LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

SINCE 1911

MESTRADO

ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

COMPORTAMENTO *ONLINE* DOS PORTUGUESES E IDENTIFICAÇÃO DOS FACTORES DE CONVERSÃO

JOÃO PEDRO PIRES DOS SANTOS

OUTUBRO 2018

LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

MESTRADO EM

ECONOMETRIA APLICADA E PREVISÃO

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

COMPORTAMENTO *ONLINE* DOS PORTUGUESES E IDENTIFICAÇÃO DOS FACTORES DE CONVERSÃO

JOÃO PEDRO PIRES DOS SANTOS

ORIENTAÇÃO:

PROFESSOR DOUTOR NUNO SOBREIRA
DOUTOR NUNO SANTOS

OUTUBRO 2018

Abstract

The present study's goal is identifying the factors that most explain the conversion of internet users in a *website* related to the banking and insurance sectors.

From a sample containing online behaviour data about a panel of portuguese users, a descriptive analysis was done, followed by the estimation of two logistic regression models for each sector of analysis.

The descriptive analysis of data showed that demographic variables such as age, sex, and the individual's work situation, were relevant to distinguish between converted and unconverted individuals.

The first estimated *logit* model, relative to banking, points the use of search engine and a combination of behavioral indicators with demographic characteristics, as the variables that best explains the conversion in the sector.

For the second logistic regression, about the insurance sector, we conclude the last click immediately before the conversion, is the most relevant to the event. There are bigger odds an internet user gets converted in the insurance sector, by jumping directly from a *website* regarded to a big corporation. This phenomenon is linked to employees that benefits from a health insurance service, provided by the corporation.

Keywords: *Logit,* Conversion, Forecast

Resumo

O presente estudo tem como objectivo a determinação de indicadores avançados que permitam identificar indivíduos susceptíveis de se converterem a um sítio online relativo aos sectores da banca e seguros. Para tal, a partir de uma amostra que contém informação comportamental online de um painel de utilizadores portugueses, foi feita uma análise descritiva aos dados, seguida da estimação de dois modelos de regressão logística respeitante a cada sector alvo de análise.

A análise descritiva aos dados demonstra que variáveis demográficas, tais como idade, região, sexo e situação laboral do indivíduo, foram relevantes na distinção dos indivíduos convertidos e não convertidos.

Do primeiro modelo *logit* estimado, relativo à banca, retiramos o motor de busca e a combinação de indicadores comportamentais com características demográficas, como as variáveis que melhor prevêem a conversão de um indivíduo no sector. Jovens com idades entre os 18 e 30 anos, recentemente pais, que passam muito tempo em casa e não estão habituados à internet, constituem um dos perfis mais destacados na probabilidade de ocorrência do evento conversão em banca.

Relativamente à segunda regressão estimada, para os seguros, conclui-se que o último clique antes da conversão é o mais relevante. Um indivíduo é mais passível de se converter ao sector de seguros, partindo directamente da página de uma grande empresa. Este fenómeno está implicitamente relacionado com os funcionários que beneficiam de seguro de saúde de uma dada empresa.

Palavras-chave: ***Logit,*** **Conversão, Previsão**

Agradecimentos

Quero agradecer a todas as pessoas que me apoiaram na realização deste trabalho, nomeadamente família e amigos mais próximos.

A toda a equipa envolvida da PSE e à minha colega Mariana, agradeço pelo conhecimento, disponibilidade e boa companhia que sempre demonstraram.

Também uma palavra de apreço ao professor doutor Nuno Sobreira e doutor Nuno Santos, que na qualidade de tutores se disponibilizaram a ajudar ao longo dos últimos meses.

Por fim, obrigado ao ISEG, pela excelência de ensino com que se pautou durante todo o meu percurso académico. Recordarei com saudade todos estes anos vividos como aluno desta instituição.

Índice

Índice de Tabelas

Índice de Figuras

Lista de Abreviaturas

ACEPI *Associação da Economia Digital*

ADSE *Assistência na Doença aos Servidores Civis do Estado*

AIC *Akaike information criterion*

ATM *Automated Teller Machine*

BBVA *Banco Bilbao Vizcaya Argentaria*

BCP *Banco Comercial Português*

BPI *Banco Português de Investimento*

BIG *Banco de Investimento Global*

CGD *Caixa Geral de Depósitos*

CHAID *Chi-square automatic interaction detection*

ETL *Extraction, Transformaton and Load*

IDC *International Data Corporation*

1. Introdução

Tem havido uma mudança de paradigma no que respeita ao comportamento *online* dos portugueses. Se antes, o peso de serviços digitais na economia era inócuo, actualmente revela-se de extrema importância para o funcionamento da economia e suas relações comerciais, existindo uma clara canalização de recursos e investimento no âmbito do *online* por parte das empresas portuguesas, cujo objectivo passa por combater a concorrência exterior e acompanhar a evolução global que se regista ao nível tecnológico no meio empresarial.

A facilidade de acesso à internet fruto de melhores e mais infraestruturas, o natural envolvimento tecnológico das gerações vindouras, originam uma exigência nunca antes vista para a oferta digital em qualquer serviço e sector.

O estudo Anual da Economia e da Sociedade Digital em Portugal indica que metade dos portugueses prefere comprar no exterior, seja pelo preço mais competitivo, maior diversidade de produto ou pelo melhor e mais cómodo serviço prestado *online*, sendo que a dificuldade do transporte e complexidade do produto também influenciam negativamente a escolha por esta opção.

É na área do *Entretenimento e Media* que se observa o maior consumo *online* da população portuguesa, indicação dada pelo enorme tempo despendido em *site*s incidentes, tais como jornais e jogos.

Dada a tendência de exploração *online* em sectores sobejamente conhecidos, importa perceber agora como e onde actuar com o intuito de captar novos clientes em sectores estratégicos não tão explorados *online*, longes do limiar de utilidade mas com o potencial em termos de crescimento de negócio.

Assim, o foco deste trabalho inserir-se-á na determinação de indicadores avançados que permitam identificar indivíduos susceptíveis de se converterem a um sítio *online* dos sectores da banca e seguros.

Para além do estudo dos factores que despoletam a conversão dos portugueses, importa perceber mais aprofundadamente os sectores em análise: quais os tipos de serviços bancários/seguros que mais são procurados digitalmente e que se adequam e respondem às necessidades dos clientes; diferenciar, de um universo de várias instituições do

mesmo subsector financeiro, as que melhores resultados obtêm na internet junto dos portugueses, e as que pecam em relação à concorrência.

No final deste trabalho, serão retiradas as devidas conclusões, tendo como objectivo capturar informações relevantes, potencialmente alvos de decisões estratégicas a tomar no seio do sector da banca e seguros.

O presente trabalho encontra-se organizado em 8 secções cuja explicação se resume ao seguinte:

Primeiramente é feita uma introdução ao tema em estudo. Segue-se um enquadramento teórico do que é o mundo *online* e em que prisma se encontram os sectores estudados. Também se vai definir o conceito da conversão e indicar estudos e modelos realizados no passado sobre este tema.

Posteriormente, na secção relativa ao âmbito do projeto, expõe-se uma breve contextualização do tema no âmbito do estágio realizado na empresa PSE – Produtos e Serviços de Estatística. É descrito o planeamento a seguir, qual o tratamento necessário aos dados e a estruturação em base de dados, tendo em vista a operacionalização das estatísticas e modelação a realizar.

Na parte empírica são apresentadas estatísticas descritivas dos dados e vai-se explicar a metodologia utilizada para a estimação dos modelos de previsão da variável resposta. Findo o anterior, temos a interpretação dos resultados e a execução de testes de especificação e qualidade de ajustamento. Finalmente, são retiradas as devidas conclusões, seguidas de recomendações para trabalhos futuros.

2. Enquadramento teórico

Historicamente, o sector bancário é considerado pioneiro em Portugal na relação entre cliente e fornecedor pela via eletrónica. Remontando ao ano de 1985, em Portugal foram introduzidas as primeiras “caixas mágicas” de Multibanco, que permitiam efectuar diversas operações sem ser necessário estar presente numa sucursal bancária (Cheta, 2007).

Ao longo dos anos seguintes é notório o crescimento da oferta de produtos digitais bancários – com a introdução do *Internet banking* no mercado em fins dos anos 90 - e a correspondente adesão por parte dos clientes.

Neste contexto, o primeiro objectivo dos bancos, para além de acompanhar o avanço natural da tecnologia e de seus concorrentes, passa por penetrar em novos nichos de mercado, sendo que os jovens em início de carreira, em particular os *millenials*[1], representam o grande *target* pela sua familiaridade e facilidade de manuseamento de qualquer meio tecnológico.

De acordo com Figueiredo (2014) é também fulcral para o sucesso de um banco saber reter e converter os clientes existentes a produtos novos, estratégia esta que é ainda mais relevante pela dificuldade em captar novos clientes.

A poupança de recursos decorrentes da menor necessidade em empregar gestores de balcão e da diminuição da utilização de ATM 's constituiu outra razão para explorar e aumentar a oferta de serviços digitais.

No sector das seguradoras, a inovação impera como meio de obter vantagem sobre a concorrência. Observa-se uma maior diversidade de produtos personalizados com vista à captação do cliente através de descontos, parcerias e, naturalmente, da aposta num serviço de comunicação mais eficaz e eficiente, via meio digital (Lacerda, 2015).

O foco deixou de estar no mundo físico e passou para todo o movimento inerente ao *online*, cuja real medida de sucesso em negócio, se dá pelo conceito de conversão.

Quando um internauta visita um *site* e tem uma determinada acção específica, dá-se a conversão. A acção específica pode ser ou não, de fácil mensuração, e varia consoante o objectivo e tipologia do negócio.

Halchuk (2014) através de um exemplo prático sugere que, num *site e-commerce*, se calcule a conversão através da relação entre o número de visitas e o número de vendas da loja.

Já se considerarmos um determinado *site* de fornecimento de conteúdos - por exemplo, um jornal de notícias - o conceito conversão baseia-se na frequência de visita do indivíduo, divergindo do significado referido anteriormente.

Para a análise realizada sobre os sectores de banca e seguros e, dada a tipologia dos dados disponíveis, consideramos para o evento conversão todo o indivíduo que visitou pela primeira vez um *site* incidente no sector de interesse.

Kim e Prabhakar (2004) afirmaram que a conversão em banca *online* está intimamente ligada à relação existente entre a internet e o indivíduo em causa, à confiança gerada pela instituição bancária e à partilha de testemunhos positivos em prol do serviço digital

bancário, feito por pessoas próximas ao indivíduo. "Se o nível de confiança ultrapassa o valor de risco percecionado" o utilizador é mais facilmente convertível à banca no *online*.

O estudo "Nielson Global Survey, consumer confidence section, (2017)" considera a publicidade em dispositivos móveis, as recomendações e opiniões - comentários *online* ou de familiares - como exemplos de formas eficazes de atrair clientes no geral, estendendo-se ao sector dos seguros. O seguro de vida é um dos produtos mais procurados via *online* de acordo com um estudo realizado em 2013 pela organização LIMRA [1].

A literatura sugere alguns estudos do âmbito da identificação de indicadores avançados na previsão de um determinado evento acontecer.

Koon e Petscher (2015) apontam que parte das análises que visam estudar relações causa-efeito entre um conjunto de variáveis, cujo foco é a identificação de um certo padrão ou característica específica na variável dependente, devem ser feitas parametricamente através da regressão logística, sobrepondo-se a abordagens alternativas tais como árvores de decisão ou de classificação. Tal razão prende-se pela facilidade de estimação dos efeitos directos das variáveis independentes sobre a variável dependente, e os seus resultados dicotómicos.

Não obstante, os métodos não paramétricos também poderão assumir relevância na análise pré-modelar, na identificação de eventuais significâncias em variáveis e suas interações, bem como na explicação e especificação do modelo.

Ritta *et al.* (2015) estudam a probabilidade de um indivíduo não cumprir com as obrigações relativas ao crédito, tendo estimado um modelo *Logit* cuja variável dependente é a previsão do indivíduo pertencer ou não, ao grupo de risco de não pagamento. O modelo final ajustado obteve 70% de acerto das classificações de indivíduos, constituindo um método consistente e viável ao suporte na avaliação do risco de crédito e identificação dos indicadores que o influenciam.

Cabral (2013) através de uma regressão logística, procurou analisar a compra de produtos de saúde de uma determinada marca em relação à concorrência, usando como variável dependente binária o consumo ou não dos produtos da marca, dado um

[1] LIMRA é uma organização de consultoria, investigação e desenvolvimento sediada nos Estados Unidos da América.

conjunto de variáveis. Através da análise da percentagem de compradores de produtos da marca pela primeira vez, concluiu-se que a taxa de conversão à marca em estudo foi de 60%.

Mudiwa (2011), para estudar a adopção de uma dada tecnologia por parte dos agricultores no Zimbabwe, estimou a partir de um conjunto de dados recolhidos em forma de inquérito, um modelo de regressão logística cujos resultados demonstraram que, factores como a experiência agrícola, experiência com a tecnologia, acesso às vendas de mercado e a posse de ativos, influenciavam positivamente a adopção da tecnologia por parte dos agricultores.

3. Âmbito do Projeto

Através de uma amostra com informação relativa ao comportamento *online* de um painel de portugueses no ano de 2017, tem-se como premissa central a estimação de um modelo de regressão logística, cuja previsão consiste na probabilidade de ocorrência do evento conversão a um domínio[2] do sector alvo de análise - Banca e Seguros.

O Projeto iniciou-se com diversas etapas intermédias essenciais para a sua execução e realizadas na empresa PSE, que se define como uma empresa consultora de serviços preditivos analíticos.

Vivendo em plena era digital, a velocidade com que se acede e recolhe dados de informação dos mais variados campos de estudo, em tudo supera o razoável, sendo por isso necessário e exigido, o correcto e competente tratamento dos dados, algo que devido ao seu dramático volume, é alcançável apenas com a ajuda de ferramentas tecnológicas (Fayyad et al., 1996).

Neste sentido, em primeiro lugar, efectuou-se o estudo e a contextualização de toda a substância teórica decorrente do tema em questão, seguida de aprendizagem e ambientação na utilização da principal ferramenta utilizada para a execução de todo o processo operacional e analítico deste trabalho, o SPSS.

Sobre o conjunto de dados em bruto de elevada dimensão, e separados em diversos ficheiros de formatos diferentes, procedeu-se ao tratamento, limpeza e organização destes, com o fim de uniformizá-los e carregá-los em uma única base de dados.

[2] Por domínio entende-se uma página da internet em sentido lato.

O Processo descrito anteriormente é conhecido por ETL e o seu objectivo é tornar a informação originalmente ruidosa e pesada, em útil, leve e manuseável, facilitando as análises e os *insights* destas retirados.

Para a produção do Banco de dados fez-se a estrutura e relacionamento em Esquema Estrela[3], tendo sido criada a tabela de factos, que contém toda a informação principal de forma redundante, conectando-a a diversas tabelas de dimensão, tal como ilustrado no diagrama da Figura 1 presente no anexo B.

Finalizadas as transformações e o posterior carregamento dos dados, são extraídos dois conjuntos de observações da Tabela de factos - respeitantes aos 2 sectores de análise - que serão alvos da modelização. São incluídos apenas os indivíduos de interesse ao estudo, i.e., os que se converteram ao sector em questão, e é feita uma divisão em dois períodos de análise: as visitas correspondentes ao intervalo entre os 30 a 15 dias anteriores à conversão, são assignadas ao período de não conversão, e as visitas entre os 15 dias anteriores e a data imediatamente precedente ao acontecimento do evento, são nomeadas como o período de pré-conversão.

Assim, através da distinção e estudo de padrões observados entre os dois períodos comportamentais em análise, é possível determinar os indicadores avançados de um internauta mais susceptível de ser convertido a um sector.

A fase de exploração dos dados é crucial para a correcta definição das variáveis independentes a incluir no modelo e constituição de bons *inputs* para a previsão da variável dependente.

Findos os passos anteriores, é feita a modelação e interpretação do modelo final, submetendo-o a um conjunto rigoroso de testes de especificação.

Que questões gostaríamos de ver respondidas relacionadas com o problema em estudo?

Existe um conjunto de variáveis clássicas que é possível identificar *à priori* como relevantes para a conversão a um *site* do sector, e que para as quais já existe uma estrutura competente do negócio para fomentar e satisfazer as necessidades desse público-alvo. Constituem exemplos concretos, as matérias relacionadas com crédito habitacional e automóvel - em termos de banca - e seguros de saúde a recém-grávidas – no caso das seguradoras.

[3]Esquema Estrela ou "Star Schema" é a abordagem mais usada para o design de construção de uma base de dados, cuja grande vantagem é a melhoria do desempenho decorrente das suas análises.

Variáveis deste tipo são importantes a incluir para dar consistência e obter corroboração ou não, de teses pré-concebidas.

Contudo, este estudo pretende ir além do anterior, na medida em que ambiciona encontrar factores comportamentais que ajudem a definir perfis muito específicos e pormenorizados potencialmente alvos de captação em termos de negócio e de crescimento nos sectores em análise.

Estarão os portugueses mais inclinados a visitarem um *site* de banca quando procuram um computador portátil em compras *online*? Se sim, quais as suas características demográficas? Serão os viciados em viagens turísticas potenciais convertidos?

Estes são os tipos de questão que se pretende obter resposta.

4. Dados em Estudo

4.1. A técnica de amostragem

Os dados utilizados para a elaboração do estudo provêm da Netquest, empresa especialista no mercado de painéis digitais em dados comportamentais, cobrindo actualmente mais de 23 países em todo o mundo. A autenticidade dos dados é garantida através de um recente e inovador método de recolha, que consiste no rastreamento digital de toda a atividade exercida por um conjunto de indivíduos representativos da população - escolhidos exclusivamente através de convite, sendo assegurado o seu anonimato - através de um sistema instalado nos respetivos dispositivos móveis (telemóvel ou *tablet*) e fixos (computador). Esta forma de recolha de dados assegura uma maior coerência e uma maior proximidade com a realidade, uma vez que capta *in loco* todos os dados comportamentais *online* em absoluto do indivíduo em questão.

Em particular, espera-se que este método seja superior a outros métodos de recolha mais tradicionais, como por exemplo, inquéritos ou entrevistas.

4.2. Descrição

Os dados usados para o presente estudo totalizam 88 890 515 de registos, que se definem como o nº total de cliques na internet, por parte dos utilizadores em painel. Estes dividem-se entre registos feitos em dispositivo móvel (8,27%) e computador fixo

(91,73%), sendo que no dispositivo móvel os registos dividem-se entre *Browser*[4] (25%) e Aplicações móveis/*Apps* (75%) tal como se pode observar na Tabela I do anexo A.

Observando a distribuição de frequências, já retirados os *outliers*, temos 88 885 147 de registos em toda a amostra.

Os portugueses ao longo do ano de 2017 visitaram um total de 254 574 domínios diferentes, distribuídos por 229 países de origem, tendo estes sido alvos de 30 723 712 visitas, cuja duração agregada perfaz 8 907 418 minutos, significando aproximadamente 148 457 horas de dados registados ao longo de todo o ano pelo conjunto de painelistas.

Uma visita pode ser feita por dois tipos de fonte, móvel ou fixo, cujos sistemas operativos podem ser *android, IOS* ou outros. O tipo de visita ao domínio pode ser em ambiente de *App* ou *Browser*, sendo que o dispositivo em uso, assume um de três: Computador, *Tablet* ou *Smartphone*.

Os *site*s estão organizados em 25 categorias, os quais se subdividem em 131 subcategorias, sendo ainda associado a cada um dos domínios, uma marca.

Por exemplo "google.pt" e "google.com" são dois domínios diferentes, sendo que ambas possuem uma só marca, "Google" e a respetiva categoria e subcategoria, "Motor de Busca", podendo ser acedida por qualquer dispositivo móvel ou fixo.

5. O Painel

Os indivíduos escolhidos para o estudo do comportamento da população na internet, caracterizam-se por um painel constituído por 1 727 indivíduos portugueses, cujo intervalo de recolha dos dados se situa entre os meses de Janeiro e Dezembro do ano 2017. Ao longo do ano o painel sofre mudanças – por exemplo, um utilizador pode ter o seu 1º registo *online* reportado em Janeiro, tendo o último sido em Setembro, enquanto outro pode ter apenas registos entre Fevereiro e Dezembro – por variadíssimas razões. A situação descrita anteriormente, fenómeno conhecido como "Panel attrition"[5] (Steven Bellman, 2000) é controlada de modo a que, no final, a eventual perda e adição de membros no painel, não resulte em impactos significativos na análise.

[4] Browser é a plataforma de navegação na internet - google chrome é um exemplo de um browser.
[5] Traduzindo para português, atrito de painel significa a perda de indivíduos da amostra ao longo do painel, correndo o risco deste perder representatividade na população.

5.1. Caracterização geral dos indivíduos

Na Tabela II, encontram-se as estatísticas descritivas de três variáveis de frequência calculadas para análise: Média mensal do nº de visitas por indivíduo, Nº absoluto de visitas por indivíduo e a Duração média por visita (em minutos).

Um indivíduo no universo dos 1 727 que constituem este painel, em média, efectua mais de 1 595 visitas por mês, sendo que cada visita dura, em média, quase 3 minutos. Em termos absolutos, em todo o ano de 2017, um individuo, em média, fez mais de 15 151 visitas à internet, perfazendo 738 horas aproximadamente, de consumo de internet em todo o ano.

Na Tabela III, observam-se as estatísticas de distribuição demográfica do painel em estudo.

A amostra é composta sensivelmente pela mesma percentagem de indivíduos do sexo masculino e feminino (49% vs 51% aproximadamente), caracterizando-se maioritariamente (56,17%) por terem habilitações universitárias, seguidos de ensino secundário (31,21%) e primário/básico (12,62%), sendo a sua distribuição regional pela seguinte ordem de percentagens: Região do Norte (36,25%), Vale do Tejo (30,69%), Centro (21,60%), Alentejo (6,08%), Algarve e Ilhas (5,39%).

Em termos de nível laboral e de classe social, estes indivíduos são maioritariamente trabalhadores ativos (73,60%), de classe alta/média alta (50% aproximadamente), com idades identicamente distribuídas, excepto para as pessoas maiores de 61 anos (constituem apenas 7,76 %).

No anexo B é possível encontrar vários[6] gráficos auxiliares para o estudo da relação entre variáveis demográficas e a frequência das visitas. Nomeadamente, observa-se o seguinte:

Os homens assumem uma média mensal de visitas superior à das mulheres. De acordo com a Figura 2, em média, um homem faz mais 173 visitas por mês, sendo que cada visita dura mais 30 segundos relativamente ao comportamento da mulher.

Ao nível de segmentos laborais, observando a Figura 3, são os estudantes os que mais visitam a internet seguidos dos desempregados, ativos e restantes, por ordem.

[6] Por motivos de espaço, nem todos os dados estatísticos que comprovam empiricamente as afirmações foram incluídos em anexo. Contudo, poderão ser disponibilizados pelo autor em caso de necessidade.

Os jovens entre os 18 e os 30 anos são os mais frequentes visitantes na internet, em contraste com os idosos que ficam em último neste capítulo, de acordo com a indicação dada pela Figura 4. Os indivíduos com mais de 60 anos são claramente os que possuem maior rácio minutos/visita em relação aos outros, demonstrando que são um perfil de internauta com pouco conhecimento ao nível tecnológico, de visita ocasional, mas duradoura.

Temos que, ao nível de classe social, os indivíduos de classe média baixa são os que mais visitam páginas *online*. Os indivíduos cujas habilitações são inferiores ao 12º ano possuem uma duração por visita superior à dos universitários, mas têm menos visitas efectuadas, em média, por mês.

Outra curiosidade interessante pode ser observada na distribuição ao nível regional, com a região do Alentejo a apresentar maior fidelidade aos *site*s visitados, facto possivelmente explicado pela pouca afluência com que clicam em novos conteúdos *online*. Já os utilizadores da região do centro, possuem um elevado nº médio de visitas mensais, assim como uma longa duração média por visita.

É ainda demonstrado que as visitas efectuadas por dispositivos móveis, mais concretamente via *Apps*, possuem uma duração média por visita muito inferior à registada em computadores fixos.

Em termos de análise ao evento conversão, nas Tabelas IV e V notamos que a proporção de convertidos no total dos indivíduos que compõem o painel é superior para o sector dos seguros comparativamente ao da banca – 36 % contra 26 % - tendo estes em comum 178 indivíduos convertidos.

De referir que 48,5% do painel em estudo, não se converteu a qualquer sector, valor algo elevado.

Os indivíduos convertidos ao sector da banca, caracterizam-se por ter um comportamento de poucas visitas mensais (ver Figura 5), mas de grande fidelidade às páginas que visitam, enquadrando-se num perfil de pouca curiosidade sobre o mundo digital, ao que demonstram terem o conhecimento e os alvos muito específicos e racionais a alvejar.

Contrariamente para as seguradoras, na Figura 6 verificamos que o padrão dos convertidos é de muita frequência ao nível de visitas médias mensais, mas pouca retenção nas páginas que visitam. Sempre em busca de novos conteúdos e de clique

fácil, a sua conversão é porventura mais fácil mas de menor valor efectivo, relativamente aos bancos.

A partir da análise do cruzamento das características demográficas com o evento conversão, é possível retirar mais algumas notas interessantes.

Os homens convertidos à banca passam pouco tempo na internet, demonstrado pela menor frequência mensal de visita e menor tempo passado, face aos não convertidos. Uma possível razão pode passar pelo pouco tempo disponível derivado das suas vidas extremamente ativas, sendo apenas consumidores *online* por necessidade.

Já as mulheres apenas se diferenciam ao nível de tempo passado por visita efectuada, sendo as convertidas à banca as que, em média, despendem mais tempo em cada visita.

Em relação às conversões em seguradoras temos um perfil comportamental completamente diferente. Quem se converte possui uma frequência mensal de visita abruptamente superior, sendo que existe uma ligeira superioridade de conversão para os indivíduos do sexo masculino. As mulheres que se convertem apesar de visitarem muitas páginas, as suas visitas têm pouca duração.

Nas seguradoras, é percetível a importância do factor idade para a conversão, sendo os indivíduos com as idades entre os 31 e 45 anos os que mais se destacam. Para todas as idades, quanto maior a frequência do nº de visitas, maior o nº de conversões, sendo que um dado curioso se prende com a duração média por visita, os quais para as idades entre os 18 e os 30 anos, os convertidos têm drasticamente uma duração inferior. São utilizadores de grande interatividade com internet, curiosos em busca de novos estímulos, que possuem uma taxa de fidelidade muito baixa.

Ao nível das regiões, no caso da banca salta apenas à vista a região do Alentejo, a qual se diferencia pela duração por visita, na comparação do evento conversão e não conversão, indo ao encontro de uma observação feita anteriormente.

Para as seguradoras, os convertidos são mais assíduos na internet, mas de pouca duração por visita, exceptuando a zona do centro. De destacar também a maior percentagem de conversão no Norte, em comparação com o Vale do Tejo.

Focando na situação laboral, destaca-se a conversão nos seguros, que se caracteriza pela maior frequência mensal de visitas dos indivíduos ativos. Para a banca não existem diferenças que mereçam destaque.

Os utilizadores de classe média que se convertem a domínios incidentes no sector da banca qualificam-se pela maior fidelidade atribuída aos conteúdos que visitam, conhecidos por serem grandes consumidores de notícias, dado o tempo passado por visita, que gostam de estar informados, e conhecedores do mundo económico e empresarial.

Foram feitas estatísticas descritivas às variáveis de frequência de visita dos utilizadores de interesse (retirados os restantes), cujo objectivo é a comparação entre o seu comportamento geral verificado na amostra total, e o seu comportamento isolado aos domínios incidentes do sector de análise em questão.

Observando as Tabelas VI e VII, presentes em anexo, os 449 convertidos possuem uma duração média por visita em *site*s de banca superior à média revelada no painel, 1,8 minutos aproximadamente, contra 1,5. Em relação aos meses de utilização nota-se pouca regularidade comportamental de visita em bancos face à amostra geral, i.e., o indivíduo visitou uma página de banca pelo menos uma vez em 4 meses diferentes aproximadamente, contra os 9,8 meses de atividade no painel total. De realçar que, em média cada indivíduo foi visitante de, pelo menos, duas marcas de bancos diferentes, demonstrando pouca diversidade. No total do painel, para cada indivíduo a duração total foi de 70 minutos em *site*s de banca, o que *a priori* é pouco face aos 50 571 minutos gastos no geral por utilizador, uma fatia ínfima de importância que os bancos representam no comportamento dos indivíduos convertidos.

É visível a maior duração por visita em *site*s de seguros comparativamente ao painel, de acordo com as Tabelas VIII e IX. No entanto, a diferença da média entre os meses de atividade em painel e os meses de utilização em seguros é considerável, 10,4 meses contra 1,8 meses aproximadamente, dando indicadores de que a taxa de retenção para os utilizadores convertidos em seguros é muito baixa.

Em média, cada indivíduo despendeu aproximadamente de 27 minutos em *site*s do sector, o que enfatiza uma utilização muito esporádica e de fraco *engagement* com as marcas seguradoras, talvez pela oferta do digital não satisfazer ou não corresponder às expetativas do internauta. O nº de marcas de seguros diferentes visitados por um indivíduo dentre os convertidos, foi inferior a 3, indiciando pouco interesse na busca de concorrentes e/ou conformismo pelo serviço de que estes dispõem.

As marcas instaladas em Portugal consideradas para os sectores em estudo são expostas na Tabela X.

Foi feito um estudo comparativo, tendo em conta duas variáveis: o nº de visitantes distintos por marca sobre todo o painel do ano 2017 e o nº de conversões alcançadas. As Figuras 7 e 8 apresentam o *top 5* das marcas de bancos no âmbito destas variáveis.

Verifica-se que a Caixa Geral Depósitos destacou-se por larga margem como a marca que mais *reach* alcançou, juntamente com a mais alta taxa de novos visitantes (convertidos). Já o banco CTT, sendo um dos 5 bancos que mais alcance atingiu no ano, não acompanhou o nível relativamente ao nº de conversões de 2017, ficando de fora do *top 5* de convertidos.

Para as seguradoras a Fidelidade teve um bom desempenho, tanto no alcance de utilizadores durante todo o ano, como em nº de novos visitantes que conseguiu converter ao seu *website*, sendo que a Logo não conseguiu aproximar o nº de visitantes distintos do painel com o nº de novos clientes, não integrando o *top 5* das marcas com mais conversões, como demonstrado pelas Figuras 9 e 10.

As conversões ao nível de fonte e tipo, indicam que para ambos os sectores, o computador fixo foi onde ocorreu o maior fluxo de utilização, indicador similar ao da amostra total.

6. Metodologia e Resultados

6.1. Regressão Logística

A regressão logística assume-se como a abordagem mais adequada para o estudo de situações em que os resultados observados seguem uma distribuição *Bernoulli*.

Caracteriza-se por ser um caso particular dos modelos lineares generalizados, sendo que tem como principal vantagem, a modelação linear da relação entre uma variável dependente dicotómica e um conjunto de variáveis explicativas discretas e/ou contínuas. Mais, não tem necessidade de estabilizar a variância ou verificar a normalidade (Pestana e Gageiro, 2009).

Para quantificar a predisposição de um internauta visitar pela primeira vez um determinado *site* relacionado com os sectores em estudo, foi considerada como variável dependente o evento conversão, cujo sucesso assume o valor um, e o oposto o valor zero. O método utilizado para a estimação da regressão logística é o da máxima verosimilhança, e o objectivo é prever se um indivíduo pertence, ou não, ao grupo dos

convertidos ao sector de interesse. Primeiramente temos que o resultado probabilístico estimado é não linear, e corresponde ao valor esperado da distribuição *bernoulli,* tal como demonstrado na equação 1.

$$\mathrm{E(Y/X)} = \mathrm{Pr(Conversão} = 1\,|\mathrm{X}) = \frac{1}{1 + \mathrm{e}^{-(\mathrm{X\beta})}} = \mathrm{G(X\beta)} \equiv p \qquad (1)$$

X é o vector de variáveis explicativas do modelo

β é um vector de parâmetros desconhecidos

$\mathrm{G: R} \in [0,1]$

Conversão ~ *Bernoulli* (p)

Através da função de ligação *logit*, dá-se a transformação linear da relação entre as variáveis explicativas e a probabilidade do acontecimento conversão, que consiste no logaritmo das probabilidades do sucesso de ocorrência do acontecimento conversão, como se observa na equação 2.

$$logit\ \mathrm{p} = \ln(\text{odds da conversão}) = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \mathrm{X\beta} \qquad (2)$$

Desta forma, é possível estudar a probabilidade de um individuo se converter online a um dos sectores de interesse. Ao exponenciar ambos os membros (ver equação 3), o valor *logit* estimado é convertido em termos de probabilidades de sucesso, facilitando a interpretação de cada variável explicativa estimada no modelo e a influência que representa para a variável resposta.

$$\frac{\mathrm{p}}{1-\mathrm{p}} = \mathrm{e}^{-(\mathrm{X\beta})} \qquad (3)$$

As hipóteses da linearidade do Modelo, homoscedasticidade e normalidade do erro podem ser relaxadas neste contexto. Em particular, as hipóteses do modelo de regressão logística que validam a análise, a interpretação das estimativas aos parâmetros e as correspondentes estatísticas de teste, são as seguintes:

1:A variável dependente é medida de forma dicotómica.

2:As variáveis independentes podem ser contínuas ou categóricas.

3:Independência entre observações.

4:Independência dos resultados.

5:Linearidade entre a função *logit* e regressores.

5:não multicolinariedade perfeita entre regressores.

6.2. Estimação

Foram ajustados dois modelos *logit,* relativos a cada um dos dois sectores de análise do presente estudo, Banca e Seguros.

Para a construção das variáveis explicativas a utilizar na estimação do modelo, utilizou-se o raciocínio económico decorrente dos problemas em questão, junto de validação e suporte de testes não paramétricos. Numa primeira fase, foram criadas variáveis comportamentais e demográficas de carácter singular. Numa segunda fase, estas foram alvo de análise sobre a existência de correlações entre estas, evitando assim perda de eficiência do modelo.

Presentes em anexo B, as figuras 11 e 12 representam dois exemplos de técnicas não paramétricas utilizadas, um a partir do comando *Apriori*[7] na forma de regra de associação e outro que corresponde a uma árvore de decisão via *CHAID*[8], ambos com o fim de aferir eventuais interações entre variáveis singulares, potenciais de inclusão na explicação do modelo.

Segundo o princípio da parcimónia (Lawal, 2003), aquando da introdução de uma variável de interação no modelo, qualquer efeito envolvido nesta, não pode ser incluído adicionalmente como singular na regressão.

Nas tabelas XI e XII do anexo A, estão descritas todas as variáveis encontradas para a estimação dos modelos, junto dos seus efeitos esperados.

As equações 4 e 5 representam os modelos iniciais 1 e 2, cujas estimativas dos coeficientes gerados encontram-se nas tabelas XIII e XIV.

Modelo 1: (4)

$$\ln(odds \text{ conversão em banca}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} Z_1 + \beta_{16} Z_2 + \beta_{17} Z_3$$

onde X_i é uma dummy comportamental e Z_i uma variável instrumental[9]

com X_8 a X_{14} variáveis de interação.

[7] É um algoritmo que extrai a partir dos dados um conjunto de regras de associação.
[8] Técnica estatística de segmentação utilizando como critério, testes de significância Qui quadrado.
[9] Apesar de não atribuírem muito sentido económico ao modelo, as variáveis instrumentais são incluídas exclusivamente com o intuito de auxiliar na limpeza estatística do modelo e branqueamento de erros.

Modelo 2: (5)

$$\ln(odds \text{ conversão em seguros}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} Z_1 + \beta_{13} Z_2 + \beta_{14} Z_3$$

onde X_i é uma dummy comportamental e Z_i uma variável instrumental

com X_{11} variável de interação.

Observando os coeficientes estimados do modelo 1 inicial, notamos que as variáveis X_1, X_3, X_6 *e* X_9 não são estatisticamente significativas, sendo que apenas X_1 é significativo a 10%. Relativamente aos efeitos esperados, o regressor X_3 é o único que não coincide com os efeitos esperados estipulados aquando da sua construção.

Para o modelo 2 estimado com todas as variáveis incluídas, verificamos a existência de um conjunto de regressores não estatisticamente significativos, que assumem valores-p muito altos, nomeadamente X_1 que possui um valor-p superior a 0,6. Para além desta, também X_6 e X_{10} não são estatisticamente significativas a 5%. Por fim, os sinais dos coeficientes estimados para os regressores X_2, X_6 e X_{11} não correspondem aos seus efeitos esperados.

Em busca de um modelo mais parcimonioso, por meio de um processo interativo, foram conduzidos vários testes de exclusão conjunta e/ou individual das variáveis independentes, utilizando os testes LR e *Wald.*

Em anexo, nas tabelas XV e XVI encontram-se resumidos os resultados dos testes à significância estatística das restrições feitas aos modelos 1 e 2 iniciais, subdivididos em vários passos. O processo de busca por um modelo mais simplificado terminou quando a H_0 da nulidade do coeficiente é rejeitada.

Com um valor-p superior a 5% em cada passo, não se rejeita a hipótese nula de insignificância das variáveis retiradas, concluindo-se que o modelo 1 estimado sem X_1, X_3 e X_9 como regressores, corresponde à versão final e ao modelo utilizado para explicar a probabilidade de um individuo se converter online ao sector de interesse.

Relativamente ao modelo 2, de acordo com as estimativas observadas superiores a 5%, para os valores-p de cada passo da análise à significância conjunta do modelo, há evidência estatística de que o modelo 2 com as variáveis X_1, X_5, X_6 e X_{10} removidas, é o modelo mais adequado na explicação da variável dependente, correspondendo à sua versão final.

Nas tabelas XVII e XVIII encontram-se os coeficientes estimados relativos aos modelos restritos (ou finais) 1 e 2, respetivamente.

Far-se-á, de seguida, a interpretação dos coeficientes estimados para cada um dos modelos, junto de respetiva análise das suas significâncias, através das estatísticas *Wald* associadas.

Todos os coeficientes estimados do modelo 1 são estatisticamente significativos a 5%, excepto as variáveis X_2 , X_{11} e X_{13} que apenas são estatisticamente significativas a 10%.

Os sinais dos coeficientes estão de acordo com o esperado *a priori*.

Para além da constante, há duas variáveis com coeficiente negativo, significando que para cada variável, se o indivíduo tiver o atributo correspondente, igual a 1, então a probabilidade da ocorrência de conversão em banca diminui, *ceteris paribus*.

Nomeadamente, estima-se que um utilizador que esteja a navegar num *site* de uma grande corporação de negócio, tenha menor probabilidade de se converter à banca, comparativamente a um indivíduo que não tenha esse comportamento, *ceteris paribus*.

Também se estima que quando um indivíduo pesquisa termos relacionados com o emprego, a probabilidade de se converter é menor, *ceteris paribus*.

Todas as variáveis de interação influenciam positivamente a probabilidade de conversão, sendo os coeficientes de X_8 , X_{10} e X_{11} os que apresentam uma maior magnitude em termos de impacto na probabilidade estimada do indivíduo se converter.

Na sequência do anterior, indivíduos com idades entre os 18 e os 30 anos, visitantes de *site*s de bebés, de adultos e de compras/vendas diversas têm 10 vezes mais de probabilidade de se converterem a um *site* de banca.

Indivíduos do sexo feminino que simultaneamente, vis*ite*m imprensa cor-rosa, literatura e apostem *online* possuem aproximadamente 11 vezes mais hipóteses de se converterem a um *site* de banca, *ceteris paribus*.

Para indivíduos que visitam *site*s governamentais ou de entidades públicas, de automóveis, gastronomia, notícias de finanças e sejam consumidores de filmes e séries, a probabilidade de se converterem são, aproximadamente, 8 vezes maior comparativamente aos outros, *ceteris paribus*.

As variáveis da pesquisa (X_4 ao X_7) por termos relacionados com o crédito, bilhetes para festivais, e portal das finanças/IRS, estão igualmente positivamente relacionados com a probabilidade de conversão em banca.

Relativamente ao modelo 2, analisando a significância individual dos coeficientes das variáveis do modelo através de estatísticas *Wald*, verificamos que são todas estatisticamente significativas a 5%.

As variáveis constante, X_2, X_7 e X_{11} têm um efeito negativo sobre a probabilidade do utilizador se converter a seguros, algo que, em relação aos coeficientes de X_2 e X_{11} não corrobora os efeitos *a priori* esperados.

Assim, um utilizador que tenha visitado num período recente, um *site* relacionado com a compra ou venda de casas, tem menor probabilidade de se converter a um *site* de seguros, *ceteris paribus*.

Os indivíduos que vi*sit*e*m site*s de entidades públicas e governamentais, em conjunto com páginas relacionadas com banca, também apresentam uma probabilidade inferior àqueles que não o fazem, *ceteris paribus*.

Ainda, a probabilidade de acontecer a conversão no clique imediatamente seguinte a uma visita em redes sociais é menor, *ceteris paribus*.

Os regressores X_3, X_4, X_8 e X_9 têm efeitos parciais positivos em relação à probabilidade de conversão.

Um indivíduo que navegue por um *site* relacionado com uma empresa de grande dimensão, é 24 vezes mais susceptível de se converter a um *site* de seguros, constituindo este, o factor que mais influencia a probabilidade do sucesso da conversão. Quando o internauta navega num motor de busca, a probabilidade de um internauta se converter aumenta 7,4 vezes, sendo este também um importante indicador na conversão *online* do indivíduo em seguradoras.

6.3. Qualidade do ajustamento

Foi feita a avaliação aos modelos finais, com recurso ao teste de significância conjunta, cujas hipóteses nula e alternativa são apresentadas na equação 6.

$$\begin{cases} H_0: \beta_1 = \cdots = \beta_{15} = 0 \\ H_1: \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \ldots \vee \beta_{15} \neq 0 \end{cases} \quad (6)$$

De acordo com o resultado do teste realizado ao modelo 1, observado na Tabela XIX, com um valor-p nulo rejeita-se a hipótese nula de não significância conjunta, existindo evidência estatística de que, pelo menos uma variável é estatisticamente significativa.

Em relação ao modelo 2 final, de acordo com a estimativa do mesmo teste, presente na tabela XX, rejeita-se claramente a hipótese nula de não significância conjunta dos regressores, dando evidência estatística de que o modelo é significante.

A medida do R quadrado de *McFadden* representa outro indicador na avaliação de uma regressão logística estimada, embora não deva ser considerado altamente convincente, de acordo com a literatura.

Para o modelo 1, observamos um R quadrado de valor 0,0806 segundo a Tabela XXI, significando que para as variáveis incluídas no modelo, apenas 8% da variabilidade da probabilidade do indivíduo se converter, é explicada por estas. Um resultado que embora aquém do desejado, não põe em causa a validade do modelo.

Já analisando o R quadrado de *McFadden* do modelo 2 (ver tabela XXII), que tem como valor 0,239, temos que aproximadamente 24% da variabilidade assumida pela variável dependente, é explicada pelas variáveis explicativas.

6.3.1. Teste *Hosmer e Lemeshow*

O teste *Hosmer e Lemeshow* testa a hipótese de existência de diferenças significativas entre os resultados previstos do modelo estimado e os observados. Na equação 7 encontram-se as hipóteses do teste.

$$\begin{cases} H_0: \ Modelo\ bem\ ajustado\ para\ previsão \\ H_1: Caso\ Contrário \end{cases} \quad (7)$$

Com base nas estimativas do *output*, presentes na Tabela XXIII, com um valor-p de 0,728, conclui-se que não há diferenças significativas entre as previsões e os valores observados pelo modelo 1.

Para o modelo 2, de acordo com a Tabela XXIV, o valor-p é de 0,710 pelo que também não se rejeita a hipótese nula, sendo a conclusão idêntica à do modelo 1.

6.3.2. Classificações

A tabela de classificações indica a percentagem de acerto das previsões do modelo, permitindo aferir a sensibilidade e especificidade, que correspondem à capacidade de previsão do evento conversão e de prever os não convertidos, respetivamente. Para os modelos 1 e 2, compararam-se as classificações entre o modelo ajustado apenas com a constante, e o modelo final estimado.

O modelo 1 com a constante apenas como regressor, obteve uma percentagem de precisão geral da previsão, no valor de 53,4%, de acordo com a Tabela XXV. Já para o modelo 1 final ajustado, de acordo com a sua tabela de classificações (Tabela XXVI), é visível uma melhoria de mais de 11% da percentagem de acerto geral da previsão, tendo sido obtida uma sensibilidade de 75%, mas apenas 47% de especificidade.

Comparando as classificações das estimações do modelo 2 ajustado só com a constante e o final, na tabela XXVIII observamos uma vantagem significativa nas classificações do modelo estimado final, com uma percentagem geral de acerto de 50,9% contra 75% do modelo final, constituindo um indicador muito positivo para as variáveis explicativas da regressão. A percentagem de acerto do sucesso da variável dependente foi de 69,6%, sendo que, como valor de especificidade, se obteve 80,6%.

6.3.3. Curva ROC

A curva ROC consiste num processo automatizado, que permite avaliar o comportamento do modelo, através das variações observadas nos valores da especificidade e sensibilidade, para diferentes pontos de corte.

A avaliação pode ser feita visualmente, analisando a diferença entre a curva azul e a diagonal do gráfico associado ao teste, e de forma quantitativa através da estimativa do teste efectuado, presente nas tabelas XXIX e XXX, cujos resultados são tanto melhores quanto maior a área associada.

$$\begin{cases} H_0: area = 0{,}5\ , a\ curva\ ROC\ coincide\ com\ a\ diagonal \\ H_1: area > 0{,}5 \end{cases} \quad (8)$$

De acordo com as figuras 13 e 14, a curva ROC é distante da diagonal, evidenciando resultados razoáveis na adequação dos modelos na classificação dos indivíduos convertidos aos sectores respetivos.

Com um valor-p igual a 0, é rejeitada claramente a hipótese nula deste teste para ambos os modelos, sendo este um indicador positivo. A área estimada entre a curva e a diagonal do modelo 2, é maior comparativamente ao modelo 1, enfatizando melhores resultados estatísticos (uma área de 0,814 contra 0,677 do modelo 1).

7. Comparação dos modelos *Logit* e *Probit*

O *Probit* é uma abordagem que, à semelhança do *Logit*, se adequa à modelação e previsão de uma variável dependente binária. O *Probit* diferencia-se do *Logit*, na distribuição assumida que, ao invés de *logistic*, segue uma normal.

Realizou-se uma breve análise comparativa entre os dois métodos de estimação para os modelos estudados, tendo sido testada a significância conjunta destes, e utilizado como critérios de avaliação, o R quadrado *McFadden*, o critério AIC, e as percentagens de acerto na classificação dos indivíduos.

Analisando o teste de hipóteses LR, exposto na tabela XXXI, com um valor-p de 0 é rejeitada a hipótese nula do teste realizado, dando evidência estatística da significância de pelo menos uma variável explicativa do modelo 1 estimado pelo *Probit*.

Com os valores de 0,0789 e 1,308 para o R quadrado *McFadden* e critério AIC, respetivamente, no modelo *Probit* observa-se uma ligeira desvantagem estatística face à estimação pelo método *Logit*, o qual possui um valor de 1,306479 para o critério AIC e 0,0806 de R quadrado *McFadden*, indicadores superiores.

A tabela XXXIII que é relativa à classificação dos indivíduos no *Probit*, indica uma percentagem de acerto geral de 60,12%, valor comparativamente inferior ao *Logit.*

Relativamente ao modelo 2 do *Probit*, de acordo com a tabela XXXII, é rejeitada a hipótese nula do teste de significância conjunta dos regressores, sendo que com um R quadrado de 0,239, conclui-se que as variáveis incluídas, explicam aproximadamente 24% da probabilidade do indivíduo se converter ao sector das seguradoras.

Comparativamente ao modelo 2 estimado pelo *Logit*, este último com um R quadrado de 0,239 e um critério AIC no valor de 1,073240, aparenta ser o que tem melhor ajuste. A excepção ocorre ao nível da classificação dos indivíduos, onde se verifica uma vantagem do *Probit*, cujo valor percentual de acerto geral foi relativamente melhor, 75,8% tal como indicado na tabela XXXIV.

Prova-se a adequação do método de estimação *Probit* ao estudo da conversão online dos dois sectores de interesse, bem como a grande semelhança revelada dos resultados relativamente ao método *Logit*. Ainda assim, este último no geral produziu melhores resultados estatísticos, tendo ainda a vantagem da interpretação linear dos coeficientes.

8. Conclusão

O objectivo central do projeto foi através da construção de duas regressões logísticas, estimar os indicadores avançados do comportamento recente dos portugueses, na explicação da conversão online nos sectores da banca e seguros.

A dimensão dos dados disponíveis revelou-se uma vantagem para a captação de padrões de estudo e respetiva execução de toda a matéria empírica realizada. Os resultados no geral, evidenciam significância para grande parte das variáveis incluídas nos modelos, mostrando que o processo de construção das variáveis explicativas foi bem sucedido, tendo sido obtidas como percentagens de acerto geral na classificação, 62,5 % para o modelo 1, e 75% relativo ao modelo 2.

O motor de busca e a combinação de indicadores comportamentais com características demográficas, foram as variáveis que melhor previram a variável dependente no modelo da banca.

É possível definir, genericamente, os perfis dos utilizadores mais susceptíveis ao sucesso do acontecimento conversão no caso da banca. Assim, destacam-se os seguintes casos:

-Jovens com idades entre os 18 e 30 anos, recentemente pais, que por vias das necessidades passam muito tempo em casa e não estão habituados à internet.

-Adultos recéns trabalhadores, que apreciam despender tempo no conforto de casa, são fãs de séries e de encomendas *online* de comida.

- Pai/Mãe de família com um filho recém-nascido à procura de um carro mais adequado às suas necessidades e seguro de saúde para o filho.

- Pessoas que não estão familiarizadas com compras *online* e, quando o fazem, têm um alvo muito específico do produto a comprar.

-Mulheres com grande sentido de moda, independentes e com filhos.

Pesquisar por termos relacionados com o crédito, bilhetes para eventos de cultura e portal das finanças/IRS são outros dos factores influenciadores da conversão no sector da banca. Já pessoas que estão à procura de emprego, são menos prováveis de se converterem a banca.

No modelo dos seguros, os factores que mais influenciam a conversão enfatizam o peso que a última página visitada assume.

No seguimento do anterior, a variável mais importante para a previsão da conversão *online* indica que o clique de conversão ocorre com maior probabilidade se partir directamente da página de internet de uma grande empresa. Este fenómeno está implicitamente relacionado com os cliques ocorridos a partir do portal do colaborador, realizados pelos funcionários da empresa, que possuem benefícios de seguro fornecidos pela entidade patronal.

Ainda, as conversões ocorridas a partir de uma página *web* de motor de busca são aproximadamente 7 vezes mais prováveis de acontecer, existindo uma clara tendência da intenção comportamental na explicação do evento conversão, cujo teor sugere que o indivíduo convertido já tem conhecimento do que quer e procura.

A partir de redes sociais, a conversão é mais difícil de acontecer, sendo esta variável a que mais influencia negativamente a conversão *online* de um indivíduo em seguros.

Com o intuito de confirmar os bons indicadores especificados no modelo *Logit*, foi feito um estudo comparativo com o método *Probit*. Constatou-se que, no geral, o *Logit* apresentou resultados mais satisfatórios embora muito similares.

O estudo sugere que se deve ter em consideração as características demográficas de um indivíduo em conjunto com as variáveis comportamentais deste, aquando da definição do público-alvo a atacar no sector da Banca, provando-se que o ajustamento da previsão do modelo melhora de forma relevante.

Recomenda-se que os bancos adotem uma política de *marketing* menos evasiva, através de introdução de anúncios em *site*s estratégicos com um alcance de pessoas alvo definidas personalizadamente.

Poderá existir uma oportunidade de mercado relativa ao segmento de pessoas desempregadas, estimulando uma certa dinâmica ao nível social e económico.

Para as seguradoras, a importância dos acordos de apólice que estas possuem com as grandes empresas revelou-se de um impacto positivo enorme sobre as conversões, política que é recomendável e produtiva de se apostar.

Os resultados relacionados com o nº de conversões via motor de busca mostram uma inequívoca intenção comportamental demonstrada por parte do indivíduo convertido, constituindo um indicador positivo para o sector ao nível de notoriedade junto do público, uma vez que o target conhece a oferta e procura-a para satisfazer as necessidades de que dispõe. No entanto, ficou provado o fraco *engagement* entre as

marcas de seguradoras e o utilizador, visível através do *gap* entre as frequências de utilização das marcas estudadas e do painel geral. A convergência destes indicadores deve ser um*a* prioridade, algo que pode eventualmente ser efectuado através da otimização do *website*, aumento de serviços digitais disponíveis, etc.

Relativamente a marcas, constata-se uma correlação entre as que alcançam regularmente mais visitantes ao longo do ano, com o nº de conversões que estas obtêm, sendo a CGD a que melhor resultado estatístico obteve neste estudo, ficando no topo em ambas as medidas. Já o banco CTT, apesar da boa prestação no global em 2017, não acompanhou o mesmo nível em matéria de conversões, e não figura no *top 5* de conversões *online*.

Das marcas de Seguradoras, a ADSE foi a seguradora que melhor resultado obteve em matéria de visitantes, pese embora se deva realçar que a ADSE enquadra-se como um seguro de saúde de natureza pública destinada aos funcionários públicos, logo a sua comparação com as demais marcas existentes não é totalmente assertiva. Excluindo a ADSE, a Fidelidade foi a que melhores indicadores obteve, tanto para visitantes distintos em absoluto em 2017, como a nível de novas captações *online*.

A AdvanceCare também figura no *top 5* de seguradoras que maior alcance tiveram em 2017, mas não transpôs o mesmo nível para o nº de conversões *online*.

Podemos ainda depreender um nicho de mercado a atacar em ambos os sectores, que se insere na aposta de desenvolvimento e otimização dos serviços oferecidos via *App*.

Uma limitação existente que trava o crescimento exponencial *online* em ambos os sectores - mais para banca - prende-se com os custos fixos associados aos balcões existentes por todo o país. Apesar da noção de que o futuro é digital, o processo decorrente deste, é demorado e de difícil gestão.

9. Recomendações

Ao longo do desenvolvimento do presente estudo surgiram algumas limitações.

Comparando as frequências de visita dos indivíduos convertidos entre o painel total e o sector de interesse, observamos um *gap* relativamente grande.

Em investigação futura, perceber que factores explicativos para o churn [10] aqui verificado, e descobrir quais as soluções mais adequadas que ajudariam à sua convergência, seriam importantes contributos complementares ao tema.

Pese embora hajam dados que permitam saber a data e duração de uma visita num determinado *site* e, quais os *site*s antecedentes e subsequentes a este, não existe meio de averiguar como se deram os cliques, i.e., se foi a partir de um anúncio publicitário, se foi direccionado automaticamente ou através de *url* inserido, etc.

Apesar do carácter abstrato associado à motivação comportamental do indivíduo, a existência de informação relativa ao conteúdo efectivo do clique, tornaria mais claras as estimativas assumidas a este. Sabendo com rigor as motivações de cada clique, teremos uma conclusão mais completa suportada por análises mais consistentes.

Referências Bibliográficas

ACEPI & IDC. " Estudo Anual da Economia e da Sociedade Digital em Portugal" (2017).

Cheta, Rita. "BANCA DIGITAL: do Multibanco ao Homebanking." (2007).

Figueiredo, José Miguel Barbosa. "Classificação e identificação do cliente churner-O caso da banca de retalho em Portugal." (2014).

Lacerda, Flávia Gomes de. "Inovação como vantagem competitiva no Mercado Segurador: um Estudo de Caso na Empresa Porto Seguro." (2015).

Halchuk, José Pedro. "UM ESTUDO SOBRE TAXA DE CONVERSÃO EM LOJAS VIRTUAIS."NO E-COMMERCE: As implicações no processo da decisão de compra no varejo de moda B2C. (2014)"

Kim, Kyung Kyu, and Bipin Prabhakar. "Initial trust and the adoption of B2C e-commerce: The case of internet banking." ACM SIGMIS Database: the DATABASE for Advances in Information Systems 35.2 (2004): 50-64.

[10] Churn é uma medida do número de indivíduos que saem de um grupo coletivo durante um período específico.

Nielson Global Survey, consumer confidence section, 2017

Koon, Sharon, and Yaacov Petscher. "Comparing Methodologies for Developing an Early Warning System: Classification and Regression Tree Model versus Logistic Regression. REL 2015-077." Regional Educational Laboratory Southeast (2015).

de Oliveira Fernandes, Luana, Anatália Saraiva Martins Ramos, and Anatália Saraiva. "Intenção de compra *online*: aplicação de um Modelo adaptado de aceitação da tecnologia para o comércio eletrônico." Revista Eletrônica de Sistemas de Informação 11.1 (2012).

Cabral, Cleidy Isolete Silva. Aplicação do Modelo de regressão logística num estudo de mercado. Diss. 2013.

Mudiwa, Benjamin. "A *Logit* estimation of factors determining adoption of conservation farming by smallholder farmers in the semi-arid areas of Zimbabwe." Agricultural and -Applied Economics, Department of Agricultural Economics and Extension Faculty of Agriculture University of Zimbabwe (2011).

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37.Lohse, Gerald, Steven Bellman, and Eric J. Johnson. "Consumer buying behavior on the Internet: Findings from panel data." (2000).

Lani, James. "Assumptions of logistic regression." *Statistics Solutions* (2010).

Bayo Lawal. "Categorical Data Analysis with SAS and SPSS Applications.", 2003.

Maria Helena Pestana & João Gageiro “Análise Categórica, Árvores de decisão e Análise de conteúdo”,2009.

Anexo A – Tabelas

Tabela I - Descrição Geral dos Dados

	Apps	***Browser***	**Total de Observações**
MÓVEL	5523320	1827614	7350934
	75,07%	24,84%	
PC FIXO	-	81533287	81527924
	-	100,00%	
TOTAL	5523320	83360901	**88878858**

Distribuição de Frequências

Variáveis	**Domínios**	**Fonte**	**Tipo visita**	**Visitas**	**Duração**
Contagem	88885147	88885147	88885147	88885147	88885147
Min	1	1	1	1	0,017
Max	254574	2	3	30723712	8907418
Alcance	254573	1	2	30723711	8907418
Variância	5,32E+08	0,076	0,059	7,76E+13	1799947
Desvio padrão	23065,37	0,276	0,242	8808662	1341,621
Erro padrão da média	2,447	0	0	934,319	0,142

Variáveis	**Tipo dispositivo**	**Sistema Operativo**
Contagem	88885147	88885147
Min	1	1
Max	3	3
Alcance	2	2
Variância	0,096	0,09
Desvio padrão	0,31	0,3
Erro padrão da média	0	0

Variáveis	**Categoria**	**Subcategoria**	**Marca**	**País *url***
Contagem	88885147	88885147	88885147	88885147
Min	1	1	1	1
Max	25	131	3993	229
Alcance	24	130	3992	228
Variância	11,629	226,991	40635,74	10,379
Desvio padrão	3,41	15,066	201,583	3,222
Erro padrão da média	0	0,002	0,021	0

Tabela II - Frequências de utilização mensal por indivíduo

Estatística	Média Mensal de Visitas	Nº Visitas	Média minutos/visita
Contagem	1727	1727	1727
Média	**1595,12**	**15151,232**	**2,924**
Min	1	1	0,017
Max	8686	98493	476,827
Alcance	8685	98492	476,81
Variância	2512189,731	298428907,1	246,637
Desvio padrão	1584,989	17275,095	15,705
Erro padrão da média	38,14	415,694	0,378

Tabela III - Distribuição Demográfica do Painel

	Descrição	Percentagem	Frequência
Género	Mulher	48,52%	838
	Homem	51,48%	889
Nível de Escolaridade			
	Ensino Primário ou Básico	12,62%	218
	Ensino Secundário	31,21%	539
	Ensino Superior	56,17%	970
Situação Laboral			
	Activo	73,60%	1271
	Desempregado	8,22%	142
	Doméstico/a	1,97%	34
	Estudante	9,44%	163
	Pensionista/ Reformado	6,77%	117
Região			
	Alentejo	6,08%	105
	Algarve e Ilhas	5,39%	93
	Centro	21,60%	373
	Norte	36,25%	626
	Vale do Tejo	30,69%	530
Classe Social			
Classe alta/média alta	AB	48,47%	837
Classe média	C1	39,78%	687
Classe média baixa	C2	9,90%	171

Classe baixa	D	1,85%	32
Classe de idades			
	18-30	32,08%	554
	31-45	32,43%	560
	46-60	27,74%	479
	>61	7,76%	134

Tabela IV - Frequências das conversões aos dois sectores de análise

	a Banca		**a Seguradoras**	
Conversão	Indivíduos	Percentagem	Indivíduos	Percentagem
Sim	449	26%	619	35,84%
Não	1278	74%	1108	64,16%
Total	**1727**	**1**	**1727**	**1**

Tabela V - Matriz de convertidos entre Banca e Seguradoras

	Banca					
	Conversão	Sim	Percentagem	Não	Percentagem	**Total**
	Sim	178	10,31%	441	25,54%	**619**
Seguradoras	Não	271	15,69%	837	48,47%	**1108**
	Total	**449**	**26%**	**1278**	**74%**	**1727**

Tabela VI - Estatísticas descritivas da frequência em todo o painel (apenas indivíduos convertidos a Banca)

	duração total painel	**duração média visitas painel**	**meses activo painel**	**média visitas mensais painel**
Contagem	449	449	449	449
Média	50571,010	1,567	9,800	1496,763
Min	29,200	0,348	1,000	3,455
Max	6263300,767	103,749	12,000	8686,444
Alcance	6263271,567	103,401	11,000	8682,990
Variância	142217574104,790	23,982	7,830	2404608,156
Desvio padrão	377117,454	4,897	2,798	1550,680
Erro padrão da média	17797,273	0,231	0,132	73,181

Tabela VII - Estatísticas descritivas da frequência só em domínios Banca (apenas indivíduos convertidos a Banca)

	duração total banca	duração média por visita banca	meses visitou banca	média visitas mensais banca	visitas banca	marcas banca visitadas
Contagem	449	449	449	449	449	449
Média	70,872	1,892	4,076	7,816	51,978	2,347
Min	0,017	0,017	1,000	1,000	1,000	1
Max	887,750	11,217	12,000	84,875	1167,000	12
Alcance	887,733	11,200	11,000	83,875	1166,000	11
Variância	12570,388	2,994	9,771	96,657	10164,651	2,799
Desvio padrão	112,118	1,730	3,126	9,831	100,820	1,673
Erro padrão da média	5,291	0,082	0,148	0,464	4,758	0,079

Tabela VIII- Estatísticas descritivas da frequência em todo o painel (apenas indivíduos convertidos a seguradoras)

	duração total painel	duração média visitas painel	meses activo painel	média visitas mensais painel
Contagem	619	619	619	619
Média	58164,441	1,376	10,468	2069,654
Min	16,033	0,342	1,000	3,500
Max	8964345,917	103,749	12,000	8351,111
Alcance	8964329,883	103,407	11,000	8347,611
Variância	191977320556,751	17,405	5,298	2742794,338
Desvio padrão	438152,166	4,172	2,302	1656,138
Erro padrão da média	17610,822	0,168	0,093	66,566

Tabela IX- Estatísticas descritivas da frequência só em domínios de seguros (apenas indivíduos convertidos a Seguradoras)

	duração total seguros	duração média por visita seguradoras	meses visitou seguros	média visitas mensais seguradoras	visitas seguros	marcas seguros visitadas
Contagem	619	619	619	619	619	619
Média	27,933	1,832	1,964	5,634	18,491	2,596
Min	0,017	0,017	1,000	1,000	1,000	1
Max	542,717	43,267	12,000	112,000	527,000	15
Alcance	542,700	43,250	11,000	111,000	526,000	14
Variância	2507,177	7,283	2,474	67,761	1378,422	4,681
Desvio padrão	50,072	2,699	1,573	8,232	37,127	2,164
Erro padrão da média	2,013	0,108	0,063	0,331	1,492	0,087

Tabela X - Marcas de Bancos e Seguradoras em estudo

Bancos	Seguradoras
Cgd	Fidelidade
Bankinter	Ok Teleseguros
Santander	Adse
Banco Ctt	Logo
Bcp	Seguro Directo
Bpi	Ageas
Montepio	Nseguros
Best	Tranquilidade
Euro Bic	Mapfre
Novo Banco	Multicare
Banco Atlântico	Advance Care
Banco Invest	Allianz Seguros
Activobank (Bcp)	Medis (Ocidental)
Big	Lusitania
Deutsche Bank	Liberty Seguros
BBVA	Acoreana Seguros
Caixa Agricola	Zurich
Finantia	Real Vida Seguros
Citibank	Europ Assistance

Tabela XI - Variáveis iniciais consideradas para o Modelo 1 e os seus efeitos esperados

Y	Igual a 1 se o indivíduo se converteu a um sector da Banca	
X 1	Igual a 1 se o indivíduo visitar 5 ou mais categorias diferentes de *site*s	+
X 2	Igual a 1 se o último clique anteriormente à conversão for a um *site* relacionado com uma grande corporação de negócios (por exemplo Sonae ou Meo)	-
X 3	Igual a 1 se o indivíduo visitar um *site* de venda de casas (por ex. Remax)	+
X 4	Igual a 1 se o indivíduo pesquisar termos relacionados com crédito	+
X 5	Igual a 1 se o individuo pesquisar termos relacionados com serviços públicos (por exemplo "como receber o IRS")	+
X 6	Igual a 1 se o individuo tiver pesquisado termos relacionados com bilhetes para eventos culturais ou festivais	+
X 7	Igual a 1 se o indivíduo pesquisar termos relacionados com a procura de emprego	-
X 8	Igual a 1 se o indivíduo com idade entre os 18 e 30 anos, visita *site*s relacionados com bebés e nascimentos, de adultos e de compra/venda diversas	+
X 9	Igual a 1 se o indivíduo com idade entre 31 e 45 anos, visitou *site*s relacionados com turismo	+
X 10	Igual a 1 se o indivíduo é mulher, visitou *site*s de imprensa cor-rosa, literatura e apostas *online*	+
X 11	Igual a 1 se o indivíduo visita *site*s governamentais/entidades públicas (por ex. autoridade tributária, *site*s da câmara municipal), de carros e motores no geral, gastronomia, notícias de finanças e de filmes	+
X 12	Igual 1 se o indivíduo é mulher, frequenta *site*s de bebés, fashion/moda, e governamentais	+
X 13	Igual a 1 se o indivíduo pesquisar termos relacionados com bebés, compra de carro e seguros saúde	+
X 14	Igual a 1 se o indivíduo pesquisar termos sobre compras *online* e serviços de pagamento *online* (por ex. MBway ou MBnet)	+
Z 1	Instrumento 1	
Z 2	Instrumento 2	
Z 3	Instrumento 3	

Tabela XII - Variáveis iniciais consideradas para o Modelo 2 e os seus efeitos esperados

Y	Igual a 1 se o indivíduo se converteu a um sector de Seguros	
X 1	Igual a 1 se o indivíduo visitar *site*s de categorias relacionadas com bebés, maternidade e nascimentos	+
X 2	Igual a 1 se o indivíduo visitar *site*s de categorias relacionadas com compra habitação	+
X 3	Igual a 1 se o indivíduo visitar *site*s de encomenda alimentar *online*	-
X 4	Igual a 1 se o indivíduo visitar *site*s relacionados com o turismo	+
X 5	Igual a 1 se o indivíduo visitar *site*s de compra e venda de artigos diversos (por ex. OLX)	+
X 6	Igual a 1 se o indivíduo visitar *site*s de interesse do mundo automóvel	+
X 7	Igual a 1 se o último clique anteriormente à conversão for a uma rede social	-
X 8	Igual a 1 se o último clique anteriormente à conversão for num motor de busca	+
X 9	Igual a 1 se o último clique anteriormente à conversão for a um *site* relacionado com uma grande corporação de negócios	+
X 10	Igual a 1 se o indivíduo pesquisar termos relacionados com a economia e sua actualidade	+
X 11	Igual a 1 se o indivíduo visitar banca e *site*s governamentais	+
Z 1	Instrumento 1	
Z 2	Instrumento 2	
Z 3	Instrumento 3	

Tabela XIII - Estimação do Modelo 1 inicial

Variável	**B**[11]	**S.E.**[12]	***Wald***	**df**[13]	**Sig.**[14]	**Exp(B)**[15]
Constante	-0,663	0,164	16,385	1	0,000	0,515
X 1	0,314	0,180	3,032	1	0,082	1,369
X 2	-0,627	0,307	4,152	1	0,042	0,534
X 3	-0,316	0,256	1,518	1	0,218	0,729
X 4	1,259	0,519	5,886	1	0,015	3,521
X 5	0,424	0,203	4,394	1	0,036	1,529
X 6	0,727	0,446	2,652	1	0,103	2,068
X 7	-0,624	0,273	5,207	1	0,022	0,536
X 8	2,185	1,070	4,173	1	0,041	8,894

[11] parâmetro β
[12] erro padrão
[13] graus liberdade
[14] valor-p da estatística *wald*
[15] exponencial da estimativa β

X 9	0,302	0,284	1,131	1	0,287	1,353
X 10	2,310	1,076	4,611	1	0,032	10,078
X 11	2,046	1,109	3,404	1	0,065	7,733
X 12	0,529	0,316	2,799	1	0,094	1,698
X 13	1,514	0,794	3,634	1	0,057	4,546
X 14	0,948	0,417	5,164	1	0,023	2,581
Z 1	0,745	0,184	16,475	1	0,000	2,107
Z 2	0,816	0,208	15,411	1	0,000	2,261
Z 3	-0,519	0,267	3,782	1	0,052	0,595

Tabela XIV - Estimação do Modelo 2 inicial

Variável	**B**	**S.E.**	***Wald***	**df**	**Sig.**	**Exp(B)**
Constante	-1,305	0,167	61,157	1	0,000	0,271
X 1	0,101	0,200	0,256	1	0,613	1,106
X 2	-0,416	0,175	5,670	1	0,017	0,660
X 3	0,401	0,151	7,029	1	0,008	1,493
X 4	0,310	0,160	3,769	1	0,052	1,363
X 5	0,242	0,144	2,835	1	0,092	1,274
X 6	-0,194	0,176	1,212	1	0,271	0,824
X 7	-1,317	0,225	34,405	1	0,000	0,268
X 8	2,007	0,152	174,808	1	0,000	7,441
X 9	3,179	1,060	8,997	1	0,003	24,025
X 10	0,596	0,399	2,233	1	0,135	1,815
X 11	-0,335	0,187	3,204	1	0,073	0,715
Z 1	0,618	0,206	9,021	1	0,003	1,855
Z 2	0,460	0,192	5,742	1	0,017	1,584
Z 3	0,970	0,271	12,833	1	0,000	2,638

Tabela XV - Conjunto de testes de significância LR sobre o Modelo 1

	variáveis retiradas	**Estatística Qui quadrado compara dois Modelos**	**graus liberdade**	**valor -p**
Modelo inicial	-	-	-	-
Passo 2	X 9	-1,147	1	0,284
Passo 3	X 3	-1,459	1	0,227
Passo 4	X 1	-2,563	1	0,109

Tabela XVI - Conjunto de testes de significância LR sobre o Modelo 2

	variáveis retiradas	Estatística Qui quadrado compara dois Modelos	graus liberdade	valor-p
Passo 1	-	-	14	0,000
Passo 2	X 1	-0,256	1	0,613
Passo 3	X 6	-1,175	1	0,278
Passo 4	X 10	-2,286	1	0,131
Passo 5	X 5	-2,685	1	0,101

Tabela XVII - Estimação do Modelo 1 Final

Variável	B	S.E.	*Wald*	Df	Sig.	Exp(B)
Constante	-0,557	0,148	14,197	1	0,000	0,573
X 2	-0,588	0,305	3,717	1	0,054	0,555
X 4	1,321	0,516	6,551	1	0,010	3,747
X 5	0,466	0,201	5,369	1	0,021	1,594
X 6	0,794	0,441	3,234	1	0,072	2,212
X 7	-0,549	0,270	4,128	1	0,042	0,578
X 8	2,321	1,070	4,702	1	0,030	10,182
X 10	2,359	1,069	4,865	1	0,027	10,576
X 11	2,111	1,102	3,668	1	0,055	8,254
X 12	0,639	0,308	4,292	1	0,038	1,894
X 13	1,499	0,792	3,577	1	0,059	4,476
X 14	0,893	0,413	4,676	1	0,031	2,441
Z 1	0,687	0,178	14,949	1	0,000	1,988
Z 2	0,853	0,206	17,133	1	0,000	2,347
Z 3	-0,538	0,265	4,114	1	0,043	0,584

Tabela XVIII - Estimação do Modelo 2 Final

Variável	B	S.E.	*Wald*	df	Sig.	Exp(B)
Constante	-1,190	0,152	61,427	1	0,000	0,304
X 2	-0,362	0,170	4,550	1	0,033	0,697
X 3	0,419	0,148	8,002	1	0,005	1,521
X 4	0,275	0,150	3,388	1	0,066	1,317
X 7	-1,319	0,224	34,722	1	0,000	0,267
X 8	2,003	0,151	175,377	1	0,000	7,413
X_9	3,188	1,056	9,119	1	0,003	24,228
X 11	-0,297	0,178	2,777	1	0,096	0,743
Z 1	0,525	0,152	11,893	1	0,001	1,690
Z 2	0,523	0,179	8,565	1	0,003	1,687
Z 3	0,826	0,224	13,619	1	0,000	2,283

Tabela XIX - Teste de Hipóteses para a significância conjunta do Modelo 1

Estatística Qui quadrado LR	valor-p
92,46548	0,000

Tabela XX - Teste de Hipóteses para a significância conjunta do Modelo 2

Estatística Qui quadrado LR	valor-p
416,8107	0,000

Tabela XXI - Estatística R quadrado e critério AIC do Modelo 1

Modelo 1	
R quadrado McFadden	Critério AIK
0,080626	1,306479

Tabela XXII - Estatística R quadrado e critério AIC do Modelo 2

Modelo 2	
R quadrado McFadden	Critério AIK
0,239436	1,073240

Tabela XXIII - Teste de ajustamento *Hosmer e Lemeshow* da previsão para o Modelo 1

Teste Hosmer and Lemeshow		
Estatística Qui quadrado compara dois Modelos	df	Sig,
3,621	6	0,728

Tabela XXIV - Teste de ajustamento *Hosmer e Lemeshow* da previsão para o Modelo 2

Teste Hosmer and Lemeshow		
Estatística Qui quadrado compara dois Modelos	df	Sig,
5,436	8	0,710

Tabela XXV - Classificações do Modelo 1 só com constante

Tabela de Classificação só com constante				
Observado	Previsão			
		Conversão		Percentagem sucesso
		0,0	1,0	
Conversão	0,0	0	387	0,0
	1,0	0	443	100,0
Percentagem Overall				53,4

Valor de corte=0,5

Tabela XXVI - Classificações do Modelo 1 final

Tabela de Classificação				
Observado	Previsão			
		Conversão		Percentagem sucesso
		0,0	1,0	
Conversão	0,0	183	204	47,28682
	1,0	107	336	75,8465
Percentagem Overall				62,5

Valor de Corte=0,5

Tabela XXVII - Classificações do Modelo 2 só com constante

Tabela de Classificação				
Observado		Previsão		
		Conversão		Percentagem sucesso
		0,0	1,0	
Conversão	0,0	0	617	0,0
	1,0	0	639	100,0
Percentagem Overall				50,9

Valor de Corte=0,5

Tabela XXVIII - Classificações do Modelo 2 final

Tabela de Classificação				
Observado		Previsão		
		Conversão		Percentagem sucesso
		0,0	1,0	
Conversão	0,0	497	120	80,6
	1,0	194	445	69,6
Percentagem Overall				75

Valor de Corte=0,5

Tabela XXIX - Teste da Curva ROC do Modelo 1

Resultados teste		
Área	Erro padrão	Valor-p
0,677	0,018	0,000

Tabela XXX - Teste da curva ROC do Modelo 2

Resultados teste		
Área	Erro padrão	Valor-p
0,814	0,012	0,000

Tabela XXXI - Estatísticas Modelo 1 estimado pelo *Probit*

Estatística Qui quadrado LR	**valor-p**
90,56095	0,000
Modelo 1	
R quadrado McFadden	Critério AIC
0,078965	1,308774

Tabela XXXII - Estatísticas Modelo 2 estimado pelo *Probit*

Estatística Qui quadrado LR	**valor-p**
416,2214	0,000
Modelo 2	
R quadrado McFadden	Critério AIC
0,239098	1,073709

Tabela XXXIII- Classificações do Modelo 1 final pelo *Probit*

Tabela de Classificação				
Observado	Previsão			
		Conversão		Percentagem sucesso
		0,0	1,0	
Conversão	0,0	296	91	76,49
	1,0	240	203	45,82
Percentagem Overall				60,12

Valor de Corte=0,5

Tabela XXXIV- Classificações do Modelo 2 final pelo Probit

Tabela de Classificação				
Observado	Previsão			
		Conversão		Percentagem sucesso
		0,0	1,0	
Conversão	0,0	496	121	80,39
	1,0	192	447	69,95
Percentagem Overall				75,8

Valor de Corte=0,5

Anexo B – Figuras

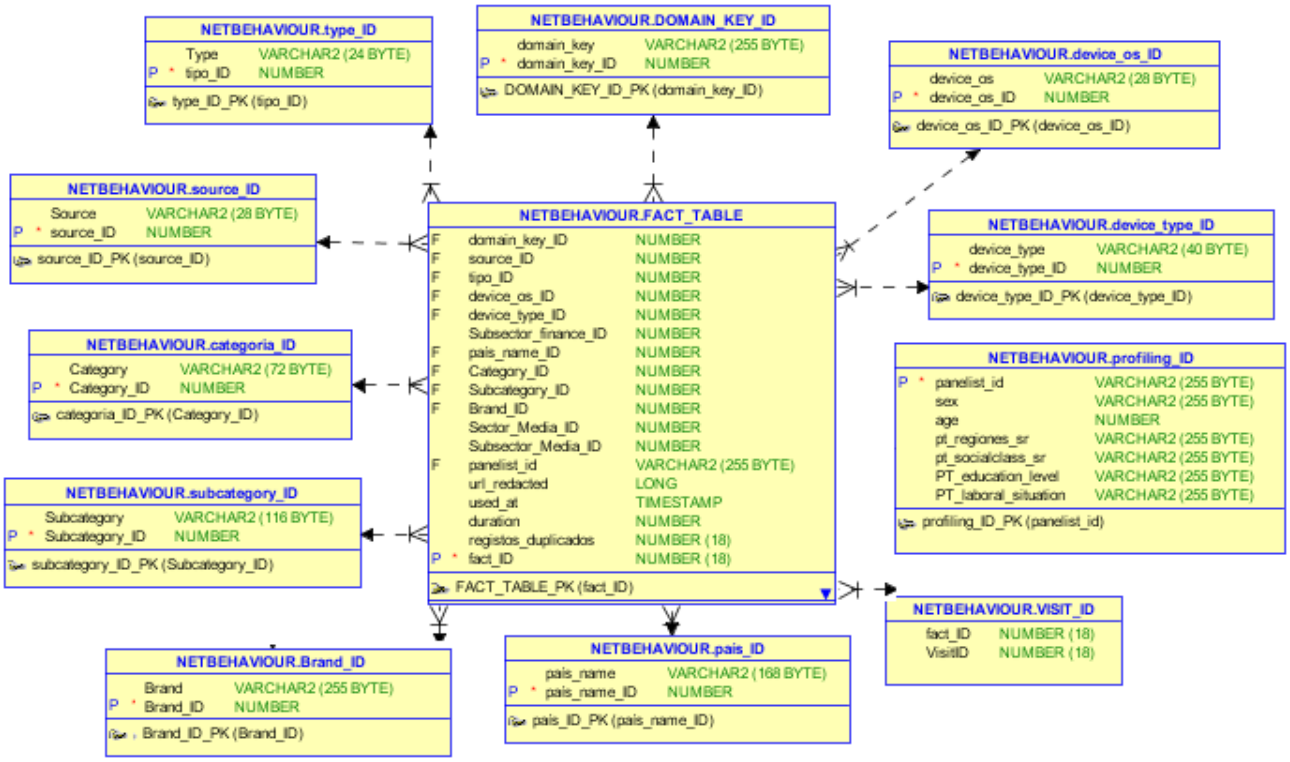

Figura 1 - Diagrama representativo da estrutura relacional da Base de Dados

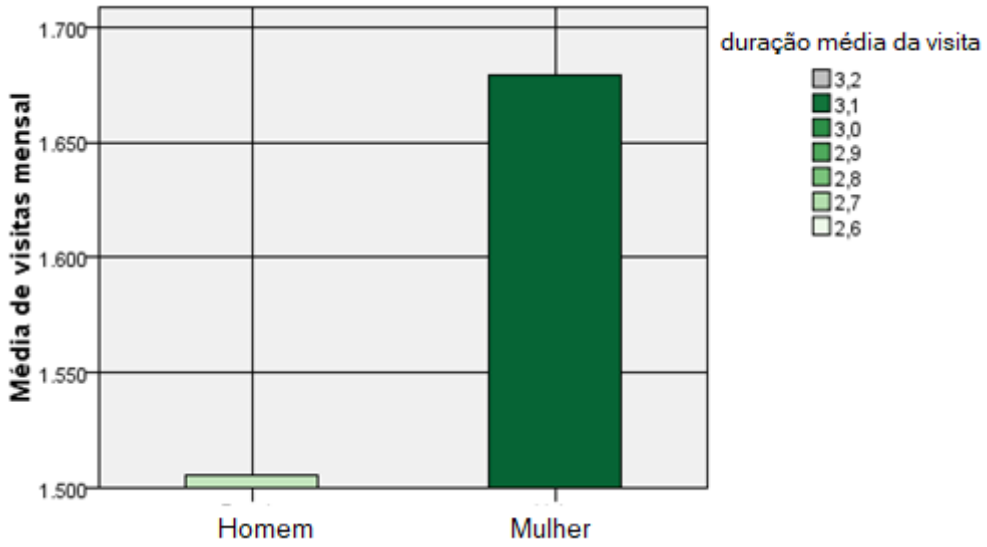

Figura 2 - Distribuição de nº visitas e duração por Géneros

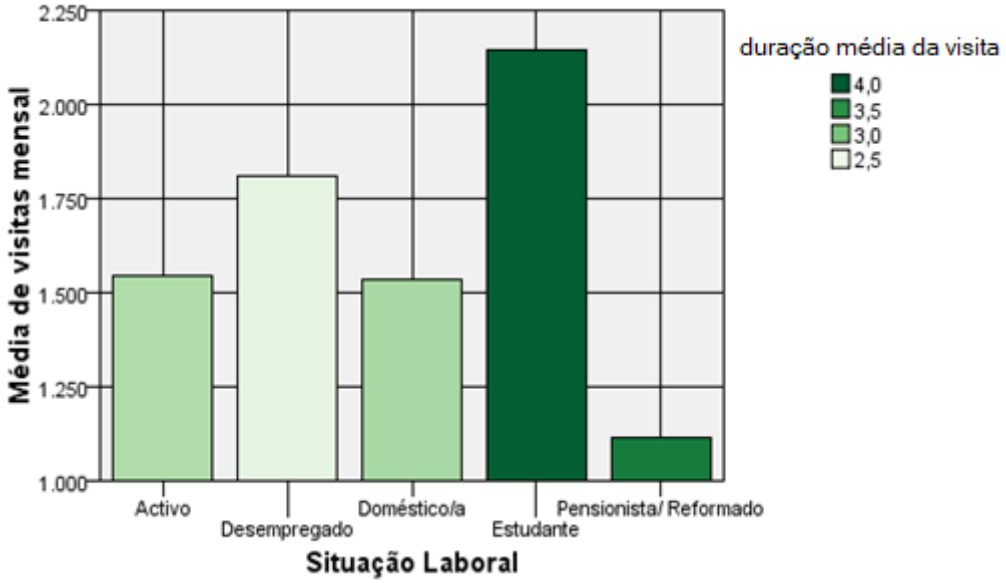

Figura 3 - Distribuição de nº visitas e duração por Situação Laboral

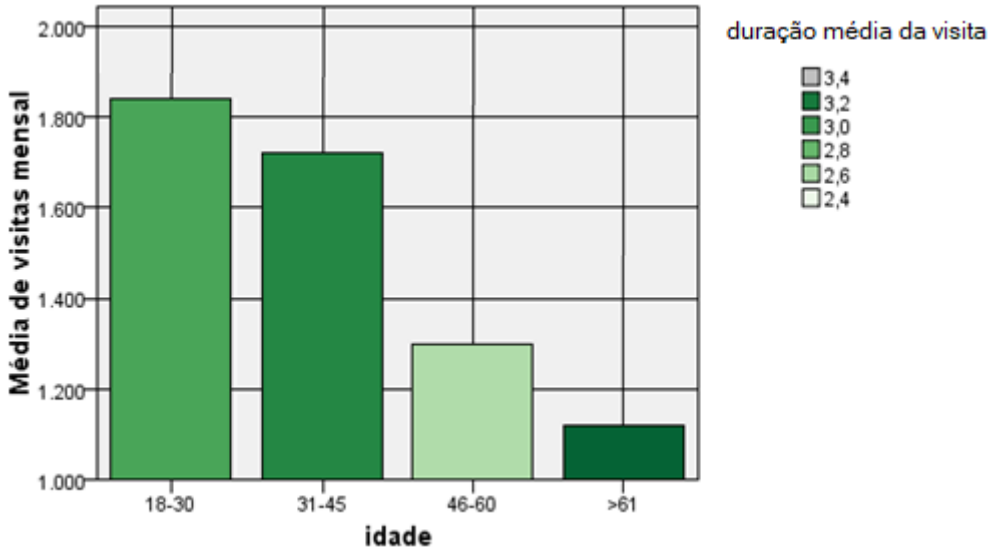

Figura 4 - Distribuição de nº visitas e duração por classes de idade

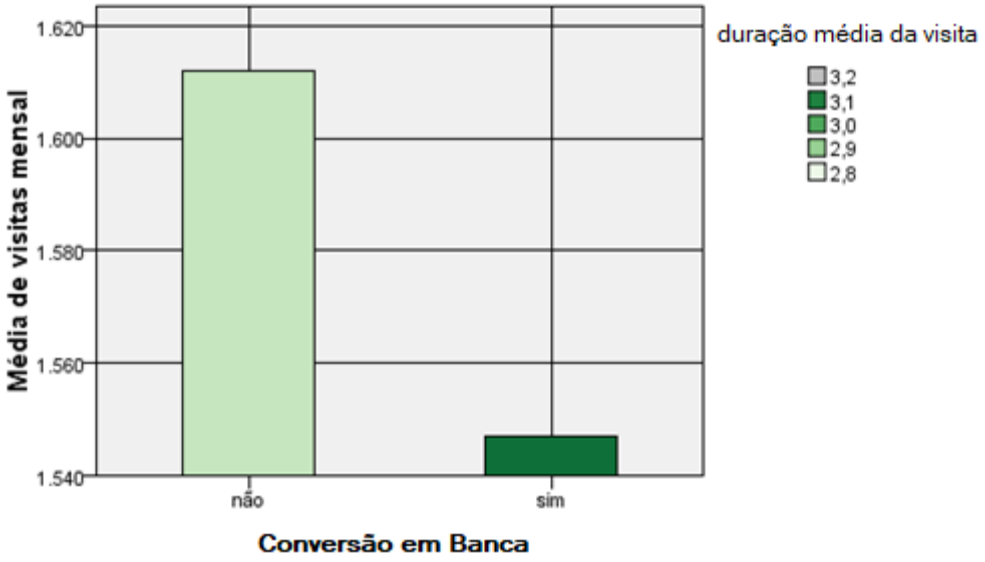

Figura 5 - Distribuição de nº visitas e duração, por Conversão e Não Conversão em Banca

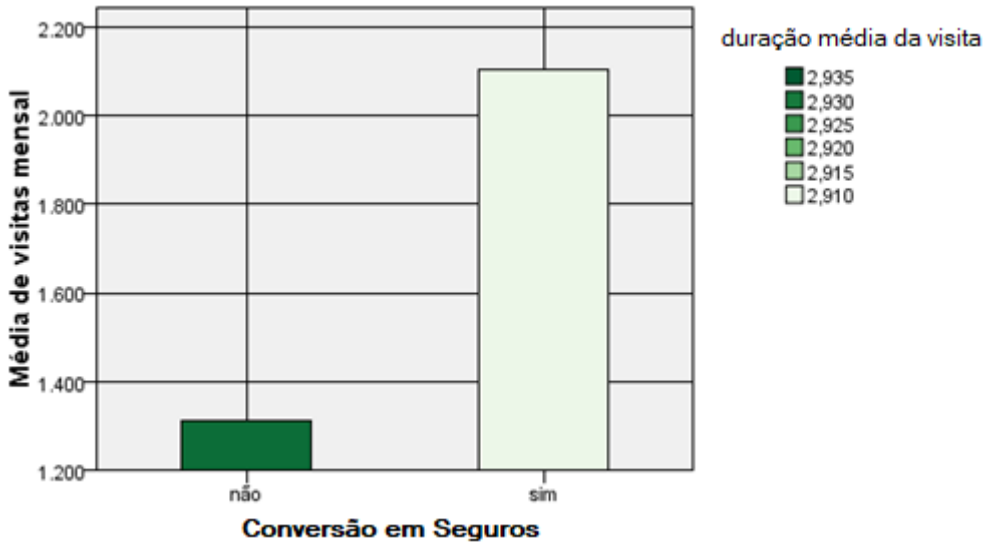

Figura 6 - Distribuição de nº visitas e duração, por Conversão e Não Conversão em Seguradoras

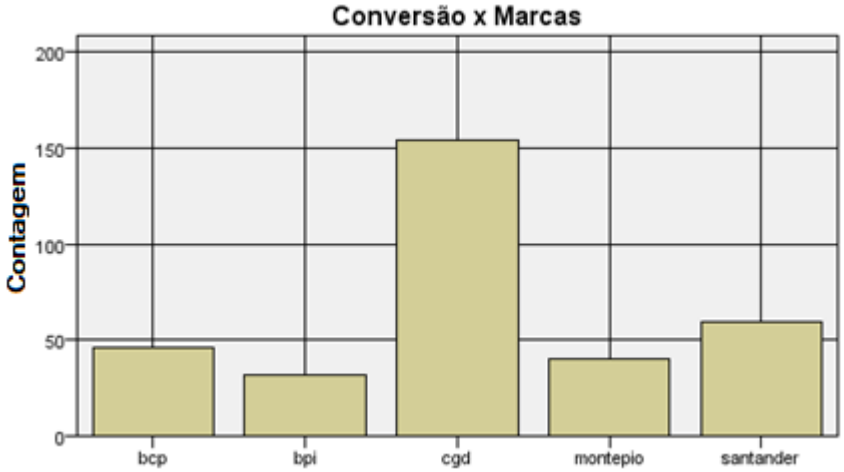

Figura 7 - Conversões alcançadas por Marcas de Bancos

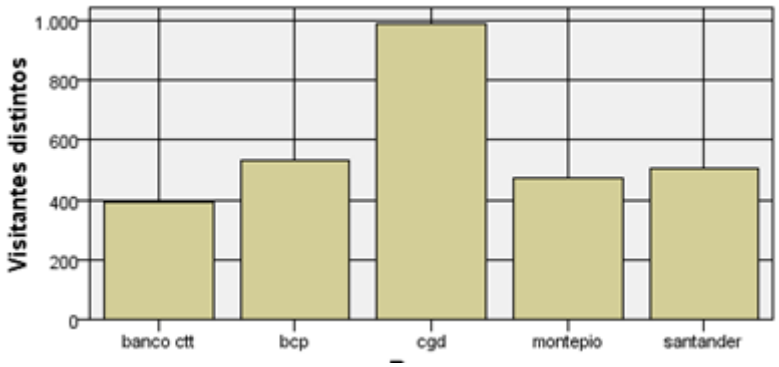

Figura 8 - Visitantes distintos total por Marcas de Bancos

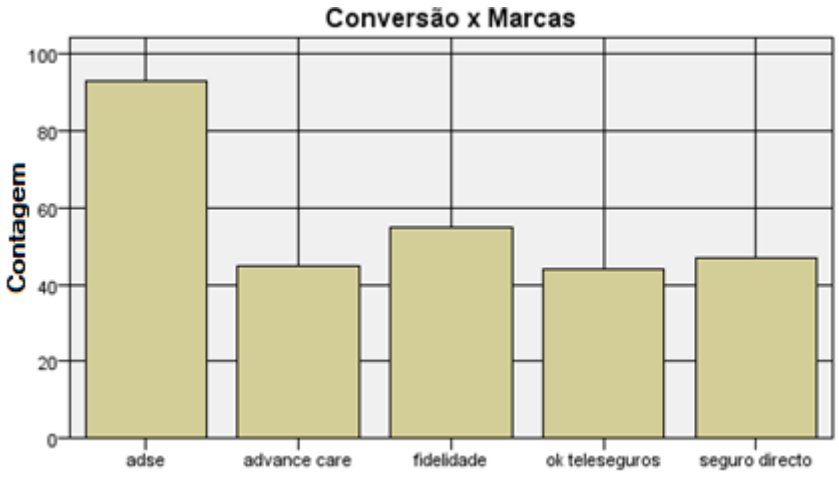

Figura 9 - Conversões alcançadas por Marcas de Seguradoras

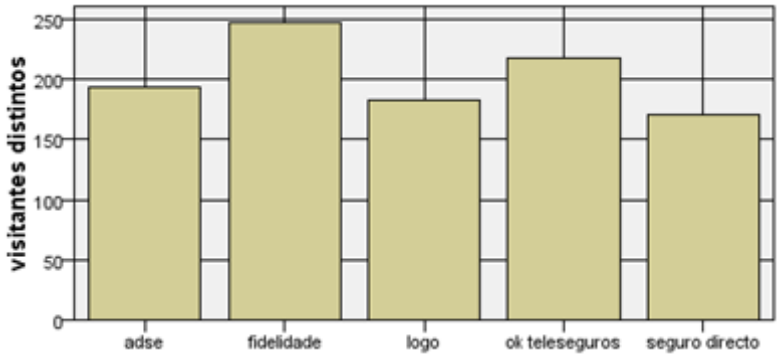

Figura 10 - Visitantes distintos total por Marcas de Seguradoras

Consequente	Antecedente	Suporte %	%Confiança
Conversão	Adultos	1,48%	92,30%
	bebés		
	18-30 anos		
	compra e vendas diversas		

Figura 11 - Exemplo de regra de associação estimada pelo algoritmo APRIORI

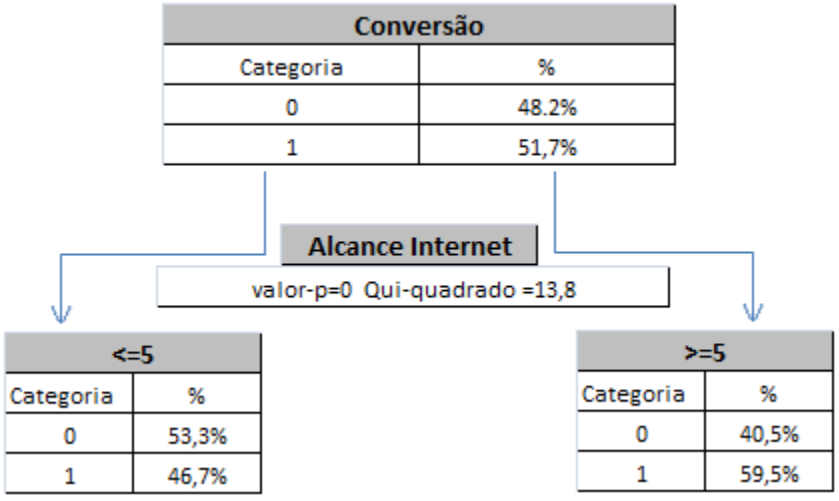

Figura 12 - Exemplo de uma árvore de decisão CHAID estimada

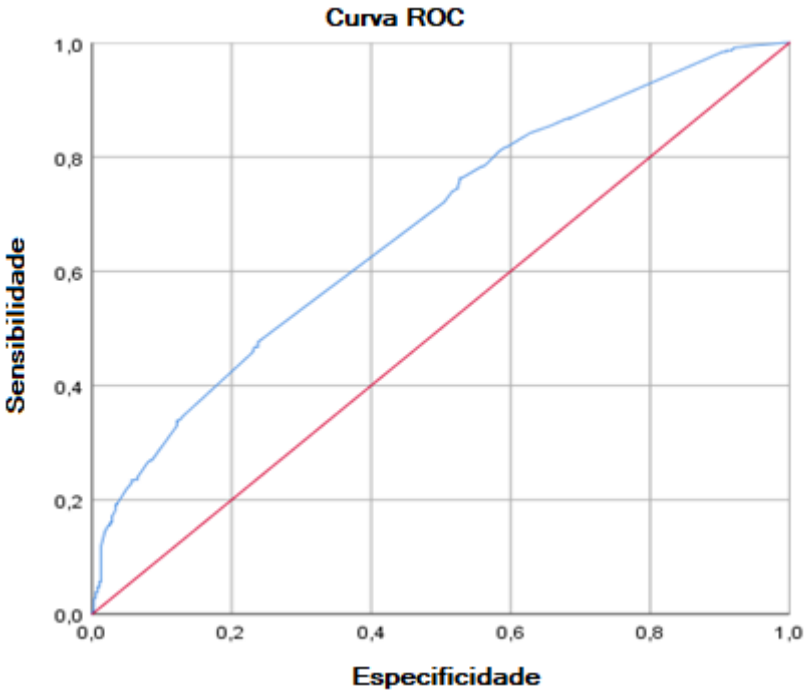

Figura 13 - Curva ROC associada ao Modelo 1

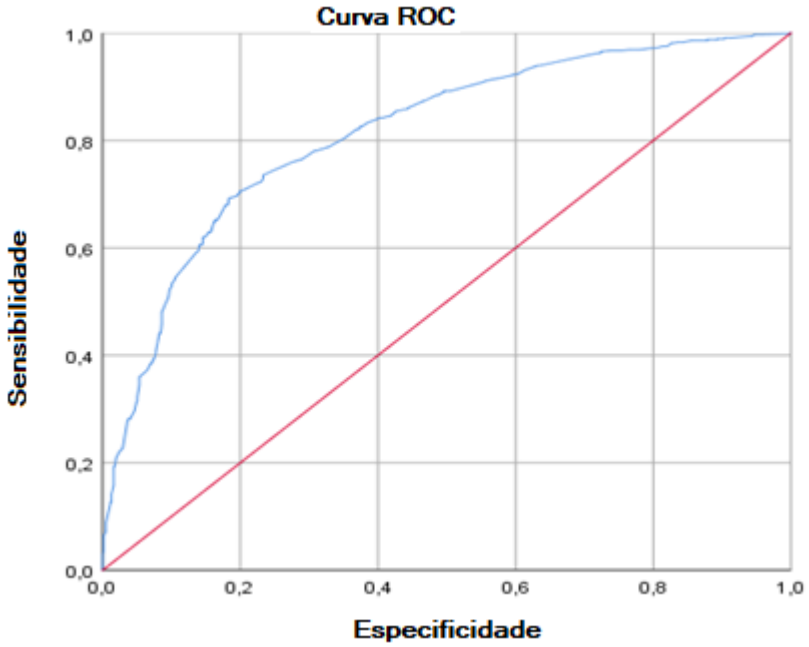

Figura 14 - Curva ROC associada ao Modelo 2