenas de variáveis. Consequentemente, os modelos têm sido tradicionalmente construídos utilizando amostras em detrimento da população total.

2.2.1 Composição da carteira de crédito

De forma a modelar o risco de crédito de uma carteira de clientes, recorreu-se a dados de uma carteira de crédito ao consumo.
A carteira de crédito utilizada neste trabalho consiste numa carteira de crédito ao consumo e é composta por 33.781 clientes dos quais 3635 entraram em incumprimento.
No presente estudo, a variável de interesse é a natureza de crédito dos clientes, é uma variável nominal, em que são atribuídos os valores 0 para clientes cumpridores e 1 para clientes incumpridores.
Agruparam-se os dados de forma a termos uma base de dados pronta para utilizar no desenvolvimento do modelo de Regressão Logística e modelo probit.

2.2.2 Tratamento dos dados da carteira de crédito

Segundo Gujarati [17], sucesso de qualquer análise econométrica depende basicamente da disponibilidade de dados apropriados.
Portanto, é essencial que despendamos algum tempo examinando a natureza, as fontes e as limitações dos dados com que poderemos nos deparar na análise impírica.
Segundo Vale [32], o tratamento de dados numa amostra de uma carteira de crédito real consistiria, com base nos dados recolhidos, na identificação das variáveis estatisticamente significativas e no desenvolvimento dos modelos.

2.2.3 Definição de cliente incumpridor

Aqui vamos explicar os conceitos de clientes incumpridores e clientes cumpridores, baseada no número de dias em atraso no pagamento das suas prestações, tais conceitos que utilizaremos no desenvolvimento do nosso trabalho.
Segundo Fernandes [12] cada instituição tem a sua própria política de crédito e estes conceitos, de cliente cumpridor ou incumpridor podem mudar dependendo da instituição, das conjecturas económicas ou do próprio país.
Segundo Vale [32] no desenvolvimento de um trabalho sobre risco de crédito, no qual se estudam modelos de *credit scoring*, é muito importante a definição de natureza de crédito, isto é, o conceito de cliente cumpridor e cliente incumpridor.
Siddiqi [26], diz que um cliente é considerado incumpridor (”mau”ou em ”default”), se se atrasar no pagamento de alguma das prestações do contrato por um período superior a 90 dias nos primeiros doze meses da vigência do contrato.
Para o mesmo autor, uma vez definidos os critérios que classificam um cliente como incumpridor, clientes incumpridores, considerar-se um cliente cumpridor quando este tem no máximo 30 dias em atraso no pagamento das suas prestações. Num esquema usual de prestações mensais, significa que o cliente não tem nenhuma prestação em atraso.
Nesta dissertação optou-se por classificar um cliente como incumpridor quando esteve, pelo menos uma vez durante o contrato, com mais de noventa dias de incumprimento.

2.2.4 As Variáveis independentes utilizadas

Segundo Anderson [2] e Siddiqi [26], a maioria dos modelos de *credit scoring* desenvolvidos, independentemente do tipo, tem um grande número de variáveis independentes que poderiam ser utilizadas. No entanto, regra geral, há apenas entre 6 a 15 características que melhor explicam o comportamento do cliente. A maioria das bases de dados utilizadas na construção de modelos de *credit scoring* apresenta um grande número de variáveis que, normalmente, são uma mistura de variáveis contínuas, semi-contínuas e categóricas. Estas podem ser categorizadas antes da sua utilização, fazendo seleção das que fazem parte do modelo final.
O modelo de *credit scoring* que propomos desenvolver é Regressão Logística, neste caso a variável dependente é a natureza de crédito (incumprimento ou cumprimento). Para classificar as observações de acordo com a natureza de crédito, foram selecionadas variáveis independentes, que pudessem influenciar a situação dos clientes nas suas obrigações enquanto credores.
Inicialmente, foi construída uma base de dados que agrega um conjunto de possíveis variáveis independentes pré-selecionadas para utilização na construção dos modelos.
Na tabela 2.1 apresentamos as variáveis independentes pré-selecionadas para o desenvolvimento deste estudo e algumas características da mesma.

Tabela 2.1: Lista de variáveis independentes

Variáveis	Código da variável	Natureza da variável	Tipo
Variáveis sociodemográficas			
Idade	Idade	Quantitativa	Numérica
Género	Genero	Qualitativa	Categórica
Estado civil	Civil	Qualitativa	Categórica
Habilitações	Habilitacoes	Qualitativa	Categórica
Atividade profissional	ActProfissional	Qualitativa	Categórica
Entidade patronal	EntPatronal	Qualitativa	Categórica
Agência	Agencia	Qualitativa	Numérica
Prazo	Prazo	Quantitativa	Numérica
Variáveis de Relação Cliente Banco			
Valor do empréstimo	ValorEmprest	Quantitativa	Numérica
Tipo de garantia	TipoGarantia	Quali-Quantitativa	Categórica
Taxa nominal	TxNominal	Quantitativa	Numérica
Prestações pagas	PrestPagas	Quantitativa	Numérica
Valor de prestação	ValorPrest	Quantitativa	Numérica

Fonte: Adaptação de Vale [32]

A escolha dessas variáveis independentes, foi com base nos outros trabalhos de Fernandes [12], Vale [32] , Esquivel et al. [11] e Silva [27].

A variável dependente, neste modelo, será, como referido anteriormente, a natureza de crédito de cada cliente, medida como a ocorrência ou não de incumprimento.
Como referido anteriormente, a metodologia utilizada neste trabalho pressupõe a análise de dados históricos da carteira, de forma a poder inferir sobre as variáveis independentes de interesse no fenómeno de ocorrência de incumprimento.
A escolha das variáveis independentes consideradas com mais relevância na explicação da ocorrência de incumprimento baseou-se nos estudos anteriores já referidos, e na sensibilidade a este assunto reportada por especialistas da área de crédito.

2.2.5 Análise estatística das variáveis

A Estatística descritiva tem como objetivo descrever e analisar a informação que nos é fornecida, caracterizando assim o conjunto de dados de que se dispõe.
Utilizando o software R-studio podemos tratar estatisticamente a base de dados da carteira de crédito.
A tabela 2.2 resume, para o conjunto de dados que temos, os principais resultados da análise da base dos dados relativamente a variável dependente.
As tabelas 2.3 e 2.4 ilustram o resultado da categorização das variáveis independentes

Tabela 2.2: Informação da variável dependente

	Cumpridor	Incumpridor	Total
População	30146	3635	33.781
%	89.2	10.8	100

Fonte: Fernandes [12]

utilizadas no estudo, o número de processos existentes e as respetivas percentagens em cada categoria.
Segundo Fernandes [12], categoriza-se as variáveis, por um lado, para evitar categorias com poucas observações, pois tal pode conduzir a estimativas pouco robustas dos parâmetros associados. Por outro lado tem a ver com a eliminação de parâmetros desnecessários para o desenvolvimento do modelo.

Tabela 2.3: Descrição das categorias de variáveis independentes

Variáveis	Categoria	Grupo	Número de processos	%
Idade	1	Inferior a 27 anos	2617	7.7
	2	Entre 27 e 31 anos	16210	48.0
	3	Entre 32 e 49 anos	6271	18.6
	4	Superior a 49 anos	8683	25.7
Género	1	Masculino	19922	59.0
	2	Feminino	13859	41.0
Estado civil	1	Divorciado, Separado	24922	73.8
	2	Solteiro, união de facto, viúvo	8719	25.8
	3	Casado	140	0.4
Habilitações	1	Habilitação desconhecida	6116	18.1
	2	Escolaridade obrigatória	17884	52.9
	3	Ensino Secundário	8020	23.7
	4	Curso médio e Formação superior	1761	5.2
Atividade profissional	1	Empregado de escritório, Comércio, Serviço e quadro médio	17628	52.2
	2	Estudante, Liberal, Quadro Superior e Outras	13074	38.7
	3	Atividade desconhecida, Doméstico, Pequenas e médias empresas	3079	9.1
Entidade patronal	1	Ministérios e aposentados, Pensionistas	13918	41.2
	2	Câmara Municipal, Grandes empresas, Instituição financeira e Institutos públicos	8845	26.2
	3	Hotel, Restaurante, Não declarada	6991	20.7
	4	Conta própria e Outras	4027	11.9
Agência	1	4,7,14,18,28 e 29	812	2.4
	2	5 e 10	5414	16.0
	3	6, 23, 26, e 30	13372	39.6
	4	1, 2, 8,9, 11 e 24	7837	23.2
	5	3, 12, 19, 22, 25, 27, 31 e 32	6346	18.8
Prazo	1	Inferior a 13	5947	17.6
	2	Entre 13 e 48	16329	48.3
	3	Superior a 48	11505	34.1

Fonte: Adaptação de Silva [27]

Tabela 2.4: Descrição das categorias de variáveis independentes

Variáveis	Categoria	Grupo	Número de processos	%
Valor do empréstimo	1	Inferior a 97.280$00	4221	12.5
	2	97.280$00 a 200.000$00	8716	25.8
	3	200.000$00 a 350.000$00	8318	24.6
	4	350.000$00 a 1.000.000$00	10641	31.5
	5	Superior a 1.000.000$00	1885	5.6
Tipo de Garantia	1	Hipoteca sem imóveis para habitação...Outras hipotecas	1876	5.6
	2	Outras cauções	23794	70.4
	3	Outras entidades	8111	24.0
Taxa Nominal	1	Inferior a 12,5%	4198	12.4
	2	Superior ou igual a 12,5%	29583	87.6
Prestações pagas	1	Inferior a 18	9608	28.4
	2	Entre 18 e 21	4623	13.7
	3	Entre 22 e 27	6756	20.0
	4	Entre 28 e 37	8014	23.7
	5	Superior a 37	4780	14.1
Valor de prestação	1	Inferior a 2952$00	2668	7.9
	2	De 2952$00 a 5687$00	6261	18.5
	3	De 5687$00 a 8908$00	6317	18.7
	4	De 8908$00 a 28384$00	16601	49.1
	5	Superior ou igual a 28384$00	1934	5.7

Fonte: Adaptação de Silva [27]

2.3 Desenvolvimento e Validação de Modelos

No desenvolvimento de um modelo de *credit scoring*, a amostra em estudo poderá ser dividida, dependendo da técnica utilizada, em amostra de treino, validação e teste e após a decisão da proporção de cumpridores e incumpridores a incluir na amostra final, (Siddiqi [26]).
Existem várias formas de dividir o conjunto de dados da amostra de treino (amostra em que o modelo de *credit scoring* é desenvolvido) e amostra de validação (amostra em que o modelo é validado). Normalmente, 70% a 80% da amostra é utilizada para treinar os modelos e os restantes 20% a 30% são reservados para a fase de validação dos modelos.
Segundo Gestel et al. [14] para a validação de um modelo de *credit scoring*, três requisitos fundamentais são considerados:

- *Estabilidade*: Um modelo estável exige coeficientes bem determinados e com grande nível de confiança e resultados semelhantes em características de desempenho, se testados dentro e fora da amostra.
- *Legibilidade*: Um modelo é legível quando os seus coeficientes têm uma interpretação fácil.
- *Poder discriminatório*: Esta característica é definida pelo Comité de Supervisão Bancária de Basileia (2005) como a capacidade de classificar corretamente as observaões sobre a base de probabilidade de incumprimento através da atribuição de pontuações.

Segundo Fernandes [12], ao testar o poder discriminatório e a comparabilidade dos modelos, várias medidas de desempenho são utilizadas: Percentagem dos Corretamente Classificados (PCC), Sensibilidade (SENS), Especificidade (Esp), Kolmogorov-Smirnov (KS), Índice de Gini, AUC e Rácio de Precisão (AR).
Nos modelos de *credit scoring*, essas medidas permitem-nos averiguar a discriminação ideal entre clientes cumpridores e incumpridores. Onde destacamos: a curva ROC e o índice ou coeficiente de Gini.

2.3.1 Curva ROC

A curva *ROC* (*Receiver Operating Characteristic*), também conhecida como curva de Lorenz é baseada nos conceitos de sensibilidade e especificidade estatísticas (medidas da taxa de classificações corretas) que podem ser obtidas a partir da construção de matriz de classificação (2×2), obtidas do resultado da classificação dos indivíduos gerada pelo modelo estimado.
Tendo o modelo ajustado, a partir de uma amostra de n clientes, atribui-se um score S a cada indivíduo. O i-ésimo indivíduo será classificado como incumpridor se $S_i \leq P_c$, (em que P_c é um ponto de corte para o score S_i, pré-determinado) e como cumpridor, caso contrário. Para um determinado P_c é possível determinar a matriz de classificação, também conhecida

pela matriz de confusão ou tabela de contingência, como apresentada na tabela 2.5.
em que:

Tabela 2.5: Matriz de classificação

	Previsto		
Observado	Cumpridor	Incumpridor	Total
Cumpridor	n_{cc}	n_{ci}	$n_{c\bullet}$
Incumpridor	n_{ic}	n_{ii}	$n_{i\bullet}$
Total	$n_{\bullet c}$	$n_{\bullet i}$	$n_{\bullet\bullet}$

Fonte: Adaptação de Fernandes [12]

n_{cc} - Número de clientes "cumpridores"classificados como "cumpridores- classificação correta;
n_{ci} - Número de clientes "cumpridores"classificados como "incumpridores- classificação incorreta;
n_{ic} - Número de clientes "incumpridores"classificados como "cumpridores- classificação incorreta;
n_{ii} - Número de clientes "incumpridores"classificados como "incumpridores- classificação correta.
Através da matriz de classificação, é possível determinar as taxas de classificações corretas, que correspondem a medidas de especificidade (proporção dos clientes "incumpridores", classificados corretamente por terem score menor que um ponto de corte) e de sensibilidade (proporção de clientes "cumpridores", classificados corretamente por terem score igual ou superior a um ponto de corte), ou seja:

$$sensibilidade = \frac{n_{cc}}{n_{c\bullet}} \quad e \quad especificidade = \frac{n_{ii}}{n_{i\bullet}} \tag{2.18}$$

Pode-se calcular também a precisão do modelo, que será dada pela proporção total da classificação correta, ou seja:

$$precisão = \frac{n_{ii} + n_{cc}}{n} \tag{2.19}$$

A curva ROC é construída a partir da união dos pontos formados pelos valores da sensibilidade e (1-especificidade), calculadas a partir de todas as matrizes de classificação, geradas pelas observaçõs da amostra, considerando-se diferentes pontos de corte do modelo.

2.3.2 Coeficiente de Gini

Segundo Thomas et al. [30], o cálculo de coeficiente de Gini resulta diretamente da utilização da curva ROC. Pode ser definido como sendo o quociente da área entre a reta e a curva sobre a área total acima da diagonal. Quanto mais a curva se afasta da reta, maior será o coeficiente de Gini e maior será a discriminação entre os clientes cumpridores

e incumpridores, ou seja,

$$Gini = \frac{área\ \ entre\ \ a\ \ reta\ \ e\ \ a\ \ curva}{área\ \ total\ \ acima\ \ da\ \ diagonal} \tag{2.20}$$

Como área entre a reta e a curva é a diferença entre a área acima da diagonal e a área acima da curva e como toda a área acima da diagonal é igual a metade da área do quadrado, pode-se obter o coeficiente de Gini da seguinte forma:

$$\begin{aligned} Gini &= \frac{área\ \ acima\ \ da\ \ diagonal - área\ \ acima\ \ da\ \ curva}{\frac{1}{2}} \\ \Leftrightarrow Gini &= 2 \times (área\ \ acima\ \ da\ \ diagonal - área\ \ acima\ \ da\ \ curva) \end{aligned}$$

ou ainda, diretamente do valor obtido da curva ROC, como:

$$Gini = 2 \times (ROC - 0.5) \tag{2.21}$$

sendo ROC, neste caso, o valor obtido do cálculo da área sob a curva ROC.
A tabela 2.6, apresenta-se os valores intervalares para avaliação do resultado da área sob a curva ROC, aplicadas em modelos de regressão linear.

Tabela 2.6: Valores de referência da curva ROC

Valor da Curva ROC	Níveis de Discriminação
[0; 0.7[	Baixa
[0.7; 0.8[	Aceitável
[0.8; 0.9[	Bom
[0.9; 1[	Excelente

Fonte: Adaptação de Hosmer [19]

2.4 Aplicação e Resultados

Nesta seção, recorrendo ao software R-studio, aplicam-se os métodos estatísticos estudados para a modelação da carteira de crédito e os respetivos resultados.
Será aplicada a Regressão Logística para ajustar os modelos e os métodos *Stepwise* e *AIC* para a escolha das variáveis que melhor explicam a variável independente e que constituem o modelo final.

2.4.1 Ajustamento do modelo utilizando regressão logística

É importante relembrar que um modelo linear tem a forma:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon, \tag{2.22}$$

Sendo $\beta_i, i = 1,\ldots,p$ os coeficientes discriminantes para as variáveis $x_i, i = 1,\ldots,s$, β_0 o intercepto e ε um vetor de erros aleatórios.
Considera-se, inicialmente, o seguinte conjunto de variáveis independentes: Idade, Género, Estado civil, Habilitação, Atividade profissional, Entidade patronal, Agência, Prazo, Valor do empréstimo, Tipo de garantia, Taxa nominal, Prestações pagas e Valor de prestações.
De forma a não alongar a descrição dos procedimentos até se encontrar o melhor modelo, todos os passos serão descritos de forma sucinta. Convém afirmar que para o desnvolvimento do modelo vamos utilizar 80% dos dados da carteira correspondente a 27025, sendo os restantes 20% reservado para a validação.

2.4.1.1 Modelo completo

O modelo completo é o modelo com todas as variáveis independentes, escolhidas para ajustar o modelo final.
Nesta primeira análise é excluída a variável independente prestações pagas, codificado (PrestPagas), uma vez que o interesse é estudar a probabilidade de incumprimento para um cliente novo que pretende contratar um crédito ao consumo, pelo que, a priori esta informação não estará disponível.
A seguir temos o resultado de estimação por máxima verosimilhança do modelo completo.

```
> Modcomp<-glm(Default~Prazo+TxNominal+ValorPrest+ValorEmprest+Idade+
Agencia+ActProfissional+Genero+EntPatronal+Civil+Habilitacoes+
TipoGarantia, data = Treino, family=binomial)
> summary( Modcomp)

Call:
glm(formula = Default ~ Prazo + TxNominal + ValorPrest + ValorEmprest +
    Idade + Agencia + ActProfissional + Genero + EntPatronal +
    Civil + Habilitacoes + TipoGarantia, family = binomial, data = Treino)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3532  -0.5130  -0.3750  -0.2598   3.1585

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.47007    0.26352 -20.758  < 2e-16 ***
Prazo            0.03910    0.04196   0.932  0.35135
TxNominal        0.49529    0.10623   4.663 3.12e-06 ***
ValorPrest       0.20415    0.04421   4.618 3.88e-06 ***
ValorEmprest     0.12197    0.04386   2.781  0.00542 **
Idade           -0.26300    0.02457 -10.703  < 2e-16 ***
Agencia          0.31732    0.02098  15.123  < 2e-16 ***
```

```
ActProfissional  0.27870    0.03264   8.539  < 2e-16 ***
Genero          -0.24702    0.04280  -5.772 7.84e-09 ***
EntPatronal      0.44174    0.01956  22.583  < 2e-16 ***
Civil           -0.01339    0.05122  -0.261  0.79376
Habilitacoes    -0.19938    0.02999  -6.648 2.96e-11 ***
TipoGarantia     0.04380    0.04101   1.068  0.28545
---
Null deviance: 18701  on 27024  degrees of freedom
Residual deviance: 16798  on 27012  degrees of freedom
Signif. codes:  ***= 0.001,    ** =0.01, *= 0.05, .= 0.1
AIC: 16824
```

O método de *stepwise* da Regressão Logística selecionou as variáveis da tabela 2.1 como explicador da variável *Default*. É de referir que a idade é uma variável utilizada na maioria dos modelos de risco de crédito por ser considerada um indicador da etapa de ciclo de vida do cliente.

2.4.1.2 Modelo nulo

Realizou-se a mesma análise mas para o modelo nulo (modelo que não contém nenhuma variável independente) e ainda, através do estatística da razão de verosimilhança, comparou-se qual dos dois modelos seria o que melhor explicaria a variável dependente.

```
> ModNulo<-glm(Default~1,family = binomial(),data=Data)
> summary(ModNulo)
Call:
glm(formula = Default ~ 1, family = binomial(), data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4772  -0.4772  -0.4772  -0.4772   2.1115

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.11544    0.01756  -120.5   <2e-16 ***
---
Null deviance: 23071  on 33780  degrees of freedom
Residual deviance: 23071  on 33780  degrees of freedom
Signif. codes:  ***= 0.001,    ** =0.01, *= 0.05, .= 0.1
AIC: 23073
```

2.4.1.3 Comparação entre modelos

Considere-se o modelo nulo encaixado no modelo completo. No software R, a função anova, com indicação de teste test = "Chisq" realiza a estatística da razão de verosi-

milhança, testando se o modelo com menos variáveis independentes é significativamente melhor que o modelo com mais variáveis independentes. Na prática, está-se a analisar:
H_0 : Modelo completo é melhor que o modelo nulo
versus
H_1: Modelo nulo é melhor que o modelo completo

```
> anova(ModNulo,Modcomp, test = "Chisq")
Analysis of Deviance Table

Model 1: Default ~ 1
Model 2: Default ~ Prazo + TxNominal + ValorPrest + ValorEmprest + Idade +
    Agencia + ActProfissional + Genero + EntPatronal + Civil +
    Habilitacoes + TipoGarantia
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     27024      18701
2     27012      16798 12   1903.3 < 2.2e-16 ***
Signif. codes:  ***= 0.001,    ** =0.01, *= 0.05, .= 0.1
```

Como valor-p do teste Wilks é menor do que nível de significância $\alpha = 0.001$, concluimos que o modelo completo com todos os parâmetros irrestrito oferece uma qualidade de ajuste significativamente melhor do que o modelo nulo dos dados amostrais. Ou seja deve-se rejeitar a hipótese de nulidade de todos os parâmetros do modelo, isto é, deve-se rejeitar a hipótese de que o modelo nulo é melhor que o modelo completo. Deste modo, é aceitável considerar que os coeficientes acrescidos são significativamente diferentes de zero, sendo assim pelos menos uma delas será estatisticamente significativo na modelação da variável de interesse.
Além da estatística de *Wilks*, podem-se observar os valores do Critério de Informação de Akaike (AIC) para cada um dos modelos. Verifica-se que o AIC para o modelo completo (16824) é menor que a do modelo nulo (23073) donde, mais uma vez se rejeita a hipótese de que o modelo nulo é significativamente melhor que o modelo completo.

2.4.1.4 Selecão das covariáveis através do método stepwise - backward

O modelo completo definido anteriormente é o modelo que se utiliza para inicializar o método *stepwise - backward*.
Observando os resultados do ajustamento do modelo completo, é possível identificar algumas variáveis independentes que a estatística de **Wald** considera menos significativas.
De acordo com o método *stepwise*, as variáveis a ser excluida do modelo são as que tem maior *p-value*, excluir-se-ão do modelo todas as variáveis com *p-value* superior a um nível de significância de 0.01, ou seja são insignificativas as variáveis independentes com *p-value* superior a 0.01.
Logo opta-se por analisar o modelo após a exclusão das variáveis: Prazo, Estado civil(Civil) e Tipo de garantia (TipoGarantia), tal modelo designamos por modelo3.

```
  > modelo3<-glm(Default~TxNominal+ValorPrest+ValorEmprest+Idade+Agencia+ActProfissio
  EntPatronal+Habilitacoes, data = Treino, family=binomial)
> summary(modelo3)

Call:
glm(formula = Default ~ TxNominal + ValorPrest + ValorEmprest +
    Idade + Agencia + ActProfissional + Genero + EntPatronal +
    Habilitacoes, family = binomial, data = Treino)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3341  -0.5134  -0.3746  -0.2609   3.1589

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.34668    0.24099 -22.186  < 2e-16 ***
TxNominal        0.51488    0.10495   4.906 9.29e-07 ***
ValorPrest       0.18004    0.03518   5.118 3.08e-07 ***
ValorEmprest     0.15027    0.03020   4.975 6.52e-07 ***
Idade           -0.26776    0.02244 -11.932  < 2e-16 ***
Agencia          0.32442    0.02000  16.218  < 2e-16 ***
ActProfissional  0.27692    0.03254   8.509  < 2e-16 ***
Genero          -0.24729    0.04271  -5.790 7.05e-09 ***
EntPatronal      0.44006    0.01951  22.551  < 2e-16 ***
Habilitacoes    -0.20185    0.02992  -6.746 1.52e-11 ***
---
Signif. codes:  ***= 0.001, **=0.01, *= 0.05, .= 0.1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 18701  on 27024  degrees of freedom
Residual deviance: 16800  on 27015  degrees of freedom
AIC: 16820
```

Observando os resultados no software R, verificamos que no modelo3:

- As variáveis tem *p-value* inferior a nível de significância de 0.01, ou seja são significativas todas as variáveis independentes que compõem esse modelo.

- Apresenta melhor performance do que o modelo completo, pois, verifica-se que o AIC para o modelo3 (16820) é menor que o do modelo completo (16824).

Assim, considera-se o modelo3 que designamos por modelo ModStp como o modelo que melhor explica a variável dependente (*Default*).

Mais concretamente, o melhor modelo encontrado através do método *Stepwise-backward* e AIC é constituído pelas covariáveis : Taxa nominal (TxNominal), Valor das prestações (ValorPrest), Valor do empréstimo (ValorEmprest), Idade, Agência (Agencia), Actividade profissional (ActProfissional), Genero, Entidade patronal (EntPatronal) e Habiltações (Habilitacoes).
Seguidamente compara-se o modelo completo com o ModStp, recorrendo a estatística da razão de verosimilhança.
No que se segue, interessa testar:
H_0 : Modelo completo é melhor que o ModStp
versus
H_1: ModStp é melhor que o modelo completo

```
 > anova(ModStp, Modcomp, test = "Chisq")
Analysis of Deviance Table

Model 1: Default ~ TxNominal + ValorPrest + ValorEmprest + Idade +
 Agencia +  ActProfissional + Genero + EntPatronal + Habilitacoes
Model 2: Default ~ Prazo + TxNominal + ValorPrest + ValorEmprest + Idade +
Agencia + ActProfissional + Genero + EntPatronal + Civil +
Habilitacoes + TipoGarantia
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     27015      16800
2     27012      16798  3   2.0818   0.5556
```

Observando a estatística de *Wilks*, verificamos que existe evidência para rejeitar a hipótese nula, ou seja, rejeitar a hipótese de que o melhor modelo é o modelo completo.
Observa-se que o valor do *p-value* pela estatística de *Wilks* não é tanto pequeno (0.5556) logo, pode dizer-se que se deve rejeitar, com alguma evidência, a hipótese de que o modelo completo é melhor que o ModStp, ou seja, pelo menos um das variáveis adicionadas será significativamente igual a zero. Desta forma podemos ajustar o modelo Logit ao modelo ModStp.

2.4.2 Estimação da probabilidade de incumprimento

Nesta subseção pretende-se ajustar um modelo de Regressão Logística para estimar a probabilidade de incumprimento de um cliente, selecionando o melhor modelo de ajustamento. Para tal, recorrer-se-á aos métodos descritos nos capítulos anteriores e ao software R.
Após o estudo estatístico dos dados e aplicação de métodos de seleção de variáveis independentes encontrou-se o modelo final (ModStp) que melhor explica a variável *default*. Este modelo contempla as covariáveis: Taxa nominal (TxNominal), Valor das prestações

(ValorPrest), Valor do empréstimo (ValorEmprest), Idade, Agência (Agencia), Actividade profissional (ActProfissional), Genero, Entidade patronal (EntPatronal) e Habiltações (Habilitacoes).
Do ajustamento do modelo final resultam os coeficientes $\beta_i^{'}s$ que serão utilizados, nesta seção, para a estimação da probabilidade de incumprimento.
A tabela 2.7 contém os valores dos coeficientes, $\beta_i s$, a considerar. Então, pela equação 2.22

Tabela 2.7: Coeficientes do modelo ajustado

Característica		Coeficientes
Intercept	β_0	-5.34668
TxNominal	β_1	0.51488
ValorPrest	β_2	0.18004
ValorEmprest	β_3	0.15027
Idade	β_4	-0.26776
Agencia	β_5	0.32442
ActProfissional	β_6	0.27692
Genero	β_7	-0.24729
EntPatronal	β_8	0.44006
Habilitacoes	β_9	-0.20185

Fonte: Adaptação de Vale [32]

tem-se:

$$\begin{aligned} Y &= \beta_0 + \beta_1.TxNominal + \beta_2.ValorPrest + \beta_3.ValorEmprest + \beta_4.Idade + \\ &+\beta_5.Agencia + \beta_6.ActProfissional + \beta_7.Genero + \beta_8.EntPatronal + \beta_9.Habilitacoes \end{aligned}$$

Finalmente, para estimar a probabilidade de incumprimento de cada cliente, p_i, utiliza-se a equação 2.7, definida na seção 2.1.2. Esta equação, pode ser reescrita da seguinte forma:

$$p_i = \frac{e^{y_i}}{1 + e^{y_i}} \tag{2.23}$$

Na tabela 2.8 estão as probabilidades de incumprimento estimadas, conforme se tem vindo a descrever, para clientes com diferentes características.
Apartir da tabela 2.8, podemos afirmar que um cliente com as características do cliente1 tem a probabilidade de incumprimento de 20,1%. Dependendo da política da instituição e do ponte corte estabelicido pode-se classificar o cliente como bom ou mau pagador.
O modelo de Regressão Logítica, como foi referido nas seções anteriores, estima a probabilidade de incumprimento de um cliente individualmente no instante da concessão do crédito, auxilia os gestores de crédito na tomada de decisão, ou seja, ajuda os gestores a decidir conceder ou não o crédito a um novo credor.

Tabela 2.8: Probabilidades de incumprimento estimadas - Exemplos

			Características			
Cliente	TxNominal	ValorPrest	ValorEmprest	Idade	Agencia	ActProfissional
1	≥ 12.5	De 8908 a 28384	De 350000 a 1000000	Entre 27 e 31	23	Liberal
28	≥ 12.5	De 5687 a 8908	De 200.000 a 350000	Inferior a 27	8	E. de escritório
450	≥ 12.5	De 2952 a 5687	De 97280 a 200000	Superior a 49	30	E. de Escritório
573	≥ 12.5	De 2952 a 5687	De 97280 a 200000	Superior a 49	26	A. Desconhecida
2713	≥ 12.5	Sup. ou igual a 28384	Sup. a 1000000	Entre 27 e 31	11	Quadro Sup.
3012	≥ 12.5	De 8908 a 28384	De 200.000 a 350000	Superior a 49	5	E. de escritório
23428	< 12.5	De 2952 a 5687	Inferior a 97280	Entre 27 e 31	10	Liberal
32415	≥ 12.5	De 2952 a 5687	De 97280 a 200000	Superior a 49	23	A. desconhecida

			Características		p_i
Cliente	Genero	EntPatronal	Habilitacoes	$\hat{y}_i$	$\frac{e^{\hat{y}_i}}{1+e^{\hat{y}_i}}$
1	Femenino	Grandes Emp.	Hab. Desconhecida	$-1,38$	$0,201$
28	Masculino	Inst. Financeira	Curso médio	$-1,95$	$0,124$
450	Masculino	Aposentado	Esc.obrigatória	$-3,23$	$0,0374$
573	Femenino	Grandes Emp.	Hab. Desconhecida	$-2,30$	$0,0912$
2713	Femenino	Inst. Financeira	Ensino secundário	$-2,20$	$0,0997$
3012	Femenino	Aposentado	Esc.obrigatória	$-2,05$	$0,1146$
23428	Masculino	Grandes Emp.	Ensino secundário	$-3,19$	$0,0397$
32415	Femenino	Inst. Financeira	Hab. Desconhecida	$-3,18$	$0,0399$

Fonte: Adaptação de Vale [32]

2.4.3 Avaliação da capacidade preditiva do modelo

Para validar os resultados obtidos, antes da aplicação dos modelos, dividimos a carteira de crédito em duas amostras, a amostra de desenvolvimento e a amostra de validação. O objetivo é avaliar se o modelo construído a partir da amostra de desenvolvimento, classifica corretamente as observações que não foram utilizadas no processo de estimação. Desta forma, a carteira de clientes (33781 clientes) foi dividida numa amostra de desenvolvimento (27025 clientes que corresponde a 80% dos dados) utilizada nas seções anteriores e numa amostra de teste (6756 clientes os restantes 20%) que é utilizada nesta seção para avaliar a capacidade preditiva do modelo Modstp estimado.
A matriz de classificação estudada na seção 2.3.1, nos ajuda a avaliar a capacidade preditiva do modelo ajustado levando em conta o ponto corte e o score atribuido a cada cliente. Para a construção da matriz de classificação consideraram-se que os clientes com a probabilidade de incumprimento superior ou igual a 40% é considerado imcumpridor, caso contrário é considerado cumpridor.
Desta forma contabilizamos os clientes da amostra de teste que eram incumpridores (1) e os clientes cumpridores (0) e para cada um destes dois grupos, os clientes cuja probabilidade de incumprimento estimada era superior ou igual a 40% e inferior a 40%, como se pode observar na tabela 2.9.

Tabela 2.9: Contabilização de clientes cumpridores e incumpridores

	Previsto		
	≤ 40(Cumpridor)	> 40(Incumpridor)	Total
0 (Cumpridor)	6043	44	6087
1 (Incumpridor)	642	27	669
Total	6685	71	6756

Fonte: Adaptação de Vale [32]

em que:
6043 - Número de clientes "cumpridores"classificados como "cumpridores- classificação correta.
44 - Número de clientes "cumpridores"classificados como "incumpridores- classificação incorreta.
642 - Número de clientes "incumpridores"classificados como "cumpridores- classificação incorreta.
27 - Número de clientes "incumpridores"classificados como "incumpridores- classificação correta.
Observando as tabelas 2.9 e 2.10 pode dizer-se que o modelo utilizado para a estimação das probabilidades de incumprimentos está a classificar corretamente cerca de 99.3% dos clientes cumpridores e 4.1% dos clientes incumpridores, se se tiver em conta que apenas se consideram clientes incumpridores os clientes com probabilidade de incumprimento estimada superior a 40%. Donde, em média, o modelo classifica corretamente 51.7% dos clientes da amostra de teste.

Tabela 2.10: Matriz de classificaçao do modelo de aprovação de crédito

Matriz de Classificaçao		
Observado	0	1
0	0.993	0.007
1	0.959	0.041

Fonte: Adaptação de Vale [32]

Observa-se também, que o modelo é mais eficiente a classificar clientes cumpridores.
Verifica-se, de uma forma geral, que o modelo desenvolvido utilizando a Regressão Logística (ModStp) obteve bons resultados na classificação de clientes como cumpridores e incumpridores.

2.4.4 Regressão probit

Nesta subseção vamos desenvolver de forma breve o modelo probit, dado que queremos focar apenas nos valores obtidos através da nossa base de dados utilizado nas seções anteriores, para posteriormente fazer a comparação deste com o modelo de Regressão logística. A idéia da função probit foi publicada por Chester Ittner Bliss (1899-1979) em um artigo de 1934 da Science sobre como tratar dados como a porcentagem de uma praga morta por um pesticida . Bliss propôs transformar a porcentagem de mortos em ”pro- bability un it ”(ou ”probit”) que estava linearmente relacionado à definição moderna. Ele incluiu uma tabela para ajudar outros pesquisadores a converter suas porcentagens de morte em seu probit, que poderiam então traçar contra o logaritmo da dose e, assim, esperava-se, obter uma linha mais ou menos direta. Esse modelo chamado probit ainda é importante na toxicologia, assim como em outros campos. A abordagem justifica-se, em particular, se a variação da resposta puder ser racionalizada como uma distribuição lognormal de tolerâncias entre indivíduos em teste, onde a tolerância de um indivíduo em particular é a dose apenas suficiente para a resposta de interesse.
Gujarati [17] afirma que para explicar o comportamento de uma variável dependente dicotômica teremos de usar uma FDA (Função Densidade Acumulada) escolhida apropriadamente. O modelo logit usa a função logística acumulada. Este não é a única FDA que pode-se usar, em algumas aplicações, a FDA normal revelou-se útil. O modelo de estimativa que emerge da FDA normal é conhecido como **modelo probit** que também pode ser dito **modelo normit**.
Vamos supor que a decisão de i-ésimo cliente estar ou não em incumprimento depende de um índice de utilidade não observável I_i, que é determinado por uma ou mais variáveis explicativas, $\mathbf{X}$. Expressamos o índice de utilidade I_i como

$$I_i = \beta_0 + \sum_{i=1}^{n} (\beta_i x_i) \tag{2.24}$$

Em que x_i são as variáveis explicativas do i-ésimo cliente e βs é o vetor de parâmetros das variáveis explicativas consideradas.

A FDA normal padronizada é dada por:

$$p_i = p_r(Y = 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-\frac{t^2}{2}} dt = F(I_i) = F\left(\beta_0 + \sum_{i=1}^{n} (\beta_i x_i)\right) \tag{2.25}$$

em que t é uma variável normal padronizada, isto é, $t \sim N(0,1)$ e p_i representa a probabilidade de um cliente escolhido ao acaso entrar em incumprimento.
Para obtermos informação sobre I_i, assim como sobre βs pegamos no inverso de 2.25, para obter

$$F^{-1}(p_i) = I_i = \beta_0 + \sum_{i=1}^{n} (\beta_i x_i) \tag{2.26}$$

em que F^{-1} é inverso da FDA normal padrão.
Na linguagem da análise probit, o índice de utilidade não observável I_i é simplesmente conhecido como *desvio equivalente normal* (d.e.n) ou simplesmente **normit**.
Como o d.e.n ou I_i será negativo sempre que $p_i < 0,5$, na prática se adiciona o número 5 ao d.e.n e o resultado é chamado de um **probit**. Isto é

$$probit(p_i) = d.e.n + 5 = I_i + 5 \tag{2.27}$$

2.4.5 Ajustamento do modelo utilizando regressão probit

Nesta seção utilizamos as covariáveis selecionadas no modelo logístico para a regressão probit. De recordar que o modelo contém as seguintes covariáveis:
Taxa nominal (TxNominal), Valor das prestações (ValorPrest), Valor do empréstimo (ValorEmprest), Idade, Agência (Agencia), Actividade profissional (ActProfissional), Genero, Entidade patronal (EntPatronal) e Habiltações (Habilitacoes). Com estas covariáveis construi-se-ão o modelo de Regressão probit.

```
  > myprobit <- glm(Default~TxNominal+ValorPrest+ValorEmprest+Idade+
  Agencia+ActProfissional+Genero+EntPatronal+
  Habilitacoes, binomial("probit"),data =Treino)
> summary(myprobit)
Call: glm(formula = Default ~ TxNominal + ValorPrest + ValorEmprest +
    Idade + Agencia + ActProfissional + Genero + EntPatronal +
    Habilitacoes, family = binomial("probit"), data = Treino)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2367  -0.5245  -0.3793  -0.2511   3.3496
Coefficients:
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.85333    0.11830 -24.119  < 2e-16 ***
TxNominal        0.24068    0.05036   4.779 1.76e-06 ***
ValorPrest       0.08695    0.01831   4.750 2.04e-06 ***
ValorEmprest     0.08212    0.01603   5.124 2.99e-07 ***
Idade           -0.14082    0.01176 -11.975  < 2e-16 ***
Agencia          0.17429    0.01058  16.474  < 2e-16 ***
ActProfissional  0.14665    0.01733   8.464  < 2e-16 ***
Genero          -0.13735    0.02242  -6.126 9.03e-10 ***
EntPatronal      0.23020    0.01038  22.172  < 2e-16 ***
Habilitacoes    -0.10000    0.01554  -6.435 1.23e-10 ***
---
Signif. codes:   ***=0.001; **=0.01; *=0.05; .=0.1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 18701  on 27024  degrees of freedom
Residual deviance: 16812  on 27015  degrees of freedom
AIC: 16832
```

Do ajustamento do modelo resultam os coeficientes $\beta_i^{'}s$ que serão utilizados.
A tabela 2.11 contém os valores dos coeficientes, $\beta_i^{'}s$, a considerar.

Tabela 2.11: Coeficientes do modelo ajustado

Característica		Coeficientes
Intercept	β_0	-2.85333
TxNominal	β_1	0.24068
ValorPrest	β_2	0.08695
ValorEmprest	β_3	0.08212
Idade	β_4	-0.14082
Agencia	β_5	0.17429
ActProfissional	β_6	0.14665
Genero	β_7	-0.13735
EntPatronal	β_8	0.23020
Habilitacoes	β_9	-0.10000

Fonte: Adaptação de Vale [32]

Então, pela equação 2.24 tem-se:

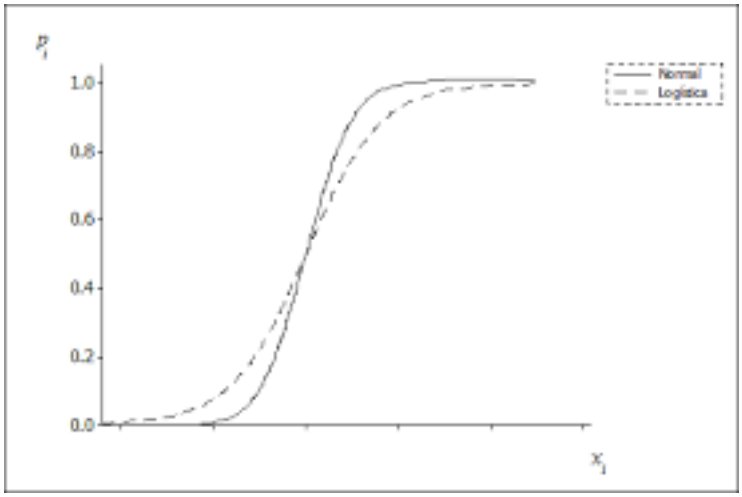
$$I_i = \beta_0 + \beta_1.TxNominal + \beta_2.ValorPrest + \beta_3.ValorEmprest + \beta_4.Idade + \beta_5.Agencia + \\ + \beta_6.ActProfissional + \beta_7.Genero + \beta_8.EntPatronal + \beta_9.Habilitacoes \quad (2.28)$$

Sendo assim, das equações 2.27 e 2.28 podemos obter os valores de $probit\,(p_i)$.

2.4.6 Regressão logística versus regressão probit

A regressão probit pode ser usada para resolver problemas de classificação binária, assim como a regressão logística.
Enquanto a regressão logística utilizou uma função logística cumulativa, a regressão probit usa uma função de densidade cumulativa normal padrão para o modelo de estimativa.
Para Filho et al. [13], uma maneira adequada de utilizar modelo baseados em escolhas qualitativas é pelas probabilidades, desse modo existem funções de ligações específicas como a logit e probit que com a utilização de funções de distribuições podem realizar o cálculo.
Os modelos de regressão logística e de regressão probit utilizam como funções de ligação logit e probit respetivamente.
De acordo com Barros [5] a escolha da função de ligação logit assim como a probit é determinada por simples conveniência matemática e computacional.
De acordo com a abordagem realizada por Cordeiro e Demétrio [10], a função de ligação logit assim como a probit têm em comum o fato de a variável dependente ser uma variável qualitativa com dois possíveis valores, assim, as funções de ligação logit e probit são dadas respetivamente pelos inversos das distribuições acumuladas logística e normal.

Figura 2.1: Distribuições acumuladas logística e normal.



Segundo Gujarati [17], do ponto de vista teórico, a diferença entre o modelo logístico e probit é conforme mostrado na figura 2.1, ou seja, essas duas formulações são bem comparáveis, sendo que a principal diferença está no fato de a logística ter caudas ligeiramente mais achatadas, ou seja, a curva normal (ou probit) se aproxima do eixo mais rapidamente do que a curva da logística. Deste modo a escolha entre os dois modelos é uma questão de conviniência (matemática) e de pronta disponibilidade de programas computacionais.Neste aspeto, o modelo logístico é em geral preferido ao probit.
De acordo com Hair et al. [18] , a regressão linear utiliza dos métodos dos mínimos quadrados ordinários para realizar a estimação de seus coeficientes, esse método consiste em minimizar a soma de quadrados das diferenças entre os valores observados e os previstos. Na regressão não linear o método da máxima verossimilhança é utilizado de forma iterativa para que sejam encontradas as estimativas mais prováveis dos parâmetros. Ao invés de minimizar os desvios quadrados, a regressão não linear maximiza a probabilidade de que um evento ocorra.
Em Gujarati [17], analisou que embora os modelos logit e probit dêem resultados qualitativamente similares, as estimativas dos parâmetros dos dois modelos não são diretamente comparáveis. Isso podemos ver na tabela 2.12.

Tabela 2.12: Regression outputs

	Dependent variable:	
	Default	
	probit	*logistic*
	(1)	(2)
TxNominal	0.241***	0.515***
	(0.050)	(0.105)
ValorPrest	0.087***	0.180***
	(0.018)	(0.035)
ValorEmprest	0.082***	0.150***
	(0.016)	(0.030)
Idade	−0.141***	−0.268***
	(0.012)	(0.022)
Agencia	0.174***	0.324***
	(0.011)	(0.020)
ActProfissional	0.147***	0.277***
	(0.017)	(0.033)
Genero	−0.137***	−0.247***
	(0.022)	(0.043)
EntPatronal	0.230***	0.440***
	(0.010)	(0.020)
Habilitacoes	−0.100***	−0.202***
	(0.016)	(0.030)
Some control	Y	Y
Observations	27,025	27,025
Note:	*p<0.1; **p<0.05; ***p<0.01	

Capítulo 3

Estimação da Probabilidade de Imcumprimento de Uma Carteira de Crédito ao Longo do Tempo

Introdução

Processos de Markov tem uma forma simples de dependência e é bastante utilizada em modelação de problemas encontrados na prática, por exemplo, dada uma particular população, cada elemento desta população pode encontrar-se em cada instante num determinado estado. A evolução da população pelos diferentes estados será modelada através de cadeias de markov. O processo de markov mais concretamente cadeias de markov é a base para este estudo.

A cadeia de Markov tem inúmeras aplicações na área financeira e não só, por exemplo nos estudos de Fernandes [12], mostra como determinar a proporção de elementos num determinado estado via cadeias de Markov, também o mesmo autor, analisou a evolução do número de indivíduos em cada estado da cadeia de Markov.

Para análise da probabilidade de incumprimento da carteira de crédito ao consumo, numa perspectiva de longo prazo, utilizamos um modelo para populações abertas sujeitas a reclassificações periódicas designado por **Vórtices Estocásticos** , a partir do qual se poderão obter estimativas pontuais e por intervalo de confiança para o número total e proporção de clientes nas diferentes classes de risco, em qualquer instante do tempo.

O objetivo principal deste capítulo é estimar a probabilidade de imcumprimento ao longo do tempo de uma carteira de crédito ao consumo de uma instituição financeira Cabo-Verdiana.

Este capítulo está dividido em duas seções. Na primeira seção descrevemos o modelo Vórtices Estocásticos, esta seção está dividida em cinco subseções: matriz de transição, fluxos de entrada na população, Vórtices baseado em estados transientes, função de verosimilhança na ausência de restrições e a forma funcional sigmoidal e na segunda seção aplicamos o

modelo Vórtices Estocásticos para estimação da evolução temporal da probabilidade de incumprimento de uma carteira de crédito ao consumo.

3.1 Modelo Vórtices Estocásticos

O modelo Vórtices Estocásticos assenta numa formulação que tem por base um modelo de Cadeias de Markov, mas considerando populações abertas.
A base do modelo que aqui apresentaremos é uma cadeia de markov finita, homogénea, com s estados transientes e 1 estado absorvente, correspondente às saídas da população.
Segundo Guerreiro e Mexia [16], **dimensão absoluta de uma sub-população** é o número total de elementos que, num dado instante, integram essa sub-população e a **dimensão relativa de uma sub-população** é a proporção de elementos que, num dado instante, integram essa sub-população. A dimensão relativa é calculada como o quociente entre o número de elementos nessa sub-população e o número de elementos que compõem a totalidade da população.
Quando a dimensão relativa de um conjunto de sub-populações estabiliza, ou seja, converge para um valor finito, diz-se-á que essas sub-populações estão integradas num **vórtice estocástico**. Esse vórtice estocástico estará sediado num conjunto maximal de estados para os quais as dimensões relativas são estáveis.
Segundo Fernandes [12], a entrada de novos elementos na populaçao, que se assume seguir uma distribuição de Poisson, está sujeita a uma classificação inicial, a qual determina a sub-população onde cada novo elemento irá ser inicialmente colocado. Esta sub-população será o ponto de partida e, após um período de tempo, o elemento será sujeito a uma reclassificação podendo manter a classificação anterior ficando, portanto, no mesmo estado da cadeia ou transitar para outra sub-população, de acordo a avaliação que dele é feita nesse instante.
Os fluxos de entrada na população, para Guerreiro e Mexia [16] têm um papel preponderante na forma como estas irão evoluir ao longo do tempo e na possibilidade de existência de uma estabilidade assimptótica ao nível das dimensões absolutas e/ou relativas das sub-populações. Os fluxos de entrada para a população têm sido modelados considerando que, em cada instante i, o número esperado de novos elementos que integram a população pode ser estimado a partir de $\lambda_i = \left(a + be^{-\theta i}\right)^{-1}, i \in \mathbb{N}, a, b, \theta \in \Theta_2$. Esta forma funcional para o número esperado de novas entradas foi desenvolvido no trabalho de Fernandes [12].

3.1.1 Matriz de transição

Consideremos que a carteira de crédito é constituída por s sub-populações em que aqui denotamos por classes de risco, correspondendo a s estados transientes de uma cadeia de markov. Considere-se ainda um estado absorvente suplementar, que será considerado como classe de saída da carteira, podemos facilmente verificar que a matriz de transição num passo pode ser definida por:

$$\mathbf{P} = \begin{pmatrix} \mathbf{K} & \mathbf{q} \\ \mathbf{0} & 1 \end{pmatrix}$$

onde:

- **K**- matriz de dimensão $s \times s$, cujos elementos representam as probabilidades de transição entre as classes transientes (as classes de risco).
- **q**-vetor coluna, de dimensão $s \times 1$, cujos elementos representam as saídas dos clientes das classes transientes.

Como se pode observar:

- a última linha da matriz representa o estado de saída dos clientes da carteira.
- a soma dos elementos de cada linha da matriz **K** com a correspondente componente do vetor **q**, será sempre igual a 1.

O seguinte lema, em que a prova é facilmente obtida por indução matemática e pode ser consultada em Guerreiro [15], nos permite determinar a matriz de transição em n passos.

Lema 3.1. *A matriz probabilidade de transição em n passos, será da forma:*

$$P^n = \begin{pmatrix} \boldsymbol{K}^n & \boldsymbol{qn} \\ \boldsymbol{0} & 1 \end{pmatrix}$$

em que: $\boldsymbol{qn} = \sum_{i=1}^{n-1} \left(\boldsymbol{K}^i q \right), n \in \mathbb{N}$

3.1.2 Fluxos de entrada na população

Sejam $E_i, i \in \mathbb{N}$, o número total de novos clientes aos quais é concedido um crédito, no período i.
Asumiremos que o número total de entradas para a carteira, em cada período, ocorre no início desse período e, sem perda de generalidade, tomaremos os períodos como meses.
Consideraremos, que o número de novos clientes que integram a carteira no mês i segue uma distribuição de Poisson de valor esperado λ_i, isto é,
$E_i \sim P(\lambda_i)$.

Neste estudos vamos considerar que a intensidade de entrada de novos clientes, isto é, que o fluxo de entradas dos clientes, λ_i, é modelada pela expressão 3.1:

$$\lambda_i = \left(a + be^{-\theta i} \right)^{-1} a, b, \theta \in \Theta_2, i \in \mathbb{N} \tag{3.1}$$

em que o espaço de parâmetros Θ_2, correspondente aos possíveis valores de a, b e θ, é definido por:

$$\Theta_2 = \left\{(a,b,\theta) : a \in \mathbb{R}^+, b, \theta \in \mathbb{R}, a + be^{-\theta i} > 0\right\} \setminus \left\{(a,b,\theta) : a + be^{-\theta i} = 0, i \in \mathbb{N}\right\} \tag{3.2}$$

A adequabilidade destes fluxos de entrada aos dados da carteira pode ser testada recorrendo a vários testes de hipóteses que poderão ser consultados de forma detalhada em [Fernandes [12]].
A modelação da evolução da carteira recorrendo a novos fluxos de entrada deixa em aberto a análise da estabilidade da carteira a longo prazo, ou seja, a análise da existência de Vórtices Estocásticos nos estados transientes da cadeia de Markov.
Esta formulação pertencerá a uma classe de funções para as quais se garante a existência de Vórtices Estocásticos na cadeia, o que nos permitirá estimar a proporção de clientes em cada classe de risco, numa perspetiva de longo prazo.

Considerando que, no início de cada contrato, os novos clientes estão sujeitos a uma classificação inicial, afim de serem distribuídos nas diferentes classes transientes do sistema, seja $\mathbf{c}_i$ o vetor de classificação inicial para o mês i, $i \in \mathbb{N}$, com

$$\mathbf{c}_i^T = \left[\mathbf{t}_i^T | 0\right] \tag{3.3}$$

em que:

- $\mathbf{t}_i$ representa um vetor cujas componentes são as probabilidade de entrada de um novo cliente em cada uma das classes de risco.
- a última componente indica que a probabilidade de um novo cliente ser imediatamente colocado no estado de saída é nula.

Considerando que poderão haver alterações nas classificações iniciais de cada mês, $\mathbf{c}_i$ poderá ser representado por:

$$\mathbf{c}_i = \mathbf{c} + \omega\gamma^i = \left[c_{ij}\right], j = 1, ..., s, 0 < \gamma < 1 \tag{3.4}$$

sendo:

- $\mathbf{c}$ - vetor fixo de probabilidades de classificação inicial cujas componentes verificam $\sum_{j=1}^{s}\left(c_j\right) = 1$;
- ω - vetor fixo cujas componentes verificam $\sum_{j=1}^{s}\left(\omega_j\right) = 0$.

É importante notar que a formulação 3.4 se traduz no seguinte $\lim_{i \to +\infty} (c_i) = \mathbf{c}$, o que equivale a considerar um vetor de classificação inicial convergente.

Suponhamos que um conjunto de N elementos, com N uma variável aleatória com distribuição de Poisson de parâmetros λ, são distribuídos por s classes, de acordo com a distribuição multinomial com vetor de probabilidades $\mathbf{c} = (c_1, ..., c_s)$.
Após a distribuição, os números $N_j, j = 1, ..., s$, de elementos na classe j, serão variáveis aleatórias independentes com distribuição de Poisson com parâmetros $\lambda c_1, \lambda c_2, ..., \lambda c_s$.

O Teorema seguinte é de extrema importância no modelo que iremos aplicar pois permite-nos estimar os parâmetros das leis de Poisson das classes da população, em qualquer data $n, n \in \mathbb{N}$.

Teorema 3.1. *Considere-se um sistema com s classes. Suponha-se que:*

- *N_i seja o número de elementos que entra na população na data $i \in \{1, ..., T\}$, com N_i seguindo uma distribuição de Poisson de parâmetro λ_i, e são imediatamente distribuídos pelas classes, segundo o vetor $\boldsymbol{c}_i^T = (c_{i1}, ..., c_{is})$.*

- *Em cada data $i \in \{1, ..., T\}$, e simultaneamente com a entrada de novos elementos no sistema, os elementos de cada classe evoluem de acordo com a lei de uma cadeia de Markov com matriz de transição $\boldsymbol{P} = [p_{ij}](i, j) \in \{1, ..., s\}^2$. Esta evolução traduz as probabilidades de reclassificação dos elementos pertencentes à população.*

- *Após n períodos, $N_{i,n}$, o número de elementos na data n, dado que entraram na data i, terão sido sujeitos a $n - i$ reclassificações e terá distribuição de Poisson com parâmetro dado por,*

$$\lambda_{i,n}^T = \lambda_i c_i^T \ \boldsymbol{P}^{(n-i)}$$

Ao fim da data n, com $n > i$, os elementos que entraram no sistema na data i, foram reclassificados $n - i$ vezes, portanto, utilizando o teorema anterior, tem-se que o vetor dos respetivos efetivos, terão distribuição Poisson de parâmetro $\lambda_{i,n}^T = \lambda_i \mathbf{c}_i^T \mathbf{P}^{(n-i)}$.

Considerando $N_n^{++} = \sum_{i=1}^{n} N_{i,n}$ ou seja, N_n^{++}, número total de elementos no sistema, no início da data n, dado que já foram contabilizadas as entradas da data n, tem-se então que N_n^{++} segue distribuição de Poisson com parâmetro dado por,

$$\lambda_n^{++T} = \sum_{i=1}^{n} \lambda_{i,n}^T = \sum_{i=1}^{n} \lambda_i \mathbf{c}_i^T \mathbf{P}^{(n-i)} \tag{3.5}$$

Este teorema permite-nos estimar a dimensão esperada de cada uma das sub-populações correspondentes às classes de risco da carteira, em qualquer instante de tempo. Poderemos, ainda, desenvolver técnicas estatísticas adequadas à estimação destes parâmetros, nomeadamente ao nível da estimação por intervalo de confiança e o desenvolvimento de testes de hipótese.
A demonstração do teorema 3.1, pode ser consultada em [Fernandes [12]].
A seguir apresentemos uma proposição cuja demonstração pode consultar-se em [Guerreiro [15]].

Proposição 3.1.1. *Após n períodos de tempo, o vetor médio do número de clientes nas sub-populações será dado por:*

$$\lambda_n^{++T} = \left(\sum_{i=1}^{n} \lambda_i \boldsymbol{t}_i^T \boldsymbol{K}^{n-i} | \sum_{i=1}^{n} \lambda_i \boldsymbol{t}_i^T Kqn - 1 \right) \tag{3.6}$$

Sendo a primeira componente do vetor, que designaremos por λ_n^{+T} , a dimensão estimada das sub-populações correspondentes aos estados transientes (as classes de risco) e a segunda componente do vetor correspondente ao número estimado de clientes que saíram da carteira de crédito.

3.1.3 Vórtices baseado em estados transcientes

Pretende-se provar aqui a existência do vórtice nas classes transientes, dado que o fluxo de entrada de novos clientes na carteira é dado por $\lambda_i = \left(a + b.e^{-\theta i}\right)^{-1}$.
Sob a hipótese das dimensões relativas das sub-populações, caracterizar-se pela existência de um vórtice estocástico estabelecido nos estados transientes.
Supõe-se que a matriz de transição entre estados trasientes, $\mathbf{K}$, é dada por,(Fernandes [12]):

$$\mathbf{K} = \sum_{j=1}^{s} \eta_j \alpha_j \beta_j^T,$$

Com $\eta_j, j = 1, ..., s$, os valores próprios da matriz $\mathbf{K}$, $\alpha_j, j = 1, ..., s$, os vetores próprios à esquerda e $\beta_j, j = 1, ..., s$ os vetores próprios à direita da matriz $\mathbf{K}$.
Notamos que $j \in 1, ..., s$ corresponde a um estado transiente se e só se $|\eta_j| < 1$.
De uma forma mais geral temos:

$$\mathbf{K}^n = \sum_{j=1}^{s} \eta_j^n \alpha_j \beta_j^T, \tag{3.7}$$

e, como consequência da expressão 3.6, para o vetor médio das dimensões das sub_populações correspondentes aos estados transientes, $\left(\lambda_n^+\right)^T$, tem-se:

$$\lambda_n^{+T} := \sum_{i=1}^{n} \lambda_i \mathbf{t}_i^T \mathbf{K}^{n-i} = \sum_{j=1}^{s} \sum_{i=1}^{n} \lambda_i \eta_j^{n-i} \mathbf{t}_i^T \alpha_j \beta_j^T \tag{3.8}$$

Supõe-se que o vetor de probabilidades de classificação inicial é constante, ou seja, para $i \geq 1$, $\mathbf{t}_i = t_0 \neq 0$. Dai reescrevemos 3.8 da seguinte forma:

$$\lambda_n^{+T} = \sum_{j=1}^{s} \left(\sum_{k=1}^{n} \lambda_k \eta_j^{n-k} \right) t_0^T \alpha_j \beta_j^T \tag{3.9}$$

Mas os resultados obtidos podem ser estendidos ao caso geral em que o vetor de classificação inicial pode não ser constante, mas convergente para um vetor não nulo, ou seja

$$\lim_{n \to +\infty} \mathbf{t}_i = t_\infty \neq 0$$

Aqui, analisamos o caso em que os parâmetros das v.a.'s de Poisson, referentes ao número de novas entradas na população, corresponde a uma sucessão convergente.
De acordo com a expresssão 3.9, o comportamento assimptótico de λ_n^{+T} é determinado pelo comportamento assimptótico de

$$I_n = \sum_{i=1}^{n} \lambda_i \eta^{n-i} \tag{3.10}$$

com $|\eta| < 1$. Como $|\eta| < 1$, e observando o segundo membro de 3.10, leva-nos a deduzir que, de uma forma geral, assimptoticamente, teremos:

$$\lim_{n \to +\infty} I_n = \lambda_n \tag{3.11}$$

Utilizaremos uma técnica standard de análise assimptótica, mais precisamente, a conhecida transformação de Abel, esta transformação é bastante útil sempre que o comportamento das somas com termos a_k são conhecidos e as oscilações dos termos b_i são controláveis,(Zorich [33]). A transformação de Abel, afirma que:

$$\sum_{i=1}^{n} a_i b_i = \left(\sum_{k=1}^{n} a_k \right) b_n + \sum_{i=1}^{n-1} \left(\sum_{k=1}^{i} a_k \right) (b_i - b_{i+1})$$

Com $b_i = \lambda_i$ e $a_i = \eta^{n-i}$, para $i = 1, ..., n+1$, tem-se:

$$I_n = \sum_{i=1}^{n} \lambda_i \eta^{n-i} = \left(\sum_{k=1}^{n} \eta^{n-k} \right) \lambda_n + \sum_{i=1}^{n-1} \left(\sum_{k=1}^{i} \eta^{n-k} \right) (\lambda_i - \lambda_{i+1})$$

Somando os n primeiros termos das progressões geométricas, obtém-se a seguinte expressão:

$$I_n = \frac{1 - \eta^n}{1 - \eta} \lambda_n + \sum_{i=1}^{n-1} \left(\eta^{n-i} \frac{1 - \eta^i}{1 - \eta} \right) (\lambda_i - \lambda_{i+1}) \tag{3.12}$$

Que se revela de grande importância para o teorema que se segue.

Teorema 3.2. *Se* $\lim_{n\to+\infty} \lambda_n = \lambda$, *com* $\lambda \in \mathbb{R}_+$ *então,* $\lim_{n\to+\infty} I_n = \dfrac{\lambda}{1-\eta}$.

Demonstração pode ser encontrado em Fernandes [12] e com base no resultado do teorema 3.2 obtemos o seguinte teorema.

Teorema 3.3. *Consideramos que um sistema, é modelado por uma cadeia de Markov em que a matriz de transição num passo entre estados transientes é diagonalizável.*
Suponhamos que as entradas no sistema são realizações de v.a independentes com intensidades $\lambda_i, i \geq 1$ *e que o vetor de classificção inicial nos estados transientes é convergente para um valor fixo, isto é,* $\lim_{i\to+\infty} \boldsymbol{t}_i = t_\infty \neq 0$. *Então, com* λ_n^{+T} *o vetor dos parâmetros de Poisson, correspondentes à dimensão das sub-populações, à data* $n \geq 1$, *teremos:*

$$\textit{Se } \lim_{n\to+\infty} \lambda_n = \lambda, \textit{ com } \lambda \in \mathbb{R}_+ \textit{ então, } \lim_{n\to+\infty} \lambda_n^{+T} = \sum_{j=1}^{s} \frac{\lambda}{1-\eta_j} t_\infty^T \alpha_j \beta_j^T.$$

Este teorema 3.3 cuja a demonstração, pode ser consultada em [Fernandes [12]] mostra que, a longo prazo, existe estabilidade da dimensão relativa das classes do sistema, característica da existência de vórtices estocásticos.

A proporção de clientes $\pi_{n,j}$, que à data n, se encontra na classe de riso $j \in \{1, 2, \cdots s\}$ é determinado por:

$$\pi_{n,j} = \frac{\lambda_{n,j}}{\sum_{k=1}^{s} \lambda_{n,k}}$$

Onde o seu estimador é

$$\hat{\pi}_{n,j} = \frac{\hat{\lambda}_{n,j}}{\sum_{k=1}^{s} \hat{\lambda}_{n,k}}$$

3.1.4 Função de verosimilhança na ausência de restrições

Nesta subseção vamos estimar, pelo método da máxima verosimilhança, os parâmetros dos fluxos de entrada de novos clientes na população. Os fluxos de entrada de novos clientes é dada pela forma sigmoidal, proposta e aplicada nos estudos de [Fernandes [12]].
Será aqui apresentada como uma proposta viável de aplicação aos dados da carteira de crédito ao consumo que iremos estudar adiante.

Consideremos uma amostra recolhida ao longo das datas $i \in \{1, ..., T\}$. Em cada data i é observado o resultado das v.a.'s independentes, correspondentes ao número de indivíduos que, nessa data, foram colocados em cada uma das classes do sistema.
Seja N_{ij} a v.a. correspondente ao número de indivíduos que, à data i, entram na carteira como novos clientes e são inicialmente colocados na classe de risco j.
$N_{ij} \frown \mathcal{P}\left(\lambda_i c_j\right), i = 1, ..., T, j = 1, ..., s.$
Sejam $N_i = [N_{ij}], i = 1, ..., T, j = 1, ..., s$, e n_{ij} as realizações das v.a.'s N_{ij}, ou seja, os elementos que, no início do período i são colocados na sub-população j. Consideremos ainda $\lambda_{ij} = \lambda_i c_j, i = 1, ..., T, j = 1, ..., s$ as componentes de $\lambda_i \mathbf{c}$.
Diremos assim, que $N_{ij} \frown \mathcal{P}\left(\lambda_{ij}\right), i = 1, ..., T, j = 1, ..., s$ e $N_i \frown \mathcal{P}(\lambda_i \mathbf{c})$.
Na ausência de restrições sobre os parâmetros dos fluxos de entrada, a função verosimilhança para $N = [N_i] = [N_{ij}]$, será dada por

$$L_\Omega(\lambda) = \Pi_{i=1}^{T} \Pi_{j=1}^{s} e^{-\lambda_{ij}} \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!} \tag{3.13}$$

com $\lambda = [\lambda_{ij}], i = 1, ..., T, j = 1, ..., s$.
Tomando $n = [n_{ij}], i = 1, ..., T, j = 1, ..., s$, a função de log-verosimilhança será dada por

$$\ell_\Omega(\lambda) = a(n) - \sum_{i=1}^{T} \sum_{j=1}^{s} \lambda_{ij} + \sum_{i=1}^{T} \sum_{j=1}^{s} n_{ij} log\left(\lambda_{ij}\right) \tag{3.14}$$

com $a(n) = -\sum_{i=1}^{T} \sum_{j=1}^{s} log\left(n_{ij}!\right)$

O máximo da função de log-verosimilhança, na ausência de restrições, obtém-se a partir da equação
$\frac{\partial \ell_\Omega(\lambda)}{\partial \lambda_{ij}} = -1 + \frac{n_{ij}}{\lambda_{ij}}, i = 1, ..., T, j = 1, ..., s$
pelo que, os estimadores de máxima verosimilhança para λ_{ij}, na ausência de restrições, são dados por

$$\hat{\lambda}_{ij,\Omega} = N_{ij}, i = 1, ..., T, j = 1, ..., s \tag{3.15}$$

Assim, pelas equações 3.14 e 3.15, o máximo da função de log-verosimilhança, na ausência de restrições, será dado por

$$\hat{\ell}_\Omega(\lambda) = a(n) - \sum_{i=1}^{T} \sum_{j=1}^{s} n_{ij} + \sum_{i=1}^{T} \sum_{j=1}^{s} n_{ij} log\left(n_{ij}\right) \tag{3.16}$$

3.1.5 Forma sigmoidal $\lambda_i = \left(a + b.e^{-\theta i}\right)^{-1}$

Aqui obteremos os estimadores de máxima verosimilhança para os parâmnetros da forma sigmoidal e desenvolveremos um teste de hipóteses para esta formulação.

$$\mathcal{H}_0 : \lambda_i = \left(a + b.e^{-\theta i}\right)^{-1} c_j \quad vs \quad \mathcal{H}_1 : \lambda_i \neq \left(a + b.e^{-\theta i}\right)^{-1} c_j, i \in \{1, ..., T\}, j \in \{1, ..., s\}$$

A função verosimilhança, sob a hipótese H_0, é dada por:

$$L_{\omega 0}(a, b, \theta, \mathbf{c}) = \prod_{i=1}^{T} \prod_{j=1}^{s} \frac{\left(\frac{c_j}{a+b.e^{-\theta i}}\right)^{n_{ij}}}{n_{ij}!} e^{-\frac{c_j}{a+b.e^{-\theta i}}} \tag{3.17}$$

A função de log-verosimilhança correspondente, sob a hipótese $\mathcal{H}_0$, é dada por

$$\ell_{\omega 0}(a, b, \theta, \mathbf{c}) = a(n) - \sum_{i=1}^{T} \sum_{j=1}^{s} \frac{1}{a + b.e^{-\theta i}} c_j + \sum_{i=1}^{T} \sum_{j=1}^{s} n_{ij} \left[log\left(\frac{1}{a + b.e^{-\theta i}}\right) + log\left(c_j\right) \right] \tag{3.18}$$

com $a(n) = -\sum_{i=1}^{T} \sum_{j=1}^{s} log\left(n_{ij}!\right)$

Introduzindo a restrição $\sum_{j=1}^{s} c_j = 1$ e recorrendo aos multiplicadores de Lagrange, obteremos a seguinte função objetiva

$$\ell^{\nu}_{\omega 0}(a, b, \theta, \mathbf{c}) = \ell_{\omega 0}(a, b, \theta, c) + \nu \left(\sum_{j=1}^{s} c_j - 1 \right)$$

Determinando o máximo da função log-verosimilhança obtém-se os estimadores de máxima verosimilhança $(\hat{a}, \hat{b}, \hat{\theta})$, para (a, b, θ), por resolução das equações seguintes:

$$\begin{cases} \sum_{i=1}^{T} \frac{1}{\left(\hat{a} + \hat{b}.e^{-\hat{\theta} i}\right)^2} = \sum_{i=1}^{T} \frac{\left(\sum_{j=1}^{s} N_{ij}\right)}{\hat{a} + \hat{b}.e^{-\hat{\theta} i}} \\ \sum_{i=1}^{T} \frac{e^{-\hat{\theta} i}}{\left(\hat{a} + \hat{b}.e^{-\hat{\theta} i}\right)^2} = \sum_{i=1}^{T} \frac{\left(\sum_{j=1}^{s} N_{ij}\right) e^{-\hat{\theta} i}}{\hat{a} + \hat{b}.e^{-\hat{\theta} i}} \\ \sum_{i=1}^{T} \frac{\hat{b} i e^{-\hat{\theta} i}}{\left(\hat{a} + \hat{b}.e^{-\hat{\theta} i}\right)^2} = \sum_{i=1}^{T} \frac{\left(\sum_{j=1}^{s} N_{ij}\right) \hat{b} i e^{-\hat{\theta} i}}{\hat{a} + \hat{b}.e^{-\hat{\theta} i}} \end{cases} \tag{3.19}$$

O estimador de máxima verosimilhança para $\mathbf{c}$, $\hat{\mathbf{c}} = (\hat{c}_1, ..., \hat{c}_s)$ é obtido para $\nu = 0$, a partir de:

$$\nu + \frac{1}{T}\sum_{i=1}^{T}\left(-\frac{1}{\hat{a}+\hat{b}.e^{-\hat{\theta}i}} + \frac{N_{ij}}{\hat{c}_j}\right) = 0, \tag{3.20}$$

Resolvendo a equação 3.20 em ordem a c_j, Obtendo-se

$$\hat{c}_j = \frac{\frac{1}{T}\sum_{i=1}^{T}\left(\frac{1}{\hat{a}+\hat{b}.e^{-\hat{\theta}i}}\right)}{\frac{1}{T}\sum_{i=1}^{T} N_{ij}} = \frac{\sum_{i=1}^{T}\left(\frac{1}{\hat{a}+\hat{b}.e^{-\hat{\theta}i}}\right)}{\sum_{i=1}^{T} N_{ij}}, \forall j \in \{1,...,s\}$$

Com esta análise estatística, estamos em condições de estimar os parâmetros da forma funcional sigmoidal através dos dados de entrada na carteira, posteriormente, com a forma funcional obtida podemos prever a evolução da população em cada uma das classes de risco.

3.2 Aplicação

O objetivo principal deste capítulo é estimar a probabilidade de imcumprimento ao longo do tempo de uma carteira de crédito ao consumo de uma instituição financeira Cabo-Verdiana.
Nesta seção vamos apresentar uma aplicação de acordo com as abordagens desenvolvidas e propostas nas seções anteriores. A modelação baseia-se na forma funcional sigmoidal, iremos estimar os parâmetros da forma funcional sigmoidal levando em conta os dados da entrada de clientes na carteira.
Modelo este que foi desenvolvido e aplicado nos estudos de Fernandes [12].
Para a obtenção dos resultados será utilizado o software Mathematica 9.0 da Wolfram. Com base nas análises e resultados obtidos, serão feitos os respetivos comentários.

3.2.1 Descrição da carteira

Por falta de dados por causa do sigilio bancário, foi também necessário simulação dos dados correspondesnte a entrada de clientes na carteira utilizado.
A carteira utilizada é composto por 17079 clientes, imaginemos que esses clientes entraram na carteira durante um perído de 61 meses. Ou seja consideramos que carteira utilizada é constituída por dados históricos de todos os clientes cujo contratos foram financiados entre Janeiro de 2012 e Janeiro de 2017, de uma carteira de empréstimos bancários para crédito ao consumo de um banco comercial Cabo-verdiano.
Neste conjunto de clientes e neste período temporal verificam-se entradas e saídas, e consideramos que as probabilidades de transição entre as classes de risco, num mês, correspondem a uma cadeia de Markov homogénea com cinco estados transientes e um estado

recorrente (absorvente), correspondente ao estado de saída da carteira.
Os clientes da carteira serão classificados em diferentes classes de risco, de acordo com o número de dias de incumprimento após a data da primeira prestação. Consideremos cinco clases de risco:

- $C_1 = [0, 30]$-contém os clientes sem prestações em dívida por um período entre 0 e 30 dias;
- $C_2 = [31, 60]$-contém os clientes com prestações em dívida por um período entre 31 e 60 dias;
- $C_3 = [61, 90]$-contém os clientes com prestações em dívida por um período entre 61 e 90 dias;
- $C_4 = [91, 120]$-contém os clientes com prestações em dívida por um período entre 91 e 120 dias;
- $C_5 = [120, +[$-contém os clientes com prestações em dívida por um período superior a 120 dias.

A tabela 3.1 seguinte ilustra as classes de risco por sub-populações.
Convém referir que todos os clientes entram diretamente na primeira classe de risco. Os

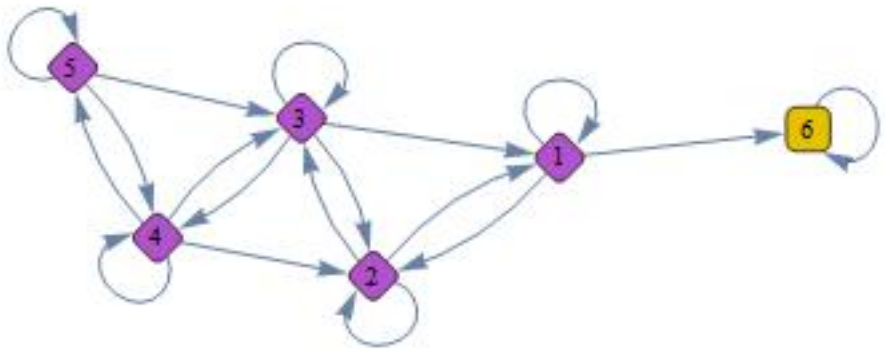
Tabela 3.1: Sub-populações-classes de risco

Sub-população	Número de dias em incumprimento
1	$0 - 30$
2	$31 - 60$
3	$61 - 90$
4	$91 - 120$
5	> 120
6-saída	-

Fonte: Fernandes [12]

clientes são, em seguida, reclassificados, a cada mês, transitando para a classe de risco correspondente de acordo com o atraso, ou não, dos seus planos de reembolso.
Sempre que um cliente termina o seu contrato na classe $[0, 30]$ dias, ele é classificado na sexta classe de risco, que corresponde a saída da carteira, indicando que o seu contrato terminou uma vez que o cliente efetuou todos os pagamentos necessários ao reembolso do empréstimo, de acordo com as condições do contrato.
Sempre que um cliente tem mais de 30 dias de atraso até a data do término do seu contrato, ele é considerado cliente da carteira mantendo-se na classe correspondente ao número de dias de incumprimento, até que a última das prestações em dívida seja paga. De acordo com esta definição, apenas serão possiveis saídas da carteira, a partir da classe 1. Os clientes em incumprimento manterão na carteira até à liquidação da sua dívida.

Figura 3.1: Grafo de transição entre as classes da cadeia.



A figura 3.1 ilustra as possibilidades de transição entre as classes de risco. Todos os clientes entram diretamente na primeira classe de risco. Seguidamente, os clientes, são reclassificados, a cada mês, transitando para a classe de risco correspondente de acordo com o atraso, ou não, dos seus planos de reembolso. Sempre que um cliente termina o seu contrato na classe $[0, 30]$ dias, ele é classificado na sexta classe de risco, que corresponde a saída da carteira, indicando que o seu contrato terminou uma vez que o cliente efectuou todos os pagamentos necessários ao reembolso do empréstimo. Sempre que um cliente tem mais de 30 dias de atraso até a data do término do seu contrato, ele é considerado cliente da carteira mantendo-se na classe correspondente ao número de dias de incumprimento, até que a última das prestações em dívida seja paga.

3.2.2 Matriz de transição

Definidas as classes de risco, construiremos a matriz de probabilidades de transição num passo para a carteira em causa. Nesta aplicação, um passo corresponderá a um mês, uma vez que, regra geral, os planos de pagamentos de crédito ao consumo tém uma periodicidade mensal. A matriz de transição num passo, que se segue ilustra as probabilidades das migrações futuras, estimadas a partir dos dados históricos da carteira de crédito, no final de cada mês.
Tal matriz num passo das classes de risco definidas como na seção 3.1.1, será dada por:

Na matriz de transição da tabela 3.2 temos:

- a soma dos elementos de cada linha será sempre igual a 1. Com efeito, um cliente, ao fim de um período (mês), ou permanece na carteira ocupando uma dada classe,

Tabela 3.2: Matriz de transição a um passo

Classes de Risco	$[0,30]$	$[31,60]$	$[61,90]$	$[91,120]$	$]120,+[$	Saídas
$[0,30]$	0.847766	0.035130	0	0	0	$0,117102$
$[31,60]$	0.437653	0.488998	0.073349	0	0	0
$[61,90]$	0.130435	0.42029	0.160869	0.288406	0	0
$[91,120]$	0	0.102916	0.411664	0.348199	0.137221	0
$]120,+[$	0	0	0.082781	0.165563	0.751656	0
Saídas	0	0	0	0	0	1

Fonte:Adaptação de Fernandes [12]

ou, se o pagamento corresponde à última prestação, o cliente transita para a classe de saída.

- a última coluna da matriz representa o estado de saída dos clientes da carteira de crédito.

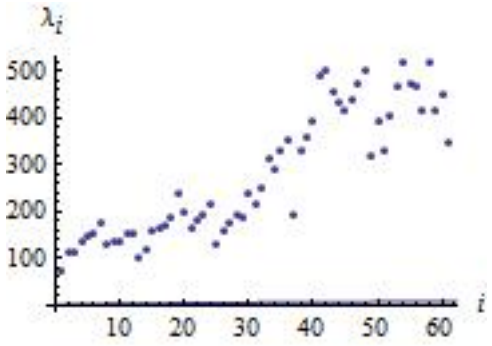
- analisando a primeira linha da matriz de transição podemos notar, por exemplo, que existe uma probabilidade de 84.8% dos clientes sem pagamentos em atraso continuarem na mesma classe de risco, 5.27% transitarem para a classe de risco C_3 e 11.7% saírem da carteira no mês seguinte.

- Podemos observar que os clientes da quinta clase de risco, têm 75.51% de probabilidade de se manter na mesma classe. Isso indica que, esses clientes tem pouca chance de efetuar o pagamento em dívida, ou seja, estão financeiramente com condições agravosas. Neste caso a probabilidade de um cliente com muitas prestações em atraso vir recuperar do incumprimento é muito baixa.

3.2.3 Entradas no sistema

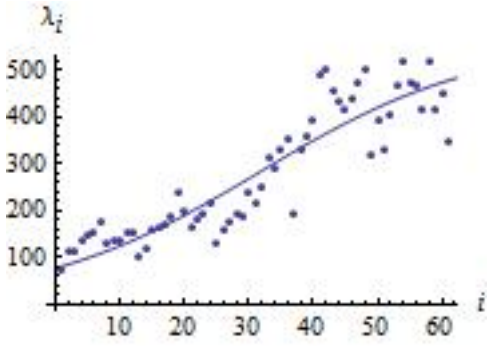
Vamos modelar o número de novos clientes que entram na carteira de crédito ao consumo de acordo com a forma funcional sigmoidal.
A figura 3.2 mostra o diagrama de dispersão correspondente à entrada de clientes na carteira durante o período de tempo referido anteriormente.

Figura 3.2: Entrada de clientes na carteira



Podemos analisar que o número de novos clientes que entraram na carteira tende a crescer com o tempo. Na verdade logicamente os bancos esperam esse comportamento ao longo do tempo, ou seja acreditam no crescimento da entrada dos novos clientes.
A figura 3.3 seguinte, ilustra a forma sigmoidal ajustado aos dados da carteira em estudo.

Figura 3.3: Ajustamento da sigmoidal



Analisando a figura 3.3, podemos ver que a forma funcional sigmoidal se ajusta bem aos dados da carteira, pois os valores estão concentrados à volta da linha do gráfico, o que nos permite afirmar que o modelo adoptado pode ser representativo da realidade da carteira.

Os parâmetros da forma funcional sigmoidal é obtido pelos estimadores de máxima verosimilhança, apresentados na seção 3.1.5 e obtidos através da resolução das equações normais 3.19.
Os novos indivíduos são inicialmente classificados na primeira classe de risco, pois nenhum cliente terá inicialmente prestações em atraso, para isso da definição de classe de risco temos que $\mathbf{c}^T = (1, 0, ..., 0)$. A tabela 3.3 mostra os detalhes do cálculos do ajustamento, o intervalo de confiança para os parâmetros e os respetivos erros do ajustamento.

Tabela 3.3: Parâmetros estimados da forma sigmoidal

Parâmetros	$f(x) = \left(a + b.e^{-\theta x}\right)^{-1}$	Intervalo de confiança	Standard Error
$\hat{a}$	0.0017648	$]0.00137142, 0.00215818[$	0.000235335
$\hat{b}$	0.011185	$]0.00750018, 0.0148698[$	0.00220443
$\hat{\theta}$	0.0582251	$]0.0403002, 0.0761501[$	0.0107235
$\lim_{x \to +\infty}$	566.637		

Fonte: Adaptação de Fernandes [12]

Estimados os parâmetros do fluxo de entrada dos clientes na carteira, e considerando a matriz de transição 3.2, vamos agora estimar as dimensões das sub-populações ao longo do tempo, e fazer estimativa temporal da probabilidade de incumprimento.

3.2.4 Estimação temporal da probabilidade de incumprimento

Aqui consideramos que a matriz, $\mathbf{K}$, de transição entre os estados transcientes é dado por:

$$\mathbf{K} = \begin{bmatrix} 0.847766 & 0.035130 & 0 & 0 & 0 \\ 0.437653 & 0.488998 & 0.073349 & 0 & 0 \\ 0.130435 & 0.42029 & 0.160869 & 0.288406 & 0 \\ 0 & 0.102916 & 0.411664 & 0.348199 & 0.137221 \\ 0 & 0 & 0.082781 & 0.165563 & 0.751656 \end{bmatrix}$$

A transição entre as cinco classes de riscos definidas anteriormente é determinado através da matriz $\mathbf{K}$.
Considerando que cada novos elementos entram diretamente na primeira classe, isto é, admitimos que todos os clientes ao entrarem pela primeira vez na carteira crédito não está com qualquer dívida. Ou seja, o vetor de classificação inicial constante para cada instante da tempo será:

$$\mathbf{t}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Lembrando que a entrada de novos clientes na carteira a cada data i é modelada através da forma funcional sigmoidal:

$$\lambda_i = \left(0.0017648 + 0.011185.e^{-0.0582251.i}\right)^{-1}$$

Estimamos a probabilidade de incumprimento ao longo de um determinado período temporal, utilizando o modelo vórtices estocásticos, estudado na seção anterior.

A tabela 3.4 mostra a probabilidade de incumprimento no primeiro mês, verificamos que 100% dos clientes entram na primeira classe, isto significa que não há clientes com prestação em atraso logo no primeiro mês da entrada na carteira. Neste caso temos 0% de clientes em incumprimento.

Tabela 3.4: Vórtices Estocástico nas classes de risco para forma funcional sigmoidal-mês 1

	Classe1	Classe2	Classe3	Classe4	Classe5
#	82	0	0	0	0
%	100	0	0	0	0

Tabela 3.5: Vórtices estocástico nas classes de risco para forma funcional sigmoidal-mês 3

	Classe1	Classe2	Classe3	Classe4	Classe5
#	222	7	1	0	0
%	96.52	3.04	0.43	0	0

Na terceira linha da tabela 3.5 mostra que no terceiro mês cerca de 96.52 % de clientes estão em cumprimento, enquanto que cerca de 0.4464% de clientes estão com mais de 60 e menos de 90 dias de prestações em atraso.
Pode ser estimado a probabilidade de incumprimento a qualquer instante do tempo aplicando o modelo Vórtices Estocásticos.

Tabela 3.6: Vórtices estocástico nas classes de risco para forma funcional sigmoidal-mês 20

	Classe1	Classe2	Classe3	Classe4	Classe5
#	1142	77	8	4	2
%	92.62	6.24	0.65	0.32	0.16

Na terceira linha da tabela 3.6 mostra que no mês 20 cerca de 92.62 % de clientes estão em cumprimento, enquanto que cerca de 0.48% (Classe4+Classe5) de clientes estão em incumprimento.
Convém reafirmar que o modelo Vórtices Estocástico estima com extrema facilidade a probabildade de incumprimento a qualquer instante do tempo, neste caso referimos que cada instante do tempo corresponde a um determinado mês.

Capítulo 4

Conclusão e Recomendações

Conclusão

Esta dissertação de mestrado assumiu sempre como objetivo estimar a probabilidade de incumprimento dos clientes no momento da concessão do crédito e de uma carteira de crédito ao consumo ao longo de um determinado período temporal.
Realizou-se uma revisão de literatura, sobre os principais conceitos relacionados com o crédito e risco de crédito na ideologia de diversos autores e também alguns conceitos sobre os principais modelos do risco de crédito onde destacamos o modelo de *credit scoring*.
Neste trabalho apresentou-se a metodologia utilizando o modelo de Regressão Logística e a metologia utilizando o modelo Vórtices Estocásticos, para estimar a probabilidade de incumprimento.
A metodologia apresentada para estimar a probabilidade de incumprimento de clientes no momento da concessão do crédito, utilizando Regressão Logística, pode servir de grande ajuda para tomadas de decisão dos gestores do crédito, pois, anticipa o resultado de bom ou mau pagador, permitindo deste modo reduzir o incumprimento.
A metodologia apresentada para estimar a probabilidade de incumprimento de uma carteira de crédito ao consumo, utilizando o modelo Vórtices Estocásticos é de extrema importância, pois, pode ajudar os gestores de crédito a classificar a carteira de crédito ao longo de determinado período temporal.
Relativamente ao modelo de Regressão Logística utilizou-se as técnicas *Stepwise-backward* e AIC para a seleção das variáveis mais significativas no que respeita a explicar a ocorrência de incumprimento.
Ambas as técnicas de seleção aplicadas, Stepwise e AIC, revelaram que as variáveis que melhor explicam a ocorrência de incumprimento, de entre todas as que foram analisadas, são: Taxa nominal, Valor das prestações, Valor do empréstimo, Idade, Agência, Actividade profissional, Genero, Entidade patronal e Habilitações.
Os resultados mostram que o modelo Regressão Logística consegue prever, em média, acertadamente cerca de 51.7% dos casos, quando se define que os clientes com as probabilidades acima de 40% são considerados em incumprimentos. Desta forma, concluimos que tendo

em conta a razoável taxa de previsibilidade do modelo ajustado, considera-se que este modelo poderá ser muito eficiente para estimar a probabilidade de incumprimento para novos clientes solicitantes de um crédito ao consumo.
Relativamente ao modelo Vórtices Estocásticos, verificou-se que a cadeia de Markov é a base, onde a matriz de tranzição a um passo desempenha um papel essencial para a transição e classificação dos clientes a quando das novas entradas.
Cada cliente ao entrar na carteira, a priori é colocado na primeira classe de risco de acordo com o vetor de classificação inicial e futuramente com as novas entradas será reclassificada e transitada para nova classe de risco de acordo com a matriz de transição.
A entrada dos clientes na carteira foi modelada através da forma funcional sigmoidal e os parâmetros deste modelo estimados através dos estimadores da máxima verossimilhança.
Numa perspetiva prática, definiu-se cinco classes de risco de acordo com o número de dias em incumprimento, de seguida com as informações da carteira apresentamos a matriz de transição entre as diferentes classes de risco.
Dos resultados obtidos através do modelo Vórtives Estocásticos constatou-se por exemplo que no mês 20 cerca de 92.62 % de clientes estão em cumprimento, enquanto que cerca de 0.48% (Classe4+Classe5) de clientes estão em incumprimento.
Por fim este trabalho é um guia prático que contribui de uma forma ou outra para académicos que investiguem áreas científicas do *credit scoring*, bem como para especialistas em gestão e análise de risco do crédito bancário.

Recomendações

As recomendações estão divididas em dois grupos: às instituições financeiras e aos gestores de crédito.
As recomendações às instituições financeiras têm como objectivo realçar alguns aspetos importantes na gestão de crédito e cobrança. Alguns destes aspetos já estarão a ser uma prática normal nas empresas. Neste sentido servirão então para reforçar a importância dos mesmos. Em relação às recomendações dirigidas aos gestores de crédito o objetivo será mencionar os pontos que ainda podem ser melhorados. Em ambos os casos, o objetivo primordial é sempre o de minimizar as perdas resultantes da política de crédito e cobrança.

- **Às instituições bancárias**

  Por vezes as instituições financeiras permitem aos clientes pagamentos tardios, estão a criar dificuldades financeiras a si próprias e qualquer empresa estaria falida se vendesse e não cobrasse ou, se cobrasse apenas parte do valor do bem vendido. Aliás, os pagamentos tardios e os incobráveis são uma das barreiras mais importantes do comércio financeiro, com custos adicionais anuais de alguma forma elevado. A situação dos incobráveis, deveu-se, na maioria dos casos não só, ao aumento das vendas de crédito, como também a uma abordagem não adequada da política de crédito da

instituições bancárias. Portanto, uma das recomendações é antever e prevenir situações de incumprimentos, acompanhando e monitorizando adequadamente os clientes e os riscos que os associam, no sentido de evitar processos de cobrança desnecessárias e, por vezes, difíceis.

As instituições financeiras lutam para evitar uma possível inadimplência ou impedir que o seu aumento interfira no seu resultado ou ainda para manter a margem de lucro em tempos de retração da economia, nestas situações precisam rever seus instrumentos de análise de risco. Os riscos financeiros representam para as instituições financeiras uma ameaça à sua sobrevivência, ao mesmo tempo que os riscos estratégicos e operacionais, precisam ser desvendados nas análises de concessão de crédito para clientes.
Instituições financeiras possuem em sua raiz uma grande preocupação com a inadimplência, tendo em vista que a liberação de recursos para seus clientes é a principal fonte de renda e motivo original de sua existência. A análise de crédito precisa ser mais criteriosa visando reduzir, mitigar ou até mesmo eliminar o risco de uma operação.
Devido ao cenário em que nos encontramos, recomenda-se novas pesquisas na área, mais especificamente quanto a análise de crédito, pesquisas que possam apresentar novas formas de análise de crédito ou então novas ferramentas para a análise de crédito para que se possa ter instituições mais fortalecidas e um sistema financeiro de alta performance.
As instituições financeiras devem desenvolver modelos internos de gestão de risco de crédito com o intuito de oferecer feramentas mais eficientes para a valorização da carteira, medição de riscos, apreçamento de novos créditos, potencializando os ganhos dos capitais emprestados adequando-os ao montante de capital que estes devem manter como parte da sua estrutura de capital.

- **Aos gestores de crédito**

  Toda e qualquer instituição financeira tem uma ou mais pessoas que lidam diretamente na decisão do crédito a clientes ou a empresas, os gestores de crédito. É fundamental compreendermos, antes de mais, o importante papel desempenhado pelos gestores de crédito no seio dos departamentos de crédito das instituições bancárias. Compete aos gestores de crédito avaliar a capacidade de um cliente servir a dívida, ou seja, pagar o devido de acordo com os prazos, montantes e condições acordados. Recomenda-se aos gestores de crédito:

  - a avaliar a capacidade creditícia do potencial cliente quer na fase preparatória do crédito como na fase posterior ao início da relação de crédito;
  - procurar a informação adequada para que possa fazer um trabalho produtivo, tal informação é-lhe facultada por fontes internos ou externos.
  - verificar o estado financeiro dos principais clientes que fazem parte da carteira

periodicamente. Existem alguns sinais que podem alertar para uma revisão do estado financeiro dos clientes e, consequentemente dos limites de crédito, nomeadamente a solicitação para o aumento dos prazos de pagamentos, os limites de créditos frequentemente excedidos em relação aos acordados, os atrasos nos pagamentos mais do que o habitual, questões relacionadas com as faturas como forma de atrasar os pagamentos e a redução das suas compras médias habituais. Portanto, se alguns dos clientes começarem a apresentar estes comportamentos deve ser, então, o momento para apurar as razões para tal mudança de comportamentos e proceder em conformidade.

Referências Bibliográficas

[1] Almeida, F. T. (2002). *Financiamento e Crédito Bancário I.Departamento Pedagógico do ISGB, Portugal*. E.Santos-Artes gráficas, Lda, 6.ª edition.

[2] Anderson, R. (2007). *Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.Oxford University Press*. Oxford.

[3] Araújo, E. (2006). Modelagem de risco de crédito: Aplicação de modelos credit scoring no fundo rotativo de ação da cidadania-cred cidadania. *Dissertação de mestrado, Universidade Federal de Pernambuco, Brasil*.

[4] Baptista, A. S. (2004). *A gestão do crédito como vantagem competitiva*. Editora Vida Económica, 3.ª edition.

[5] Barros, G. (2008). Modelos de previsão da falência de empresas: aplicação empírica ao caso das pequenas e médias empresas portuguesas. (dissertação) - instituto superior de ciências do trabalho e da empresa - departamento de economia - lisboa, portugal.

[6] Brito, G. A. S. & Neto, A. A. (2008). Modelos de classificação de risco de crédito de empresa. *Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo.Brasil*.

[7] Caoutte, J. B., Altman, E. I., Narayanan, P., and Nimmo, R. (1998). *Managing Credit Risk: The Next Great Financial Challenge*. Hardcover.

[8] Caoutte, J. B., Altman, E. I., Narayanan, P., and Nimmo, R. (2008). *Managing Credit Risk, The Great Challenge for the Global Financial Markets, United States of America*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2.ª edition.

[9] Chaia, A. (2003). Modelos de gestão do risco de crédito e sua aplicabilidade no mercado brasileiro. *Dissertação de mestrado, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo. Brasil*.

[10] Cordeiro, G. and Demétrio, C. (2007). Modelos lineares generalizados. in: Simpósio de estatística aplicada à experimentação agronômica ; reunião anual da região brasileira da sociedade internacional de biometria . santa maria.brasil.

[11] Esquível, M. L., Guerreiro, G. R., Fernandes, J. M., and Silva, A. F. (2014). On a spread model for portfolio credit risk modeling. *Proceedings of Conference ICNAAM, Greece.*

[12] Fernandes, J. M. L. (2012). Estudo de uma carteira de crédito ao consumo de um banco de cabo verde. *Dissertação apresentada como requisito parcial para obtenção do grau de Doutor em Estatística e Gestão de Informação pelo Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa, Portugal.*

[13] Filho, R. B., P.Bagolin, I., and V.Comim, F. (2010). Determinantes da permanência na condição de pobreza crônica: aplicação do modelo logit multinomial. texto para discussão. porto alegre.

[14] Gestel, V., Baesens, B., Dijcke, P. V., Garcia, J., J.Suykens, and Vanthienen, J. (2006). *A process model to develop an internal rating system: Sovereign credit ratings. Decision Support Systems.*

[15] Guerreiro, G. (2001). Uma abordagem alternativa para bonus malus. *Tese de mestrado, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa, Portugal.*

[16] Guerreiro, G. and Mexia., J. (2008). Stochastic vortices in periodically reclassified populations. *Discussiones Mathematica e Probability and Statistics. Department of Mathematics, FCT - New University of Lisbon Campus da Caparica, Portugal.*

[17] Gujarati, D. N. (2000). *Econometria Básica.* Makron Books, 3.ª edition.

[18] Hair, J., W.Black, and Anderson, R. (2009). Multivariate data analysis. wiley, hardcover.

[19] Hosmer, D. and Lemeshow, S. (1989). Applied logistic regression. *New York: Wiley.*

[20] Lewis, E. (1992). An introduction to credit scoring. *An Introduction to Credit scoring, California, E.U.A.*

[21] McCullagh, P. and Nelder, J. (1989). Generalized linear models. chapman and hall/crc, usa.

[22] Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A(135). New York: Wiley .*

[23] Pereira, J. (2009). Credit risk.

[24] Rodrigues, E. V. T. (2011). Uma abordagem comparativa e analÍtica de dois sistemas de bonus malus em cabo verde: O sistema actual e a proposta da garantia. *Mestrado em Matemática e Aplicações. Universidade Nova de Lisboa, Portugal.*

[25] Saunders, A. and Allen, L. (2002). *Credit Risk Measurement: New Approaches To Value-At-risk And Other Paradigms.* Hardcover.

[26] Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* New Jersey.

[27] Silva, A. F. A. V. (2014). Modelação do risco de crédito numa carteira de crédito ao consumo. *Dissertação para obtenção do Grau de Mestre em Matemática e Aplicações, no ramo de Actuariado, Estatística e Investigação Operacional.Universidade Nova, Lisboa.*

[28] Silva, J. (2000). Gestão e análise do risco de crédito. são paulo, atlas: Atlas.

[29] Sousa, L. S. D. (2012). Análise e avaliação do risco de crédito bancário nas pmes-utilização do modelo de rating. *Licenciatura em Contabilidade e Administração. Instituto Superior de Ciências Económicas e Empresariais. Mindelo.*

[30] Thomas, L., Edelman, D., and J.N, C. (2002). Credit scoring and its applications. siam, philadelphia.

[31] Turkman, M. and Silva, J. (2000). Modelos lineares generalizados-da teoria à prática.fct - praxis xxi- feder. *Lisboa.*

[32] Vale, C. A. L. (2010). Modelação e estimação do risco de crédito-estudo de uma carteira. *Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Matemática e Aplicações - Actuariado, Estatística e Investigação Operacional, Portugal.*

[33] Zorich, V. A. (2009). Mathematical analysis i. *Springer-Verlag, Berlin.*